Science & Education
https://doi.org/10.1007/s11191-026-00747-3

ARTICLE

A New Perspective on the Evaluation of Pupils' Inquiry Skills Using Four-tier Test

Dominik Šmida[1] · **Anna Drozdíková**[1] · **Ráchel Nechajová**[1]

Received: 11 September 2025 / Revised: 16 March 2026 / Accepted: 8 April 2026

Abstract
In biology education, it is essential to systematically develop pupils' inquiry skills, with a particular focus not only on their procedural component but also on understanding their scientific nature. However, conventional tests do not provide a comprehensive view of the pupils' acquired inquiry skills. For this reason, it is necessary to explore new approaches to assess them and provide deeper insights into pupils' thinking when solving various items through scientific methods. One perspective instrument can be a four-tier test, which has become increasingly popular among researchers. Nevertheless, it has more often been applied to assess conceptual understanding of knowledge rather than inquiry skills. However, the results of our research indicate that the four-tier test represents a valid and reliable instrument that provides more objective results in contrast to the one-tier and two-tier multiple-choice tests, which significantly overestimate the pupils' mean score, and in the case of the two-tier test, also overestimate the mean frequency of pupils' misconceptions associated with the application of inquiry skills. In addition, using the four-tier test, we identified frequent misconceptions among pupils associated with skills such as designing experiments, identifying variables, and formulating research questions. These findings can contribute to preventing or eliminating these misconceptions in biology education. Overall, the results show that the four-tier test can provide more detailed information about pupils' difficulties and problems, which can help us find more effective strategies for their development in biology education.

Keywords biology education · diagnostic test · inquiry skills · misconceptions · primary school pupils

✉ Dominik Šmida
smida8@uniba.sk

Anna Drozdíková
anna.drozdikova@uniba.sk

Ráchel Nechajová
nechajova4@uniba.sk

[1] Department of Didactics in Science, Psychology and Pedagogy, Faculty of Natural Sciences, Comenius University Bratislava, Ilkovičova 6, 842 15 Bratislava, Slovak Republic

Published online: 20 May 2026

1 Introduction

Many countries have focused on the systematic implementation of inquiry into school practice as part of their curriculum reforms (Wang et al., 2015) to increase the level of pupils' scientific literacy (Iskandar et al., 2019; Wen et al., 2020; Yuliati et al., 2021). However, if pupils want to be able to conduct inquiry activities, it is essential that they acquire inquiry skills that will enable them to discover the world of science and scientific work (Harrison, 2014; Wang et al., 2015), understand it and make relevant, informed and scientifically based decisions that they will be able to apply in future professions and in solving problems in daily life (Harrison, 2014). According to Kruit et al. (2018a), however, precisely defining the term inquiry skill is challenging because it is associated with various related terms, such as *science skills* (Kruit et al., 2018a), *investigation skills* (Harlen & Qualter, 2009), or *intellectual skills* (Wu & Hsieh, 2006). Some authors also equate inquiry skills with *science process skills*, considering both terms as synonyms (e.g., O'Connor & Rosicka, 2020; Feyzioğlu, 2019; Song, 2016). Aslan (2017) even equates these skills to *the twenty-first century skills* because their acquisition is essential for pupils to succeed not only in the academic environment but also in broader societal contexts. The literature also presents various classifications of inquiry skills. Reiss and Abrahams (2015) differentiate these skills into two categories according to their cognitive demands: *process skills* (e.g., observation, classification, planning, prediction, experimentation) and *practical skills* (e.g., titration, microscopy). van den Berg (2013), Fradd et al. (2001), Tamir and Lunetta (1981), and Fuhrman (1978) classified skills according to the inquiry cycle (e.g., asking questions, planning, conducting experiments, and analysing and interpreting data). On the other hand, Wenning (2010) categorized inquiry skills into six categories based on the cognitive level of pupils (rudimentary, basic, intermediate, integrated, culminating, and advanced skills). In general, inquiry skills can be considered to arise from the inquiry cycle and reflect the procedures and methods of scientific work (Arnold et al., 2013; van den Berg, 2013; Fradd et al., 2001; Tamir & Lunetta, 1981; Fuhrman, 1978), which enable pupils (similarly to scientists) to solve various research problems (Indri et al., 2020) by engaging in authentic inquiry activities (Kruit et al., 2018a) that are appropriate to their age and cognitive abilities (Wenning, 2005, 2010). However, the findings of numerous studies point to a relatively low level of pupils' inquiry skills across various educational levels (e.g., Čipková et al., 2026; Indri et al., 2020; Prahani et al., 2021; Sholihah et al., 2020a; Šmida et al., 2024; Tanti et al., 2020), which cannot be considered satisfactory. At the same time, it is necessary to acknowledge that pupils are not born with inquiry skills, nor do they acquire them spontaneously. Therefore, formal education should place systematic emphasis not only on their development but also on the identification of potential learning difficulties and pupils' errors that may hinder or even constrain their meaningful acquisition and subsequent practical application.

1.1 Inquiry Skills in Biology Education

Biology education offers a relatively broad range of topics in which various inquiry-based activities can be implemented with pupils (Kremer et al., 2014). Through these activities, pupils explore biological phenomena and processes, collect data, test hypotheses, and formulate conclusions. For this reason, it is essential to systematically develop pupils' inquiry skills in biology lessons (Nunaki et al., 2020), beginning already at the primary school level

(Çakır & Sarıkaya, 2010; Wenning, 2010). This requirement is also reflected in numerous curricular documents that shape educational content not only in science education in general (e.g., Education Scotland, 2018; The New Zealand Curriculum, 2017; Alberta's Curriculum, 2014; NGSS Lead States, 2013; NRC, 2012), but also specifically in biology education (e.g., MEYS, 2023; SEP, 2023), which is taught as a separate subject in some countries. In addition, many other researchers have addressed this topic in biology education (e.g., Artayasa et al., 2021; Asilevi et al., 2024; Bónus et al., 2024; Delgado-Iglesias et al., 2024). Asilevi et al. (2024) found that inquiry-based activities have a more positive effect on pupils' perception of inquiry skills than the traditional teacher-centred approach, which remains a relatively dominant teaching method in primary schools. Likewise, Delgado-Iglesias et al. (2024) point out that, with appropriate support and suitable pedagogical intervention, systematic development of inquiry skills can occur even among younger primary school pupils. Bónus et al. (2024) also report a positive impact of digital game-based and inquiry-based learning on the development of selected inquiry skills among Hungarian primary school pupils. Despite these positive findings, various challenges related to the development of inquiry skills persist (e.g., Farooq & Islam, 2023; Šmida et al., 2024; Subali et al., 2019). Subali et al. (2019) note that inquiry skills are still not optimally developed among pupils because: teachers tend to focus primarily on pupils' performance in standardized tests, while other aspects of education are marginalized; or teachers lack sufficient competencies or experience necessary for implementing inquiry-based activities. Furthermore, Farooq and Islam (2023) report that science teachers continue to emphasize content knowledge rather than the development of pupils' inquiry skills. These factors may consequently have a negative impact on the development of pupils' inquiry skills. Similarly, it is important to consider that their development cannot be limited only to the procedural component, because their application also depends on individual cognitive processes (Kruit et al., 2018a). The implementation of inquiry activities requires the acquisition of certain knowledge about inquiry skills, which are important for their implementation into practice (Seeratan et al., 2020; Lou et al., 2015; Harlen, 2014). Students may be able to make observations, but they often lack a clear understanding of the rules or purpose of observation in scientific research (Shahali et al., 2017), not only at the primary school level but even at university (Emereole, 2008). As a result, their recorded observations may not be accurate. The same approach can be applied to other skills as well. For instance, if we want pupils to effectively formulate a research question or hypothesis, they must first understand the rules and methods for doing so. Additionally, they need to learn the meanings of terms such as dependent variable, independent variable, and constant variable.

1.2 Assessment of Inquiry Skills

The literature reports a relatively wide range of instruments designed to identify and assess pupils' inquiry skills at different educational levels (e.g. Čipková & Karolčík, 2018; Gormally et al., 2012; Sarıoğlu, 2023; Shahali & Halim, 2010; Šmida et al., 2024; Temiz, 2020; Tosun, 2019; Wenning, 2006, 2007; Zeidan & Jayosi, 2015). Among the most frequently used instruments for assessing the level of inquiry skills of primary and secondary school pupils are various written tests (Dahsah et al., 2017), as they are generally less time-consuming to administer and score than interviews (Milenković et al., 2016). As early as 1965, Walbesser developed *The Process Instrument*, which was intended for primary school pupils (Dillashaw & Okey, 1980). A few years later, Tannenbaum (1969) constructed the *Test of Science Processes*, which assessed the skills of students in Grades 7–9,

including observing, comparing, classifying, quantifying, measuring, experimenting, inferring, and predicting. McLeod et al. (1975) subsequently developed a test measuring four skills (controlling variables, interpreting data, formulating hypotheses, and defining operationally), which was again intended for primary school pupils. Dillashaw and Okey (1980) constructed the *Test of Integrated Process Skills (TIPS)* for students in Grades 7–12, measuring skills associated with planning, conducting, and interpreting the results of investigations. Two years later, Tobin and Capie (1982) developed the *Test of Integrated Science Processes (TISP)* for students in Grades 6–8, focusing on skills related to planning and conducting investigations. Burns et al. (1985) later created the *Test of Integrated Process Skills II (TIPS II)*, again for pupils in Grades 7–12, with the aim of developing an additional set of test items targeting the same set of skills assessed by Dillashaw and Okey (1980). However, the items in these tests were typically formulated within the context of science education in general rather than specifically within the context of biology. An exception was the *Processes of Science Test* developed in 1962, which was based on the Biological Sciences Curriculum Study (BSCS) and whose items were oriented to biology (Dillashaw & Okey, 1980). However, the individual test items were subsequently further developed and modified and became the basis for many contemporary research instruments used at the primary and secondary school levels, among which one-tier tests using multiple-choice items still prevail. Nevertheless, these items are again formulated predominantly within the context of general science (e.g., Shahali & Halim, 2010; Temiz, 2020; Zeidan & Jayosi, 2015), chemistry (e.g., Tosun, 2019; Feyzioglu et al., 2012), or physics (e.g., Wenning, 2006, 2007), rather than within the context of biology itself. This may complicate their application in countries where biology is taught as a separate subject at the primary school level. Additionally, one-tier tests with multiple-choice items are not very suitable due to the nature of inquiry skills (Sarıoğlu, 2023) and their use also increases the risk of guessing the correct answer (Habiddin & Page, 2019; Milenković et al., 2016). These types of tests typically assess whether pupils have acquired inquiry skills but usually do not assess whether pupils understand the nature of those skills, because one-tier tests do not require pupils to justify their answers, which can lead to challenges in interpreting the results accurately (Çil, 2015; Gurel et al., 2015). Therefore, to obtain more detailed data, some authors (e.g., Koksal & Berberoglu, 2014; Kruit et al., 2018a; Sarıoğlu, 2023) have decided to combine one-tier multiple-choice and open-ended items, but their analysis and scoring are quite problematic (Sarıoğlu, 2023; Yan & Subramaniam, 2018). Their use is also limited by pupils' reluctance to write their answers and the longer time required to administer them, which is why multiple-choice items are still preferred (Habiddin & Page, 2019; Milenković et al., 2016).

The limitations of a one-tier multiple-choice test can be eliminated by using a two-tier test that examines not only whether pupils can choose the correct answer regarding the inquiry skill in the first tier, but also whether they can justify their choice in the second tier. In this way, it is possible to assess whether the pupils have really acquired the skill or whether their answer is just the result of coincidence or guessing (Kurniawati, 2021). This type of test was also used by Çil (2015) to assess the conceptual understanding of variables and their application in experiments. Similarly, some other two-tier tests (e.g., Sholihah et al., 2020b; Singamurti et al., 2017) require not only the correct answer in the first tier (the indicator of inquiry skill) but also its justification in the second tier. On the other hand, it is important to note that even two-tier tests have limitations. One of these limitations is the inability of test designers to clearly distinguish whether pupils' incorrect answers are identified misconceptions or are simply the result of a lack of knowledge, which can distort the interpretation of results and subsequently the selection of appropriate teaching strategies (Yan & Subramaniam, 2018; Milenković et al., 2016; Gurel et al., 2015). For this reason, there was a need to add one more

tier (confidence tier), which would eliminate this limit (Gurel et al., 2015). However, when evaluating a three-tier test, we cannot be sure whether pupils are equally confident in the answer in the first tier and the reasoning in the second tier, because both levels require different demands on pupils' thinking operations (Caleon & Subramaniam, 2010a; Gurel et al., 2015; Habiddin & Page, 2019; Yan & Subramaniam, 2018). This deficiency can subsequently lead to distorted results (Gurel et al., 2015), and for this reason, an additional confidence tier was added to increase the reliability of the reasoning, turning the three-tier test into a four-tier test (Banawi et al., 2022; Fakhriyah & Masfuah, 2021).

A four-tier test thus enables researchers to assess not only the correctness of pupils' selected answers and their justifications, but also their level of confidence in both choices. This approach may help to examine more deeply the strength of pupils' understanding and to identify the potential presence of misconceptions (Samsudin, 2023). Consequently, a research instrument constructed in this way can provide more relevant and objective information not only about whether a pupil is able to apply individual inquiry skills, but also about whether and to what extent the pupil understands them, or which specific aspects remain unclear. On this basis, it is subsequently possible to facilitate the process of conceptual change among students (Espinosa et al., 2024). On the other hand, this process may also be hindered, particularly when pupils confidently select incorrect answers, which may indicate the presence of various misconceptions. For this reason, it is necessary to provide pupils with objective and relevant feedback (Butterfield & Metcalfe, 2006; Espinosa et al., 2024). Based on such feedback, pupils should be able to identify and resolve discrepancies between their own misconceptions and scientifically correct concepts, leading to the restructuring of incomplete or incorrect conceptual frameworks (Leonard et al., 2014). According to Espinosa et al. (2024), fostering well-calibrated confidence in pupils' judgments is essential for the further development of their metacognitive skills, which underscores the need to implement four-tier tests in biology education. Moreover, results obtained through four-tier tests can be more precisely differentiated and analysed using cognitive diagnostic models (Mi et al., 2023). Within these models, assessment focuses more on the mastery of individual concepts rather than on pupils' overall performance (Hunsu et al., 2022), while simultaneously enabling a more detailed and accurate diagnosis of pupils' strengths and weaknesses (Helm et al., 2022; Im, 2025).

