Chemistry Education Research and Practice

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5rp00372e

Layers of competence: analyzing 8th graders' perceptions of visual and verbal assessment questions on chemical processes *vis-à-vis* their performance

Einat Ben-Eliyahu and Elon Langbeheim *

The design of assessments shapes what we learn about students' conceptual understanding. In the context of chemistry education, visual representations are important components of learning and assessment. To examine the role of representations in assessing students' reasoning about chemical and physical changes, we developed two multiple choice questionnaires: one that represents the choice options in verbal form and one that represents isomorphic options – in pictures. The questionnaires included a second-tier rating scale of the questions' general comprehensibility and the clarity of the pictures. The questionnaires were distributed among 374 eighth graders in two phases. In the first phase we found that on average, students performed slightly better on the verbal version, and some verbal items were significantly easier than their visual counterparts, but one item showed the opposite trend. Interviews revealed that visual representations exposed a wider range of ideas among respondents, and in some cases, created confusion. The second phase focused on the visual version and revealed significant positive correlations between clarity judgements and performance in questions with visuals such as graphs that represent the change in mass and molecular structures that represent the chemical formula, and no correlations on others. The analysis of the interviews, together with the clarity ratings, indicates that in these questions, visuals can be conceived as an additional layer of challenge, while other questions entail conceptual misunderstandings that are either exposed or concealed by cues in the external, visual layer.

Received 7th October 2025,
Accepted 17th December 2025

DOI: 10.1039/d5rp00372e

rsc.li/cerp

Introduction

Visual representations, such as illustrations (Nussbaum and Novick, 1982) and computer simulations (*e.g.*, Levy, 2013; Langbeheim and Levy, 2018), are effective and commonly used instructional tools for learning science in general and chemistry in particular (Kozma *et al.*, 2000; Vojíř and Rusek, 2022). Specifically, ball-and-stick illustrations that visualize molecular structures and diagrams that show their interactions are essential tools for explaining macroscopic phenomena, using underlying chemical mechanisms (Eilam and Gilbert, 2014; Talanquer, 2022). Indeed, such visual representations help students to understand the movement of particles and interactions between them (Nussbaum and Novick, 1982; Adadan *et al.*, 2009; Adadan, 2013).

Representations – whether visual or verbal – are ''external'', since they refer to core concepts or ''referents'' (Rau, 2017) such as interparticle interactions, conservation laws or terms such as ''reactants'' or ''mixtures''. Scientific competence is based on understanding these concepts, but since concepts are related to representations, it is crucial to be also competent in using representations (Kozma and Russell, 2005). Representational competence (RC) is demonstrated by the ability to identify and analyze key features of a representation or to select an alternative representation that conveys the same information as the original (Daniel *et al.*, 2018; Küchemann, *et al.*, 2021). Furthermore, the visual representations themselves often include symbolic components that require prior acquaintance or specific explication (Rau, 2017; Tonyali *et al.*, 2023). Thus, developing competence in a scientific domain involves learning to relate representational features to domain-relevant concepts. For example, assessments that entail visuals of the shells and electron occupancy depicted in the Bohr model and ask students to infer the identity of the atom (Rau, 2015) evaluate both students' competence in applying core chemistry concepts and their representational competence (Gkitzia *et al.*, 2020; Ralph and Lewis, 2020).

Domain relevant concepts – conservation of mass

Competence in chemistry relies on a coherent understanding of core concepts such as chemical change and the law of conservation of mass. Recognizing that chemical change

School of Education, Ben-Gurion University of the Negev, Beer-Sheva, Southern, Israel. E-mail: einatbe@post.bgu.ac.il, elonlang@bgu.ac.il

involves the rearrangement of atoms to form new substances and that atoms are neither created nor destroyed in the process is essential for developing this competence. However, conservation of matter, though fundamental, is not intuitive for students. Piaget and Inhelder (1974) showed that children often struggle with conservation: when observing sugar ''disappearing'' in water, many incorrectly predicted that the mass of the sugar solution would be lower than the combined mass of the solid sugar and water. That reliance on the perceptual experience of disappearance prevented young children from fully grasping the conservation principle. Stavy (1990) found similar issues regarding the process of evaporation, noting that students perceive evaporation as disappearance unless they see the vapor. Özmen and Ayas (2003) found that many tenth-grade students believed that the mass of products in a chemical reaction changes according to the phase of the substances involved. They reached the erroneous conclusion that mass changed even though the chemical reaction happens in a closed system, because the products or reactants were gases (Özmen and Ayas, 2003). Similarly, many students failed to include the mass of oxygen when calculating the mass of exhaust gas produced from burning petrol and oxygen (Barker and Millar, 1999). These misunderstandings persisted despite repeated teaching efforts.

Misconceived reasoning can reflect either naïve or intuitive ideas or fragmented knowledge. According to Vosniadou's (2013) framework theory, when students learn new scientific ideas (*e.g.*, conservation of mass in chemical reactions), they attempt to integrate this new information into their existing knowledge structures, and thereby may develop fragmented, ''synthetic'' concepts. In our case, fragmented knowledge is evident when students state that ''matter is conserved in chemical reactions'' but fail to apply conservation when the reaction produces gas that seems to ''disappear'', even though the system is closed. Synthetic models differ from ''misconceptions'' since the latter are held with high confidence, while synthetic ideas appear less coherent and students are less confident in stating them (Planinic *et al.*, 2006). One way to identify such fragmented knowledge is by using isomorphic versions of questions that address the same concepts and share the same solution processes (Simon and Hayes, 1976). Although their conceptual core is the same, isomorphic questions use different representations for the choice options. For example, when assessing the concept of Newton's 3rd law in Susac *et al.* (2023), one version included images of two cars that collide and the arrows representing the forces the cars exert, whereas the graph version addressed the same collision scenario, but did not illustrate the cars, nor the direction of the forces. Susac *et al.* (2023) did not find significant differences in performance between the visual, verbal or graphical versions, but another study found that performance on items that represent forces using vectors was lower than that on isomorphic items representing forces using bar charts or verbal descriptions (Nieminen *et al.*, 2010). Such variations in performance due to changes in representations indicate either fragmented knowledge or limited representational competence.

The roles of representations in eliciting reasoning

Visual and symbolic representations may differ in the type and amount of detail they convey. Visual representations may attract attention to superficial features, while verbal or symbolic notations often require acquaintance with symbols and scientific vocabulary. Thus, the relative complexity of each format depends on context, task demands, and learners' prior knowledge (Talanquer, 2022). Indeed, several studies comparing the performance of physics students on isomorphic multiple-choice items with different external representations have shown that performance varies significantly depending on the type of representation used (Kohl and Finkelstein, 2005; Meltzer, 2005; Nieminen *et al.*, 2010).

The inclusion of pictures of familiar situations in test items can reduce response time and improve accuracy, particularly when the task requires applying relationships between elements, especially among younger children (Sass *et al.*, 2012). However, this effect is not universal: in more complex items with less familiar visuals, additional pictorial information may increase processing time, and the benefits depend on factors such as task complexity and the familiarity of the picture. There are several examples where features of visuals in assessments can mislead students (*e.g.*, Kozma *et al.*, 2000) and decrease performance (Ralph and Lewis, 2020). For example, Bruner *et al.* (1966) showed that children's predictions of the amount of liquid after pouring from narrow to wide beakers were significantly better when the beakers were covered with a screen than when they were on display. They explained that children tend to rely on misleading features of visual displays – the height of the water in a narrow/wide beaker, rather to the logic of conservation, and defined this tendency as 'perceptual seduction' (Bruner *et al.*, 1966). Thus, inclusion of visual representations in assessments has a potentially important role in revealing fragmented conceptual understandings. To conclude, responses to assessment items reflect both (mis)understanding of the conceptual core, and (mis)understanding the representations that refer to it.

