frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 13 February 2026
DOI 10.3389/feduc.2026.1749760

OPEN ACCESS

EDITED BY
Joaquin Marc Veith,
Leipzig University, Germany

REVIEWED BY
Yasemin Tas,
Atatürk University, Türkiye
Valentina Nachtigall,
Ruhr University Bochum, Germany

*CORRESPONDENCE
Deniz C. Senel
✉ senel@physik.rwth-aachen.de

RECEIVED 19 November 2025
REVISED 13 January 2026
ACCEPTED 14 January 2026
PUBLISHED 13 February 2026

CITATION
Senel DC, Schüttler T, Ertl B and Watzka B (2026) Development of a multidimensional questionnaire on students' perceived authenticity in science learning contexts. *Front. Educ.* 11:1749760. doi: 10.3389/feduc.2026.1749760

Development of a multidimensional questionnaire on students' perceived authenticity in science learning contexts

Deniz C. Senel[1]*, Tobias Schüttler[2], Bernhard Ertl[3] and Bianca Watzka[1]

[1]Didactics of Physics and Technology, RWTH Aachen University, Aachen, Germany, [2]DLR_School_Lab, German Aerospace Center (DLR), Weßling, Germany, [3]Learning and Teaching with Media, Institute of Education, Universität der Bundeswehr München, Neubiberg, Germany

Authentic learning environments are recognized as effective for increasing students' interest and engagement in science education. However, the development and improvement of such environments are constrained by the absence of robust and comprehensive measurement tools for students' perceptions of authenticity. Existing instruments often do not capture the multidimensional nature of authenticity. More recently, researchers have developed questionnaires that operationalize authenticity as a multidimensional construct in educational contexts. Yet, even these instruments do not address all theoretically relevant aspects of authenticity identified in established models. In this study, we therefore develop and validate a new questionnaire designed to measure students' perceived authenticity in learning settings. The questionnaire comprises six dimensions: *location*, *instructor*, *innovation*, *methods*, *materials*, and *content*. Data were collected from $N = 146$ secondary school students at an out-of-school laboratory in Germany and analyzed using confirmatory factor analysis. The results indicate a six-factor solution aligned with the hypothesized model, with internal consistencies exceeding 0.75 for five of the six scales. The resulting instrument provides a more precise tool for assessing students' perceptions of authenticity, facilitating a more accurate evaluation of authentic learning environments in science education and supporting the development of interventions to enhance authenticity in instructional design.

KEYWORDS

confirmatory factor analysis, disciplinary authenticity, instrument development, perceived authenticity, questionnaire validation, science communication

1 Introduction

Authenticity has become a central concept in science education, serving both as a guiding principle for designing learning environments and as a means to enhance students' interest and engagement in science (Habig and Gupta, 2021). In response to critiques of traditional instruction, which often features tasks with procedural clarity and predetermined outcomes (Chinn and Malhotra, 2002), researchers have advocated for more authentic learning experiences that are intended to bridge the gap between science-based school activities and scientific practice (Schwartz et al., 2004).

Although authenticity is a widely used term in science education, it remains complex and multifaceted (Anker-Hansen and Andrée, 2019). Various research domains highlight different aspects, including disciplinary realism, professional practice, real-world authenticity, and personal authenticity (Lee and Butler, 2003; Watkins et al., 2012; van Vorst et al., 2015; Kapon et al., 2018; Schriebl et al., 2023).

A domain-oriented perspective conceptualizes authenticity (so-called "disciplinary authenticity") as the extent to which learning environments mirror the epistemic, procedural, and contextual characteristics of actual scientific practice (Betz, 2018; Sommer et al., 2018). Building on this framework, Finger et al. (2022) conceptualize authenticity as a continuum from simplified tasks to complex, open-ended representations of professional practice. This study adopts this perspective and draws on the definition provided by Nachtigall et al. (2022, p. 1482), who, based primarily on seminal work by Herrington and Oliver (2000), describe authentic learning settings as "characterized by learners, who collaboratively try to solve a complex and ill-structured real-world problem through self-directed inquiry and investigation [...] together with practitioners or experts, in a real-world or professional setting, [...] using materials and tools that are either typically also applied by practitioners or are used in daily life."

While this definition sets a high-level ideal of authentic learning, it is important to clarify that authenticity in science education does not require direct participation in professional research or immersion in real scientific institutions. Rather, authenticity can be meaningfully achieved through pedagogically designed representations of scientific practice—by approximating core features such as epistemic openness, procedural reasoning, and the use of authentic tools or data within the constraints of school and outreach settings (Betz et al., 2016; Roth, 1995). Students do not need to generate novel scientific knowledge to experience learning as authentic. In this sense, learning is considered authentic when it reflects the routines, tools, and complexity of scientific work (Braund and Reiss, 2006).

However, authenticity is not only a property of learning environments but also emerges through the interaction between the learner and the learning context itself (Honebein et al., 1993; Nachtigall et al., 2022). As Habig et al. (2018) argue, students' perceptions are not only influenced by features of the learning environment but are also shaped by their prior knowledge, beliefs and representations of science. This implies that even well-designed environments may not be perceived as authentic by all learners.

1.1 Authenticity as a multidimensional construct

To conceptually capture the interaction between learner characteristics and environmental design features, Betz et al. (2016) proposed a multidimensional model of authenticity in science communication (see Figure 1). The model outlines how perceived authenticity emerges through a process called *authentication*, in which objectively authentic features of a learning environment are interpreted through the lens of the learner's individual characteristics, such as prior knowledge, interest, epistemological beliefs or experience. The model consists of five elements: (a) *personal characteristics*; (b) *characteristics of the learning setting*; (c) *authentication through interaction*; (d) *an individual feeling of authenticity*; and (e) *its effects*. These are organized into three conceptual parts: input (Part 1), interaction (Part 2), and output (Part 3).

In the first part (input), a learner with their *personal dispositions and characteristics* (a) interacts with a *learning setting and its design features* (b). This interaction determines whether a student perceives the situation as authentic. The model identifies key dimensions of instructional design that act as cues for perceived authenticity, including *location*, *instructors*, *methods*, *materials*, *content*, and *innovation*. These features can be varied to convey more or less authenticity depending on the domain. Empirical studies show that physical materials (Peacock, 1997; Mierwald et al., 2018) and the learning location (Schüttler et al., 2021) often play particularly salient roles. Recent findings by Hohrath et al. (2024) further support these dimensions: In interviews following an experimental project day, students referred to authenticity cues such as location, instructor and materials when describing what they considered authentic research. In the second part (interaction), the two inputs (a) and (b) converge into the element *authentication* (c), describing how learners evaluate and interpret authenticity through their own dispositions, resulting in *an individual perception of authenticity* (d). In the final part (output), this perception may lead to various *affective*, *cognitive*, and *behavioral effects* (e), such as situational interest, motivation, changes in beliefs and the development of domain-specific competencies.

1.2 Relevance of assessing learners' perceived authenticity

Beyond its conceptual relevance, learners' perceived authenticity plays a central role in explaining the effects of authentic learning environments. A comprehensive review by Nachtigall et al. (2022), synthesizing 50 experimental and quasi-experimental studies, demonstrates that authentically designed learning settings frequently yield positive motivational and cognitive outcomes, such as increased situational interest, motivation, engagement, and learning performance. The results of this review suggest that authentic learning environments tend to foster motivational outcomes—particularly situational interest—more consistently than cognitive learning outcomes. For example, rich real-world materials and contexts are effective in triggering learners' curiosity and engagement, as they are perceived as meaningful and relevant. At the same time, these same features may affect cognitive learning if the complexity of authentic materials exceeds learners' processing capacities or if instructional guidance is insufficient. Consequently, authenticity-related gains in interest do not automatically translate into improved conceptual understanding. Rather, cognitive outcomes appear to depend on how different aspects of authenticity are realized and supported through instruction. This further underscores the importance of distinguishing between different dimensions of perceived authenticity when examining their effects on motivational vs. cognitive learning outcomes.

Empirical studies further substantiate this differentiation by demonstrating that specific dimensions of learners' perceived authenticity are associated with motivational and cognitive learning outcomes in distinct ways (Nachtigall and Rummel, 2021). With regard to the authenticity of the method, Hohrath et al. (2024) showed that learners' perceived method authenticity was positively related to knowledge acquisition, indicating that this dimension can directly support cognitive learning under appropriate instructional