The need to develop an objective diagnostic instrument is also consistent with documents by OECD (2019) and NGSS Lead States (2013), which emphasize that science learning should include not only content knowledge involving various theories, information, or facts, but also knowledge of standard methods and procedures used in scientific inquiry, which are essential for critically examining evidence supporting relevant claims (OECD, 2019). However, four-tier test is usually used to identify pupils' understanding of scientific concepts or misconceptions regarding the content of science subjects (e.g., Firdaus et al., 2021; Habiddin & Page, 2019; Habiddin et al., 2020; Istiyono et al., 2023; Wu et al., 2025; Zhao et al., 2021) rather than to test pupils' inquiry skills, but their potential can also be assumed in this area, and therefore it is important to examine their use in practice.

2 Objectives and Research Questions

The aim of this research is to design and empirically verify a four-tier test to enable a more comprehensive assessment of selected inquiry skills among primary school pupils, and to compare the results with those obtained from other types of tests commonly used to assess inquiry skills. Based on this aim, we formulated two research questions (RQs):

- **RQ1**: *What is the difference between pupils' mean scores identified through one-tier, two-tier and four-tier tests measuring selected inquiry skills*?
- **RQ2**: *What is the difference between the frequency of pupils' misconceptions associated with inquiry skills identified through two-tier and four-tier tests*?

Furthermore, Griffiths and Thompson (1993) identified as many as 63 different misconceptions related to inquiry skills. For example, some pupils believed that a hypothesis or prediction was merely a random guess about the outcome of an experiment; that the independent variable did not affect the outcome of an experiment; or that interpreting data also included hypothesising about what would happen during the experiment. Such findings are concerning, as misconceptions can negatively affect the acquisition of other inquiry skills and hinder the effective implementation of inquiry activities. For this reason, we focused on a more detailed identification of pupils' misconceptions related to selected inquiry skills using a four-tier test, thereby highlighting the diagnostic potential of this instrument. Based on this objective, we formulated the third research question:

- **RQ3**: *What misconceptions associated with inquiry skills have primary school pupils when solving items in the four-tier test?*

3 Material and Methods

3.1 Construction of a Four-tier Test

In constructing the four-tier test, we drew on previous research (Šmida et al., 2024) in which we developed a valid and reliable multiple-choice test to measure primary school pupils' inquiry skills using a single tier (henceforth the one-tier test). This test measured 10 selected inquiry skills through 20 items (each skill was measured through two items). However, administering a four-tier test is relatively time-consuming (Alan & Akbaş, 2025; Caleon & Subramaniam, 2010a). Therefore, it was necessary to reduce the number of items (as well as the number of tested inquiry skills) in the one-tier test so that, after conversion into a four-tier test, it could be administered to pupils within one lesson (45 min). From the original test, we retained the skills of formulating research question, identifying variables, predicting, designing experiment, and describing relationships between data/variables, many of which Wenning (2010) recommends developing among primary school pupils. Furthermore, some research (e.g., Glazer, 2011; Šmida et al., 2024) indicates that pupils often struggle to interpret data presented in graphs or tables, which in turn hinders their ability to explain relationships between variables (Glazer, 2011). For this reason, we added two items assessing the skill of processing data into a table or graph, and two items assessing the skill of analysing data presented in a table or graph.

The basic version of the test consisted of 14 items, each requiring pupils to select one correct answer from five options at the first tier. We then expanded the test to include a second tier, in which pupils indicated whether they were confident in their first-tier answer. A third tier was also added, in which pupils selected one of five options representing the reasoning behind their first-tier response. If pupils disagreed with the provided reasoning, they could write their own. This approach is also recommended by Gurel et al. (2015), as pupils often have their own ideas that may not match the given options, which could otherwise lead them to select an answer at random. In the fourth tier, pupils were again asked to

indicate whether they were confident in their reasoning. Individual distractors were formulated based on frequent errors and misconceptions related to selected inquiry skills (Griffiths & Thompson, 1993; Hodosyová et al., 2015; NRC, 1996; Šmida & Čipková, 2021; Šmida et al., 2024), as recommended by Gierl et al. (2017) and Lai et al. (2016). Examples of the items are provided in Appendix A (full version of the four-tier test is available in Online Resource 1).

3.2 Validity and Reliability of the Four-tier Test

We established the validity of the test through expert assessment (Heale & Twycross, 2015; Kiray & Simsek, 2021) by four experts in the field of biology education. The experts independently assessed each item in terms of its clarity, suitability for 7th-grade primary school pupils (12–13 years old), and whether the items adequately measured the inquiry skills. The experts were invited to provide detailed comments on individual items in the event of any concerns and to propose revisions to the wording of the items, where appropriate. Based on their feedback, some items were revised and sent back to the experts for re-evaluation. Subsequently, the same experts were asked to assess the relevance of each item for measuring the specified inquiry skills of pupils in this age group using a 4-point scale (1 = not relevant; 4 = very relevant). Based on their ratings, we first calculated the item-level content validity index (I-CVI) for each item and then the scale-level content validity index using the average method (S-CVI/Ave), which represents the average proportion of relevance judgments provided by all four experts (Yusoff, 2019). The resulting value of this index was 1.00 indicating unanimous agreement among the experts regarding the relevance of the items. For a small number of experts (fewer than five), such unanimity is considered essential (Lynn, 1986).

Subsequently, we conducted think-aloud cognitive interviews (Peterson et al., 2017) with a smaller group of pupils ($n = 12$) to assess their understanding of the test items, identify any difficulties in completing them, and determine whether the allotted time (45 min) was sufficient.

After administering the test, we examined construct validity through confirmatory factor analysis (Prosser & Trigwell, 2006). In the constructed hypothetical model, we assumed the presence of seven latent variables (factors), namely: formulating research questions, identifying variables, predicting, designing experiments, describing relationships between variables, processing data into tables and graphs, and analysing data from tables or graphs. Each factor was represented by two test items (Table 1). Both items corresponding to the same factor were designed so that their resolution was not limited to a single specific

Table 1 Distribution of test items across the individual latent variables (factors)

Factors		Items
1	Formulating research questions	1, 5
2	Identifying variables	2, 6
3	Predicting	3, 7
4	Designing experiments	4, 8
5	Describing relationships between variables	9, 10
6	Processing data into tables or graphs	11, 12
7	Analysing data from tables or graphs	13, 14

biological topic and so that they were of equivalent difficulty level. The values of the TLI (0.90), CFI (0.93), RMSEA (0.05), and SRMR (0.04) indices indicate a relatively good fit between the data and the hypothetical model (Heubeck & Neill, 2000).

Some authors (e.g., Hestenes & Halloun, 1995; Kiray & Simsek, 2021; Özveren et al., 2025; Taban & Kiray, 2022; Taslidere, 2016) note that, for a valid four-tier test, the frequency of false positive (correct answer, incorrect reasoning) and false negative responses (incorrect answer, correct reasoning) should be below 10%. The lower these frequencies, the more valid the test is considered to be (Özveren et al., 2025; Taslidere, 2016). The analysis showed that pupils had an average frequency of false positive misconceptions related to the inquiry skill of 7.5% and an average frequency of false negative misconceptions related to the inquiry skill of 3.7%. These results indicate that the research instrument can be considered valid also from this perspective.

Subsequently, we examined the plausibility of individual distractors. Distractors should be plausible to pupils with lower ability while being implausible to those with higher ability (Haladyna et al., 2019); therefore, they should be formulated to maximize their attractiveness to pupils. One of the fundamental methods for analysing distractors is the distractor choice frequency (Gierl et al., 2017; Haladyna et al., 2019), which was also applied in the analysis of distractors in a four-tier test by Asih et al. (2022). Forthmann et al. (2020) report that if a distractor is chosen by fewer than 5% of pupils, it is necessary to consider whether it is sufficiently attractive to pupils or requires revision. In the fourteen items (levels 1 and 3), we formulated 112 distractors, of which 95.5% were selected by pupils with a frequency higher than 5% (Table 2). Only five distractors were chosen with a frequency lower than 5% (in items 2, 6, 13, and 14); however, none of these frequencies (minimum 4.49%) did not fall substantially below the recommended threshold. Moreover, we consider that these distractors still have a diagnostic function, and therefore we retained them in their original form.

Table 2 Distractor choice frequency analysis

Inquiry skill	Item	Frequency of pupils' answers in the 1st tier [%]					Frequency of pupils' answers in the 3rd tier [%] [a]				
		a	b	c	d	e	a	b	c	d	e
Formulating research question	1	9.29	7.37	57.7	5.77	19.9	18.6	16.7	5.45	22.1	29.8
	5	15.4	9.62	34.6	10.3	30.1	14.4	19.9	32.7	14.1	13.5
Identifying variables	2	43.3	10.9	36.9	4.49[b]	4.49[b]	8.01	27.2	16.0	30.8	13.1
	6	13.5	25.3	16.7	40.1	4.49[b]	26.6	13.8	16.7	23.7	14.7
Predicting	3	21.8	16.3	6.73	43.9	11.3	21.8	11.5	13.8	31.7	16.0
	7	29.2	9.94	8.01	47.1	5.77	33.0	17.9	18.6	15.4	8.33
Designing experiment	4	14.7	29.2	20.8	25.0	10.3	13.5	17.9	29.8	17.6	16.3
	8	21.2	27.9	32.4	9.94	8.65	16.0	22.8	27.2	16.7	9.94
Describing relationships between variables	9	17.3	50.3	16.0	7.69	8.75	41.3	18.3	9.29	11.5	12.5
	10	19.6	18.6	18.3	30.8	12.7	15.1	18.9	18.9	21.5	18.6
Processing data into tables and graphs	11	51.3	12.2	12.5	11.9	12.2	38.5	18.6	15.1	11.5	10.6
	12	10.6	14.1	8.43	53.8	13.1	6.41	10.9	15.4	43.3	17.0
Analysing data from tables and graphs	13	66.3	8.01	13.8	7.37	4.49[b]	60.6	10.9	8.33	5.77	6.09
	14	4.49[b]	8.97	10.3	70.5	5.77	11.9	5.13	9.94	60.9	7.05

[a] The frequency of pupils selecting the "other" option was not included in the analysis because this option did not represent a distractor (pupils could provide a response in their own words). [b] Distractors chosen by less than 5% of pupils. Grey colour: correct answer

The reliability of the four-tier test was calculated using the Kuder and Richardson Formula No. 20. Based on the obtained value ($KR_{20} = 0.78$), the test can be considered a reliable research instrument (Wahyuni et al., 2021).

3.3 Research Sample

The test was administered to 312 seventh-grade pupils from eleven primary schools (ISCED 2) across various regions of Slovakia. The sample consisted of 53.5% boys and 46.5% girls, aged 12–13, with an average grade in biology of $M = 1.64$ on their last school report card.

3.4 Method of Evaluating the Mean Score in One-tier, Two-tier and Four-tier Tests

Initially, only pupils' answers in the first tier of the four-tier test were considered, ignoring the reasoning tier and both confidence tiers. For this reason, the resulting assessment corresponds to a one-tier test. Pupils received 1 point for a correct answer in the first tier, with a maximum possible score of 14 points, and 0 points for an incorrect answer (Table 3). The reliability of the test was determined using the Kuder and Richardson Formula No. 20. The obtained value ($KR_{20} = 0.74$) indicates that, even after reducing the number of tiers from four to one, the test remains a reliable research instrument (Wahyuni et al., 2021).

Similarly, we converted the four-tier test into a two-tier test. In this evaluation, we considered pupils' answers in the first tier and their reasoning in the third tier (which corresponds to the second tier in the two-tier version), while ignoring both confidence tiers. In evaluating the test, we followed the approach of previous studies (e.g., Fadillah & Salirawati, 2018; Gurel et al., 2015; Gürsel & Akçay, 2022; Kaniawati et al., 2019). Pupils received 1 point if they chose the correct answer in the first tier and the correct reasoning in the second tier, with a maximum possible score of 14 points. Any other combination of answers was scored 0 points (Table 3). The reliability of the test was determined using the Kuder and Richardson Formula No. 20. The obtained value ($KR_{20} = 0.78$) indicates that the test is a reliable research instrument (Wahyuni et al., 2021).

In evaluating the four-tier tests, we considered pupils' responses at all four tiers, following the recommendations of Prayitno and Hidayati (2022) and Gurel et al. (2015). If pupils chose the correct answer in the first tier and were confident in their choice and simultaneously selected the correct reason in the third tier with confidence, we can conclude that they have mastered the selected inquiry skill. Pupils received 1 point for each correctly completed task in this manner, with a maximum possible score of 14 points.

Table 3 Criteria for correct answer in the one-tier, two-tier and four-tier test

Tiers	One-tier test	Two-tier test	Four-tier test
Answer in the 1 st tier	Correct answer	Correct answer	Correct answer
Confidence for the answer in the 1 st tier	-	-	Confident answer
Reasoning of answer	-	Correct reasoning	Correct reasoning
Confidence for the reasoning	-	-	Confident reasoning
Points per item (total score)	1 (max. 14)	1 (max. 14)	1 (max. 14)

All other answer combinations were scored 0 points (Table 3). A detailed scoring key for the test is provided in Appendix B.