Despite extensive use of visual representations in science education, studies of their role in assessments of basic chemistry concepts are limited (Ralph and Lewis, 2020; Langbeheim *et al.*, 2022, 2023). But even within this limited context, findings reveal that visual representations have mixed influences on performance. Items asking middle schoolers to choose a visual molecular representation that matches a macro phenomenon were generally easier than verbal descriptions of the same molecular mechanism (Langbeheim *et al.*, 2023). However, college students' performance on questions that addressed mole ratios was lower when visual, particulate representations replaced verbal statements with symbolic representations (Ralph and Lewis, 2020). Furthermore, when relating symbolic representations to visual ones (*e.g.*, selecting molecular pictures that represent a chemical formula) the distracting visual options increase the difficulty of the items, compared to items where the picture is shown, and the selection is between symbolic options (Lin *et al.*, 2016; Gkitzia *et al.*, 2020). Does this ''asymmetric'' gap in performance indicate fragmented

knowledge of core concepts, or limited representational competence? These mixed results and open questions regarding the role of visual representations in assessing student understanding of chemical reactions, especially among young learners – require further investigation.

Methodological frameworks: classical test theory and metacognitive rating scales

Classical test theory (CTT) can be used to determine the extent to which assessment items measure their intended constructs (Doran, 1980; Ding and Beichner, 2009). Measures such as item difficulty and discrimination can be used to compare whether isomorphic questions elicit different reasoning patterns. Another way to evaluate assessment items is through two-tier multiple-choice (MC) questions that ask the respondents to rate their confidence in the correctness of their response (Lin and Singh, 2013). Confidence judgements in two-tier MC items measure metacognitive monitoring that reflects students' awareness of the limitations of their knowledge (Stankov *et al.*, 2012). Confidence measures reveal the Dunning–Kruger effect, where individuals with limited knowledge or expertise in a particular field tend to overestimate their abilities (Kruger and Dunning, 1999). They are also used for identifying multiple choice questions with high ratings of confidence related to choosing specific *incorrect* responses that represent persistent misconceived reasoning (Planinic *et al.*, 2006; Brandriet and Bretz, 2014). While confidence ratings are the most common measures of metacognitive monitoring, measures of the perceived *difficulty* of the assessment content are also important for selection decisions, such as switching from the initial choice to the correct response (Tiffin-Richards *et al.*, 2022). Such judgements of difficulty and complexity of learning tasks are commonly used in studies of reading and text comprehension (Ozuru *et al.*, 2012; Follmer and Clariana, 2022) and are shown to be positively correlated with the test-takers' performance on several types of visual and conceptual questions (Hoch *et al.*, 2023).

Another way to interpret perceived difficulty/complexity ratings of learning or assessment tasks is through the lens of cognitive load theory. Cognitive Load Theory (CLT) examines cognitive processing by differentiating intrinsic load, determined by the complexity of the content *vis-à-vis* the level of the students, and extraneous load, which is related to the presentation of information in the instructional or assessment design (Sweller *et al.*, 2011; Hoch *et al.*, 2023). For example, when instructional visuals are too detailed or contain redundant, or irrelevant information, they induce extraneous load and may hinder performance (Butcher, 2006; Joo *et al.*, 2021). To conclude, including visual representations in multiple choice chemistry assessments may impart extraneous cognitive load or can reveal flawed or fragmented reasoning. However, no study examined whether judgements of the question content and visuals can be used to separate challenges related to the conceptual core of the questions and those related to the cognitive load imparted by the details of the representations.

Scope of the current study

In the current study, we examine students' responses to conceptual questions in which information is represented by different representations: symbolic (chemical equations), visual/pictorial (molecular and macroscopic illustrations), and graphical-symbolic (*e.g.*, change-over-time graphs). Specifically, we focus on the law of conservation of matter in physical and chemical processes, and on the identification of reactants and products in a chemical reaction equation. For example, we examine whether students believe that the total mass of materials changes when gases are produced in a chemical reaction, even in a closed system (Özmen and Ayas, 2003) and whether they respond differently when the response options are presented visually in pictures, rather than as verbal statements.

We view the responses to verbal and visual items as reflecting two potential sources of difficulty: conceptual understanding and representational issues that affect the comprehension of the question. To disentangle the two, we used both interviews in which students explain their reasoning, and a rating scale in which they assess their comprehensibility of the question and the visuals in the response options. Unlike confidence rating scales that address the entire response process (comprehending the question, interpreting the visuals, and applying conceptual knowledge) as one combined judgment, the clarity-of-question scale focuses specifically on students' perceptions of how clear or comprehensible the item is. By separating students' perceptions of the clarity-of-question from answer correctness, the scale allowed us to identify the source of errors more precisely: items with high clarity ratings by many students that responded incorrectly suggest that students understood the representation but struggled with the underlying concepts. Clarity ratings that correlate with performance suggest that the question's representation imposed additional difficulty, due to low representational competence, or overly complex visuals. Thus, the clarity scale can help to determine whether students' difficulties arise from conceptual misunderstanding, representational challenges, or a combination of both. Triangulating the correlational analysis with interview data helped us to determine the sources of difficulty in different items.

Research questions

1. What are the similarities and differences in 8th graders' performance on and clarity ratings of isomorphic verbal and visual assessment items addressing chemical and physical processes in matter?

2. How are 8th graders' clarity ratings (of questions and illustrations) correlated with their performance on specific items, and what are the aspects that characterize items with significant correlations?

3. What are the main reasons for large performance gaps between verbal and visual versions, and the correlations between perceived clarity ratings and performance in some of the assessment items, according to students' reasoning in interviews?

Methodology

Design and evaluation of questionnaires

The research entailed two data collection phases: Phase 1: two questionnaires were developed with isomorphic verbal and visual items (see Appendices A1 and A2). The isomorphic items shared the same stem (*i.e.*, the introductory statement of the question) but differed in the format of the response options. In one format, the response options were verbal statements; in the other, they were pictorial illustrations. The stem itself included either a verbal description accompanied by a symbolic chemical equation or a visual representation of a phenomenon. All items also included a second-tier rating component, in which students evaluated the clarity of the question. In addition, we conducted interviews with a sample of 11 students, focusing on assessment items with large differences between the visual and verbal formats.

Phase 2: We re-administered the visual version of the questionnaire, this time with an additional rating scale specifically addressing the clarity of the illustrations in the response options (see Fig. 2). This addition allowed us to distinguish the comprehensibility of the visual representations and the clarity of the question prompt that pertained to its conceptual core.

In developing the assessment items, we sought to represent the same ideas visually and verbally in the response choices but acknowledge that the two formats cannot be entirely equivalent. We use the term *isomorphic* to indicate that each format addressed the same underlying conceptual framework, consistent with other studies comparing performance across representational formats (Kohl and Finkelstein, 2005; Nieminen *et al.*, 2010; Susac *et al.*, 2023). As is common in chemistry assessments, some items asked students to relate a macro-level representation to a molecular or particle-level one, while others presented symbolic chemical equations and asked for the appropriate submicro/nano representation (Gkitzia *et al.*, 2020). In this study, items were classified according to the level of representation addressed in the response options. Macro-level items depicted observable, everyday phenomena (*e.g.*, boiling water and melting butter). Nano-level items involved submicroscopic representations of particles and molecular structures (*e.g.*, diagrams of particle motion or spacing in different states of matter). Symbolic items presented chemical equations that required interpreting the reactants and products in the conventional chemical notation. This classification follows common frameworks in chemistry education research (*e.g.*, Johnstone, 1991; Gkitzia *et al.*, 2020).