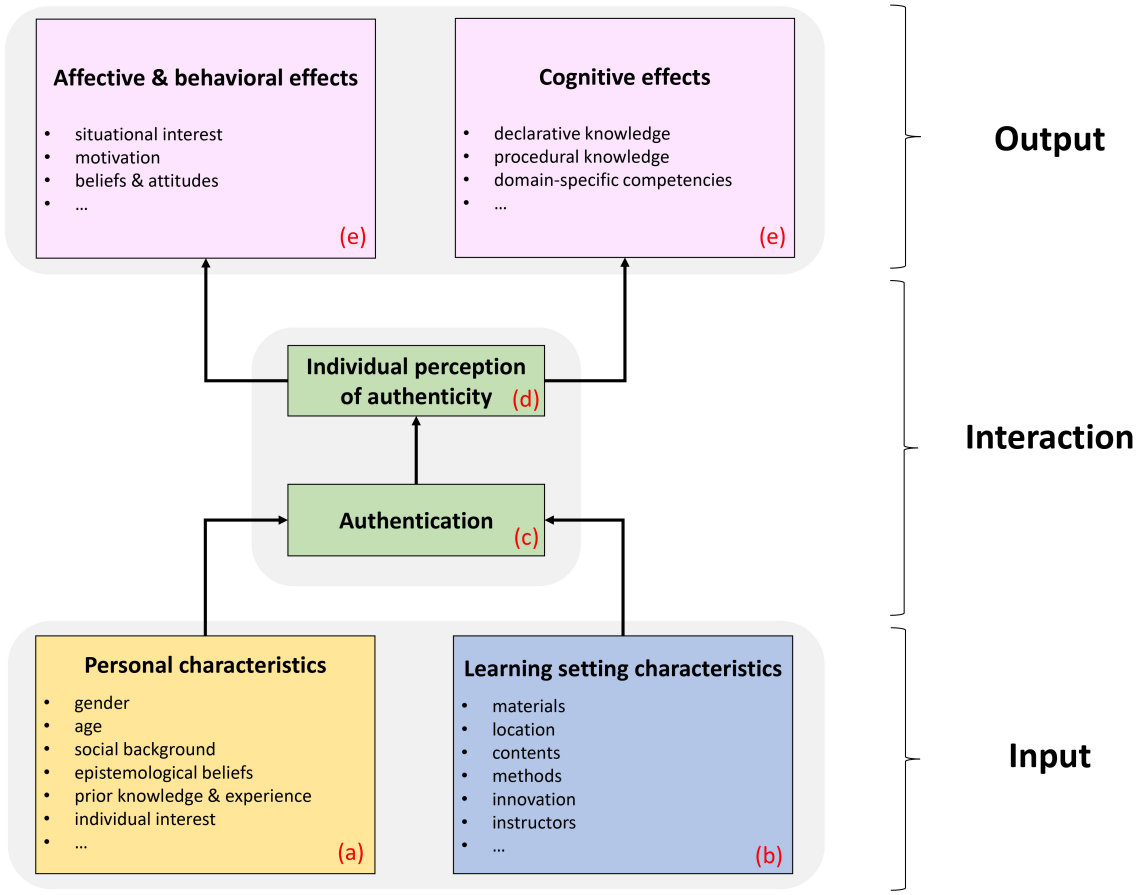

FIGURE 1
Conceptual model of perceived authenticity in science education, based on the theoretical framework by Betz et al. (2016). It illustrates how authenticity emerges through the interaction between individual learner characteristics **(a)** and features of the learning environment **(b)**. This interaction **(c)** results in an individual perception of authenticity **(d)**, which in turn may affect motivational, cognitive, and behavioral outcomes **(e)**.

conditions. In contrast, the authenticity of the instructor appears to primarily affect motivational outcomes: Hagenkötter et al. (2024) demonstrated that perceived instructor authenticity was associated with triggered situational interest, whereas no significant effects on knowledge outcomes were observed. A similar pattern emerges for material-related authenticity: Zachrich et al. (2024) reported that perceived authenticity of materials in the context of history education was associated with increased engagement and interest, while effects on cognitive learning outcomes were more variable. More evidence is provided by Schüttler et al. (2021), who showed that the authenticity of laboratory equipment selectively affected epistemic components of situational interest and learners' perceived relevance of the learning content, whereas emotional and value-related interest components were less consistently influenced. With regard to location-related authenticity, Schüttler et al. (2021) demonstrated that perceived authenticity of the learning site positively influenced students' situational interest and relevance perceptions during visits to out-of-school science laboratories, a finding consistent with earlier work by Betz (2018). In contrast, findings regarding knowledge acquisition are inconsistent: Scharfenberg et al. (2007) reported a positive effect of the learning location, whereas no positive effects were found by Itzek-Greulich et al. (2015). When it comes to the perceived authenticity of content and innovation in particular, empirical evidence remains scarce, as these aspects have rarely been examined as learners' perceptions and instead often remain subsumed under broader design characteristics.

Conclusively, while empirical studies have begun to disentangle the effects of different dimensions of authenticity, empirical evidence on the effects of perceived authenticity remains fragmented, thereby motivating a multidimensional assessment approach.

This differentiated relationship between learners' perceived authenticity and motivational vs. cognitive learning outcomes is also consistent with the previously described theoretical model of authenticity by Betz et al. (2016). In this model, perceived authenticity is conceptualized as a central mediating variable through which design elements of authentic learning environments influence learning processes and outcomes. In particular, the model assumes that authenticity primarily operates by enhancing learners' engagement, relevance perceptions, and motivation, which may subsequently, but not necessarily, support cognitive learning. The empirically observed pattern of robust motivational effects but more variable cognitive outcomes aligns with this assumption (Nachtigall et al., 2022). It suggests that perceived authenticity may constitute a necessary but not sufficient condition for cognitive

learning, which additionally depends on instructional support, task structure, and cognitive load. From this perspective, the present focus on learners' perceived authenticity and its multidimensional assessment provides an important empirical basis for further examining and refining the mechanisms proposed in the model by Betz et al. (2016).

1.3 Instruments for measuring authenticity

Despite the growing consensus on the multidimensionality of authenticity, as demonstrated by recent frameworks (e.g., Betz, 2018; Nachtigall et al., 2024; Sommer et al., 2018), this perspective has only gradually been reflected in empirical research. Earlier studies typically employed one-dimensional measures, such as global ratings or single items, to assess perceived authenticity (e.g., Engeln, 2004; Pawek, 2009; Nicaise et al., 2000). More recent studies have aimed to develop instruments that better reflect the multidimensional nature of authenticity by linking it to design elements of learning environments, such as *location* (Schriebl et al., 2021; Schüttler et al., 2021), *instructional format and design* (Svärd et al., 2024), and *tools or materials* (Peacock, 1997; Schüttler et al., 2021). Yet, many of these instruments assess only isolated aspects of authenticity or merge distinct dimensions into broader factors, thus failing to capture its full complexity.

A first validated multidimensional attempt was made by Finger et al. (2022), who developed a questionnaire to capture perceived authenticity in out-of-school settings. Their instrument, also known as FEWAW[1], is based on the framework by Betz et al. (2016) and serves as the starting point for the present study. However, their confirmatory factor analysis did not support the proposed five-factor model. Instead, *materials* and *methods* collapsed into one factor (*approach*, "Vorgehen" in FEWAW). Furthermore, the instrument excluded the theoretically relevant dimension of *content*. Taken together, these findings motivated an extension of the original FEWAW instrument to differentiate six theoretically grounded dimensions of perceived authenticity, namely *location*, *instructor*, *materials*, *methods*, *innovation*, and *content* (see Figure 1).

1.4 Aims and research questions

Based on theoretical considerations and previous empirical findings, the aims of this study are to develop and validate an extended multidimensional instrument for assessing perceived authenticity in science education. To address this, the following questions (RQs) guide our research:

- RQ1 (factorial structure): Does the adapted and extended version of the FEWAW questionnaire reflect a six-factor structure corresponding to the dimensions of *location*, *instructor*, *materials*, *methods*, *innovation*, and *content*?
- RQ2 (psychometric properties): To what extent does the revised instrument demonstrate psychometric quality across key dimensions of scale validation?
  - RQ2.1 (internal consistency): Do all the scales show sufficient internal consistency (e.g., McDonald's $\omega \geq 0.70$), indicating reliable measurement of each construct?
  - RQ2.2 (discriminant validity): Are all dimensions empirically distinguishable, as indicated by latent correlations below 0.80 (Cheung et al., 2024) and Heterotrait-Monotrait (HTMT) ratios of correlations below 0.85 (Henseler et al., 2015)?
  - RQ2.3 (convergent validity): Are the scales positively and significantly correlated with theoretically related constructs such as general authenticity and is the Average Variance Extracted (AVE) for each dimension 0.50 or greater (Fornell and Larcker, 1981)?
  - RQ2.4 (measurement invariance): Does the instrument yield comparable scores across different learner groups, as shown by measurement invariance?

2 Methods

In the following, we will first describe the development of the instrument, then outline the sample and the setting in which this study took place and finally present the procedures used for data analysis.

2.1 Instrument development

2.1.1 Approach for a new multidimensional instrument on authenticity

In this study, we aim to develop and evaluate a refined set of scales that distinguishes key components of disciplinary authenticity, drawing on the authenticity framework proposed by Betz et al. (2016). The resulting instrument comprises six theoretically derived dimensions: *location*, *instructor*, *materials*, *methods*, *innovation*, and the newly introduced dimension *content*. To address issues of conceptual overlap and insufficient discrimination, the definitions for the *materials* and *methods* scales were revised and a new definition was created for *content*, while the original definitions for *location*, *instructor*, and *innovation* were mostly retained. In addition, each dimension is explicitly differentiated from other related dimensions to ensure conceptual clarity and to avoid overlap in interpretation. As the original definitions were formulated in German, the following section provides English translations of all six dimensions, including the revised and newly developed ones.

Location refers to the physical context surrounding the learning setting. This dimension captures how authentically students perceive the buildings, rooms and places visited. It also includes locations beyond student laboratories, such as botanical gardens, excavation sites or research vessels (Braund and Reiss, 2006). A location is considered authentic if it is actually or potentially used for scientific work (Schüttler et al., 2021). This dimension differs from *instructor*, *materials*, and *methods* in that it focuses solely on the spatial characteristics of the learning setting rather than on the people, artifacts or procedures involved in scientific work.

1 FEWAW is the German acronym for *Fragebogen zur Erfassung der Wahrnehmung von Authentizität in der Wissenschaftsvermittlung* (Questionnaire for Assessing the Perception of Authenticity in Science Communication).

Instructor refers to individuals who take on a teaching or guiding role within the learning activity. This dimension reflects how authentically students perceive these individuals. Instructors are considered authentic if they work professionally as scientists and offer direct access to real scientific practices within their discipline (Nicaise et al., 2000).