3.5 Identifying the Frequency of Misconceptions in the Two-tier and Four-tier Tests

According to Yusrizal and Halim (2017), a one-tier test is not suitable for identifying pupils' misconceptions. Therefore, when analysing the frequencies of misconceptions, we converted the four-tier test into a two-tier test. However, with this type of test, it remains difficult to determine with certainty whether pupils have a misconception, lack of knowledge, or are simply guessing (Milenković et al., 2016; Gurel et al., 2015; Xiao et al. 2018) describe several approaches to evaluating two-tier tests: (a) the first and second tiers can be evaluated separately (e.g., pupils receive 1 point for a correct answer in the first tier and 1 point for a correct justification). However, this approach increases the risk of distorted results due to the possibility of guessing, similar to a one-tier test; (b) a paired evaluation, in which each item is evaluated together. In this case, pupils receive 1 point only if they correctly answer both the first and second tiers, and 0 points for all other responses. This approach reduces the risk of guessing to some extent. However, this evaluation method is limited for identifying pupils' misconceptions, as it does not allow us to determine with certainty the reasons behind incorrectly answered items that receive a score of 0 points. For this reason, some authors (e.g., Chu et al., 2009; Hilton et al., 2013; Loh et al., 2014; Sreenivasulu & Subramaniam, 2013) pair pupils' answers in the first tier with their reasoning in the second tier using a frequency table. They then analyse only those alternative ideas that appear in at least 10% of the responses and have the potential to be considered misconceptions. However, this evaluation method is quite difficult when statistical comparisons between different data sets are required. To simplify the evaluation of two-tier tests, some authors (e.g., Hutahaean et al., 2024; Lengkong et al., 2021; Verma & Choudhuri, 2025) categorize answers as correct or incorrect according to pupils' response combinations. In the case of an incorrect answer, they further distinguish whether the error is due to guessing, a lack of pupils' knowledge, or the presence of a misconception (Table 4).

Subject: *CHE* Chemistry, *PHY* Physics, *BIO* Biology, *MAT* Mathematics. Decision: *UC* Understanding the concept, *PU* Partial understanding the concept, *MSC* Misconception, *NC* Not understanding the concept.

The authors' interpretations listed in Table 4 differ slightly. However, all of them agree that if pupils choose the correct answer in the first tier but an incorrect answer in the second tier, this indicates a misconception due to selecting the wrong reasoning (Kaltakci & Didis, 2007). This premise is based on the study by Hestenes and Halloun (1995). According to Sıbıç et al. (2022), the second tier is diagnostically more significant for identifying misconceptions. For this reason, we followed a similar approach when evaluating the two-tier test for the presence of misconceptions related to inquiry skills. We then evaluated the four-tier test. According to Gurel et al. (2017), an individual with a misconception is deeply convinced that their misconception represents the correct scientific idea. Therefore, in the four-tier test, a misconception related to the application of inquiry skills can be defined as a combination of responses in which pupils select an incorrect answer in the first tier, provide incorrect reasoning in the third tier, and are confident in both choices (Gurel et al., 2015; Kiray & Simsek, 2021; Prayitno & Hidayati, 2022). If pupils select the correct answer in the first tier with confidence but provide incorrect reasoning in the third tier, also with confidence, they can be identified as having a false positive misconception. Conversely, if pupils select an incorrect answer in the first tier with confidence

Table 4 Analysis of different combinations of pupils' answers in two-tier tests

Research by…		Fadillah and Salirawati (2018)	Lengkong et al. (2021)	Verma and Choudhuri (2025)	Hutahaean et al. (2024)	Kaniawati et al. (2019)	Kaltakci & Didis (2007)	Myanda et al. (2020)	Humaidi et al. (2023)	Yamtinah et al. (2019)
Subject		CHE	PHY	BIO	CHE	PHY	PHY	BIO	MAT	CHE
1 st tier	**2nd tier**	**decision**								
Correct	**Correct**	UC	UC	UC	UC	UC	UC	UC	UC	UC
Correct	**Incorrect**	MSC	MSC	MSC	MSC	MSC	MSC	MSC	MSC	MSC
Incorrect	**Correct**	NC	MSC	MSC	PU	NC	Error	MSC	Guess	MSC
Incorrect	**Incorrect**	NC	NC	NC	NC	MSC	MSC	NC	NC	NC

but provide correct reasoning in the third tier, also with confidence, they can be identified as having a false negative misconception (Gurel et al., 2015; Kiray & Simsek, 2021; Prayitno & Hidayati, 2022). According to Prayitno and Hidayati (2022), false positive and false negative misconceptions are likely the result of guessing or inattention. Therefore, it is recommended to treat them as separate categories during evaluation (Kiray & Simsek, 2021; Taban & Kiray, 2022). If pupils are unconfident about an answer, it should not be considered either a misconception or a correct response (Kiray & Simsek, 2021).

3.6 Data Analysis

For each type of test, we determined basic descriptive characteristics (e.g., mean, mode, median, standard deviation) and calculated the mean values of the difficulty index and the discrimination (upper–lower) index (Pande et al., 2013). In addition, the Shapiro–Wilk test showed that the data from the one-tier ($W = 0.95$; $p < 0.05$), two-tier ($W = 0.89$; $p < 0.05$), and four-tier ($W = 0.83$; $p < 0.05$) tests were not normally distributed (Yap & Sim, 2011). Therefore, we used the Spearman correlation coefficient (Neideen & Brasel, 2007) to examine correlations between pupils' mean scores. To compare the mean scores of the same group of pupils across three repeated measurements, we applied the Friedman test (Cleophas & Zwinderman, 2016) followed by the Durbin–Conover post hoc test (Campelo et al., 2023) to identify statistically significant differences.

To compare the average frequency of pupils' misconceptions in the two-tier and four-tier tests, we recoded the results so that the presence of each identified misconception was scored as 1 point, while its absence was scored as 0 points. The Shapiro–Wilk test showed that the data from the two-tier ($W = 0.94$; $p < 0.05$) and four-tier ($W = 0.81$; $p < 0.05$) tests did not follow a normal distribution (Yap & Sim, 2011). Therefore, we used the Spearman correlation coefficient (Neideen & Brasel, 2007) to examine correlations and the Wilcoxon signed-rank test (Xia, 2020) to determine statistically significant differences in the frequency of misconceptions for the same group of pupils across two repeated measurements.

4 Results

We analysed the results of the one-tier, two-tier, and four-tier tests. Given the context of the research and the formulated research questions, we focused primarily on pupils' acquired inquiry skills, the frequency of misconceptions, and the identification of misconceptions that occurred most frequently among pupils. The summary results of the four-tier test are presented in Appendix C.

4.1 Difference in Pupils' Mean Score Identified Through One-tier, Two-tier, and Four-tier Tests

The pupils achieved a mean score of 6.42 points ($SD = 3.24$) in the one-tier test, representing 45.9% (Table 5). However, based on this result, it is not possible to determine with certainty whether the pupils truly mastered the inquiry skills or whether they merely arrived at the correct answers through logical deduction or guessing (the probability of guessing correctly in an item with one correct and four incorrect options is 20%). For this reason, we also analysed the results of the two-tier test, which considered not only the pupils' answers in the first tier but also their reasoning in the second tier.

Table 5 Comparison of one-tier, two-tier, and four-tier tests regarding acquired inquiry skill

Characteristics	One-tier test	Two-tier test	Four-tier test
Number of pupils	312	312	312
Mean	6.42	3.56	2.50
Median	7.00	3.00	2.00
Mode	0.00	0.00	0.00
Variance	10.5	8.75	6.78
Standard deviation	3.24	2.96	2.61
Minimum	0.00	0.00	0.00
Maximum	14.0	11.0	11.0
Range	14.0	11.0	11.0
Standard skewness	0.20	4.65	9.20
Standard kurtosis	−3.53	−1.82	5.11
Coefficient of variation	50.5%	83.0%	104.1%
Reliability	$KR_{20}=0.74$	$KR_{20}=0.78$	$KR_{20}=0.78$

In this test format, the pupils achieved a lower mean score of 3.56 points ($SD=2.96$), corresponding to 25.4% (Table 5). In the four-tier test, pupils achieved the lowest mean score of all test types, namely 2.5 points ($SD=2.61$), corresponding to 17.9% (Table 5). The Friedman test confirmed a significant difference between pupils' results across the different types of tests (χ^2 (2) = 538; $p<0.001$). The Durbin-Conover post hoc analysis further revealed that pupils obtained significantly higher mean scores in the one-tier test compared to both the two-tier ($p<0.001$) and four-tier tests ($p<0.001$), and significantly higher scores in the two-tier test compared to the four-tier test ($p<0.001$). These findings indicate that when the evaluation of acquired inquiry skills is based solely on pupils' answers in the first tier, or when pupils' confidence levels are not considered, the mean score (mean frequency of acquired inquiry skills) is significantly overestimated. From these findings, we can conclude that in assessing inquiry skills, a four-tier test provides more objective and accurate results compared to the one- or two-tier tests.

We also observed differences between the one-, two-, and four-tier tests in the mean values of the difficulty and discrimination indices (Table 6). The results show that adding tiers to the test decreases both indices, indicating that the items in the four-tier test had lower discriminatory power and higher difficulty than those in the one- or two-tier tests. On the other hand, most test items demonstrated good discrimination indices ($M=0.43$), except for items 2 and 10, which exhibited poor discrimination ability (Pande et al., 2013).

On the other hand, we found positive correlations between the results of the individual tests (Table 7), suggesting that pupils who achieved higher mean scores in the one-tier test also tended to achieve higher mean scores in the two- and four-tier tests, and vice versa.

A more detailed analysis of the results revealed that the pupils' mean score was consistently overestimated compared to the four-tier test for each of the assessed inquiry skills. This difference was confirmed by the Friedman test and subsequently by the Durbin–Conover post hoc analysis (Table 8). The largest discrepancy was observed in the skill of formulating research questions, where the mean score in the one-tier test was overestimated by up to 35% compared to the four-tier test. In contrast, the smallest difference was observed in the skill of designing experiment (16%), although it should be noted that pupils achieved the lowest mean score for this skill in the one-tier test among all the tested skills.

Table 6 Values of difficulty and discriminations indices

Inquiry skill	Item	One-tier test		Two-tier test		Four-tier test	
		Difficulty index [%]	Discrimination index	Difficulty index [%]	Discrimination index	Difficulty index [%]	Discrimination index
(I)	1	57.7	0.60	22.4	0.29	12.8	0.31
	5	34.6	0.45	15.1	0.39	9.62	0.26
(II)	2	36.9	0.29	14.4	0.25	8.65	0.19
	6	40.1	0.67	14.7	0.31	8.97	0.23
(III)	3	43.9	0.61	21.8	0.55	14.1	0.40
	7	47.1	0.63	18.9	0.39	11.9	0.35
(IV)	4	26.0	0.50	21.5	0.55	13.8	0.37
	8	32.4	0.51	17.0	0.49	12.5	0.37
(V)	9	50.3	0.63	28.5	0.68	15.7	0.44
	10	30.8	0.39	10.3	0.27	5.77	0.18
(VI)	11	51.3	0.80	30.8	0.75	24.0	0.67
	12	53.8	0.69	33.0	0.60	21.2	0.51
(VII)	13	66.3	0.68	55.1	0.92	48.1	0.92
	14	70.5	0.67	52.6	0.77	43.3	0.81
Mean value		45.9	0.58	25.4	0.51	17.9	0.43

(I) formulating research question; **(II)** identifying variables; **(III)** predicting; **(IV)** designing experiment; **(V)** describing relationships between variables; **(VI)** processing data into tables and graphs; **(VII)** analysing data from tables and graphs.

Table 7 Correlations between the results of a one-tier, two-tier, and four-tier test

	One-tier test	**Two-tier test**	**Four-tier test**
One-tier test	-	$r_s = 0.85$; $p < 0.05$	$r_s = 0.77$; $p < 0.05$
Two-tier test	$r_s = 0.85$; $p < 0.05$	-	$r_s = 0.89$; $p < 0.05$
Four-tier test	$r_s = 0.77$; $p < 0.05$	$r_s = 0.89$; $p < 0.05$	-

4.2 Difference Between the Frequency of Pupils' Misconceptions Associated with Inquiry Skills Identified Through Two-tier test and Four-tier Test

In the two-tier test, the mean frequency of pupils' misconceptions was higher ($M = 2.85$; $SD = 1.71$) than in the four-tier test ($M = 2.14$; $SD = 2.41$). The Wilcoxon signed-rank test also revealed a statistically significant difference ($W = 4.65$; $p < 0.05$) in favour of the two-tier test. These results indicate that pupils' misconceptions associated with inquiry skills were overestimated in the two-tier test. Moreover, the Spearman correlation coefficient did not show a statistically significant correlation between the frequency of pupils' misconceptions in the two-tier and four-tier tests ($r_s = -0.11$; $p > 0.05$), suggesting a lack of relationship between the results obtained through these two tests. Descriptive characteristics of the tests are presented in Table 9.