In each isomorphic pair, the stem of the question contained text and sometimes also a picture was identical across the verbal and visual formats. Fig. 1 illustrates such an item: the stem presents a description and picture of the decomposition of mercuric oxide. In the visual format, the prompt asked: *"Which of the following illustrations correctly depicts the mass of the products at the end of the experiment?"* followed by four pictorial options comparing the mass of products to that of the reactants. The verbal format presented the same stem but differed in the prompt ("What is correct to say about the total mass of the products?"), with response options expressed in text.

The questionnaires were based on the chemistry curriculum for grades 7 and 8 in Israel that addresses the law of conservation of mass, phase change, diffusion, mixing, and chemical reactions. Each questionnaire initially consisted of 15 items, some of which were derived from previous research (Hadenfeldt *et al.*, 2016; Langbeheim *et al.*, 2023) and others (*e.g.*, the NH_3 item, Fig. 4) from national tests for the 8th grade. Both visual and verbal versions of the items were reviewed for evidence in support of content validity by a panel of science teachers with over 10 years of teaching experience and three experts in science education. The reviewers were asked to verify the alignment between the verbal and visual response options and to comment on the quality of the items in terms of their alignment with 8th-grade curricular expectations. After reviewing, three items were removed since they required knowledge beyond the 8th grade curriculum, which reduced the questionnaire to a total of 12 items. A detailed table specifying the source of each questionnaire item is provided in Appendix A4.

To assess the comprehensibility of the items, we added a "clarity" rating scale after each item. The clarity scale ranged from 1 (unclear) to 5 (very clear), as shown in Fig. 2.

Data collection

The study adopted a mixed-methods approach that combined quantitative and qualitative strands. The quantitative analyses first identified differences between representation format, and relationships between performance and clarity ratings. These statistical patterns guided the selection of specific items for the qualitative phase, in which interviews explored students' reasoning in light of the representations in these specific items. The qualitative insights were then used to interpret and clarify the sources of the trends revealed by the quantitative results.

Phase 1: The verbal and visual questionnaires were administered to 200 eighth-grade students from schools with similar socioeconomic backgrounds. The questionnaires were distributed online using the Qualtrics platform. Each class responded to either the verbal or the visual version. To ensure comparability of the student samples across formats, the survey included background rating items about the frequency of use of different teaching practices (*e.g.*, *"My teacher uses computer simulations to show the movement of particles"*). The scale ranged from 1 ("almost every lesson") to 5 ("never"). Entire classes were assigned to one questionnaire format, so that all students within a class received either the verbal or the visual version. The assignment of classes to questionnaire formats was random and not based on students' prior achievement, nor socioeconomic background. Socioeconomic similarity across schools was determined based on the Ministry of Education classifications of school catchment areas.

Phase 2: The revised visual version (with an additional clarity scale for illustrations) was administered to 174 eighth-grade students from two schools. The rationale for Phase 2 was to disentangle potential sources of difficulty in the visual format. In Phase 1, the clarity scale referred to the question

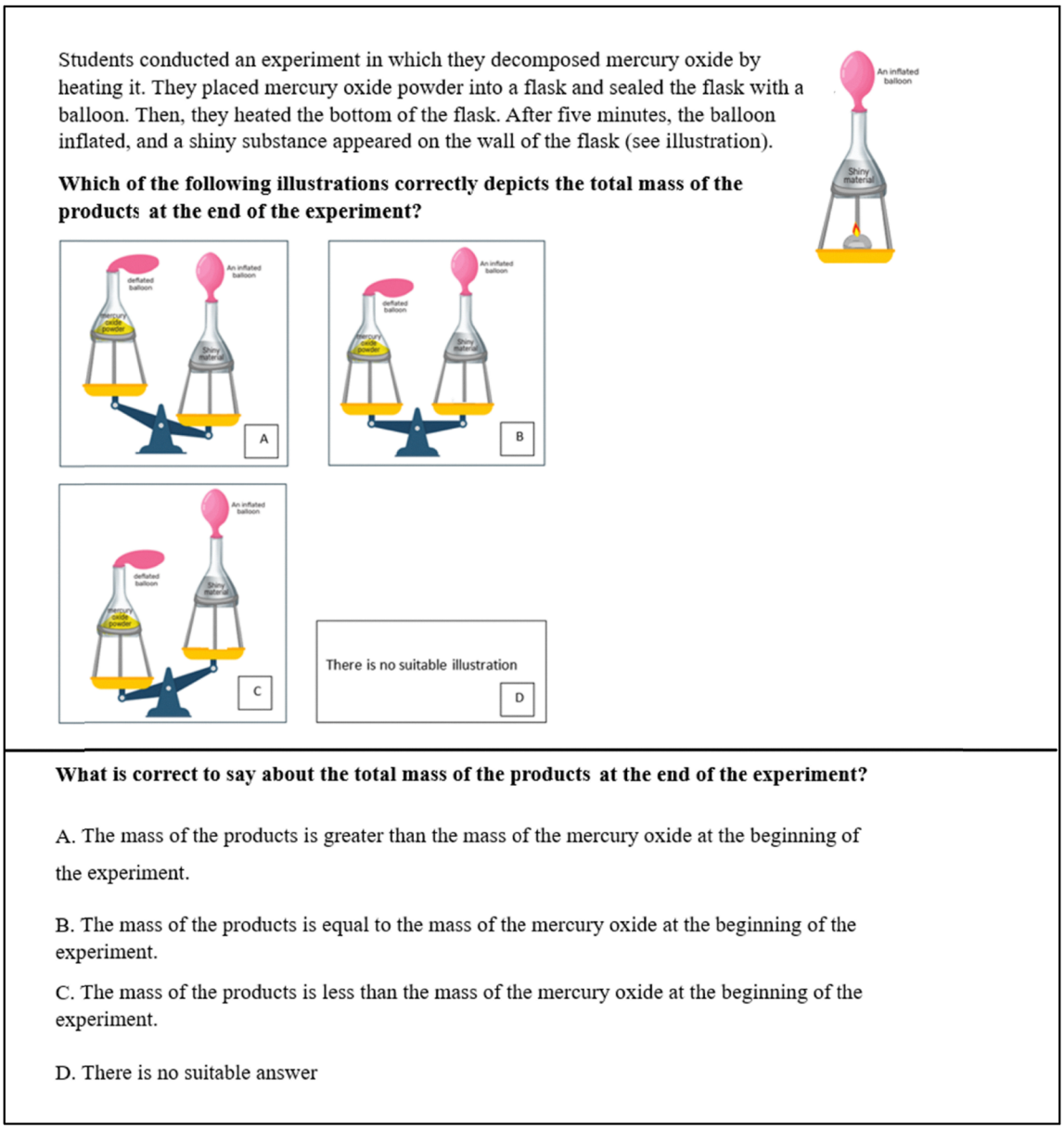

Fig. 1 Decomposition of mercuric oxide by heating – an item requiring students to apply conservation of mass. Visual version (top) and verbal version (bottom).

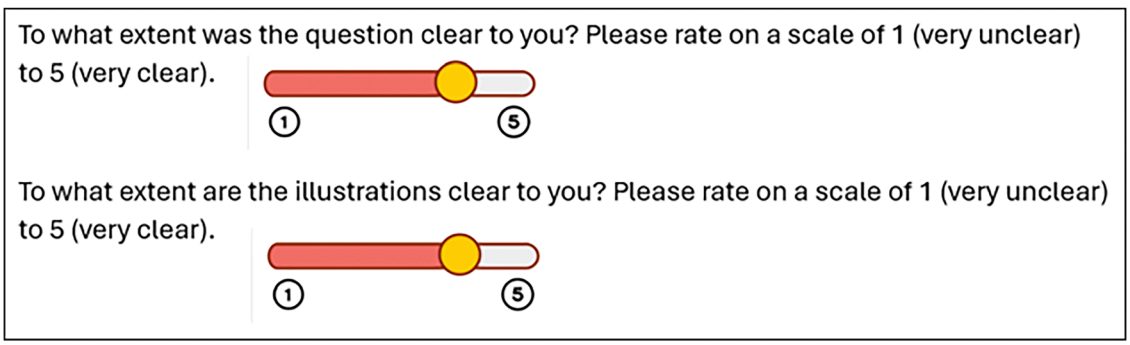

Fig. 2 Clarity rating scales used to evaluate both questions and illustrations (1 = unclear, 5 = very clear).

as a whole, making it impossible to distinguish whether low clarity ratings reflected the phrasing of the prompt or the comprehensibility of the illustration. By adding a separate rating scale in Phase 2, we were able to differentiate between these two aspects and to identify whether students' difficulties stemmed from the wording of the item or from interpreting the visual representation. This provided additional insight into the specific role of illustrations in shaping students' performance. The assessments were administered during regular science classes at school, under direct supervision of one of the researchers. Students completed the Qualtrics test individually on a voluntary basis; no incentives or grade implications were involved. Prior to participation, the importance of the study was explained, and students were encouraged to provide thoughtful responses. Students who did not wish to participate were given an alternative quiet classroom activity prepared by the teacher.