Innovation refers to the novelty of the research question or problem addressed. A question is considered authentic if it is embedded in ongoing, relevant research and allows students to contribute to its resolution. The more central the inquiry into something genuinely new is and the less existing knowledge is merely received or replicated, the more innovative the project is perceived to be (Sommer et al., 2020). This dimension differs from *content* in that it focuses on the epistemic novelty of the task rather than on the scientific relevance of the subject matter itself.

Methods refer to the specific scientific procedures used to investigate a research question. This dimension reflects the extent to which students perceive their working process as scientific. A method is considered authentic when it aligns with the standards of disciplinary inquiry and is embedded in a systematic process of knowledge generation. Typical examples include experimentation, modeling, simulation, data analysis or observation (Bolger et al., 2021). Although there is a pluralism of methods within scientific disciplines, in most cases, only one of these methods is focused on in educational contexts. This dimension differs from *materials* in that it concerns the scientific procedures themselves rather than the physical artifacts used to carry them out.

Materials refer to all objects that students interact with during the activity, excluding instructional materials. This dimension reflects how authentic students perceive these materials. Materials are considered authentic when they originate from scientific research or closely resemble such materials in form, function and use, with minimal didactic simplification (Sommer et al., 2018). This includes both objects of investigation (e.g., soil samples, texts, interview transcripts) and instruments or visualizations that support real research (e.g., microscopes, measuring devices, simulations or lab software; Smetana and Bell, 2012). This dimension differs from *methods* in that it focuses on the physical artifacts used in scientific work and not on the procedural steps through which knowledge is generated.

Content refers to the scientific topics, concepts or questions addressed during the learning activity. This dimension reflects how authentically students perceive the subject matter. Content is considered authentic when it relates to a real research topic in the discipline and has not been overly simplified or artificially constructed for instructional purposes (Sommer et al., 2018). This dimension differs from *innovation* in that it focuses on the relevance of the subject matter itself, not on the epistemic novelty or potential for generating new knowledge.

To clarify the conceptual refinements made to the original FEWAW instrument, it is necessary to specify how the present definitions differ from and expand upon those previously proposed by Finger et al. (2022).

In their version, *materials* were broadly defined as any non-instructional objects that students interact with, including objects of investigation (e.g., soil samples, texts) and research instruments (e.g., microscopes, indicator solutions). We mostly retained this two-part structure but made the criteria for authenticity more explicit by emphasizing the origin of the materials (i.e., from authentic scientific research) and their similarity in form, function and use to those typically found in scientific practice. Additionally, the definition was extended to include scientific visualizations and digital tools (e.g., simulations or laboratory software), which are widely used in science (Smetana and Bell, 2012).

Similarly, the original *methods* dimension referred to the scientific character of students' activities and their participation in systematic inquiry. While the core idea was preserved, concrete examples of disciplinary procedures (e.g., experimenting, modeling and data analysis) were added to reflect the methodological diversity of scientific practice and to help learners identify what constitutes authentic scientific work (Bolger et al., 2021).

2.1.2 Item generation and refinement

The initial item pool was developed based on theoretically derived dimensions of perceived authenticity, using the FEWAW questionnaire by Finger et al. (2022) as the primary reference. Definitions for the dimensions *location*, *instructor*, and *innovation*, which showed empirical distinctiveness and acceptable mean values in the original study, were adopted with minor linguistic modifications for clarity and consistency. The original item sets for these three dimensions were retained to preserve continuity with the established scale structure. In contrast, new items were created for the *methods*, *materials*, and *content* dimensions according to the extended theoretical definitions. The *content* dimension, which was not included in the original instrument, was added to capture the perceived authenticity of scientific topics and disciplinary concepts (Betz et al., 2016).

To refine and validate the item pool, several iterative rounds of theoretical and linguistic review were conducted. A group of eight experts in science education, including professors, teachers, and doctoral students, as well as a group of five secondary school students, independently classified each item into one of the six dimensions of perceived authenticity. For this purpose, all participants were provided with the item pool and the corresponding dimensions. In addition to the classification task, students were asked to provide feedback on the linguistic comprehensibility of the items using an optional open-response format. Experts were additionally asked to evaluate item unidimensionality, that is, whether each item clearly reflected only one dimension or could plausibly be assigned to multiple dimensions. This qualitative feedback was incorporated into subsequent item revisions. Despite substantial inter-rater reliability (Fleiss' $\kappa > 0.85$; Landis and Koch, 1977), several items required revision due to ambiguous or overlapping content.

Subsequent internal discussions focused on items that potentially reflected more than one dimension. For instance, several items in the *materials* and *methods* categories included implicit references to scientific instruments, also associated with the *location* dimension. To minimize anticipated side and cross-loadings, wording was refined to better represent the intended constructs. The conceptual distinction between *content* and *innovation* was also revisited, given the difficulty students may have in distinguishing novel scientific topics from those simply new to them.

Unlike the original FEWAW questionnaire, which used a response scale based on the perceived correctness of statements, the present version was designed to assess students' subjective perceptions, aligning with the interpretative nature of perceived authenticity. Most items were formulated in the first-person perspective and used active constructions to encourage participants to draw on personal experiences rather than evaluate externally verifiable facts. A five-point Likert scale was implemented, ranging from "strongly disagree (1)" to "strongly agree (5)."

2.1.3 External validation criteria

To gather initial evidence for construct validity, selected external criteria theoretically related to specific aspects of perceived authenticity were included. Most validation scales were adapted from the original FEWAW study by Finger et al. (2022) with minor linguistic adjustments for stylistic consistency. These included measures of general authenticity (Damerau, 2012), authenticity of tasks (Nachtigall et al., 2018), perceived instructor expertise (Huber et al., 2009) and epistemological beliefs regarding the origin (Kremer, 2010) and changeability of knowledge (Moschner and Gruber, 2017).

Two additional short scales were included to reflect the extended model structure. To further validate the *methods* scale, an adapted version of a measure on participation in scientific practices based on the PISA framework and revised according to Chi et al. (2018) was used. It assesses engagement in core elements of scientific inquiry such as planning, conducting, and reflecting on experiments. For the *materials* scale, we included a scale adapted from Schüttler et al. (2021) capturing the perceived authenticity of laboratory equipment and scientific tools used during the program. A summary of all validation scales ($N = 39$ items) and descriptive statistics is presented in Table 1.

2.2 Sample and setting

A total of $N = 146$ secondary school students participated in this study. The sample included 55 students (37.7%) from 10th grade and 68 (46.6%) from 11th grade. In terms of gender, 79 participants (54.1%) identified as male, 65 (44.5%) as female and 2 (1.4%) as diverse. The average age was 16.42 years (SD = 1.36).

TABLE 1 Overview of external validation scales including the number of items, means (*M*), standard deviations (SD), and internal consistencies (α).

Source	Scale	Items	*M*	SD	α
Damerau (2012)	General authenticity	3	3.37	0.97	0.78
Nachtigall et al. (2018)	Task and material authenticity	9	3.03	0.69	0.78
Huber et al. (2009)	Instructor expertise	5	4.15	0.86	0.89
Kremer (2010)	Epistemological beliefs: origin	5	2.56	0.94	0.85
Moschner and Gruber (2017)	Epistemological beliefs: changeability	7	3.56	0.68	0.74
Chi et al. (2018)	Participation in scientific practices	6	3.40	0.74	0.70
Schüttler et al. (2021)	Perceived authenticity of lab equipment	4	2.98	0.96	0.83

Data collection took place between June and July 2025 at the DLR_School_Lab, an out-of-school science lab in Germany. The full-day extracurricular program engaged school classes in activities related to aerospace research, including flight simulation, infrared technology, robotics programming and satellite data analysis. The program followed a non-linear, station-based structure and covered a broad spectrum of topics under the overarching theme of "aeronautics and space."

To ensure standardized administration, instructors received written guidelines outlining the procedure. Students were informed that the questionnaire was part of a scientific study investigating their impressions of authentic learning environments in the out-of-school lab. Participation was voluntary and students were assured that their responses would be anonymized and used exclusively for research and quality-improvement purposes. As part of the general admission process for the out-of-school program, written parental consent for participation in research-related evaluation activities had already been obtained. In addition, students received an information sheet explaining the purpose of the study. The study was conducted in accordance with the ethical standards approved by the Ethics Committee of Otto-von-Guericke University Magdeburg.

The questionnaire was administered in paper-and-pencil format and completed within approximately 30 minutes, directly after the final session of the program.

2.3 Data analysis

2.3.1 Data preprocessing

All statistical analyses were conducted in R (RStudio version 2025.05.0, Build 496) using psychometric packages including `lavaan` (Rosseel, 2012). As the data were collected using a pencil-and-paper format, responses were digitized independently by two researchers. To assess data entry accuracy, each researcher independently re-digitized a subset of 15 questionnaires originally entered by the other, resulting in a total of 30 double-coded questionnaires. This procedure yielded excellent inter-rater agreement (Cohen's $\kappa = 0.998$). Questionnaires with extensive missing data or implausible response patterns—such as uniformly selecting the same option across scales or other typical indicators of careless responding (Meade and Craig, 2012)—were excluded, resulting in a cleaned dataset of $N = 130$ valid cases. The full R script used for all data preprocessing steps and statistical analyses is provided as Supplementary material to this article.

Overall, 1.67% of item responses were missing and considered missing at random. Prior to model estimation, univariate and multivariate distributions were examined. Most items exhibited acceptable skewness and kurtosis, yet Mardia's test indicated significant multivariate skewness ($b_{1p} = 851.72$, $p < 0.001$) but no excess kurtosis ($b_{2p} = 1838.57$, $p = 0.44$), suggesting moderate deviations from normality (Finney and DiStefano, 2006). Therefore, all confirmatory factor analyses (CFA) were estimated using full information maximum likelihood (FIML) with a robust

maximum likelihood estimator (MLR), which accounts for both non-normality and missing data (Little and Rubin, 2002).