A more detailed analysis of the results from both test types revealed statistically significant differences in the mean frequency of pupils' misconceptions for the individual tested inquiry skills. The frequency of misconceptions was overestimated for all skills except

Table 8 Pupils' mean score per item and per skill

Inquiry skill	Item	One-tier test		Two-tier test		Four-tier test		Difference [%]	Durbin-Conover post hoc analysis
		Mean score per item [%]	**Mean score per skill [%]**	**Mean score per item [%]**	**Mean score per skill [%]**	**Mean score per item [%]**	**Mean score per skill [%]**		
(I)	1	57.7	46.2	22.4	18.8	12.8	11.2	−7.6 [a]	$p<0.001$ [a]
	5	34.6		15.1		9.6		−35.0 [b]	$p<0.001$ [b]
Friedman test		$\chi^2(2)=257; p<0.001$						−27.4 [c]	$p<0.001$ [c]
(II)	2	36.9	38.5	14.4	14.6	8.7	8.9	−5.7 [a]	$p<0.001$ [a]
	6	40.1		14.7		9.0		−29.6 [b]	$p<0.001$ [b]
Friedman test		$\chi^2(2)=263; p<0.001$						−23.9 [c]	$p<0.001$ [c]
(III)	3	43.9	45.5	21.8	20.4	14.1	13.0	−7.4 [a]	$p<0.001$ [a]
	7	47.1		18.9		11.9		−32.5 [b]	$p<0.001$ [b]
Friedman test		$\chi^2(2)=267; p<0.001$						−25.1 [c]	$p<0.001$ [c]
(IV)	4	26.0	29.2	21.5	19.3	13.8	13.2	−6.1 [a]	$p<0.001$ [a]
	8	32.4		17.0		12.5		−16.0 [b]	$p<0.001$ [b]
Friedman test		$\chi^2(2)=129; p<0.001$						−9.9 [c]	$p<0.001$ [c]
(V)	9	50.3	40.6	28.5	19.4	15.7	10.8	−8.6 [a]	$p<0.001$ [a]
	10	30.8		10.3		5.8		−29.8 [b]	$p<0.001$ [b]
Friedman test		$\chi^2(2)=261; p<0.001$						−21.2 [c]	$p<0.001$ [c]
(VI)	11	51.3	52.6	30.8	31.9	24.0	22.6	−9.3 [a]	$p<0.001$ [a]
	12	53.8		33.0		21.2		−30.0 [b]	$p<0.001$ [b]
Friedman test		$\chi^2(2)=251; p<0.001$						−20.7 [c]	$p<0.001$ [c]
(VII)	13	66.3	68.4	55.1	53.9	48.1	45.7	−8.2 [a]	$p<0.001$ [a]
	14	70.5		52.6		43.3		−22.7 [b]	$p<0.001$ [b]
Friedman test		$\chi^2(2)=181; p<0.001$						−14.5 [c]	$p<0.001$ [c]

[a] difference between four-tier test and two-tier test. [b] difference between four-tier test and one-tier test. [c] difference between two-tier test and one-tier test. **(I)** formulating research question; **(II)** identifying variables; **(III)** predicting; **(IV)** designing experiment; **(V)** describing relationships between variables; **(VI)** processing data into tables and graphs; **(VII)** analysing data from tables and graphs

Table 9 Comparison of two-tier and four-tier tests regarding frequency of misconception

Characteristics	Two-tier test	Four-tier test
Number of pupils	312	312
Mean frequency of misconceptions	2.85	2.14
Median	3.00	1.00
Mode	3.00	0.00
Variance	2.94	5.82
Standard deviation	1.71	2.41
Minimum	0.00	0.00
Maximum	9.00	11.0
Range	9.00	11.0
Standard skewness	3.36	10.2
Standard kurtosis	−0.46	6.38
Coefficient of variation	60.2%	112.9%

for the skill of designing experiment (Table 10). For this skill, the four-tier test captured a higher frequency of misconceptions than the two-tier test, which may be attributed to pupils' inability to select the correct answer from the options provided in the first tier. This interpretation is further supported by the one-tier test results, where pupils achieved the lowest mean score (i.e., the lowest mean frequency of correctly acquired inquiry skills) for this skill among all tested skills (Table 8).

Table 10 Pupils' mean frequency of misconception per item and per skill

Inquiry skill	Item	Two-tier test		Four-tier test		Difference [%] [a]	Wilcoxon signed-rank test
		Mean score per item [%]	**Mean score per skill [%]**	**Mean score per item [%]**	**Mean score per skill [%]**		
(I)	1	34.6	27.1	15.7	16.5	−10.6	$W=3.55$; $p<0.05$
	5	19.6		17.3			
(II)	2	22.4	23.9	20.8	18.8	−5.1	$W=1.94$; $p<0.05$
	6	25.3		16.7			
(III)	3	21.8	25.0	17.6	15.7	−9.3	$W=3.22$; $p<0.05$
	7	28.2		13.8			
(IV)	4	4.5	10.0	25.6	23.4	13.4	$W=5.36$; $p<0.05$
	8	15.4		21.2			
(V)	9	21.8	21.2	11.2	15.1	−6.1	$W=2.30$; $p<0.05$
	10	20.5		18.9			
(VI)	11	20.5	20.7	10.3	10.8	−9.9	$W=3.72$; $p<0.05$
	12	20.8		11.2			
(VII)	13	11.2	14.6	6.7	6.6	−8.0	$W=4.15$; $p<0.05$
	14	17.9		6.4			

[a] difference between four-tier test and two-tier test. (I) formulating research question; (II) identifying variables; (III) predicting; (IV) designing experiment; (V) describing relationships between variables; (VI) processing data into tables and graphs; (VII) analysing data from tables and graphs

4.3 Misconceptions Associated with Inquiry Skills Identified Through the Four-tier Test

To highlight the diagnostic function of the four-tier test, we also focused on identifying common misconceptions associated with the measured inquiry skills. The highest mean frequencies of misconceptions were observed for the skills of designing experiment (items 4 and 8; $M = 23.4\%$), identifying variables (items 2 and 6; $M = 18.8\%$), and formulating research questions (items 1 and 5; $M = 16.5\%$).

In item 4, pupils were asked to select the appropriate experimental design to demonstrate that temperature and humidity are crucial factors influencing mold growth. However, the most common response ($n = 16$) indicated that pupils believed it was sufficient to place the bread in a single warm and humid environment. This reasoning reflects the misconception that an experiment requires only one condition in which both independent variables change simultaneously, while all other variables remain constant (Table 11). In this case, however, pupils would not be able to determine which variable influences mold growth, and the experimental design also lacked a control group. In item 8, pupils were asked to select the most appropriate modification of the experimental design to verify the effect of water on pea seed germination. The most common response ($n = 23$) suggested that pea seeds should be watered regularly and provided with sufficient heat and oxygen, based on the belief that all seeds must be kept under the same conditions necessary for germination (Table 11). In both items, it appears that pupils' frequent misconceptions related to designing experiment primarily concerned the manipulation of the independent variable. Some

Table 11 Common misconceptions associated with designing experiment skill

Item 4: *Which experimental procedure would best verify the statement given in the text for item 4: „Penicillium notatum is a type of mold that forms characteristic green coatings on food. The key factors influencing its growth are the temperature and humidity of the environment (independent variables). "*	
Common pupils' misconceptions	**number of pupils (%)**
4BE: We place one slice of bread in a warm and moist place and observe how quickly mold grows. This is because, during the experiment, we should create only one environment in which both independent variables (humidity and temperature) change simultaneously, while keeping the other variables constant	16 (5.13%)
4BD: We place one slice of bread in a warm and moist place and observe how quickly mold grows. This is because, during the experiment, we should create only two environments in which both independent variables (humidity and temperature) change simultaneously, while keeping the other variables constant	11 (3.53%)
4CB: We place one slice of bread in a moist place and the other slice in a dry place, and we observe which slice grows mold faster. This is because, during the experiment, we should create only two environments in which we change the humidity, while keeping the other variables constant	10 (3.21%)
Item 8: *Read the proposed procedure that could be used to test the effect of water on the germination of pea (Pisum sativum) seeds. Which option correctly improves this procedure? Proposed procedure: We plant pea seeds, water them regularly, and observe whether they germinate*	
Common pupils' misconceptions	**number of pupils (%)**
8BB: The pea seeds are watered regularly, and sufficient warmth and oxygen must be provided. This is because, during the experiment, all seeds should be provided with the same (constant) conditions required for germination; otherwise, they would not germinate	23 (7.37%)

pupils assume that creating a single environment is sufficient (without including a control group), while others believe that the independent variable does not need to be manipulated at all during the experiment.

In item 2, pupils were asked to identify the dependent variable that could be observed as changing due to the different placement of plants. However, many pupils demonstrated misconceptions (Table 12). The most frequent response ($n=29$) indicated that they believed the intensity of light would change in this experiment because the ambient temperature is the dependent variable that changes in response to changes in the intensity of plant light. Another group of pupils ($n=10$) incorrectly assumed that plant growth represented the independent variable that changed as a result of the placement of individual plants. In this experiment, however, the dependent variable is plant growth, not light intensity (independent variable) or ambient temperature (constant variable). Moreover, the independent variable does not change as a result of another variable; rather, it is manipulated to observe its influence on the dependent variable.

In item 6, pupils were asked to identify the independent variable, i.e., the variable that should be manipulated when investigating the effect of temperature on yeast respiration. However, many pupils demonstrated misconceptions (Table 12). A frequent incorrect response ($n=23$) was that the amount of carbon dioxide should be changed, because they considered carbon dioxide to be the independent variable influenced by temperature. Other pupils ($n=9$) believed that humidity and pH should be manipulated, as if these represented dependent variables affected by changes in temperature. However, the amount of carbon dioxide is the dependent variable, while humidity and pH are constant variables. These patterns of responses indicate that pupils' misconceptions are not limited to confusing dependent and independent variables but also reflect a deeper misunderstanding of their role and function in experimental design.

Table 12 Common misconceptions associated with identifying variables skill

Item 2: *You don't know how to grow tomatoes properly. You decided to plant one plant in a pot placed in the shade on the balcony, another in a pot placed in a light spot, and third plant in a pot placed in a dark room. You regularly water each plant with the same amount of water with dissolved fertilizer. At the same time, you ensured the same temperature for all the plants. Which quantity should be observed to determine how it changes due to the different location of the plants?*	
Common pupils' misconceptions	**number of pupils (%)**
2AD: Light intensity, because in this experiment, the ambient temperature is the dependent variable which changes in response to changes in light intensity	29 (9.29%)
2AC: Light intensity, because in this experiment, plant growth is the independent variable that changes due to the location of the plants	10 (3.21%)
Item 6: *Yeasts are unicellular organisms that cannot be seen with the naked eye. The intensity of yeast respiration also depends on the temperature of the environment in which they are located. This can be demonstrated by measuring the amount of carbon dioxide released by the yeast into the environment. What would you need to change during the measurement to confirm that the intensity of yeast respiration depends on environmental temperature?*	
Common pupils' misconceptions	**number of pupils (%)**
6BA: The amount of carbon dioxide, because this product of yeast respiration in this experiment represents an independent variable that changes due to the influence of temperature	23 (7.37%)
6CC: Humidity and pH of the environment, because in this experiment, both are dependent variables that change due to different temperature	9 (2.89%)

In item 1, pupils were asked to select the most appropriate research question to formulate prior to starting the experiment, based on the description provided. However, a considerable number of pupils ($n=13$) most often chose a question that could only be answered dichotomously, mistakenly assuming that a research question should always be phrased in such a way that it can be answered with a simple "yes" or "no," ensuring absolute clarity for the researcher. A similar misconception was observed in item 5, where some pupils ($n=9$) again opted for dichotomous formulations of research questions (Table 13).

5 Discussion & Conclusion

The use of a four-tier test for assessing inquiry skills is not yet well established among researchers. Nevertheless, the results of our study indicate that one-tier and two-tier tests overestimate pupils' mean scores (i.e., the mean frequency of acquired inquiry skills) compared to the four-tier test. Similar findings were reported by Sreenivasulu and Subramaniam (2013) in their examination of pupils' conceptual understanding of selected chemical concepts, where the mean scores obtained from one-tier and two-tier tests were also overestimated. In the case of the one-tier test, such overestimation is understandable, as pupils' reasoning and the level of confidence in their responses are not considered. However, it remains questionable whether the results of such a test can be considered an objective measure of pupils' actual acquisition of inquiry skills (Ketelhut et al., 2009). Our findings show that when an additional tier is included—requiring pupils to justify their answers from the first tier—the mean score decreases significantly. This indicates that pupils experience difficulties in reasoning about the scientific nature of a given inquiry skill. One possible explanation is that selecting the correct reasoning is inherently more demanding, as it requires higher-order cognitive processes (Caleon & Subramaniam, 2010a; Gurel et al., 2015; Habiddin & Page, 2019; Yan & Subramaniam, 2018). Another explanation is that some correct answers in a one-tier test may simply result from guessing, which would also

Table 13 Common misconceptions associated with formulating research question skill

Item 1: *With a sample of 50 classmates, you used a simple experiment to observe how their heart rate changes after running from the ground floor to the third floor. You found that after this physical activity, the pupils had a higher heart rate because oxygen had to be transported to the muscles more quickly through the blood. In this way, you demonstrated the connection between the circulatory system and muscle activity*	
Common pupils' misconceptions	**number of pupils (%)**
1EA: Will the pupils' heart rate increase after running from the ground floor to the third floor? I am reasoning my answer by saying that a research question should always be answered with only "yes" or "no", making the answer completely unambiguous for the researcher	13 (4.17%)
Item 5: *Dave claimed that the body temperature of the common toad depends on the temperature of the environment in which it is currently located. Mary, on the other hand, claimed that the toad has a constant body temperature that is not influenced by the environment. They decided to conduct an experiment to find out who was right. Which research question should Dave and Mary ask before conducting the experiment?*	
Common pupils' misconceptions	**number of pupils (%)**
5EE: Does the body temperature of the common toad change due to environmental temperature? I am reasoning my answer by saying that the research question should be answered only with "yes" or "no" so that the answer is as clear as possible for the researcher	9 (2.89%)

account for the reduction in pupils' mean scores once a second tier is added (Sreenivasulu & Subramaniam, 2013; Widiyatmoko & Shimizu, 2018). Moreover, when analysing responses, it is not possible to clearly distinguish correct answers selected based on proper reasoning from those chosen due to incorrect reasoning (Caleon & Subramaniam, 2010b; Gurel et al., 2015). As a result, pupils may select the correct answer, but for the wrong reasons (Rollnick & Mahooana, 1999), which can lead to an overestimation of their performance. Research by Yusrizal and Halim (2017) also suggests that the overestimation of pupils' mean scores compared to the multi-tier test may be even more significant if the one-tier test is administered independently rather than in combination with the other tiers, since pupils tend to be more careful when selecting an answer in the first tier if they are also required to justify it. In addition, some pupils in our research may have been influenced by the available reasoning when choosing their first-tier answer (Timmermann & Kautz, 2015), which could in turn have affected their overall performance in the one-tier test. This type of test provides only a very limited insight into how pupils think about the given inquiry skills and the cognitive processes underlying their selection of the correct answers. Consequently, this may complicate the interpretation of results, lead to a distorted view of pupils' inquiry skills and make it more difficult to select the appropriate pedagogical interventions to foster their development.