Interviews

To delve deeper into the differences in response patterns between the two formats, semi-structured interviews were conducted with a sample of 11 female students from one of the schools that participated in Phase 1 (see Appendix A3). The students who volunteered had *not* completed the verbal or visual questionnaires beforehand; therefore, their responses were independent of the questionnaire data collected in Phase 1. All interviews were conducted in a single-sex school, hence only female students were included. The participants were a convenience sample, selected due to the school's accessibility and the research team's prior working relationship with the school staff. We ensured that the group represents a variety of achievement levels in science (low, medium, and high), to reflect various conceptual levels. Similar sample sizes and selection strategies have been used in previous mixed-method studies in chemistry education (*e.g.*, Gkitzia *et al.*, 2020). Participation was voluntary for both methodological and ethical reasons, to promote openness and validity of responses during extended interviews. Each interview lasted approximately 20 minutes and was conducted by the first author within a few weeks of the quantitative data collection. The interviews were audio-recorded, transcribed verbatim, and analyzed through open coding (Corbin and Strauss, 1990) to identify recurring reasoning patterns and conceptual difficulties that could clarify the quantitative trends. Informed consent was obtained from both students and their parents.

Quantitative analysis

We used classical testing methods to examine the questionnaire data of phase 1 and phase 2. Specifically, we examined test internal consistency and item discrimination values. To assess the equivalence of the student samples that responded to the verbal and visual versions, we compared the teaching rating scales using non-parametric tests. Cronbach's alpha internal consistency was acceptable for performance scores in both the verbal format $\alpha = 0.733$ and the visual format $\alpha = 0.698$. The clarity ratings were also very reliable with $\alpha = 0.91$ for the verbal version and $\alpha = 0.92$ for the visual one. Table S1 (see supplementary information) shows that the discrimination of the items (their correlation with the total score) was acceptable too. Items such as ''bubbles in boiling water'', ''egg shells'', and ''NH_3'' exhibited lower discrimination indices, with the verbal item on the composition of boiling water bubbles having a marginal discrimination value (0.185). The literature (Ding and Beichner, 2009) suggests that a discrimination index of 0.2 or higher is generally considered acceptable (Doran, 1980). Despite the marginal discrimination, we did not remove the item from the questionnaire because the corresponding visual item had a discrimination of 0.337 with a significant correlation. The relatively low discrimination value for the verbal item can be attributed to the fact that it is particularly challenging and involves a common misconception, namely, that the gas in the bubbles is air and not water vapor (Osborne and Cosgrove, 1983). The rating statements referred to the frequency of teaching practices and specifically to visual instructional tools (*e.g.*, textbook illustrations, demonstrations, and computer simulations). This brief teacher–practice survey was included solely to verify that the participating classes had comparable exposure to visual teaching approaches, to ensure the equivalence of the groups. A comparison of the average ratings revealed no significant differences in teaching practices between the classrooms that completed the verbal questionnaire and those that completed the visual questionnaire (see Appendix A4).

The datasets from phases 1 and 2 were analyzed separately. Phase 2 was conducted almost a year after phase 1 and included slight revisions to the questionnaire: items with low reliability from phase 1 were removed and clarity-rating scales were added. Because of these changes, the datasets were not merged. Phase 1 results guided the selection of questions for the qualitative interviews.

Independent-sample *t*-tests were used to compare average total scores and average clarity ratings between the verbal and visual versions. In addition, non-parametric tests were used to compare performance and clarity ratings for individual items. Finally, gamma correlations between clarity ratings and performance were examined both for total scores and for individual items. Fig. 3 is a scatter plot of the clarity ratings of performance *vs.* the illustrations' clarity (Pic_clarity) for two items. Each point represents the result for a student who assigned the clarity rating shown on the *x* axis. The scatter plots include a small amount of random ''jitter'' to the position of data points to prevent them from overlapping. A score of 1 on the vertical axis represents a correct response, and a score of zero – an incorrect response.

As shown in Fig. 3, in the CO_2 item there is a clear increase in correct responses as clarity ratings rise, whereas in the Hg item no such trend is observed. Two separate gamma correlations were computed for each item from the phase 2 dataset: one between students' performance (1 = correct, 0 = incorrect) and the perceived clarity of the question (rating 1 to 5), and another between performance and the perceived clarity of the accompanying illustration. Gamma correlations are nonparametric measures of association used when both variables are

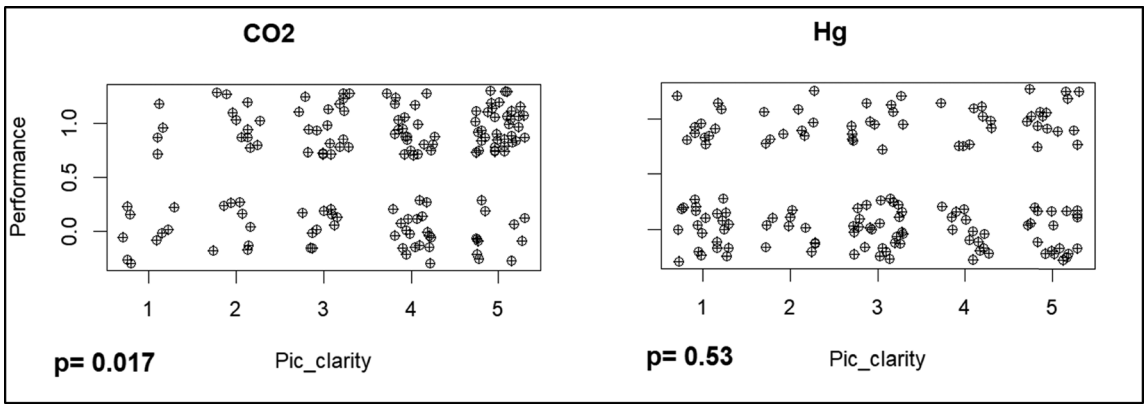

Fig. 3 "Jittered" scatter plots of clarity ratings and performance for two items. A clear increase of correct responses with higher rating appears in the CO_2 item (left), and no clear pattern in the Hg (right). The *p*-values represent the Wald-test of the gamma correlations.

ordinal, but the distances between categories are not necessarily equal—as in our case (Goodman and Kruskal, 1954).