2.3.2 Confirmatory factor analysis and model optimization

The initial CFA model included all items from the six theoretically defined dimensions. Model evaluation was based on standard fit indices: Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA, 90% CI), Standardized Root Mean Square Residual (SRMR) and the chi-square to degrees of freedom ratio (χ^2/df). Good model fit is typically indicated by CFI/TLI $>$ 0.95, SRMR $<$ 0.08, RMSEA $<$ 0.06 and $\chi^2/df \leq 2$ (Hu and Bentler, 1999; Kline, 2016; Schermelleh-Engel et al., 2003).

To optimize model fit, an iterative item reduction procedure was applied, consistent with recommended and common practice in CFA. The three established dimensions—*location*, *instructor*, and *innovation*—were locked to preserve comparability with the original FEWAW instrument (Finger et al., 2022). Items of the remaining dimensions (*methods*, *materials*, *content*) were evaluated in multiple steps using established CFA thresholds (Brown, 2015; Kline, 2016): (1) strong cross-loadings as indicated by high modification indices (MI $>$ 10); (2) low standardized loadings (λ $<$ 0.40); and (3) low explained variance (R^2 $<$ 0.30). After each modification step, the model was re-estimated and fit indices recalculated until no further exclusions were warranted.

The resulting model was then conceptually inspected for content redundancy, theoretical representativeness and comparable scale lengths. Additional items were removed where appropriate to further improve model fit while maintaining adequate construct coverage.

2.3.3 Reliability

Once satisfactory fit indices were achieved, internal consistency was evaluated using McDonald's ω as a measure of scale reliability under heterogeneous factor loadings. In addition, Cronbach's α was reported to ensure comparability with the original FEWAW instrument (Finger et al., 2022), where scale reliability was evaluated using this coefficient. Following common conventions, values of $\omega \geq 0.70$ and $\alpha \geq 0.70$ were considered indicative of acceptable reliability (Bortz and Döring, 2015; McDonald, 1999; Cheung et al., 2024).

2.3.4 Discriminant and convergent validity

Furthermore, discriminant and convergent validity were also assessed. Discriminant validity was examined by inspecting the latent inter-factor correlations among the FEWAW dimensions, with all correlations remaining below the recommended threshold of $r = 0.80$, indicating adequate separation of the constructs. Additionally, discriminant validity was assessed using the Heterotrait–Monotrait (HTMT) ratio of correlations, defined as the ratio of the average correlations between items of different constructs to the average correlations between items within the same construct. HTMT therefore reflects the extent to which two scales overlap in what they measure, with values at or below 0.85 indicating that the dimensions are empirically separable (Henseler et al., 2015; Cheung et al., 2024). Convergent validity was evaluated by correlating latent factor scores with external validation scales to confirm theoretically expected relationships (Brown, 2015). In addition, the Average Variance Extracted (AVE) was examined for each dimension, with values of at least 0.50 indicating that a latent construct explains more variance in its indicators than is due to measurement error (Fornell and Larcker, 1981).

2.3.5 Measurement invariance

Finally, to test group equivalence, measurement invariance was examined for gender and grade level using a stepwise comparison of configural, metric, and scalar models. Following Chen (2007), changes of ΔCFI $\leq$ 0.01 and ΔRMSEA $\leq$ 0.015 were used as invariance criteria.

3 Results

In the subsequent sections, we present the results of the confirmatory factor analysis addressing RQ1, followed by the evaluation of the psychometric properties addressing RQ2 and its sub-questions (RQ2.1–RQ2.4).

3.1 Confirmatory factor analysis (RQ1)

The model evaluation proceeded in three distinct stages of item refinement, each aimed at improving factorial validity while maintaining theoretical coverage, following the procedure described in Section 2.3.2. Table 2 summarizes the evolution of model fit statistics across these steps.

(1) Full item set

The initial CFA comprised all 42 items from the six theoretically derived dimensions of perceived authenticity. The model demonstrated acceptable overall fit, $\chi^2(804) = 1,200.54$, CFI = 0.84, TLI = 0.83, RMSEA = 0.062, SRMR = 0.079, with most items showing clear primary loadings and only small cross-loadings. However, the questionnaire length proved impractical for educational settings such as school-based laboratory visits. In these contexts, time and cognitive load are limited and the administration of 42 items, alongside other possible instruments, was not feasible. This test-ecological consideration prompted the need for item reduction, which was carried out not due to model misfit but to increase the instrument's practical usability.

(2) Statistical trimming

Based on the aforementioned trimming criteria, a first reduction was conducted following predefined statistical thresholds. This step resulted in the removal of 10 items. The revised 32-item model showed considerably improved fit: $\chi^2(449) = 616.83$, CFI = 0.91, TLI = 0.90, RMSEA = 0.054, SRMR = 0.073. All six dimensions remained covered and scale structures were preserved.

(3) Conceptual trimming

A subsequent manual trimming phase was conducted to enhance the conceptual clarity of the instrument. Even though the retained items were statistically reasonable, further inspection revealed cases of overlapping phrasing, marginal R^2 values (just above 0.30) or

TABLE 2 Model fit statistics for different item sets across CFA iterations.

Model version	No. of items	χ^2	df	CFI	TLI	RMSEA	SRMR
Full item set (1)	42	1,200.54	804	0.837	0.825	0.062	0.079
After statistical trimming (2)	32	616.83	449	0.907	0.898	0.054	0.073
After conceptual trimming (3)	21	214.81	174	0.962	0.954	0.042	0.070

overly narrow content. This final model achieved an excellent fit: $\chi^2(174) = 214.81$, CFI = 0.96, TLI = 0.95, RMSEA = 0.042, SRMR = 0.070. Overall, the final model explained $R^2 = 0.55$ of item variance.

The final confirmatory factor model comprised six correlated latent variables representing the theorized dimensions of perceived authenticity: *content*, *innovation*, *materials*, *methods*, *location*, and *instructor*. Each factor in the final model was measured by three to five items that had been retained through sequential statistical and conceptual trimming procedures. Table 3 displays the item codes, wording, descriptive statistics, and standardized factor loadings (λ). All indicators were specified as continuous variables and the model was estimated using MLR with FIML for missing data.

The retained items loaded clearly and substantially on their intended factors, with standardized factor loadings ranging from λ = 0.42 to 0.90. In five of the six dimensions, all items exceeded λ > 0.60. An exception to this pattern emerged within the *innovation* scale, where two of the three items (AUIN1: λ = 0.53; AUIN2: λ = 0.42) exhibited comparatively lower loadings.

Furthermore, no cross-loadings were estimated in the final model to preserve discriminant clarity. However, the modification indices indicated a small number of cross-factor relations that would have emerged if cross-loadings had been permitted in the model estimation. As shown in Table 3, despite a clear primary loading of AUL2 on *location* (AUL), modification indices indicate potential cross-loadings on *materials* (MI = 20.41, λ = 0.38), *methods* (MI = 17.96, λ = 0.33), and *content* (MI = 11.05, λ = 0.31).

3.2 Psychometric properties (RQ2)

In line with RQ2 and its sub-questions (RQ2.1–RQ2.4), we examined four central aspects of the psychometric quality of the revised instrument: Internal consistency, discriminant validity, convergent validity, and measurement invariance. The corresponding results are presented in the following subsections.

3.2.1 Reliability and validity (RQ2.1–RQ2.3)

3.2.1.1 Internal consistency

To assess the reliability of the FEWAW dimensions, internal consistencies were examined using McDonald's ω and Cronbach's α (see Table 4). McDonald's ω and Cronbach's α coefficients ranged from ω = 0.64 (*innovation*) to ω = 0.87 (*materials*) and from α = 0.63 (*innovation*) to α = 0.87 (*materials*), respectively. With both ω and α values above 0.75 for five of the six FEWAW dimensions, internal consistency was generally acceptable (see Section 2.3.3). Only the *innovation* dimension fell below commonly accepted thresholds, pointing to limited reliability of this scale (Bortz and Döring, 2015; McDonald, 1999; Cheung et al., 2024).

3.2.1.2 Discriminant validity

To examine the distinctiveness of the FEWAW dimensions, we first inspected latent inter-factor correlations (see Section 2.3.4). Following the effect size conventions proposed by Cohen (1988), correlations of $|r| \approx 0.10$ are interpreted as small, $|r| \approx 0.30$ as medium and $|r| \approx 0.50$ as large effects. In line with common CFA practices, inter-factor correlations below $r = 0.80$ are additionally considered indicative of discriminant validity (Brown, 2015). Inter-dimension correlations were largely consistent with theoretical expectations. Dimensions tied to the realism of scientific work—*content*, *materials*, and *methods*—were moderately to strongly related (e.g., $r = 0.67$ for *content–materials*; $r = 0.68$ for *materials–methods*). Particularly high correlations between *materials* and *methods* suggest that students perceived the way experiments were conducted and the materials used as closely intertwined, while mean ratings indicate that *materials* were overall perceived as more authentic than *methods*. The *innovation* dimension showed the weakest inter-dimension correlations, with associations ranging from $r = 0.02$ to 0.40 and the lowest overall mean ($M = 2.02$), whereas the highest mean was found for the *location* dimension ($M = 4.14$). As a complementary criterion, discriminant validity was further examined using the Heterotrait-Monotrait (HTMT) ratio of correlations (see Section 2.3.4). Most HTMT values remained below the recommended threshold of 0.85, indicating sufficient separation between the FEWAW dimensions (see Table 5). The highest ratios were observed between *materials* and *methods* (HTMT = 0.83) as well as between *materials* and *content* (HTMT = 0.79).