Some authors (e.g., Çil, 2015; Sholihah et al., 2020b; Singamurti et al., 2017) recommend the use of two-tier tests, which require not only the correct answer in the first tier (serving as an indicator of inquiry skills) but also its reasoning, which provides more evidence of pupils' understanding and their cognitive processes. According to Kaniawati et al. (2019), the two-tier test represents a more valuable diagnostic instrument than the one-tier test; however, when compared with the four-tier test, we also observed a significant overestimation of pupils' mean scores in this type of test. A similar finding was also reported by Sreenivasulu and Subramaniam (2013, 2014), who explained that even when both the answer and its reasoning were correct, pupils might have arrived at this combination by guessing, thereby distorting the obtained results. It appears, therefore, that even two-tier tests cannot reliably capture whether pupils truly understand the scientific nature of the given inquiry skills. This is because pupils may select their justification in the second tier based on a process of elimination or logical inference influenced by their choice in the first tier (Griffard & Wandersee, 2001), or simply because part of the statement containing the justification seems partially correct or familiar to them, which may inflate their score. For this reason, without knowledge of pupils' confidence in their responses, it is very difficult to assess the depth and nature of their reasoning. The inclusion of a confidence tier leads to a decrease in the proportion of pupils correct answers (Taslidere, 2016). Similarly, Yang (2022) as well as Renner & Renner (2001) emphasise that pupils' scores are strongly influenced by their level of confidence regarding their responses. However, we assume that if pupils have genuinely mastered a given inquiry skill, they should be able to apply it with confidence (Fariyani et al., 2017). The interpretation of the four-tier test results is therefore not limited to distinguishing between correct and incorrect responses but also captures the level of confidence associated with pupils' thinking and reasoning. Levels of confidence may thus serve an epistemic function, as in combination with pupils' answers and justifications they enable differentiation between stable and unstable conceptual structures as well as strongly held misconceptions (Orhani, 2025). The use of a four-tier test to assess pupils' inquiry skills, therefore, appears to be an appropriate choice, as its items allow a deeper examination of whether pupils truly understand different concepts (Yan & Subramaniam, 2018).

The constructed four-tier test can be regarded as a valid (Hestenes & Halloun, 1995; Heubeck & Neill, 2000; Kiray & Simsek, 2021; Özveren et al., 2025; Peterson et al., 2017;

Taban & Kiray, 2022) and reliable (Wahyuni et al., 2021) research instrument. The results of this test indicate relatively low pupils scores, suggesting difficulties in understanding the selected inquiry skills. Such understanding is inherently linked to scientific reasoning, which supports the systematic investigation of problems, argumentation, experimentation, and the evaluation of evidence leading to the formation of concepts about the world of nature (Kambeyo & Csapó, 2018). Analysis of the results from the four-tier test can therefore help identify pupils' difficulties with individual inquiry skills, which may subsequently contribute to improving their scientific reasoning in biology education.

The results of the test also demonstrated a good discrimination index for most items (Habiddin & Page, 2019), but a relatively high difficulty index compared to the one- and two-tier tests. In fact, the difficulty index of the four-tier test is below the recommended range (Habiddin & Page, 2019; Pande et al., 2013), and the excessively high difficulty of certain items (items 2 and 10) may have reduced their discrimination ability (Pande et al., 2013). Similar levels of difficulty in four-tier tests have also been reported by Özveren et al. (2025), Habiddin & Page (2019), Yan and Subramaniam (2018), and Taslidere (2016). This may be attributed to the fact that when completing a multi-tier test, pupils are required to employ higher-order thinking skills, as they must evaluate multiple response alternatives and contend with a larger number of distractors, which complicates the identification of the correct answer (Çil, 2015). Distractors are deliberately constructed to be as appealing as possible to pupils (e.g., Gierl et al., 2017; Raymond et al., 2019), and a correct response requires pupils to have truly mastered the inquiry skill in item to both identify the correct option and reject the incorrect ones. Rezigalla et al. (2024) further report that items containing distractors that are particularly attractive to pupils may exhibit lower difficulty indices and reduced discriminatory power. Although two distractors with lower selection frequency were identified in item 2, the remaining distractors in both items were selected by pupils with relatively high frequency. These findings suggest that, despite their high difficulty and lower discrimination indices, the items retain an important diagnostic function. Moreover, the correct answer is not only considered to be the selection of the correct answer in the first tier and the appropriate justification in the third tier, but also the confident selection of these answers, which may initially appear to be a relatively strict evaluation criterion. The literature, however, presents alternative approaches to scoring these types of tests. For instance, Putranta & Afifah (2025) and Kafiyani et al. (2019), in developing a four-tier test to diagnose pupils' mental models on static fluids, considered selecting the correct answer in at least one tier as indicative of partial understanding of the concept, regardless of high school pupils' confidence in that response. Gurel et al. (2015), however, argue that if there is any doubt about a pupil's response in at least one tier of a four-tier test, the response should be scored as incorrect or considered indicative of a flaw in the pupil's reasoning. Even if pupils select the correct answer, a lack of confidence likely reflects insufficient conceptual understanding (Odom & Barrow, 2007; Yang & Lin, 2015) or may indicate an attempt to guess the outcome (Odom & Barrow, 2007). On the other hand, pupils' selections may also be influenced by the desire to provide a socially acceptable answer (Caleon & Subramaniam, 2010a) or by their own level of confidence (Caleon & Subramaniam, 2010a; Renner & Renner, 2001), which is directly related to their metacognitive calibration. Pupils who answer correctly with high confidence, or who answer incorrectly and indicate low confidence, can be considered accurately metacognitively calibrated. Incorrect metacognitive calibration occurs when pupils' confidence does not align with the correctness of their response, suggesting difficulties in accurately evaluating their own knowledge. If pupils select an incorrect answer but are confident in their choice, this may reflect strongly held misconceptions or the presence of Dunning and Kruger's

metaignorance phenomenon. Conversely, if pupils select the correct answer but lack confidence, this may be associated not only with guessing but also with problematic prior experiences or an inability to clearly eliminate certain distractors (Koevoets-Beach et al., 2023). These factors can distort the results and should therefore be investigated in future research. Similarly, a stricter scoring approach may affect the difficulty of individual items; however, Putica (2023) argues that a higher level of difficulty should be expected in this type of test, as it is primarily designed as a diagnostic rather than an achievement instrument.

The results of our research also showed that in the two-tier test, there was a significant overestimation of the mean frequency of misconceptions compared to the four-tier test, which is also confirmed by Kaniawati et al. (2019). This overestimation is likely due to the inherent difficulty in evaluating a two-tier test, where it is challenging to determine with certainty whether pupils' responses reflect a genuine misconception, insufficient knowledge, or pupils just guessing answers randomly (Milenković et al., 2016; Gurel et al., 2015; Sreenivasulu & Subramaniam, 2013). Taslidere (2016) also emphasises that in two-tier tests, any incorrect response may be interpreted as a misconception, and that the addition of a confidence tier reduces the observed frequency of such misconceptions. By linking pupils' answer choices and their corresponding reasoning with tiers assessing confidence in each response, it becomes possible to more accurately distinguish genuine misconceptions from a lack of knowledge (Kaniawati et al., 2019). Consequently, the four-tier test developed in this study represents a more suitable instrument for identifying pupils' misconceptions than a two-tier test, a conclusion that is further supported by Kaniawati et al. (2019) and Sreenivasulu and Subramaniam (2013).

We identified common misconceptions among pupils in the skills of designing experiment, identifying variables, and formulating research questions. In the skill of designing experiment, some pupils believed that proper manipulation of independent variables requires creating only a single environment (without a control environment) or, in some cases, that the independent variable does not need to be manipulated at all during the experiment. Griffiths and Thompson (1993) note that this misconception is widespread because pupils often perceive the independent variable as separate from the rest of the experiment, assume it regulates itself, or change it with a constant variable. However, Çil (2015) emphasises that understanding how to work with variables during experiment is a crucial component of the scientific process, highlighting the need to develop effective strategies to develop this skill among pupils.

Similarly, we observed prevalent misconceptions among pupils in items assessing the skill of identifying variables. A common misconception involved not only confusing the terms dependent and independent variables but also misinterpreting their function and role within experimental design. The NRC (1996) similarly notes that pupils in this age group often struggle to correctly identify individual variables and to understand their respective influences in an experiment. Griffiths and Thompson (1993) report comparable findings, emphasising that pupils frequently lack sufficient familiarity with these concepts or apply them incorrectly—for instance, believing that the independent variable is termed as such because it cannot be manipulated and occurs naturally. Çil (2015) and Saat (2004) further highlight that pupils' inability to correctly identify variables undermines their capacity to plan appropriate experimental procedures.

Regarding the formulation of research questions, pupils frequently misconceive that a properly constructed question is one requiring only a dichotomous response. They often believe that a research question should always be answerable with a simple "yes" or "no," as this would provide an entirely unambiguous result for the researcher. We observed this misconception in our previous study as well (Šmida et al., 2024), where pupils selected this

distractor nearly as often as the option representing the correct answer. However, framing research questions in this manner is generally inappropriate (e.g., Dekker & van Baren-Nawrocka, 2017; Kruit et al., 2018b) because a "yes/no" response yields limited insights that do not reflect the investigative effort of the researcher.

In summary, the results of our study suggest that a four-tier test can serve as a valid and reliable diagnostic instrument, providing more objective and detailed insights into pupils' acquired inquiry skills and their related misconceptions, in contrast to one-tier or two-tier tests. Although these two versions of the test may provide reliability comparable to that of a four-tier test, this test allows teachers to gain a more detailed insight into pupils' reasoning about the selected inquiry skills. This, in turn, enables them to offer more targeted and individualized feedback on pupils' progress. Moreover, by identifying misconceptions, teachers can address and correct them, which could contribute to the improvement of biology education. Objective results from the test allow biology teachers to optimize instructional and learning strategies aimed at developing individual inquiry skills. The requirement for pupils to indicate their confidence in each response also provides an opportunity for them to reflect on their own understanding, practice self-assessment, and thereby strengthen their metacognitive skills (Koevoets-Beach et al., 2023). The four-tier test is also a valuable instrument for researchers, who can use its results to analyse the impact of instructional interventions designed to develop pupils' inquiry skills and to assess the current level and depth of pupils' understanding. Additionally, this test can be administered to relatively large samples of pupils (Gurel et al., 2017), which enhances the generalizability of the obtained results and the conclusions.

6 Limits and recommendations for future research

The use of a four-tier test offers several advantages; however, it is also associated with certain limitations. These include its inherent difficulty, more complex construction, longer administration time, and the potential distortion of pupils' results due to their level of confidence, which can be influenced by self-confidence or socially desirable responses (Caleon & Subramaniam, 2010a). Because of its complexity, the four-tier test is generally more suitable as a diagnostic instrument rather than as a tool for assessing pupils' academic achievement (Caleon & Subramaniam, 2010a; Putica, 2023).

We did not administer the one-tier or two-tier tests to the pupils separately; instead, they were derived from the four-tier test by disregarding the remaining pupils' responses during evaluation. However, such a comparison may not be ideal, as the additional tiers could have influenced pupils' responses (Timmermann & Kautz, 2015; Yusrizal & Halim, 2017), and therefore the results should be carefully interpreted. It is also necessary to consider the sample size, as it can affect the generalizability of the results and conclusions.

Despite considerable efforts to formulate all distractors to be as attractive as possible, it is appropriate to note that 5 out of 112 distractors were selected by fewer than 5% of pupils, which may indicate minor issues with their functionality. However, the frequency of these distractors was not zero; on the contrary, it was close to the recommended threshold. Therefore, these distractors were retained in the test without revision.

Although the four-tier test provides valuable information about pupils' misconceptions related to selected inquiry skills, it does not reveal the underlying reasons or causes for their emergence. Due to the size and structure of the research sample, it was not feasible to implement qualitative research methods (e.g., interviews) with the participating pupils,

which could have offered deeper insights into the difficulties pupils face with individual inquiry skills. Future research could benefit from incorporating such qualitative approaches to obtain a more comprehensive understanding of the issues involved. Similarly, in a four-tier test, pupils are required to justify their answers. When combined with the expression of their confidence for each response, can help reveal their epistemic beliefs, encompassing beliefs about the certainty, development, source, and justification of knowledge (Conley et al., 2004; Schiefer et al., 2022). However, our study did not focus on this aspect, and it would therefore be interesting to address it in future research, which could expand our understanding of the use of this type of tests in this context.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11191-026-00747-3.

Authors Contributions All authors contributed to the study conception and design. Conceptualization was performed by [Dominik Šmida, ORCID ID: 0000–0003-2551–8941], material preparation, data collection and analysis were performed by [Dominik Šmida], [Anna Drozdíková, ORCID ID: 0000–0002-8709–1050] and [Ráchel Nechajová, ORCID ID: 0009–0009-1976–8937]. The first draft of the manuscript was written by [Dominik Šmida] and all authors commented it. All authors read and approved the final manuscript.

Funding Open access funding provided by The Ministry of Education, Science, Research and Sport of the Slovak Republic in cooperation with Centre for Scientific and Technical Information of the Slovak.

Data Availability Results of the research are available in dataset (Šmida et al., 2025).

Full reference: Šmida, D., Drozdíková, A., & Nechajová, R. (2025). *Inquiry skills and four-tier test – results* [dataset]. Figshare. https://doi.org/10.6084/m9.figshare.30018943.v1.

Declarations

The authors declare that submitted article is original work which has not been submitted elsewhere for publication. All sources used are properly listed in the list of references.

Ethics We confirm that the presented research did not require any special approval from the ethics committee. According to the principles of the Belmont Report, all participants were informed about the research process as well as the possibility to withdraw from the research. All participants participated in the research voluntarily and we obtained valid informed consent from legal guardians. The results of the research have been processed with regards to the GDPR. Information about pupils and schools was strictly anonymized. We declare that our research met all the ethical, human, and legal subject requirements imposed on this type of research.