Coding of qualitative data

Guided by Vosniadou's framework theory (2013), students' interview explanations were analyzed in terms of three forms of reasoning: naïve/intuitive (consistent misunderstanding), synthetic (fragmented), and scientific. According to this framework, students develop scientific understanding by reconstructing their naïve ideas. When new information is combined with their intuitive prior knowledge, they express *synthetic models* – inconsistent reasoning that combines elements of naïve ideas with fragments of the scientific model. For example, students who stated that the mass of mercury oxide increased or decreased during heating, while acknowledging conservation of mass in a different context, demonstrated such synthetic reasoning. These three categories served as the initial, theory-driven coding scheme. The categories were then refined inductively based on recurring reasoning patterns identified in the transcripts. Thus, the coding process combined *a priori*, framework-based definitions with data-driven adjustments, allowing Vosniadou's model to guide, rather than constrain, the analysis. To illustrate the *consistent misunderstanding* (naïve/intuitive) category, we quote students' who repeated the same incorrect response across both versions: *"The mass decreased because the products are lighter. That helped the balloon inflate. It's not in the bottle, it went up."* (Visual version) *"The mass decreased because it was heated; it kind of evaporated and melted, and that's what made the balloon inflate."* (Verbal). The *fragmented* (synthetic) category captured all instances of unstable reasoning, including shifts in either direction, *i.e.*, from incorrect (visual) to correct (verbal) as well as from correct (visual) to incorrect (verbal). For example, one student initially reasoned incorrectly in the visual version of the mercury oxide item: "*Air entered the balloon, so it weighs more, the mass increased."* but changed the explanation in the verbal version: *"The mass of the products is equal (to the mass of the reactants); they sealed it, so it did not increase or decrease."* In the *scientific* category, students consistently applied a scientifically valid explanation in the visual version: *"Because the particles did not leave, the mass did not change."* and the verbal one *e.g.*, *"Since nothing was added, it stayed the same. The particles did not leave"*. The two authors of this manuscript independently coded the transcripts. Inter-rater agreement was high, and discrepancies were resolved through discussion until full consensus was reached. Finally, the study was conducted in accordance with established ethical guidelines. Informed consent was obtained from both the students and their parents, and only participants who gave consent were included.

Results

The subsequent sections are organized according to the three research questions that guide this research. The first research question addressed the overall performance and clarity rating differences between students who responded to the verbal and visual versions of the assessment. The data in Table 1 show that students' performance on the verbal questionnaire is slightly higher compared to the performance on the visual version of the questionnaire, although the difference is not significant ($p = 0.177$). The average clarity ratings of the verbal version are also higher than those of the visual version, and this difference is also not significant ($p = 0.26$). While performance and clarity were slightly lower in the visual format, the correlation between the overall accuracy of items and clarity rating was slightly higher in the visual format ($r = 0.379$, $p = 0.008$) than that in the verbal format ($r = 0.322$, $p = 0.004$).

In phase 2, the average clarity ratings of the visual version were slightly higher than those of phase 1 ($M = 3.39$) and were higher than the clarity ratings of the illustrations ($M = 3.15$).

Table 1 Average total scores and self-reported clarity of the two questionnaire formats

	Visual ($N = 106$)	Verbal ($N = 94$)	Difference (SIG)
Clarity mean (SD)	3.27 (0.98)	3.44 (0.86)	$t = 1.31$, $p = 0.26$
Performance mean (SD)	0.437 (0.23)	0.482 (0.23)	$t = 1.35$, $p = 0.177$

The correlation between clarity and performance was similar to that of phase 1, ($r = 0.32$, $p = 0.012$) and the correlation between the comprehensibility of the illustrations and overall accuracy was lower ($r = 0.22$, $p = 0.048$). Unsurprisingly, the clarity ratings of the questions, and of the illustrations, were highly correlated with each other ($r = 0.77$, $p < 0.001$).

While overall averages of clarity ratings and performance provide a general overview, item-level analyses revealed how clarity and performance patterns differed across individual items and representation types.

Table 2 shows the average performance (percentage of correct responses) and the average clarity of the items, and the differences between formats, both in terms of performance and in terms of perceived clarity. The NH_3, butter and mercury oxide items show a significant difference in performance ($p < 0.001$, $p = 0.009$ and $p = 0.011$, respectively), with the verbal version yielding substantially higher accuracy than the visual version in the Hg and NH_3 items and lower in the melting butter item.

In addition, the tea, bubble and egg shell items exhibited significant differences in clarity ratings, with clearer verbal versions than the visual ones, but the average performance on these items was similar. Half of the items (*e.g.*, metal ball, balloon, and MgCl) did not differ significantly on either measure; this indicates that, overall, the representation format affected only a subset of items.

Table 3 summarizes the relationship between the clarity ratings of the questions and the illustrations to address the 2nd research question, namely how are clarity ratings (of questions and illustrations) correlated with their performance on specific items, and what are the aspects that characterize items with significant correlations? The clarity ratings resembled those of phase 1, with the bubble, butter and flower items rated as relatively clear questions, despite low performance on these items. Gamma correlations were used to relate question clarity (Table 3 – left) and illustration clarity (Table 3 – right) to students' accuracy.

The clarity of the illustrations was significantly correlated with performance only for the CO_2 ($p = 0.017$), NH_3 ($p = 0.003$), and egg shells in vinegar ($p = 0.0002$) items and the dissolving salt ($p = 0.07$) and metal ball ($p = 0.10$) were near significance.

Table 3 Average clarity ratings of items and gamma correlations between clarity and accuracy and significance for visual items in phase 2

	Question clarity			Illustration clarity		
	Mean	Gamma	*p*_value	Mean	Gamma	*p*_value
Ball (heat)	3.68	0.22	0.06*	3.78	0.19	0.10*
Balloon	3.08	0.15	0.28	3.32	0.04	0.79
Bubble	3.30	−0.01	0.96	3.58	0.02	0.89
Butter	2.82	0.38	<0.001**	3.16	0.19	0.14
CO_2	3.10	0.21	0.08*	2.81	0.29	0.02**
Egg shells	3.38	0.24	0.03**	3.24	0.47	<0.001**
Flower	3.66	0.11	0.39	3.28	0.04	0.79
Hg	3.53	0.16	0.24	2.89	0.08	0.53
MgCl	3.49	0.11	0.31	2.89	0.05	0.66
NH_3	2.89	0.32	0.01**	3.03	0.41	<0.001**
Salt	3.33	0.41	<0.001**	3.27	0.22	0.07*
Tea	3.1	0.29	0.024**	2.81	0.07	0.58

These items asked to either relate a chemical formula to a molecular illustration (CO_2, NH_3) or to select a graph of the change in mass during a chemical or physical process (metal ball, salt and egg shells). In these two contexts, the perceived clarity of the illustration seemed to be related to students' performance.

As for the clarity of the question ratings, salt ($\gamma = 0.41$; $p = 0.000$), tea ($\gamma = 0.29$; $p = 0.024$), NH_3 ($\gamma = 0.32$; $p = 0.012$), butter ($\gamma = 0.38$; $p = 0.004$), and egg shells ($\gamma = 0.29$; $p = 0.005$) demonstrated significant correlations with performance, while CO_2 ($\gamma = 0.21$; $p = 0.078$) and the metal ball item ($\gamma = 0.22$; $p = 0.06$) were marginally significant. In two items – dissolving tea and melting butter – performance was correlated with the clarity of the question, but not the clarity of the illustration—indicating that the difficulty experienced by some of the students stemmed from understanding the scenario as a whole, rather than decoding the image. In half of the items (*e.g.*, bubble, Hg, $MgCl_2$, balloon, and flower) we found no correlations with either the clarity of the question or the illustration. The correlations of these items are near zero, indicating no connection between clarity ratings and performance as shown in the right panel of Fig. 3. The large proportion of students who chose the incorrect response, but perceived the question as clear, implies that the external visual layer of the question was not perceived as problematic, even for students' with low conceptual misunderstanding.