3.2.1.3 Convergent validity

Convergent validity was evaluated using two complementary approaches (see Section 2.3.4). First, we correlated latent factor scores with theoretically related external validation scales (see Table 6) to assess whether the FEWAW dimensions align with constructs that are conceptually linked to authenticity. The dimensions *content*, *materials*, and *methods* showed the strongest and most consistent associations across general authenticity ($r = 0.51 - 0.57^{***}$), task and material authenticity ($r = 0.47 - 0.64^{***}$), participation in scientific practices ($r = 0.39 - 0.42^{***}$) and perceived authenticity of lab equipment ($r = 0.46 - 0.67^{***}$). The *location* and *instructor* scales showed medium but consistent associations, whereas *innovation* displayed a distinct profile, including a selective association with epistemological beliefs about the *origin of knowledge* ($r = 0.32^{**}$). Second, we examined the Average Variance Extracted (AVE) for each dimension, where values at or above 0.50 indicate that a latent construct explains

TABLE 3 Final CFA model: scales, item codes, item wordings, descriptive statistics (M, SD), and standardized factor loadings (λ).

Scale	Code	Item wording	*M*	SD	λ_C	λ_{IN}	λ_{MA}	λ_{ME}	λ_L	λ_I
Content	AUC1	The content I encountered was based on real scientific research.	3.42	1.13	0.729	–	–	–	–	–
	AUC2	The content reflected what researchers actually work on.	3.59	1.04	0.764	–	–	–	–	–
	AUC6	The content seemed closely related to real scientific questions.	3.12	1.07	0.706	–	–	–	–	–
	AUC9	I had the impression that the topics were genuine research topics.	3.29	1.14	0.710	–	–	–	–	–
Innovation	AUIN1	I tried to discover something new for science.	2.17	1.24	–	0.525	–	–	–	–
	AUIN2	I helped science answer an important question.	1.70	1.14	–	0.418	–	–	–	–
	AUIN3	I contributed to answering a current research question.	2.18	1.24	–	0.824	–	–	–	–
Materials	AUMA2	I worked with instruments similar to those used in research.	3.05	1.06	–	–	0.765	–	–	–
	AUMA5	I used materials that are also used in real research.	3.11	1.17	–	–	0.818	–	–	–
	AUMA7	The instruments felt like genuine research devices in their use.	3.12	1.19	–	–	0.707	–	–	–
	AUMA9	I used materials typical of scientific research.	3.11	1.12	–	–	0.757	–	–	–
	AUMA10	The measuring instruments felt largely unmodified for teaching.	2.98	1.23	–	–	0.711	–	–	–
Methods	AUME5	I proceeded in the same way as one would in science.	2.55	1.05	–	–	–	0.763	–	–
	AUME8	I proceeded as I imagine researchers would.	2.61	1.02	–	–	–	0.654	–	–
	AUME9	The way I worked reminded me of scientific research.	2.97	1.10	–	–	–	0.772	–	–
Location	AUL1	I was in a place where real research is conducted.	4.19	1.12	–	–	–	–	0.867	–
	AUL2	I worked in a place where research takes place.	3.87	1.28	0.305[a]	–	0.378[a]	0.331[a]	0.697	–
	AUL3	I was in an environment where research is done.	4.35	1.06	–	–	–	–	0.824	–
Instructor	AUI1	I was supported by a real researcher today.	3.02	1.43	–	–	–	–	–	0.903
	AUI2	A scientist personally introduced me to research today.	2.88	1.31	–	–	–	–	–	0.701
	AUI3	The person leading the project was a scientist.	3.16	1.38	–	–	–	–	–	0.809

[a]Despite no cross-loadings being estimated in the final CFA model, modification indices for item AUL2 indicate the hypothetical magnitude of cross-loadings that would improve global model fit if cross-loadings were permitted.

TABLE 4 Internal consistencies (ω, α), means (M), standard deviations (SD), and intercorrelations of the FEWAW dimensions (lower triangle only).

Dimension	ω	α	*M*	SD	Content	Innovation	Materials	Methods	Location	Instructor
Content	0.82	0.82	3.36	0.88	–	–	–	–	–	–
Innovation	0.64	0.63	2.02	0.91	0.19*	–	–	–	–	–
Materials	0.87	0.87	3.94	0.93	0.67***	0.30***	–	–	–	–
Methods	0.78	0.78	2.71	0.87	0.52***	0.40***	0.68***	–	–	–
Location	0.83	0.83	4.14	1.00	0.45***	0.02	0.41***	0.21*	–	–
Instructor	0.85	0.84	3.03	1.21	0.32***	0.22*	0.44***	0.29**	0.31***	–

All *p*-values were adjusted using Holm's method (Holm, 1979) and significance is denoted by $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 5 HTMT ratios for the FEWAW dimensions (lower triangle only).

Dimension	Content	Innovation	Materials	Methods	Location	Instructor
Content	–	–	–	–	–	–
Innovation	0.28	–	–	–	–	–
Materials	0.79	0.42	–	–	–	–
Methods	0.64	0.58	0.83	–	–	–
Location	0.55	0.11	0.49	0.28	–	–
Instructor	0.37	0.34	0.51	0.36	0.37	–

Values below the recommended threshold of 0.85 indicate sufficient discriminant validity (Henseler et al., 2015).

TABLE 6 Correlations between FEWAW dimensions and external validation scales.

External scale	Content	Innovation	Materials	Methods	Location	Instructor
General authenticity	**0.56*****	0.27	**0.57*****	**0.51*****	**0.44*****	**0.45*****
Task and material authenticity	**0.57*****	**0.32***	**0.64*****	**0.47*****	**0.32****	**0.33****
Instructor expertise	**0.34****	−0.06	**0.35****	0.19	**0.44*****	0.15
Epistemological beliefs: origin	−0.09	**0.32****	0.08	0.06	−0.08	0.22
Epistemological beliefs: changeability	**0.36****	0.11	**0.30***	0.22	0.25	0.09
Participation in scientific practices	**0.39*****	0.13	**0.42*****	**0.42*****	**0.35****	0.18
Perceived authenticity of lab equipment	**0.54*****	0.16	**0.67*****	**0.46*****	**0.34****	**0.44*****

All p-values were adjusted using Holm's method (Holm, 1979) and significance is denoted by $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Values in bold indicate at least moderate and significant correlations with $r > 0.30$ and $^{*}p < 0.05$.

more variance in its indicators than is due to measurement error (Fornell and Larcker, 1981). The results for each scale were $AVE_{AUC} = 0.528$, $AVE_{AUMA} = 0.564$, $AVE_{AUME} = 0.539$, $AVE_{AUL} = 0.625$, $AVE_{AUI} = 0.658$, while $AVE_{AUIN} = 0.386$ fell slightly below, indicating weaker convergence within this factor.

3.2.2 Measurement invariance (RQ2.4)

To evaluate the robustness of the final six-factor structure, measurement invariance was tested across gender and grade level (see Section 2.3.5). A sequence of multiple-group CFA models was estimated using robust MLR estimation, comparing configural, metric, and scalar models following the recommendations by Chen (2007). A *post hoc* power analysis was conducted using G*Power (Version 3.1.9.7) to assess the statistical sensitivity of the test (Faul et al., 2009). Based on the sample size of $N = 130$ and assuming a medium effect size ($w = 0.30$), the analysis indicated sufficient statistical power ($1 - \beta = 0.87$) to detect deviations from measurement equivalence in two groups. Model comparisons were based on changes in CFI and RMSEA, complemented by the Satorra–Bentler scaled χ^2 difference test (Satorra and Bentler, 2001). The results are summarized in Table 7.

Across both gender and grade level, changes in fit indices were negligible ($\Delta CFI \leq 0.01$, $\Delta RMSEA \leq 0.015$) and all Satorra–Bentler χ^2 difference tests were non-significant ($p > 0.20$). These results indicate that the six-factor model is scalar invariant across groups.

Building on these results, we conducted group comparisons of observed scale means using non-parametric Mann–Whitney U-tests (Bortz and Döring, 2015). Across all six dimensions, no statistically significant differences were found between male and female students. Holm-adjusted p-values ranged from $p_{Holm} = 0.194$ to $p_{Holm} = 1.000$, with effect sizes ranging from $r = -0.102$ to $r = 0.114$, indicating negligible group differences (Holm, 1979). Comparisons between students of different grades yielded a similar pattern. Although descriptive medians suggested slightly higher authenticity ratings for lower grade students in the *methods*, *materials*, and *location* dimensions, none of these differences reached statistical significance after Holm correction ($p_{Holm} = 0.450$–1.000). The corresponding effect sizes were consistently small, with values ranging from $r = -0.130$ to $r = 0.171$. These findings suggest that, beyond measurement equivalence, perceptions of authenticity were consistent across gender and grade level in this sample.

4 Discussion

The present study aimed to validate an extended version of the FEWAW questionnaire for measuring students' perceived authenticity in educational settings. The revised instrument includes six scales: *content*, *innovation*, *materials*, *methods*, *location*, and *instructor*. The following section discusses the findings in the context of the research questions (RQ1–RQ2.4), focusing on the factorial structure, internal consistency, discriminant, and convergent validity and measurement invariance.

4.1 Factorial structure (RQ1)

Confirmatory factor analysis supported the proposed six-factor structure with acceptable-to-excellent global fit and robust loadings (see Tables 2, 3). This adds to the literature that has long argued for a multi-faceted, relational construct of authenticity in science education (Betz, 2018; Nachtigall et al., 2022). In contrast to earlier validation results in which *materials* and *methods* merged into a single factor (Finger et al., 2022), the present results empirically distinguish these subdimensions. Theoretically, this separation echoes the distinction between the *artefactual* basis of science (instruments, samples, devices) and the *procedural* orchestration of scientific inquiry (planning, measuring, analyzing)—a difference that is central to disciplinary authenticity (Watkins et al., 2012; Lee and Butler, 2003). Disentangling these aspects enables instructional designers and researchers to reason about how specific design elements (e.g., access to professional tools vs. engagement in scientific practices) affect learners' authenticity perceptions. The improved separation is likely due to iterative item clarification and tighter semantic boundaries between dimensions. By reducing wording overlaps (e.g., items that implicitly referenced both tools and procedures) and prioritizing first-person, experience-near formulations, we possibly minimized cross-loadings that otherwise threaten discriminant validity.