Originality The authors declare that submitted article is original work which has not been submitted elsewhere for publication. All sources used are properly listed in the list of references.

Competing Interest The authors declare that they have no conflict of interest.

References

Alan, Ö. G. T., & Akbaş, U. (2025). Examination of misconceptions in the field of alternative measurement and evaluation with a four-tier test. *Kalem Egitim Ve Insan Bilimleri Dergisi, 15*(1), 187–204. https://doi.org/10.23863/kalem.2024.294

Alberta's Curriculum (2014). *Science (Grades 7–9)*. Alberta Education & Childcare. Retrieved January 15, 2026, from https://curriculum.learnalberta.ca/curriculum/en/pos/SCIGEN_79

Arnold, M. E., Bourdeau, V. D., & Nott, B. D. (2013). Measuring science inquiry skills in youth development programs: The science process skills inventory. *Journal of Youth Development, 8*(1), 1–12. https://doi.org/10.5195/jyd.2013.103

Artayasa, I. P., Muhlis, M., Hadiprayitno, G., & Merta, I. W. (2021). Guided Inquiry Based Biological Teaching Materials to Improve Science Process Skills of Junior High School Students. *Advances in Social Science, Education and Humanities Research, 556*, 90–94. https://doi.org/10.2991/assehr.k.210525.052

Asih, N. F., Linuwih, S., & Fianti, F. (2022). Analysis of four-tier diagnostic test on the topic of temperature and heat in high school. *Physics Communication, 6*(1), 1–6. https://doi.org/10.15294/physcomm.v6i1.36652

Asilevi, M. N., Kärkkäinen, S., Sormunen, K., & Havu-Nuutinen, S. (2024). A comparison of science learning skills in the teacher-centered approach and inquiry-based science fieldwork: Primary school students' perceptions. *International Journal of Education in Mathematics, Science and Technology, 12*(1), 1–19. https://doi.org/10.46328/ijemst.3146

Aslan, S. (2017). Learning by teaching: Can it be utilized to develop inquiry skills? *Journal of Education and Training Studies, 5*(12), 190–198. https://doi.org/10.11114/jets.v5i12.2781

Banawi, A., Sopandi, W., Kadarohman, A., & Solehuddin, M. (2022). Five-Tier multiple-choice diagnostic test development: Empirical evidences to improve students' science literacy. In *International Conference on Madrasah Reform 2021* (pp. 131–138). Atlantis Press. https://doi.org/10.2991/assehr.k.220104.020

Bónus, L., Antal, E., & Korom, E. (2024). Digital Game-Based Inquiry Learning to Improve Eighth Graders' Inquiry Skills in Biology. *Journal of Science Education and Technology, 33*(4), 1–17. https://doi.org/10.1007/s10956-024-10096-x

Burns, J. C., Okey, J. R., & Wise, K. C. (1985). Development of an integrated process skill test: TIPS II. *Journal of Research in Science Teaching, 22*(2), 169–177. https://doi.org/10.1002/tea.3660220208

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning, 1*(1), 69–84. https://doi.org/10.1007/s11409-006-6894-z

Çakır, N. K., & Sarıkaya, M. (2010). An evaluation of science process skills of the science teaching majors. *Procedia-Social and Behavioral Sciences, 9*, 1592–1596. https://doi.org/10.1016/j.sbspro.2010.12.370

Caleon, I. S., & Subramaniam, R. (2010a). Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education, 40*, 313–337. https://doi.org/10.1007/s11165-009-9122-4

Caleon, I., & Subramaniam, R. (2010b). Development and Application of a Three-Tier Diagnostic Test to Assess Secondary Students' Understanding of Waves. *International Journal of Science Education, 32*(7), 939–961. https://doi.org/10.1080/09500690902890130

Campelo, D., Koch, A. J., & Machado, M. (2023). Caffeine, lactic acid, or nothing: What effect does expectation have on men's performance and perceived exertion during an upper body muscular endurance task? *International Journal of Health Sciences, 17*(6), 39–42.

Chu, H., Treagust, D. F., & Chandrasegaran, A. L. (2009). A stratified study of students' understanding of basic optics concepts in different contexts using two-tier multiple-choice items. *Research in Science & Technological Education, 27*(3), 253–265. https://doi.org/10.1080/02635140903162553

Çil, E. (2015). Effect of two-tier diagnostic tests on promoting learners' conceptual understanding of variables in conducting scientific experiments. *Applied Measurement in Education, 28*(4), 253–273. https://doi.org/10.1080/08957347.2015.1064124

Čipková, E., & Karolčík, Š. (2018). Assessing of scientific inquiry skills achieved by future biology teachers. *Chemistry, Didactics, Ecology, Metrology, 23*(1–2), 71–80. https://doi.org/10.1515/cdem-2018-0004

Čipková, E., Fuchs, M., Huszárová, K., & Trško, Ľ. (2026). Ready to experiment? A closer look at students' scientific skills in lower secondary education. *International Journal of Science Education*, 1–19. https://doi.org/10.1080/09500693.2026.2617918

Cleophas, T. J., & Zwinderman, A. H. (2016). Clinical data analysis on a pocket calculator: Understanding the scientific methods of statistical reasoning and hypothesis testing. *Springer*. https://doi.org/10.1007/978-3-319-27104-0_34

Conley, A. M., Pintrich, P. R., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology, 29*(2), 186–204. https://doi.org/10.1016/j.cedpsych.2004.01.004

Dahsah, C., Seetee, N., & Lamainil, S. (2017). The use of interview about events to explore children's basic science process skills. In New perspectives in science education (6th ed., pp. 497–503). Libreria Universitaria. Retrieved August 4, 2025, from https://conference.pixel-online.net/NPSE/files/npse/ed0006/FP/3399-SERA2201-FP-NPSE6.pdf

Dekker, S., & van Baren-Nawrocka, J. (2017). *Wetenschappelijke doorbraken de klas in!. Molecuulbotsingen, Stress en Taal der Zintuigen* [Scientific breakthroughs in the classroom!]. Wetenschapsknooppunt Radboud Universiteit.

Delgado-Iglesias, J., Bobo-Pinilla, J., Reinoso-Tapia, R., & Vega-Agapito, M. V. (2024). Is It Possible to Apply Inquiry in the First Level of Primary School without Hindering the Acquisition of Scientific Competencies? *Perspectives of Pupils and Their Pre-Service Teacher*. *Education Sciences, 14*(1), 96. https://doi.org/10.3390/educsci14010096

Dillashaw, F. G., & Okey, J. R. (1980). Test of the integrated science process skills for secondary science students. *Science Education, 64*(5), 601–608. https://doi.org/10.1002/sce.3730640506

Education Scotland (2018). *Curriculum for Excellence: Experiences and outcomes*. Scottish Government. Retrieved January 15, 2026, from https://education.gov.scot/media/wpsnskgv/all-experiencesoutcomes18.pdf

Emereole, H. U. (2008). Learners' and teachers' conceptual knowledge of science processes: The case of Botswana. *International Journal of Science and Mathematics Education, 7*(5), 1033–1056. https://doi.org/10.1007/s10763-008-9137-8

Espinosa, A. A., Koperová, D., Kuhnová, M., & Rusek, M. (2024). Preservice Chemistry Teachers' Conceptual Understanding and Confidence Judgment: Insights from a Three-Tier Chemistry Concept Inventory. *Journal of Chemical Education, 102*(1), 53–65. https://doi.org/10.1021/acs.jchemed.4c01146

Fadillah, A., & Salirawati, D. (2018). Analysis of misconceptions of chemical bonding among tenth grade senior high school students using a two-tier test. In Y. D. Jatmiko, R. Azrianingsih, M. A. Pamungkas, A. Safitri, Nurjannah, & C. Karim (Eds.), The 8th Annual Basic Science International Conference: Coverage of Basic Sciences toward the World's Sustainability Challenges (Vol. 2021, p. 080002). AIP Publishing. https://doi.org/10.1063/1.5062821

Fakhriyah, F., & Masfuah, S. (2021). The development of a four tier-based diagnostic test diagnostic assessment on science concept course. *Journal of Physics: Conference Series, 1842*(1), Article 012069. https://doi.org/10.1088/1742-6596/1842/1/012069

Fariyani, Q., Rusilowati, A., & Sugianto, S. (2017). Four-tier diagnostic test to identify misconceptions in geometrical optics. *Unnes Science Education Journal, 6*(3), 1724–1729. Retrieved August 8, 2025, from https://journal.unnes.ac.id/sju/usej/article/view/20396

Farooq, A., & Islam, U. M. (2023). Effect of Inquiry Method on Scientific Inquiry Skills of Elementary School Students. *Pakistan Languages and Humanities Review, 7*(2), 127–139. https://doi.org/10.47205/plhr.2023(7-ii)11

Feyzioğlu, B. (2019). The role of inquiry based self-efficacy, achievement goal orientation, and learning strategies on secondary school students' inquiry skills. *Research in Science & Technological Education, 37*(3), 366–392. https://doi.org/10.1080/02635143.2019.1579187

Feyzioglu, B., Demirdag, B., Akyildiz, M., & Altun, E. (2012). Developing a science process skills test for secondary students: Validity and reliability study. *Educational Sciences: Theory and Practice, 12*(3), 1899–1906. Retrieved August 20, 2025, from https://www.proquest.com/scholarly-journals/developing-science-process-skills-test-secondary/docview/1242000240/se-2?accountid=17229

Firdaus, N. R., Kirana, T., & Susantini, E. (2021). A four-tier test to identify students' conceptions in inheritance concepts. *International Journal of Recent Educational Research, 2*(4), 402–415. https://doi.org/10.46245/ijorer.v2i4.128

Forthmann, B., Förster, N., Schütze, B., Hebbecker, K., Flessner, J., Peters, M. T., & Souvignier, E. (2020). How much G is in the distractor? Re-thinking item-analysis of multiple-choice items. *Journal of Intelligence, 8*(1), 1–36. https://doi.org/10.3390/jintelligence8010011

Fradd, S. H., Lee, O., Sutman, F. X., & Saxton, M. K. (2001). Promoting science literacy with English language learners through instructional materials development: A case study. *Bilingual Research Journal, 25*(4), 417–439. https://doi.org/10.1080/15235882.2001.11074464

Fuhrman, M. (1978). *Development of a laboratory structure and task analysis inventory and an analysis of selected chemistry curricula* [Unpublished master's thesis]. University of Iowa, Iowa, United States of America.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of Educational Research, 87*(6), 1082–1116. https://doi.org/10.3102/0034654317726529

Glazer, N. (2011). Challenges with Graph Interpretation: A Review of the Literature. *Studies in Science Education, 47*(2), 183–210. https://doi.org/10.1080/03057267.2011.605307

Gormally, C., Brickman, P., & Lutz, M. (2012). Developing a test of scientific literacy skills (TOSLS): Measuring undergraduates' evaluation of scientific information and arguments. *CBE Life Sciences Education, 11*(4), 364–377. https://doi.org/10.1187/cbe.12030026

Griffard, P. B., & Wandersee, J. H. (2001). The two-tier instrument on photosynthesis: What does it diagnose? *International Journal of Science Education, 23*(10), 1039–1052. https://doi.org/10.1080/09500690110038549

Griffiths, A. K., & Thompson, J. (1993). Secondary school students' understandings of scientific processes: An interview study. *Research in Science & Technological Education, 11*(1), 15–26. https://doi.org/10.1080/0263514930110103

Gurel, D. K., Eryılmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics Science and Technology Education, 11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a

Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science & Technological Education, 35*(2), 238–260. https://doi.org/10.1080/02635143.2017.1310094

Gürsel, F. G., & Akçay, B. (2022). Developing two-tier diagnostic instrument to determine misconceptions on socioscientific issues. *Cumhuriyet International Journal of Education, 11*(1), 228–239. https://doi.org/10.30703/cije.1016641

Habiddin, H., & Page, E. M. (2019). Development and validation of a four-tier diagnostic instrument for chemical kinetics (FTDICK). *Indonesian Journal of Chemistry, 19*(3), 720–736. https://doi.org/10.22146/ijc.39218

Habiddin, H., Ameliana, D. N., & Suaidy, M. (2020). Development of a four-tier instrument of acid-base properties of salt solution. *Journal of Chemistry Education Research, 4*(1), 51–57. https://doi.org/10.26740/jcer.v4n1.p51-57

Haladyna, T. M., Rodriguez, M. C., & Stevens, C. (2019). Are Multiple-choice Items Too Fat? *Applied Measurement in Education, 32*(4), 350–364. https://doi.org/10.1080/08957347.2019.1660348

Harlen, W. (2014). Helping children's development of inquiry skills. *Inquiry in Primary Science Education, 1*(1), 5–19. Retrieved August 15, 2025, from https://ipsejournal.com/wp-content/uploads/2015/03/3-ipse-volume-1-no-1-wynne-harlen-p-5-19.pdf

Harlen, W., & Qualter, A. (2009). *The Teaching of Science in Primary Schools*. Routledge.