Table 2 Item analysis: Pct correct and clarity in visual and verbal formats. Comparisons are based on the Mann–Whitney test

Item	Pct correct visual	Pct correct verbal	Sig (Mann–Whitney)	Avg clarity visual	Avg clarity verbal	Sig (Mann–Whitney)
Balloon (macro to nano)	0.75	0.681	0.287	3.564 (1.268)	3.690 (1.103)	0.474
MgCl (symbolic to macro)	0.543	0.484	0.408	3.250 (1.266)	3.229 (1.151)	0.907
CO_2 (symbolic to nano)	0.472	0.462	0.896	3.383 (1.201)	3.277 (1.193)	0.558
Salt (conservation of matter − graph)	0.543	0.677	0.054*	3.370 (1.512)	3.553 (1.150)	0.381
NH_3 (symbolic to nano)	0.295	0.670	<0.001**	3.253 (1.235)	3.138 (1.241)	0.537
Flower (macro to nano)	0.387	0.383	0.956	3.484 (1.138)	3.471 (1.228)	0.943
Tea (macro to nano)	0.481	0.468	0.854	2.830 (1.234)	3.694 (1.155)	<0.001**
Bubble (macro to nano)	0.267	0.234	0.597	2.912 (1.180)	3.671 (1.106)	<0.001**
Butter (macro to nano)	0.358	0.196	0.009**	3.292 (1.205)	3.581 (1.132)	0.097*
Hg (conservation of matter)	0.324	0.489	0.012**	3.056 (1.309)	3.012 (1.171)	0.818
Metal ball (conservation of matter − graph)	0.533	0.596	0.377	3.222 (1.130)	3.409 (1.283)	0.304
Egg shells (conservation of matter + graph)	0.365	0.473	0.127	2.807 (1.153)	3.256 (1.160)	0.011**

Analysis of specific items with significant differences in performance across formats

The third research question addressed specific items with large performance gaps between verbal and visual versions. The relation between representation format and student reasoning was unpacked in semi-structured interviews with a sample of 11 eighth grade students from one of the schools that participated both in phase 1. The visual items were discussed first, followed by the verbal format of the same item. The interviewees first chose their answer, and then were asked to explain their reasoning for choosing that answer. The data collected from the interviews are summarized in Table 4, with particular attention to the middle column, which captures students who changed their responses between the visual and verbal formats. These changes, or ''shifts'' indicate that the representation itself shaped the reasoning process. Because such shifts typically occur when students hold fragmented knowledge that is sensitive to surface features of the representation, we treated these shifts as key evidence of representational impact. For instance, when a student provided an incorrect or fragmented explanation in the visual version but a scientifically consistent explanation in the verbal version (or *vice versa*), this was coded as a shift.

Table 4 summarizes students' reasoning patterns across visual and verbal formats. The central square under the 'fragmented' category represents instances where students changed their explanations when the question format shifted from visual to verbal (or *vice versa*). It should be noted that consistency or fragmentation is not a stable attribute of particular students but depends on the representational context; for instance, a student might reason consistently about the mercury oxide item but show fragmented reasoning in the NH_3 question. Some association between students' achievement levels and their reasoning patterns was observed. For example, the three students classified as high-achievers, answered both verbal and visual versions of the mercury-oxide and butter items correctly, and two of them answered both versions of the NH_3 item correctly. Middle and lower-achieving students more often displayed fragmented or synthetic reasoning. These differences suggest that general academic achievement may

Table 4 Categories of student reasoning between visual and verbal formats (interview data)

Item	Correct % (visual)	Correct % (verbal)	Consistent-incorrect category (−vis → −ver)	Fragmented category (−vis → +ver) (+vis → −ver)	Correct category (+vis → +ver)
Hg	4/11	7/11	*''Because after they heated it, it was like it became easier... the material sort of turned into something that wasn't really weighed. Here it's lighter and the mass went down, because it was as if it no longer sat on the bottle. It's not in the bottle.''* → *''The mass of the products is smaller than the mass of the mercuric oxide. It's like the substances, because they were heated, kind of evaporated or melted''* (Student 2, medium achiever)	*''I think it's answer B. Because, if the balloon inflates it means air went in, and then it weighs more. Because, like, if the balloon inflates it means air entered it, and then there is more.''* → *''I think it's answer C. The mass of the products is equal to the mass of the reactants. Because, like, if they sealed it then it didn't go up or down, it kind of stayed the same.''* (Student 4, medium-achiever)	*''No material was added or removed, so there's no reason that after the experiment (the heating) the amount would decrease or increase. I think they're supposed to be equal, like, it's the same both at the beginning and after.''* → *''It's like no material was added or removed, so there's no reason for their mass to increase or decrease.''* (Student 1, high achiever)
Butter	4/8	4/8	*''It's not answer B because it shows that it's a solid... The answer is that it's something in between, in my opinion. And it's also not (answer D) because that's, like, really a liquid. A. It's answer C, between solid and liquid (a gas state).''*→ *''Answer C: they spread out and create large spaces between them.''* (Student 10, low achiever)	*''Answer A (the correct answer), because the arrows show that it is liquid and spreading to the sides. Distractor B is a solid, and distractor C looks too separated since the spaces there seem large.''* → *''Answer B'' is correct (description of the solid), because when the butter is liquid, the particles spread out more, and when it is solid they are like an atomic lattice and connected. Answer D (the correct one) is more like the middle of the process.''* (Student 6, high achiever)	*''Because, like, they're not like air that spreads everywhere. They do move and switch places with each other.''* → *''At first (solid) they were only moving in place and were closer together, and then they started moving more and had more space between them.''* (Student 8, high achiever)
NH3	2/11	9/11	*''I think it's answer B. Like, it seems more logical to me because here it's four H, and then the N has four H around it, and those (Cl) are single because there's only one chlorine each.''* → *''I think it's B: the reactants are molecules made of a nitrogen atom connected to four hydrogen atoms, and in addition, single chlorine atoms. Because again, it's basically the same thing. Here it turns into 4H, then they are connected to the N, and there's the Cl which is kind of alone.''* (Student 5, low achiever)	*''Answer C. Like, they connected, and it was kind of one, and then it connected with the other, and then it didn't really break apart, and so they are joined together.''* → *''Answer A. Because it says it's three hydrogens and one nitrogen, and the second molecule is one chlorine and one hydrogen.''* (Student 9, medium achiever)	*''The chemical symbol fits both types of molecules in the reactants.''* → *''Because there are two types of molecules, and in the end a solid was formed – one is nitrogen with three hydrogens, and the other is chlorine with hydrogen.''* (Student 1, high achiever)

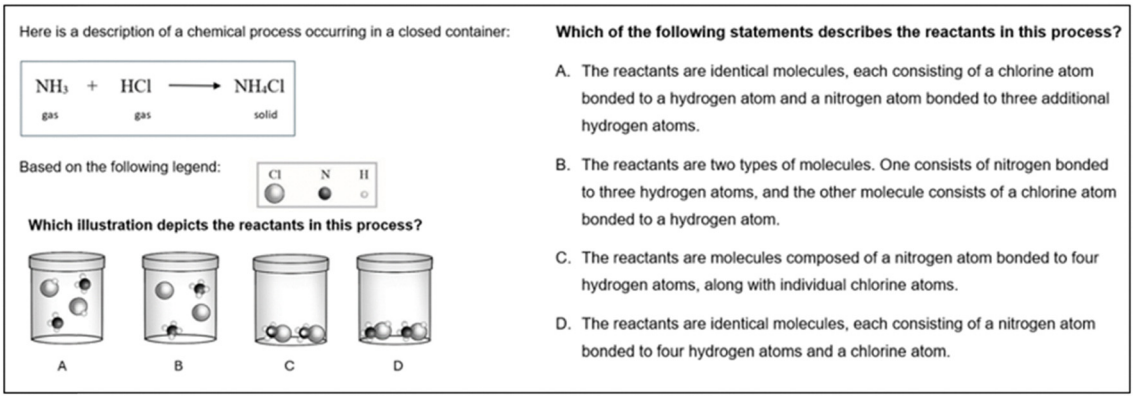

Fig. 4 Reaction of NH_3 and HCl items – visual format (left) and verbal format (right).

contribute to the stability and coherence of students' conceptual reasoning, although representational format still played a central role in shaping their responses.