TABLE 7 Model fit statistics for measurement invariance of the six-factor model across gender and grade level.

Model	χ^2	df	CFI	TLI	RMSEA	SRMR	$\Delta\chi^2$	Δdf	p
Measurement invariance (gender)									
Configural invariance	512.00	348	0.855	0.825	0.094	0.095	–	–	–
Metric invariance	528.38	363	0.854	0.831	0.093	0.105	16.38	15	0.358
Scalar invariance	548.00	378	0.850	0.834	0.092	0.107	19.21	15	0.204
Measurement invariance (grade)									
Configural invariance	539.00	348	0.814	0.775	0.109	0.100	–	–	–
Metric invariance	552.00	363	0.816	0.787	0.106	0.109	14.34	15	0.500
Scalar invariance	560.00	378	0.823	0.803	0.102	0.110	8.68	15	0.894

4.2 Psychometric properties (RQ2)

Overall, the results provide promising evidence that the extended FEWAW instrument demonstrates satisfactory psychometric quality across central aspects of scale validation. The majority of scales showed solid internal consistencies and most dimensions were empirically distinguishable despite theoretically expected associations among some constructs (e.g., *materials–methods*, *content–materials*). The patterns of correlations with external validation measures further supported the convergent validity of the dimensions, particularly for *materials*, *content*, and *methods*, which emerged as the most robust and theoretically coherent indicators of perceived authenticity in our sample. At the same time, the weaker performance of the *innovation* scale points to conceptual and psychometric limitations that warrant reconsideration in future revisions of the instrument. With respect to measurement invariance, the results tentatively suggest that the instrument functions similarly across gender and grade level. However, given the limited sample size, these findings should be interpreted with caution. Taken together, the findings indicate that the revised FEWAW represents a psychometrically robust instrument, while also recognizing specific areas where further research is needed.

In the following, we discuss the results for each aspect of RQ2 in more detail, structured according to the four sub-questions on internal consistency (RQ2.1), discriminant validity (RQ2.2), convergent validity (RQ2.3), and measurement invariance (RQ2.4).

4.2.1 Internal consistency (RQ2.1)

The final 21-item version achieved strong psychometric properties while maintaining broad construct coverage. The trimming strategy combined (a) statistical thresholds (factor loadings, item-level explained variance, and modification indices) with (b) conceptual trimming to remove redundancy and harmonize scale length. This two-step approach is consistent with recommendations in the measurement literature (Brown, 2015; Kline, 2016; Bortz and Döring, 2015) and meets demands in time-constrained environments such as out-of-school labs, where additional measurement instruments are often administered.

Internal consistencies of the six scales ranged from $\omega = 0.64$ to 0.87 (see Table 4). Five of the six dimensions exceeded the commonly recommended threshold of $\omega \geq 0.70$, indicating solid reliability even with the reduced number of items. Cronbach's α values showed a comparable pattern, ranging from $\alpha = 0.63$ to 0.87, with five of the six dimensions exceeding the commonly recommended threshold of $\alpha \geq 0.70$. The highest reliability was observed for *materials* ($\omega = 0.87$, $\alpha = 0.87$), followed by *instructor*, *location*, and *content* (all $\omega \geq 0.82$, $\alpha \geq 0.82$). In comparison to Finger et al. (2022), who reported consistently high reliability values across all scales ($\alpha \geq 0.81$), the present results are largely comparable—with the exception of the *innovation* scale, which showed a noticeably lower internal consistency ($\omega = 0.64$, $\alpha = 0.63$).

This result may be partially explained by the fact that the *innovation* scale, along with *location* and *instructor*, was retained without revision. Since these scales were locked for comparability, they could not be optimized or improved in the same way as the newly extended dimensions. We return to the consequences of this decision, especially with regard to the *innovation* dimension, in Section 4.2.3.

4.2.2 Discriminant validity (RQ2.2)

Inter-factor correlations for all factors remained below the conventional threshold of $r = 0.80$, generally indicating that the six dimensions capture distinct aspects of perceived authenticity. However, the magnitude of the intercorrelations between several dimensions, notably *content* and *materials* ($r = 0.67$) and *materials* and *methods* ($r = 0.68$), should be reviewed critically. High inter-factor correlations may raise questions regarding the discriminant validity of these dimensions and signal the potential existence of an overlying, higher-order factor. The association between *materials* and *methods*, for example, is theoretically coherent as disciplinary authenticity involves both access to the *tools* and engagement with the *practices* of the field. This overlap suggests that learners can and do differentiate between *what* they used and *how* they used it, yet the resulting correlations indicate that these facets are perceived as closely connected. Similarly, the strong link between *content* and *materials* may reflect the fact that authentic scientific content is typically embedded within the artifacts through which it is explored.

These patterns are reflected in the HTMT ratio of correlations, which provides a stricter indication of conceptual distinctiveness.

Most HTMT values remained below the recommended threshold of 0.85 (Henseler et al., 2015; Cheung et al., 2024), supporting the separability of the majority of dimensions. However, elevated values were observed for the pairs *content–materials* (HTMT = 0.79) and *materials–methods* (HTMT = 0.83), indicating particularly close empirical associations. This suggests that while the dimensions are conceptually defined as distinct, they may not be fully separated in learners' perceptions. One plausible explanation is the inherent structure of the learning activities in our sample. The tasks were designed such that the materials followed directly from the content and the methods naturally emerged from the available materials. For instance, activities on robotics (*content*) involved working with the robot itself (*materials*) and required the students to program (*methods*). This built-in sequencing may have made it difficult for learners to perceive these dimensions as clearly distinct from one another.

Future research should therefore investigate a hierarchical factor model to determine whether these dimensions are best represented as distinct yet strongly associated facets or as expressions of a broader, underlying construct. Despite these challenges, the intercorrelations and HTMT values remained below conventional thresholds, providing evidence for adequate empirical separability. The refined wording and the clear conceptual distinction between dimensions likely contributed to the overall acceptable statistical separation.

4.2.3 Convergent validity (RQ2.3)

The analyses of the correlations between FEWAW scales and external validation measures support the convergent validity of most dimensions. In particular, the scales *materials*, *content*, and *methods* showed the strongest associations with constructs that are theoretically aligned with key components of perceived authenticity such as general and task-specific authenticity, participation in scientific practices, and perceptions of lab equipment (e.g. Damerau, 2012; Nachtigall et al., 2018; Schüttler et al., 2021; Chi et al., 2018). These findings suggest that learners' authenticity judgments are closely linked to material and procedural features of the learning environment, which appear to serve as salient cues for authenticity perception (Schüttler et al., 2021). This interpretation aligns with earlier research indicating that authenticity is often anchored in tangible characteristics of science education settings (Braund and Reiss, 2006; Roth, 1995).

Among these, *materials* and *content* stood out as the dimensions with the most pronounced and consistently high correlations across multiple external measures, suggesting that these two dimensions play a particularly central role in shaping students' authenticity perceptions. From a theoretical perspective, both dimensions reflect foundational elements of scientific practice, namely the physical artifacts and the substantive focus of scientific inquiry.

The *location* and *instructor* scales were also positively associated with several external constructs, though their relationships were less pronounced. Both dimensions correlated moderately with general authenticity and perceived authenticity of lab equipment, while *location* additionally related to participation in scientific practices. These results point to a secondary role of spatial and interpersonal aspects in shaping perceived authenticity, possibly reflecting that these features are more context-sensitive and subject to variation in learners' interpretations.

In contrast, the *innovation* scale displayed lower correlations with most external validation measures. Although it was modestly associated with task-related authenticity and epistemological beliefs about the origin of knowledge, it showed no meaningful correlation with other constructs we measured. This pattern is consistent with the relatively low $AVE_{AUIN} = 0.386$, indicating weaker convergence of its items and providing an internal validity-based explanation for the selective and generally weaker external associations. Importantly, the weaker empirical performance reflects a deeper conceptual issue that arises when operationalizing *innovation* in formal learning contexts. While the present data does not allow for definitive conclusions about the underlying causes, a few considerations may help interpret the observed pattern.

Authenticity in formal learning settings typically involves a simulation of scientific practice rather than direct engagement in authentic research (Betz et al., 2016). Students' perceptions of authenticity may therefore be bounded by their role expectations in school-based learning, as many see themselves as exploring existing ideas rather than generating new knowledge. Unlike the other dimensions, which are anchored in tangible cues such as physical materials, tools, spatial settings, or even instructors, *innovation* relies on learners' subjective assessment of epistemic novelty. Such assessments draw on beliefs about who is entitled to create new scientific knowledge, making the construct inherently more complex than dimensions triggered directly by environmental features. This asymmetry is also reflected in the current item formulations, which adopt an "objective" perspective on scientific discovery (e.g., "I tried to discover something new for science," "I contributed to answering a current research question"). These formulations implicitly assume a professional scientific role and therefore may exceed what students consider realistic within a classroom or an outreach setting. This interpretation aligns with prior research emphasizing that authenticity or student perceptions on epistemological beliefs depend not only on environmental features but also on learners' epistemic role expectations (Sandoval, 2005; Stroupe, 2014; Nachtigall et al., 2024). In this view, *innovation* can be considered an outcome of how learners position themselves within these contexts.