Harrison, C. H. (2014). Assessment of inquiry skills in the SAILS project. *Science Education International, 25*(1), 112–122. Retrieved August 18, 2025, from https://files.eric.ed.gov/fulltext/EJ1022890.pdf

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence Based Nursing, 18*(3), 66–67. https://doi.org/10.1136/eb2015102129

Helm, C., Warwas, J., & Schirmer, H. (2022). Cognitive diagnosis models of students' skill profiles as a basis for adaptive teaching: An example from introductory accounting classes. *Empirical Research in Vocational Education and Training, 14*(1), 1–30. https://doi.org/10.1186/s40461-022-00137-3

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory a response to Huffman and Heller. *The Physics Teacher, 33*, 502–502. https://doi.org/10.1119/1.2344278

Heubeck, B. G., & Neill, J. T. (2000). Confirmatory factor analysis and reliability of the mental health inventory for Australian adolescents. *Psychological Reports, 87*(2), 431–440. https://doi.org/10.2466/PR0.87.6.431

Hilton, A., Hilton, G., Dole, S., & Goos, M. (2013). Development and application of a two-tier diagnostic instrument to assess middle-years students' proportional reasoning. *Mathematics Education Research Journal, 25*(4), 523–545. https://doi.org/10.1007/s13394-013-0083-6

Hodosyová, M., Útla, J., Vanyová, M., Vnuková, P., & Lapitková, V. (2015). The Development of Science Process Skills in Physics Education. *Procedia - Social and Behavioral Sciences, 186*, 982–989. https://doi.org/10.1016/j.sbspro.2015.04.184

Humaidi, M. N., Triansyah, F. A., Sugianto, R., & Laila, A. R. N. (2023). Development of a HOTS-leveled Two-Tier Multiple Choice (TTMC) test to measure student misconceptions in Islamic studies. Assyfa Journal of Islamic Studies, 1(1), 31–40. https://doi.org/10.61650/ajis.v1i1.148

Hunsu, N., Yao, K., Al Weshah, A., Olaogun, O., & Wang, S. (2022). Work in progress: The Electric Circuit Concepts Diagnostic (ECCD). In Annual Proceeding of the American Society for Engineering Education (pp. 1–8). American Society for Engineering Education. Retrieved August 17, 2025, from https://par.nsf.gov/servlets/purl/10348376

Hutahaean, E., Pardiana, P., & Hadiyati, Y. (2024). Identifying students' misconceptions on electrolysis using a two-tier diagnostic test. Journal of Research in Environmental and Science Education, 1(1), 1–11. https://doi.org/10.70232/bvc08237

Im, S. (2025). Targeted assessment of hypothesis testing skills using cognitive diagnostic models: Implications for formative practice. *International Journal of Educational Research, 134*, Article 102801. https://doi.org/10.1016/j.ijer.2025.102801

Indri, O. W., Sarwanto, & Nurosyid, F. (2020). Analysis of high school students' science process skills. *Journal of Physics: Conference Series*, *1567*(3), 032098. https://doi.org/10.1088/1742-6596/1567/3/032098

Iskandar, Sastradika, D., & Defrianti, D. (2019). Optimizing inquiry-based learning activity in improving students' scientific literacy skills. *Journal of Physics: Conference Series, 1233(1)*, 1–11. https://doi.org/10.1088/1742-6596/1233/1/012061

Istiyono, E., Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S. S. N., & Saepuzaman, D. (2023). The development of a four-tier diagnostic test based on modern test theory in physics education. *European Journal of Educational Research, 12*(1), 371–385. https://doi.org/10.12973/eu-jer.12.1.371

Kafiyani, F., Samsudin, A., & Saepuzaman, D. (2019). Development of four-tier diagnostic test (FTDT) to identify student's mental models on static fluid. *Journal of Physics: Conference Series, 1280*(5), Article 052030. https://doi.org/10.1088/1742-6596/1280/5/052030

Kaltakçi, D., & Didiş, N. (2007). Identification of pre-service physics teachers' misconceptions on gravity concept: A study with a 3-tier misconception test. In S. A. Cetin & I. Hikmet (Eds.), Sixth International Conference of the Balkan Physical Union (Vol. 899, pp. 499–500). AIP Publishing. https://doi.org/10.1063/1.2733255

Kambeyo, L., & Csapó, B. (2018). Scientific reasoning skills: A theoretical backgrounds to science education. Reform Forum: Journal for Educational Research in Namibia, 26(1), 27–36. Retrieved August 15, 2025, from https://publicatio.bibl.u-szeged.hu/18918/1/Reform_Forum_Volume26_Issue_1_Kambeyo-Csapo.pdf

Kaniawati, I., Fratiwi, N. J., Danawan, A., Suyana, I., Samsudın, A., & Suhendi, E. (2019). Analyzing students' misconceptions about newton's laws through four-tier newtonian test (FTNT). *Journal of Turkish Science Education, 16*(1), 110–122. https://doi.org/10.36681/

Ketelhut, D. J., Nelson, B. C., Clarke, J., & Dede, C. (2009). A multi-user virtual environment for building and assessing higher order inquiry skills in science. *British Journal of Educational Technology, 41*(1), 56–68. https://doi.org/10.1111/j.1467-8535.2009.01036.x

Kiray, S. A., & Simsek, S. (2021). Determination and Evaluation of the Science Teacher Candidates' Misconceptions About Density by Using Four-Tier Diagnostic Test. *International Journal of Science and Mathematics Education, 19*(5), 935–955. https://doi.org/10.1007/s10763-020-10087-5

Koevoets-Beach, C., Julian, K., & Balabanoff, M. (2023). "I guess it was more than just my general knowledge of chemistry": Exploring students' confidence judgments in two-tiered assessments. *Chemistry Education Research and Practice, 24*(4), 1243–1261. https://doi.org/10.1039/d3rp00127j

Koksal, A. E., & Berberoglu, G. (2014). The Effect of Guided-Inquiry Instruction on 6th Grade Turkish Students' Achievement, Science Process Skills, and Attitudes Toward Science. *International Journal of Science Education, 36*(1), 66–78. https://doi.org/10.1080/09500693.2012.721942

Kremer, K., Specht, C., Urhahne, D., & Mayer, J. (2014). The relationship in biology between the nature of science and scientific inquiry. *Journal of Biological Education, 48*(1), 1–8. https://doi.org/10.1080/00219266.2013.788541

Kruit, P. M., Oostdam, R. J., van den Berg, E., & Schuitema, J. A. (2018a). Assessing students' ability in performing scientific inquiry: Instruments for measuring science skills in primary education. *Research in Science & Technological Education, 36*(4), 413–439. https://doi.org/10.1080/02635143.2017.1421530

Kruit, P. M., Oostdam, R. J., van den Berg, E., & Schuitema, J. A. (2018b). Effects of explicit instruction on the acquisition of students' science inquiry skills in grades 5 and 6 of primary education. *International Journal of Science Education, 40*(4), 421–441. https://doi.org/10.1080/09500693.2018.1428777

Kurniawati, A. (2021). Science process skills and its implementation in the process of science learning evaluation in schools. *Journal of Science Education Research, 5*(2), 16–20. https://doi.org/10.21831/jser.v5i2.44269

Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A. P., & De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teaching and Learning in Medicine, 28*(2), 166–173. https://doi.org/10.1080/10401334.2016.1146608

Lengkong, M., Istiyono, E., Rampean, B. A. O., Tumanggor, A. M. R., & Nirmala, M. F. T. (2021). Development of two-tier test instruments to detect student's physics misconception. In 7th International Conference on Research, Implementation, and Education of Mathematics and Sciences (pp. 561–566). Atlantis Press. https://doi.org/10.2991/assehr.k.210305.082

Leonard, M. J., Kalinowski, S. T., & Andrews, T. C. (2014). Misconceptions yesterday, today, and tomorrow. CBE—Life Sciences Education, 13(2), 179–186. https://doi.org/10.1187/cbe.13-12-0244

Loh, A. S. L., Subramaniam, R., & Tan, K. C. D. (2014). Exploring students' understanding of electrochemical cells using an enhanced two-tier diagnostic instrument. *Research in Science & Technological Education, 32*(3), 229–250. https://doi.org/10.1080/02635143.2014.916669

Lou, Y., Blanchard, P., & Kennedy, E. (2015). Development and validation of a science inquiry skills assessment. *Journal of Geoscience Education, 63*(1), 73–85. https://doi.org/10.5408/14-028.1

Lynn, M. R. (1986). Determination and Quantification of Content Validity. *Nursing Research, 35*(6), 382–385. https://doi.org/10.1097/00006199-198611000-00017

McLeod, R. J., Berkheimer, G. D., Fyffe, D. W., & Robison, R. W. (1975). The development of criterion-validated test items for four integrated science processes. *Journal of Research in Science Teaching, 12*(4), 415–421. https://doi.org/10.1002/tea.3660120413

Ministry of Education, Youth and Sports. (2023). Rámcový vzdělávací program pro základní vzdělávání [Framework educational program for basic education]. Retrieved January 18, 2026, from https://www.edu.cz/wp-content/uploads/2023/07/RVP_ZV_2023_zmeny.pdf

Mi, S., Ye, J., Yan, L., & Bi, H. (2023). Development and validation of a conceptual survey instrument to evaluate senior high school students' understanding of electrostatics. *Physical Review Physics Education Research, 19*(1), Article 010114. https://doi.org/10.1103/physrevphyseducres.19.010114

Milenković, D. D., Hrin, T. N., Segedinac, M. D., & Horvat, S. (2016). Development of a three-tier test as a valid diagnostic tool for identification of misconceptions related to carbohydrates. *Journal of Chemical Education, 93*(9), 1514–1520. https://doi.org/10.1021/acs.jchemed.6b00261

Myanda, A. A., Riezky, M. P., & Maridi, M. (2020). Development of two-tier multiple-choice test to assess students' conceptual understanding on respiratory system material of 11th grade of senior high school. *International Journal of Science and Applied Science: Conference Series*, *4*(1), 44–55. https://doi.org/10.20961/ijsascs.v4i1.49457

Neideen, T., & Brasel, K. (2007). Understanding statistical tests. *Journal of Surgical Education, 64*(2), 93–96. https://doi.org/10.1016/j.jsurg.2007.02.001

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press. Retrieved January 15, 2026, from https://www.nationalacademies.org/read/18290/chapter/1

National Research Council. (1996). National science education standards. National Academies Press. https://doi.org/10.17226/4962

NRC (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. National Academies Press. https://doi.org/10.17226/13165

Nunaki, J. H., Siagian, S. I. R., Nusantari, E., Kandowangko, N. Y., & Damopolii, I. (2020). Fostering students' process skills through inquiry-based science learning implementation. *Journal of Physics: Conference Series, 1521*(4), Article 042030. https://doi.org/10.1088/1742-6596/1521/4/042030

O'Connor, G., & Rosicka, C. (2020). Science in the early years. Paper 2: Science inquiry skills. Australian Council for Educational Research. Retrieved August 15, 2025, from https://research.acer.edu.au/early_childhood_misc/16/

Odom, A. L., & Barrow, L. H. (2007). High school biology students' knowledge and certainty about diffusion and osmosis concepts. *School Science and Mathematics, 107*(3), 94–101. https://doi.org/10.1111/j.1949-8594.2007.tb17775.x

OECD. (2019). PISA 2018 Assessment and Analytical Framework. *OECD Publishing*. https://doi.org/10.1787/b25efab8-en

Orhani, S. (2025). From stated knowledge to certainty of thought: A study on four-level tests in mathematics teaching. *Journal of Education for Sustainable Development Studies, 2*(2), 156–174. https://doi.org/10.70232/jesds.v2i2.53

Özveren, G., Turan, B., Ulucan, M., & Tosun, C. (2025). University students' scientific knowledge levels regarding chemical reaction arrows and electron arrows. *Journal of the Serbian Chemical Society, 90*(10), 1267–1284. https://doi.org/10.2298/JSC240523053O

Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., & Agrekar, S. H. (2013). Correlation between difficulty and discrimination indices of MCQs in formative exam in physiology. *South-East Asian Journal of Medical Education, 7*(1), 45–50. https://doi.org/10.4038/seajme.v7i1.149

Peterson, C. H., Peterson, N. A., & Powell, K. G. (2017). Cognitive interviewing for item development: Validity evidence based on content and response processes. *Measurement and Evaluation in Counseling and Development, 50*(4), 217–223. https://doi.org/10.1080/07481756.2017.1339564

Prahani, B. K., Deta, U. A., Lestari, N. A., Yantidewi, M., Jauhariyah, M. N. R., Kelelufna, V. P., Siswanto, J., Misbah, M., Mahtari, S., & Suyidno. (2021). A profile of senior high school students' science process skills on heat material. *Journal of Physics: Conference Series, 1760*(1), 012010. https://doi.org/10.1088/1742-6596/1760/1/012010

Prayitno, T. A., & Hidayati, N. (2022). Analysis of students' misconception on general biology concepts using four-tier diagnostic test (FTDT). *IJORER: International Journal of Recent Educational Research, 3*(1), 1–10. https://doi.org/10.46245/ijorer.v3i1.177

Prosser, M., & Trigwell, K. (2006). Confirmatory factor analysis of the approaches to teaching inventory. *British Journal of Educational Psychology, 76*(2), 405–419. https://doi.org/10.1348/000709905X43571

Putica, K. B. (2023). Development and Validation of a Four-Tier Test for the Assessment of Secondary School Students' Conceptual Understanding of Amino Acids, Proteins, and Enzymes. *Research in Science Education, 53*(3), 651–668. https://doi.org/10.1007/s11165-022-10075-5

Putranta, H., & Afifah, F. (2025). Development of the four-tier diagnostic test to identify student misconceptions in the static fluids chapter. *Journal on Efficiency and Responsibility in Education and Science, 18*(4), 268–281. https://doi.org/10.7160/eriesj.2025.180403

Raymond, M. R., Stevens, C., & Bucak, S. D. (2019). The optimal number of options for multiple-choice questions on high-stakes tests: Application of a revised index for detecting nonfunctional distractors. *Advances in Health Sciences Education, 24*(1), 141–150. https://doi.org/10.1007/s10459-018-9855-9

Reiss, M. J., & Abrahams, I. (2015). The assessment of practical skills. *School Science Review, 357*, 40–44.

Renner, C. H., & Renner, M. J. (2001). But I thought I knew that: Using confidence estimation as a debiasing technique to improve classroom performance. *Applied Cognitive Psychology, 15*(1), 23–32. https://doi.org/10.1002/1099-0720(200101/02)15:1%3c23::aid-acp681%3e3.0.co;2-j

Rezigalla, A. A., Eleragi, A. M. E. S. A., Elhussein, A. B., Alfaifi, J., ALGhamdi, M. A., Al Ameer, A. Y., Yahia, A. I. O., Mohammed, O. A., & Adam, M. I. E. (2024). Item analysis: The impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Medical Education*. https://doi.org/10.1186/s12909-024-05433-y

Rollnick, M., & Mahooana, P. P. (1999). A quick and effective way of diagnosing student difficulties: Two tier from simple multiple choice questions. *South African Journal of Chemistry, 52*(4), 161–164.