The NH_3 + HCl reaction item (Fig. 4) showed a significant correlation between the perceived clarity of the visuals and accuracy (Table 3), and a significant difference between verbal and visual versions with 29.5% of students choosing the correct answer in the visual condition, compared to 67% on the verbal one (see Table 2). Similarly, in the interviews, only (2/11) responded correctly on the visual version and (9/11) on the verbal version. The common incorrect distractor for the visual item was a representation of the product of the reaction: two NH_4Cl molecules in a solid state (option D in Fig. 4). Students who answered correctly on the verbal version but incorrectly on the visual one revealed in the interviews that they knew that the reactants in the chemical equation are shown on the left side of the chemical equation (''It's NH_3; it has three hydrogens and HCl'' – Student 11) but still chose the distractor describing the product ''*because the gases combined into a solid; they are connected like a solid*'' (Student 7). It seems that for these students the details of the pictures representing the combined product, and the index (s) that represents a solid in the chemical equation created a confusion between reactants and products. In other words, they viewed the picture of NH_4Cl as the molecular manifestation of the expression NH_3 + HCl. This type of confusion can be interpreted as evidence of either cognitive load imparted by the visuals or limited representational competence that is elicited by the visual version.

The item concerning the decomposition of mercury oxide (Fig. 1) exhibited no correlation between perceived clarity and performance (Table 3), with 48.9% choosing correctly on the verbal version, and only 32.4% on the visual one (Table 2). This item illustrates the chemical reaction of mercury oxide powder that decomposes into mercury that gets adsorbed to the beaker's walls and gaseous oxygen that inflates a balloon. Similar to the larger sample, most interviewees (7/11) chose the correct response (that mass is conserved) on the verbal version compared to only (4/11) in the visual one. Three of the interviewees demonstrated fragmented concepts: while in the verbal version they referred to the conservation of mass correctly, in the visual version the additional details of the inflated balloon triggered the idea that the inflation of the balloon caused either an increase or a decrease in mass (the verbal format did not mention the inflated balloon). These students chose one of the pictures of the unbalanced scales, with the flask of the reactants and a deflated balloon on one side and the same vessel with the products and an inflated balloon on the other. The interviews revealed that the image of the inflated balloon elicited the idea that the mass had either decreased or increased, as stated by student 4: *''I think it's answer B. Because, like, if the balloon inflates it means air went in, and then it weighs more. . . .''*, while in the verbal version she explained: *''I think that the mass of the products is equal to the mass of the reactants. Because, like, if they sealed it then it didn't go up or down, it kind of stayed the same.''* (Student 4). Similarly, student 6 acknowledged the balloon in her response to the visual version: ''*I think the total mass is smaller because the balloon inflated; it turned into gas. So, because they heated it, it became lighter*'', and in the verbal version she chose the correct response and said: ''*It (the mass) might be equal because they just heated it and heating it doesn't make the mass heavier*''. The choice of the correct response seems to misinterpret the chemical process as mere heating that does not change the nature of the material, again revealing fragmented knowledge.

In the melting butter item (Fig. 5), students performed significantly better on the visual version (35.8% correct) than on the verbal one (19.6% correct), and the clarity of the item was correlated with students' performance (Table 3). The relatively low performance on this item indicates limited understanding of particle motion and configuration during phase change. Unlike the Hg and NH_3 questions, it also shows that the pictorial options can support and streamline students' reasoning. The interviews reveal that one reason for the difference between versions is the ambiguity regarding particle configuration in the melted butter, which is evident in the verbal statements regarding interparticle distances. For example, many students (55.2%) chose the incorrect option B: ''Butter particles move and spread out, creating large spaces between them'' although large distances between particles characterize gases not liquids. The large distances between particles were more salient in the

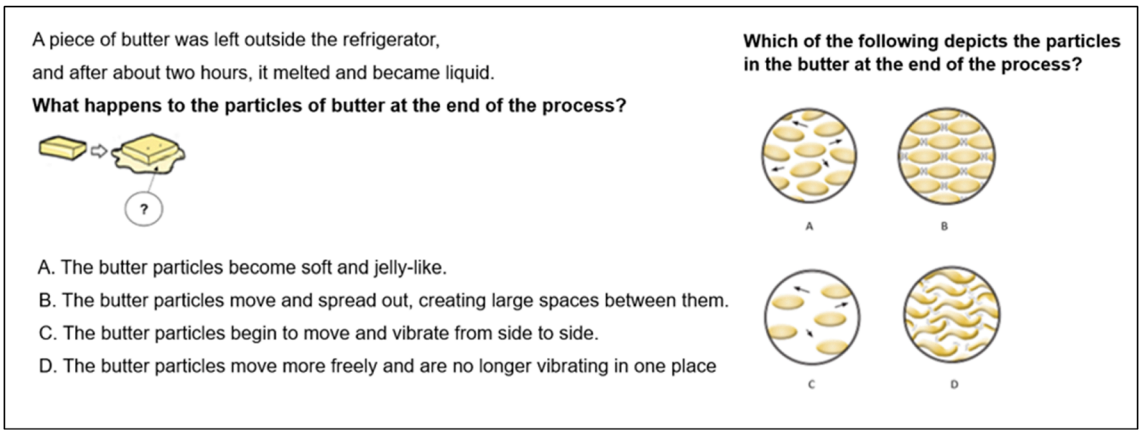

Fig. 5 Melting butter item – verbal format (left) and visual format (right).

visual representation and so most students rejected option "C" that represented spread out particles as in a gas with only 23% of the participants selecting this distractor. For example, student 6 stated: "*option C looks too separated, since the spaces there seem too large*" (Table 4). However, the interviews also reveal some hesitation regarding the actual phase of the butter that is depicted by the question scenario. This is evident in the reasoning of the same interviewee regarding the verbal version: "*It's either B or D. It's probably B because D is more like the middle of the process. When the butter is liquid, they [the particles] spread out more, and when it is solid, they're like in an atomic lattice and connected. Now [when it is liquid] they have more space and can move. Answer D [the correct one] represents the middle of the process; where it is not liquid yet.*" This excerpt shows that she knew the differences between solid and liquid, but the vagueness of the verbal distractors created a hesitation regarding which of the two options is a better representation of the phase of the macroscopic butter depicted in the question. This hesitation may indicate that the scenario of the question was unclear, leading to low performance and perhaps explains the significant correlation between question clarity and performance.

Discussion and conclusions

Impact of verbal and visual formats on student performance and perceived clarity

This study examines the role of visuals in assessing students' understanding of the structure of matter with a focus on chemical reactions. We used a novel "clarity" rating to juxtapose the ways students understood representations in assessment items and their ability to solve them; in line with our first research question, we found that for most items the visual versions of the choice options were more difficult than the verbal ones, although for two items we found the opposite trend. Unlike our previous study with the 7th graders (Langbeheim *et al.*, 2023), in the current study the proportion of correct responses in the verbal format was higher than that in the visual format. However, in items related to the particle model of matter that did not require chemical details (*e.g.*, dissolving tea, melting butter, and spread of the smell of a flower), the visual version of the assessment was easier or equal in difficulty to the verbal one. Additionally, the arrows in the pictures no longer posed an issue, possibly because the students in the current study were the 8th graders and had more experience with this symbol than the 7th graders in our prior study.

Slightly higher performance on verbal representations was also reported in the context of introductory university physics courses, where visual items (especially those that contain unfamiliar symbols such as vectors) were more cognitively demanding (Meltzer, 2005; Nieminen *et al.*, 2010; Susac *et al.*, 2023). Still, the populations of these studies were university or high school students, whereas we focus on the eighth graders – a significantly younger cohort. University and high school students have more experience in interpreting visual representations and making connections between different forms of representations than middle schoolers. Consequently, the smaller differences between verbal and visual versions among older students are probably related to greater experience with representations, (*i.e.*, representational competence) among older students (Garcia-Mila *et al.*, 2014).