The selective correlation with epistemological beliefs about the *origin of knowledge* further supports this explanation. Learners who have a broader view of who can generate scientific knowledge appear more likely to perceive their own activities as innovative. Conversely, students who believe that only professional scientists produce new knowledge may experience a mismatch between item wording and their perceived role, resulting in overall lower scores. This pattern is consistent with research showing that many students differentiate sharply between school science and professional science and attribute knowledge generation primarily to scientists rather than learners (Voitle et al., 2022). The absence of a corresponding relationship with beliefs about the *changeability of knowledge* suggests the importance of learners' assumptions about who is allowed to generate new ideas. Taken together, these findings indicate that the subjective experience of *innovation* is constrained by role-based plausibility boundaries. Students may feel they can explore or test ideas, but not "discover something new for science."

From a psychometric perspective, this conceptual misalignment helps explain the low AVE for the *innovation* scale and the lack of meaningful correlations with most external validation measures. It also clarifies why this dimension has shown weak performance in Finger et al. (2022). To improve both theoretical coherence and empirical functioning, future iterations of the scale may benefit from shifting toward a more learner-centered conceptualization of epistemic novelty, emphasizing personal discovery ("I explored something new to me") and idea generation ("I was able to generate and test my own ideas") rather than contributions to professional scientific knowledge. Such formulations would align more closely with authenticity as a subjective experience and be more consistent with the role expectations typically held by students in structured learning contexts.

Furthermore, the unexpectedly weak association between *instructor* and external ratings of instructor expertise should be noted. A plausible explanation lies in the specific wording of the items. The FEWAW scale refers to instructors as "real scientists," invoking authenticity markers rooted in scientific identity and professional status. By contrast, the external validation scale refers to "our tutors" and emphasizes perceived competence rather than professional affiliation. Particularly in outreach contexts such as student labs, where instruction is often provided by university students or early-career staff, learners may recognize expertise without perceiving the instructors as authentic representatives of science.

4.2.4 Measurement invariance (RQ2.4)

The analysis provided initial evidence for scalar invariance across gender and grade level, suggesting that the six-factor structure functions equivalently across these groups. This indicates that the instrument allows for unbiased mean comparisons between male and female students as well as across grade levels.

Nevertheless, these findings must be viewed as preliminary. Given the small sample size ($N = 130$), the absence of significant differences does not conclusively confirm full invariance. Minor variations in the pattern of loadings or intercepts might remain undetected in small samples. Furthermore, as often observed in small-to-moderate samples, a slight deterioration of overall model fit occurred during the stepwise invariance testing. This pattern is well-documented as a sample-size-sensitive artifact rather than a substantive refutation of invariance (Chen, 2007). Accordingly, future research should validate the model structure with larger and more diverse datasets encompassing different school types, age groups, and disciplinary contexts. This could help assess the robustness of the FEWAW measurement model and to determine whether the observed factorial stability generalizes across sub-populations.

5 Limitations

While the present study offers important insights into the factorial structure and validity of the revised FEWAW instrument, several limitations must be acknowledged.

First, the generalizability of the results is limited by the specific context in which the data were collected. All participants took part in an extracurricular physics lab at a single institution in Germany. As a result, the findings may not generalize to other subject domains or other science learning contexts. Moreover, the sample consisted exclusively of students from schools in grades 10 and 11, which restricts conclusions about the instrument's applicability to other age groups.

Second, our study relied on a smaller and more context-specific sample. While this reflects the narrower goal of refining and validating selected dimensions rather than developing a new instrument from scratch, the reduced sample size limits generalizability—particularly in confirmatory factor analysis and measurement invariance testing (Brown, 2015).

Third, although the *innovation* scale showed comparatively weaker psychometric performance, no qualitative or process-oriented data were collected to clarify how students interpreted these items. Consequently, the study cannot provide in-depth explanations for the low reliability and limited external associations of this scale. Future research should explore learners' interpretations through cognitive interviews or think-aloud protocols to better understand the source of this mismatch.

Taken together, these limitations underscore the need for further validation studies with more diverse and larger samples, ideally across different school types, age groups and disciplinary contexts.

6 Conclusions and outlook

The present contribution provides a theoretically grounded, multidimensional instrument for assessing students' perceived authenticity in educational settings. The six-factor structure was empirically confirmed, with good support for internal consistency, discriminant validity and scalar measurement invariance across gender and grade. The results highlight the need to view authenticity not as a unitary construct but as dependent on instructional context and learner characteristics (Betz, 2018; Nachtigall et al., 2022; Schriebl et al., 2023).

Future studies should validate the FEWAW structure in larger and more diverse samples (e.g., $N > 500$) to assess the weighting and potential hierarchy among dimensions. This would help prioritize design elements and advance theory-driven instructional development.

Further research should also explore the predictive validity of the FEWAW dimensions. Investigating how perceived authenticity relates to motivational (e.g., situational interest), epistemological (e.g., beliefs about knowledge), and cognitive outcomes (e.g., learning gains) could demonstrate the instrument's practical utility and theoretical significance.

In sum, the revised questionnaire offers a concise and theoretically reliable instrument for assessing perceived authenticity in science education. It addresses prior concerns regarding construct overlap and establishes a foundation for future research and instructional design focused on authenticity.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the Otto von Guericke University Magdeburg (approval no. 98/23). The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

DS: Conceptualization, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing, Methodology. TS: Investigation, Writing – review & editing. BE: Formal analysis, Methodology, Writing – review & editing. BW: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing, Methodology.

Funding

The author(s) declared that financial support was received for this work and/or its publication. We thank the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) for support through the Research Unit FOR 5599 on structured magnetic elastomers, project no. 511114185, via DFG grant reference no. WA 5276/1-1.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2026.1749760/full#supplementary-material

References

Anker-Hansen, J., and Andrée, M. (2019). In pursuit of authenticity in science education. *Nord. Stud. Sci. Educ.* 15, 498–510. doi: 10.5617/nordina.4723

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]. *Unterrichtswiss.* 46, 261–278. doi: 10.1007/s42010-018-0021-0

Betz, A., Flake, S., Mierwald, M., and Vanderbeke, M. (2016). "Modelling authenticity in teaching and learning contexts: a contribution to theory development and empirical investigation of the construct," in *Transforming Learning, Empowering Learners: The International Conference of the Learning Sciences (ICLS) 2016, Vol. 2*, eds. C. K. Looi, J. L. Polman, U. Cress, and P. Reimann (Singapore: International Society of the Learning Sciences), 815–818.

Bolger, M. S., Osness, J. B., Gouvea, J. S., and Cooper, A. C. (2021). Supporting scientific practice through model-based inquiry: a students'-eye view of grappling with data, uncertainty, and community in a laboratory experience. *CBE Life Sci. Educ.* 20:ar59. doi: 10.1187/cbe.21-05-0128

Bortz, J., and Döring, N. (2015). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler [Research Methods and Evaluation: For Researchers in the Human and Social Sciences*]. Heidelberg: Springer.

Braund, M., and Reiss, M. (2006). Towards a more authentic science curriculum: the contribution of out-of-school learning. *Int. J. Sci. Educ.* 28, 1373–1388. doi: 10.1080/09500690500498419

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Press.

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Struct. Equ. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., and Wang, L. C. (2024). Reporting reliability, convergent and discriminant validity with structural equation modeling: a review and best-practice recommendations. *Asia Pac. J. Manag.* 41, 745–783. doi: 10.1007/s10490-023-09871-y

Chi, S., Liu, X., Wang, Z., and Han, S. W. (2018). Moderation of the effects of scientific inquiry activities on low SES students' PISA 2015 science achievement by school teacher support and disciplinary climate in science classroom across gender. *Int. J. Sci. Educ.* 40, 1284–1304. doi: 10.1080/09500693.2018.1476742

Chinn, C. A., and Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: a theoretical framework for evaluating inquiry tasks. *Sci. Educ.* 86, 175–218. doi: 10.1002/sce.10001

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge.

Damerau, K. (2012). *Molekulare und Zell-Biologie im Schülerlabor. Fachliche Optimierung und Evaluation der Wirksamkeit im BeLL Bio (Bergisches Lehr-Lern-Labor Biologie)* [*Molecular and cellular biology in an out-of-school lab: subject-specific optimization and evaluation of effectiveness in the BeLL Bio*] (Ph.D. thesis). Bergische Universität Wuppertal, Wuppertal, Germany.

Engeln, K. (2004). *Schülerlabors. Authentische, aktivierende Lernumgebungen als Möglichkeit, Interesse an Naturwissenschaften und Technik zu wecken [Out-of-School Labs: Authentic, Activating Learning Environments as a Means to Foster Interest in Science and Technology*]. Berlin: Logos Verlag Berlin.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/BRM.41.4.1149

Finger, L., van den Bogaert, V., Fleischer, J., Raimann, J., Sommer, K., Wirth, J., et al. (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 of authentic science education? Development and validation of a questionnaire to assess the perceived authenticity of science education in out-of-school labs]. *Z. Didakt. Naturwiss.* 28:2. doi: 10.1007/s40573-022-00139-4

Finney, S. J., and DiStefano, C. (2006). "Non-normal and categorical data in structural equation modeling," in *Structural Equation Modeling: A Second Course*, eds. G. R. Hancock, and R. O. Mueller (Charlotte, NC: Information Age Publishing), 269–314.