Saat, R. M. (2004). The acquisition of integrated science process skills in a web-based learning environment. *Research in Science & Technological Education, 22*(1), 23–40. https://doi.org/10.1080/02635140420001875 20

Samsudin, A. (2023). Conceptual change based on Virtual Media (CC-VM) versus POE strategy: Analysis of mental model improvement and changes on light wave concepts. *International Journal of Technology in Education and Science, 7*(2), 230–252. https://doi.org/10.46328/ijtes.449

Sarıoğlu, S. (2023). Development of online science process skills test for 8th grade pupils. *Journal of Turkish Science Education, 20*(3), 418–432. https://doi.org/10.36681/tused.2023.024

Schiefer, J., Edelsbrunner, P. A., Bernholt, A., Kampa, N., & Nehring, A. (2022). Epistemic Beliefs in Science—A Systematic Integration of Evidence From Multiple Studies. *Educational Psychology Review, 34*(3), 1541–1575. https://doi.org/10.1007/s10648-022-09661-w

Seeratan, K. L., McElhaney, K. W., Mislevy, J., McGhee, R., Jr., Conger, D., & Long, M. C. (2020). Measuring students' ability to engage in scientific inquiry: A new instrument to assess data analysis, explanation, and argumentation. *Educational Assessment, 25*(2), 112–135. https://doi.org/10.1080/10627197.2020.1756253

Ministry of Education, Research, Development and Youth of the Slovak Republic. (2023). Štátny vzdelávací program – Človek a príroda [State educational program - Man and nature]. Retrieved January 15, 2026, from https://vzdelavanie21.sk/digitalny-statny-vzdelavaci-program/

Shahali, E. H., & Halim, L. (2010). Development and validation of a test of integrated science process skills. *Procedia – Social and Behavioral Sciences, 9*, 142–146. https://doi.org/10.1016/j.sbspro.2010.12.127

Shahali, E. H., Halim, L., Treagust, D. F., Won, M., & Chandrasegaran, A. L. (2017). Primary school teachers' understanding of science process skills in relation to their teaching qualifications and teaching experience. *Research in Science Education, 47*, 257–281. https://doi.org/10.1007/s11165-015-9500-z

Sholihah, N. A. A., Sarwanto, & Aminah, N. S. (2020b). Development of two-tier multiple-choice instrument to measure science process skill. *Journal of Physics: Conference Series, 1521*(2), 022053. https://doi.org/10.1088/1742-6596/1521/2/022053

Sholihah, N. A., Sarwanto, & Aminah, N. S. (2020a). Analysis Science Process Skills of 11th Grade of Senior High School Students. *Journal of Physics: Conference Series, 1491*(1), 012036. https://doi.org/10.1088/1742-6596/1491/1/012036

Singamurti, M., Yamtinah, S., Utomo, S., & Ashadi, M. (2017). Development of two-tier multiple choice question assessment instruments for measuring science process skills global warming. In International Conference on Teacher Training and Education 2017 (pp. 545–551). Atlantis Press. https://doi.org/10.2991/ictte-17.2017.62

Sıbıç, O., Akçay, B., & Arik, M. (2022). Review of two-tier tests in the studies: Creating a new pathway for development of two-tier tests. International Journal of Contemporary Educational Research, 7(2), 81–98. https://doi.org/10.33200/ijcer.747981

Šmida, D., & Čipková, E. (2021). Inquiry skills of primary school pupils in Slovakia. In*4th ICTLE, Proceedings of the 4th International Conference on Teaching, Learning and Education* (pp. 84–97). Retrieved August 3, 2025, from https://www.dpublication.com/wp-content/uploads/2021/08/31-6641.pdf

Šmida, D., Čipková, E., & Fuchs, M. (2024). Developing the test of inquiry skills: Measuring the level of inquiry skills among pupils in Slovakia. *International Journal of Science Education, 46*(1), 73–108. https://doi.org/10.1080/09500693.2023.2219811

Šmida, D., Drozdíková, A., & Nechajová, R. (2025). Inquiry skills and four-tier test – results. Figshare. https://doi.org/10.6084/m9.figshare.30018943.v1

Song, Y. (2016). We found the 'black spots' on campus on our own": Development of inquiry skills in primary science learning with BYOD (Bring Your Own Device). *Interactive Learning Environments, 24*(2), 291–305. https://doi.org/10.1080/10494820.2015.1113707

Sreenivasulu, B., & Subramaniam, R. (2013). University students' understanding of chemical thermodynamics. *International Journal of Science Education, 35*(4), 601–635. https://doi.org/10.1080/09500693.2012.683460

Sreenivasulu, B., & Subramaniam, R. (2014). Exploring undergraduates' understanding of transition metals chemistry with the use of cognitive and confidence measures. *Research in Science Education, 44*(6), 801–828. https://doi.org/10.1007/s11165-014-9400-7

Subali, B., Kumaidi, K., Aminah, N. S., & Sumintono, B. (2019). Student achievement based on the use of scientific method in the natural science subject in elementary school. *Jurnal Pendidikan IPA Indonesia, 8*(1), 39–51. https://doi.org/10.15294/jpii.v8i1.16010

Taban, T., & Kiray, S. A. (2022). Determination of science teacher candidates' misconceptions on liquid pressure with four-tier diagnostic test. *International Journal of Science and Mathematics Education, 20*(8), 1791–1811. https://doi.org/10.1007/s10763-021-10224-8

Tamir, P., & Lunetta, N. V. (1981). Inquiry related tasks in high school science laboratory handbooks. *Science Education, 65*(5), 477–484. https://doi.org/10.1002/sce.3730650503

Tannenbaum, R. S. (1969, February 8). *The Development of the Test of Science Processes* [Paper presentation]. National Association for Research in Science Teaching, Pasadena, CA, United States. https://files.eric.ed.gov/fulltext/ED027222.pdf

Tanti, T., Kurniawan, D. A., Wirman, R. P., Dari, R. W., & Yuhanis, E. (2020). Description of student science process skills on temperature and heat practicum. *Jurnal Penelitian Dan Evaluasi Pendidikan, 24*(1), 88–101. https://doi.org/10.21831/pep.v24i1.31194

Taslidere, E. (2016). Development and use of a three-tier diagnostic test to assess high school students' misconceptions about the photoelectric effect. *Research in Science & Technological Education, 34*(2), 164–186. https://doi.org/10.1080/02635143.2015.1124409

Temiz, B. (2020). Assessing skills of identifying variables and formulating hypotheses using scenario based multiple choice questions. *International Journal of Assessment Tools in Education, 7*(1), 1–17. https://doi.org/10.21449/ijate.561895

The New Zealand Curriculum (2017). *The New Zealand curriculum – Science*. Ministry of Education. Retrieved January 18, 2026, from https://newzealandcurriculum.tahurangi.education.govt.nz/5637209897.p?activeTab=tab:3

Timmermann, D., & Kautz, C. H. (2015). Multiple choice questions that test conceptual understanding: A proposal for qualitative two-tier exam questions. In *2015 ASEE Annual Conference & Exposition* (pp. 26–1179). American Society for Engineering Education. https://doi.org/10.18260/p.24516

Tobin, K. G., & Capie, W. (1982). Development and validation of a group test of integrated science processes. *Journal of Research in Science Teaching, 19*(2), 133–141. https://doi.org/10.1002/tea.3660190205

Tosun, C. (2019). Scientific process skills test development within the topic 'matter and its nature' and the predictive effect of different variables on 7th and 8th grade students' scientific process skill levels. *Chemistry Education Research and Practice, 20*(1), 160–174. https://doi.org/10.1039/C8RP00071A

van den Berg, E. (2013). The PCK of laboratory teaching: Turning manipulation of equipment into manipulation of ideas. *Scientia in Educatione, 4*(2), 74–92. https://doi.org/10.14712/18047106.86

Verma, P., & Choudhuri, R. (2025). Identification of Students' Misconceptions in Biology through Two-tier Diagnostic Test. *National Journal of Education, 23*(1), 1–13. Retrieved August 7, 2025, from https://bhu.ac.in/Images/files/1(20).pdf

Wahyuni, N., Bhakti, Y. B., Mutakin, T. Z., & Astuti, I. A. D. (2021). The development of four-tier diagnostic test instrument to identify the learners' misconception on circular motions. *Impulse:*

Journal of Research and Innovation in Physics Education, 1(1), 24–31. https://doi.org/10.14421/impulse.2021.11-03

Wang, J., Guo, D., & Jou, M. (2015). A study on the effects of model-based inquiry pedagogy on students' inquiry skills in a virtual physics lab. *Computers in Human Behavior, 49*, 658–669. https://doi.org/10.1016/j.chb.2015.01.043

Wen, C. T., Liu, C. H. C. H., Chang, H. Y., Chang, C. H. J.J., Chang, H., Chiang, S. H. F., Yang, C. H. W., & Hwang, K. (2020). Students' guided inquiry with simulation and its relation to school science achievement and scientific literacy. *Computers & Education, 149*, 1–14. https://doi.org/10.1016/j.compedu.2020.103830

Wenning, C. J. (2005). Levels of inquiry: Hierarchies of pedagogical practices and inquiry processes. *Journal of Physics Teacher Education Online*, *2*(3), 3–12. Retrieved August 7, 2025, from https://higherlogicdownload.s3.amazonaws.com/APS/379bd548-10a1-4054-aec2-911470db8df9/UploadedImages/Documents/levels_of_inquiry.pdf

Wenning, C. J. (2006). Assessing nature-of-science literacy as one component of scientific literacy. *Journal of Physics Teacher Education Online, 3*(4), 3–14.

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. *Journal of Physics Teacher Education Online, 4*(2), 21–24.

Wenning, C. J. (2010). Levels of inquiry: Using inquiry spectrum learning sequences to teach science. *Journal of Physics Teacher Education Online, 5*(3), 11–20.

Widiyatmoko, A., & Shimizu, K. (2018). The Development of Two-Tier Multiple Choice Test to Assess Students' Conceptual Understanding about Light and Optical Instruments. *Jurnal Pendidikan IPA Indonesia, 7*(4), 491–501. https://doi.org/10.15294/jpii.v7i4.16591

Wu, H. K., & Hsieh, C. E. (2006). Developing sixth graders' inquiry skills to construct explanations in inquiry-based learning environments. *International Journal of Science Education, 28*(11), 1289–1313. https://doi.org/10.1080/09500690600621035

Wu, M., Tian, P., Sun, D., Feng, D., & Luo, M. (2025). Evaluating students' conceptual understanding of isomers based on a four-tier diagnostic tool in upper secondary schools. *International Journal of Science and Mathematics Education, 23*(4), 907–947. https://doi.org/10.1007/s10763-024-10494-y

Xia, Y. (2020). Correlation and association analyses in microbiome study integrating multiomics in health and disease. *Progress in Molecular Biology and Translational Science, 171*, 309–491. https://doi.org/10.1016/bs.pmbts.2020.04.003

Xiao, Y., Han, J., Koenig, K., Xiong, J., & Bao, L. (2018). Multilevel Rasch modeling of two-tier multiple choice test: A case study using Lawson's classroom test of scientific reasoning. *Physical Review Physics Education Research, 14*(2), Article 020104. https://doi.org/10.1103/PhysRevPhysEducRes.14.020104

Yamtinah, S., Indriyanti, N. Y., Saputro, S., Mulyani, S., Ulfa, M., Mahardiani, L., Satriana, T., & Shidiq, A. S. (2019). The identification and analysis of students' misconception in chemical equilibrium using computerized two-tier multiple-choice instrument. *Journal of Physics: Conference Series, 1157*, Article 042015. https://doi.org/10.1088/1742-6596/1157/4/042015

Yan, Y. K., & Subramaniam, R. (2018). Using a multi-tier diagnostic test to explore the nature of students' alternative conceptions on reaction kinetics. *Chemistry Education Research and Practice, 19*(1), 213–226. https://doi.org/10.1039/c7rp00143f

Yang, D. C. (2022). Investigating the differences between confidence ratings in the answer and reason tiers in fourth graders via online four-tier test. *Studies in Educational Evaluation, 72*, Article 101127. https://doi.org/10.1016/j.stueduc.2022.101127

Yang, D.-C., & Lin, Y.-C. (2015). Assessing 10- to 11-year-old children's performance and misconceptions in number sense using a four-tier diagnostic test. *Educational Research, 57*(4), 368–388. https://doi.org/10.1080/00131881.2015.1085235

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation, 81*(12), 2141–2155. https://doi.org/10.1080/00949655.2010.520163

Yuliati, L., Yogismawati, F., Purwaningsih, E., & Affriyenni, Y. (2021). Concept acquisition and scientific literacy of physics within inquiry based learning for STEM education. *Journal of Physics: Conference Series, 1835*(1), 1–7. https://doi.org/10.1088/1742-6596/1835/1/012012

Yusoff, M. S. B. (2019). ABC of Content Validation and Content Validity Index Calculation. *Education in Medicine Journal, 11*(2), 49–54. https://doi.org/10.21315/eimj2019.11.2.6

Yusrizal, Y., & Halim, A. (2017). The effect of the one-tier, two-tier, and three-tier diagnostic test toward the students'confidence and understanding toward the concepts of atomic nuclear. *Unnes Science Education Journal, 6*(2), 1593–1600. Retrieved August 15, 2025, from https://journal.unnes.ac.id/sju/usej/article/view/15856

Zeidan, A. H., & Jayosi, M. R. (2015). Science process skills and attitudes toward science among Palestinian secondary school students. *World Journal of Education, 5*(1), 13–24. https://doi.org/10.5430/wje.v5n1p13

Zhao, C., Zhang, S., Cui, H., Hu, W., & Dai, G. (2021). Middle school students' alternative conceptions about the human blood circulatory system using four-tier multiple-choice tests. *Journal of Biological Education, 57*(1), 51–67. https://doi.org/10.1080/00219266.2021.1877777