Fornell, C., and Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *J. Mark. Res.* 18, 39–50. doi: 10.1177/002224378101800104

Habig, B., and Gupta, P. (2021). Authentic STEM research, practices of science, and interest development in an informal science education program. *Int. J. STEM Educ.* 8:57. doi: 10.1186/s40594-021-00314-y

Habig, S., Blankenburg, J., van Vorst, H., Fechner, S., Parchmann, I., Sumfleth, E., et al. (2018). Context characteristics and their effects on students' situational interest in chemistry. *Int. J. Sci. Educ.* 40, 1154–1175. doi: 10.1080/09500693.2018.1470349

Hagenkötter, R., Nachtigall, V., Rolka, K., and Rummel, N. (2024). Model authenticity in learning mathematical experimentation: how students perceive and learn from scientist and peer models. *Eur. J. Psychol. Educ.* 39, 3301–3324. doi: 10.1007/s10212-024-00843-4

Henseler, J., Ringle, C. M., and Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. Acad. Mark. Sci.* 43, 115–135. doi: 10.1007/s11747-014-0403-8

Herrington, J., and Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educ. Technol. Res. Dev.* 48, 23–48. doi: 10.1007/BF02319856

Hohrath, S., Aßmann, S., Krabbe, H., and Opfermann, M. (2024). Students' perceived authenticity and understanding of authentic research while experimenting in a non-formal learning setting. *Eur. J. Psychol. Educ.* 39, 3325–3349. doi: 10.1007/s10212-024-00810-z

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.

Honebein, P. C., Duffy, T. M., and Fishman, B. J. (1993). "Constructivism and the design of learning environments: context and authentic activities for learning," in *Designing Environments for Constructive Learning*, eds. T. M. Duffy, J. Lowyck, D. H. Jonassen, and T. M. Welsh (Berlin: Springer), 87–108. doi: 10.1007/978-3-642-78069-1_5

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Huber, F., Vollhardt, K., and Meyer, F. (2009). Helden der Werbung? Eine Untersuchung der Relevanz von Werbefiguren für das Konsumentenverhalten. *Mark. Z. Forsch. Prax.* 31, 183–196. doi: 10.15358/0344-1369-2009-3-183

Itzek-Greulich, H., Flunger, B., Vollmer, C., Nagengast, B., Rehm, M., Trautwein, U., et al. (2015). Effects of a science center outreach lab on school students' achievement—are student lab visits needed when they teach what students can learn at school? *Learn. Instr.* 38, 43–52. doi: 10.1016/j.learninstruc.2015.03.003

Kapon, S., Laherto, A., and Levrini, O. (2018). Disciplinary authenticity and personal relevance in school science. *Sci. Educ.* 102, 1077–1106. doi: 10.1002/sce.21458

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling*. New York, NY: Guilford Press.

Kremer, K. H. (2010). *Die Natur der Naturwissenschaften verstehen: Untersuchungen zur Struktur und Entwicklung von Kompetenzen in der Sekundarstufe I* [*Understanding the nature of science: investigations into the structure and development of competencies in lower secondary education*] (Ph.D. thesis). Universität Kassel, Fachbereich Mathematik und Naturwissenschaften, Kassel, Germany.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Lee, H.-S., and Butler, N. (2003). Making authentic science accessible to students. *Int. J. Sci. Educ.* 25, 923–948. doi: 10.1080/09500690305023

Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York, NY: Wiley. doi: 10.1002/9781119013563

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. London: Psychology Press.

Meade, A. W., and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychol. Methods* 17, 437–455. doi: 10.1037/a0028085

Mierwald, M., Lehmann, T., and Brauch, N. (2018). Zur Veränderung epistemologischer Überzeugungen im Schülerlabor: Authentizität von Lernmaterial als Chance der Entwicklung einer wissenschaftlich angemessenen Überzeugungshaltung im Fach Geschichte? [Changing epistemological beliefs in student labs: authentic learning materials as a chance to foster the development of academically adequate beliefs in the domain of history?]. *Unterrichtswiss.* 46, 279–297. doi: 10.1007/s42010-018-0019-7

Moschner, B., and Gruber, H. (2017). "Erfassung epistemischer Überzeugungen mit dem FEE [Assessing epistemic beliefs using the FEE]," in *Wissen und Lernen*, eds. A. Bernholt, H. Gruber, and B. Moschner (Bedford Heights, OH: Waxmann), 17–38.

Nachtigall, V., and Rummel, N. (2021). Investigating students' perceived authenticity of learning activities in an out-of-school lab for social sciences: a replication study. *Instr. Sci.* 49, 779–810. doi: 10.1007/s11251-021-09556-3

Nachtigall, V., Rummel, N., and Serova, K. (2018). 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]. *Unterrichtswiss.* 46, 299–319. doi: 10.1007/s42010-018-0020-1

Nachtigall, V., Shaffer, D. W., and Rummel, N. (2022). Stirring a secret sauce: a literature review on the conditions and effects of authentic learning. *Educ. Psychol. Rev.* 34, 1479–1516. doi: 10.1007/s10648-022-09676-3

Nachtigall, V., Shaffer, D. W., and Rummel, N. (2024). The authenticity dilemma: towards a theory on the conditions and effects of authentic learning. *Eur. J. Psychol. Educ.* 39, 3483–3509. doi: 10.1007/s10212-024-00892-9

Nicaise, M., Gibney, T., and Crane, M. (2000). Toward an understanding of authentic learning: student perceptions of an authentic classroom. *J. Sci. Educ. Technol.* 9, 79–94. doi: 10.1023/A:1009477008671

Pawek, C. (2009). *Schülerlabore als interessefördernde außerschulische Lernumgebungen für Schülerinnen und Schüler aus der Mittel- und Oberstufe* [*Out-of-school labs as interest-promoting extracurricular learning environments for lower and upper secondary students*] (Ph.D. thesis). Universität Kiel, Kiel, Germany.

Peacock, M. (1997). The effect of authentic materials on the motivation of EFL learners. *ELT J.* 51, 144–156. doi: 10.1093/elt/51.2.144

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Roth, W.-M. (1995). *Authentic School Science: Knowing and Learning in Open-Inquiry Science Laboratories*. Dordrecht: Springer. doi: 10.1007/978-94-011-0495-1

Sandoval, W. A. (2005). Understanding students' practical epistemologies and their influence on learning through inquiry. *Sci. Educ.* 89, 634–656. doi: 10.1002/sce.20065

Satorra, A., and Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* 66, 507–514. doi: 10.1007/BF02296192

Scharfenberg, F.-J., Bogner, F. X., and Klautke, S. (2007). Learning in a gene technology laboratory with educational focus: results of a teaching unit with authentic experiments. *Biochem. Mol. Biol. Educ.* 35, 28–39. doi: 10.1002/bmb.1

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods Psychol. Res. Online* 8, 23–74.

Schriebl, D., Müller, A., and Robin, N. (2023). Modelling authenticity in science education. *Sci. Educ.* 32, 1021–1048. doi: 10.1007/s11191-022-00355-x

Schriebl, D., Müller, A., Robin, N., and Henrich, B. (2021). An instrument to measure students' perception of the authenticity of an out-of-school learning place. *Prog. Sci. Educ.* 4, 66–74. doi: 10.25321/prise.2021.1072

Schüttler, T., Watzka, B., Girwidz, R., and 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]. *Z. Didakt. Naturwiss.* 27, 109–125. doi: 10.1007/s40573-021-00128-z

Schwartz, R. S., Lederman, N. G., and Crawford, B. A. (2004). Developing views of nature of science in an authentic context: an explicit approach to bridging the gap between nature of science and scientific inquiry. *Sci. Educ.* 88, 610–645. doi: 10.1002/sce.10128

Smetana, L. K., and Bell, R. L. (2012). Computer simulations to support science instruction and learning: a critical review of the literature. *Int. J. Sci. Educ.* 34, 1337–1370. doi: 10.1080/09500693.2011.605182

Sommer, K., Firstein, A., and Rothstein, B. (2020). "Authentizität (der Wissenschaftsvermittlung) im Schülerlabor [Authenticity (of science education) in out-of-school labs]," in *Handbuch Forschen im Schülerlabor*, eds. K. Sommer, J. Wirth, and M. Vanderbeke (Münster: Waxmann), 21–30.

Sommer, K., Wirth, J., and Rummel, N. (2018). Authentizität der Wissenschaftsvermittlung im Schülerlabor—Einführung in den Thementeil [Authenticity of instruction about natural and social sciences in out-of-school labs—introduction to the special issue]. *Unterrichtswiss.* 46, 253–260. doi: 10.1007/s42010-018-0022-z

Stroupe, D. (2014). Examining classroom science practice communities: how teachers and students negotiate epistemic agency and learn science-as-practice. *Sci. Educ.* 98, 487–516. doi: 10.1002/sce.21112

Svärd, J., Schönborn, K. J., and Hallström, J. (2024). Students' perceptions of authenticity in an upper secondary technology education innovation project. *Res. Sci. Technol. Educ.* 42, 467–487. doi: 10.1080/02635143.2022.2116418

van Vorst, H., Dorschu, A., Fechner, S., Kauertz, A., Krabbe, H., Sumfleth, E., et al. (2015). Charakterisierung und Strukturierung von Kontexten im naturwissenschaftlichen Unterricht—Vorschlag einer theoretischen Modellierung [A theoretical framework for categorizing context-based tasks in science education]. *Z. Didakt. Naturwiss.* 21, 29–39. doi: 10.1007/s40573-014-0021-5

Voitle, F., Heuckmann, B., Kampa, N., and Kremer, K. (2022). Assessing students' epistemic beliefs related to professional and school science. *Int. J. Sci. Educ.* 44, 1000–1020. doi: 10.1080/09500693.2022.2059821

Watkins, J., Coffey, J. E., Redish, E. F., and Cooke, T. J. (2012). Disciplinary authenticity: enriching the reforms of introductory physics courses for life-science students. *Phys. Rev. Phys. Educ. Res.* 8:010112. doi: 10.1103/PhysRevSTPER.8.010112

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