Relationship between clarity ratings and performance on visual items

Our second research question asked how students' clarity ratings (of questions and illustrations) are related to their performance on items with different types of visual representations. Clarity ratings and interviews provide insight into the reasoning concerning visual and verbal representations in the conceptual questions. For example, in the NH_3 item, student performance was lower in the visual version than that on the verbal one, and the clarity ratings of the question and illustrations were correlated with performance. The difficulty of selecting the visual representation that matches the symbolic representation was also reported in Gkitzia *et al.* (2020). Successful performance on these items requires understanding the details of the symbols of the chemical equation, and the visual depiction of the molecules. In both items of this sort

(NH_3 and the similar CO_2), we found significant correlations between illustration clarity and students' actual performance, indicating that higher perceived comprehensibility of the visual representation was associated with likeliness to choose the correct answer. In addition, correlations were also found in items with mass *vs.* time graphs such as the dissolved salt, egg shells and the metal ball items. These tasks entailed graphs of changes in mass *vs.* time during chemical or physical processes, and students with higher clarity rating of these graphs were more likely to answer correctly. Together with the NH_3 and CO_2 items, these findings suggest that correlations between clarity and performance emerge especially when the solution depends directly on interpreting visual – symbolic correspondences. This correlation extends the findings of studies that show that perceived complexity and other mental load ratings of visual items are correlated with performance (Hoch *et al.*, 2023). Our study shows that these judgements are especially evident on items with technical visual details that most likely stem from limited representational competence.

Still, for almost half of the items, the correlations between clarity and performance were very low. One explanation for this relatively low correlation is that the middle school students in this study often overlooked details that were pertinent to the question scenario and still believed they understood it, hence the higher-than-average ratings even among low performers. This is similar to ''meta-ignorance'' – an unwarranted high sense of confidence among students with low scores, in Brandriet and Bretz (2014). For example, in the bubble item, despite high clarity ratings, students commonly believed that the bubbles in boiling water are made of air (rather than water vapor) demonstrating a persistent misconceived idea (*e.g.*, Osborne and Cosgrove, 1983; Johnson, 1998). Similarly, Potgieter *et al.* (2010) showed that confidence judgments of items with (incorrect) distractors that most students found very convincing did not correlate with performance. Thus, the low correlations between clarity and performance on these items stem from students who misunderstood the concepts and performed poorly, yet believed they understood the questions and the visuals rather well.

Insights from items with large performance gaps between formats

Our third research question addressed the large performance gaps between verbal and visual versions in some of the assessment items. Interviews revealed that differences in performance between verbal and visual versions in some of the items (*e.g.*, NH_3, butter, and Hg) originated from several factors. In the mercury oxide item (Fig. 1), it originated from misunderstanding the core concepts: the belief that gases have no mass or that the appearance of new materials increased the mass. Some students expressed this misunderstanding in both formats, while students with fragmented knowledge – misapplied conservation only in the visual version. According to Vosniadou's (2013) framework, students' explanations reflect ''synthetic'' mental models that combine elements of prior knowledge with new ideas that are activated by the specific context. Visualizations can activate productive and unproductive knowledge elements that characterize such fragmented knowledge (Langbeheim, 2015). For example, the belief that conservation of mass holds only in processes that do not produce gases may emerge in the context of questions that visualize the gaseous products. That is why, in the mercury oxide item with the image of the inflated balloon difficulties originated from the visuals that attracted their attention and blurred the underlying conservation principle. The inflated balloon can be interpreted as ''seductive imagery'' (Bruner *et al.*, 1966) – learners' tendency to attend to perceptually salient but conceptually irrelevant visual details that distract from the underlying scientific principle. In our case, the inflated balloon diverted students' attention from the conservation principle, leading them to infer a change in mass, but most visual details did not reveal such a misleading effect.

In a couple of cases such as the melting butter item (Fig. 4), performance on the visual version was better than that on the verbal one, although the clarity ratings of the verbal version were higher. On the verbal version, most students selected the option that stated that particles ''spread out'' and occupy larger spaces in the liquid; about a quarter of students viewed the particles of the liquid as ''soft and moving like jelly.'' Both descriptions reflect a mental model that endows individual particles with resemblance to the macroscopic appearance of the material (Johnson, 1998). The visual format helped students reject some of these ideas when representations revealed that large spaces resembled a model of gas rather than a liquid, or when the shape of jelly-like particles seemed too odd to represent particles. This demonstrates how visual representations cued information that ''masked'' fragmented knowledge. The butter item shows that visual representations can also ''flag'' problematic answers so that children *avoid* selecting them thereby superficially improve performance. Interestingly, this item also exhibited a correlation between students' performance and their clarity ratings of the question – indicating that students who selected the incorrect responses also rated the clarity of the question lower. This may indicate that the difficulty for low performers was not rooted in the pictures of the choice options, but rather in the scenario of the question as whole.

Finally, the interview data corroborated the findings that awareness of the complexity of visuals may reflect cognitive load, as in the NH_3 item. In this case, many of the students knew that reactants are the materials shown on the left side of the chemical reaction, and the products – on its right, but when confronted with the detailed molecular, space-filling pictures alongside the chemical equation, they got confused by the abundant information and chose the wrong answer. This shows that the cognitive demand created by the visuals might have blurred students' ability to apply the core concepts, which explains the significantly lower performance on the visual version of this item, when compared to the verbal one. These findings can also clarify the reciprocal yet distinct relationship between conceptual understanding and representational competence (Kozma and Russell, 2005; Rau, 2017).

To conclude, visual representations can be viewed as the external layer mediating activation of the core concepts, and clarity ratings serve as a diagnostic lens for this mediation. The external layer consists of pictorial and symbolic cues that activate students' core concepts. For instance, in the mercury oxide item, the image of the inflated balloon drew students' attention away from the conservation principle, illustrating how the visual layer can activate conceptual misunderstanding. When students understand the core concept but struggle with the external layer, their clarity ratings are directly tied to the accuracy of their responses (as in the NH_3 item). When the external, visual layer exposes flawed reasoning that students are not aware of, clarity ratings are detached from performance (as in the Hg item). And when perceptual cues in the visual layer help some of the students avoid incorrect choice options, clarity ratings of the question are correlated with performance, but the clarity of the illustrations is not (as in the melting butter item). Thus, visual representations can expose, mask, or overload conceptual reasoning, and clarity ratings help distinguish between these modes of interaction.

Distinguishing students' reactions to external representations and their understanding of core concepts resembles the separation between *conceptual sense-making* and *perceptual fluency* in learning with multiple external representations (Rau *et al.*, 2015). Rau (2015) showed that students who learn by building their perceptual fluency *via* seemingly superficial engagement with representations improve their chemistry knowledge but only when they have prior conceptual knowledge. This too supports the layered view: perceptual learning with visual representations is based on an initial layer of conceptual knowledge, and in order to learn from external representations, students need an initial knowledge structure for noticing variations and patterns in these representations (Bransford and Schwartz, 1999; Rau, 2015). Our study adds that visual representations can reduce performance compared to the verbal version either because they expose conceptual misunderstandings or because they blur the clarity of the question, due to limited representational competence. The two influences can be distinguished using an analysis of clarity ratings: flawed performance due to limited representational competence or fluency is characterized by correlations between question clarity and performance and cases of fragmented conceptual understanding are not correlated with clarity. In rare cases, visual representations can also mask conceptual misunderstanding and increase performance, when familiar images flag incorrect distractors.

Conclusions

Defining and measuring competence require models that bridge learning theory and classroom practice (Ufer and Neumann, 2018). Competence in the domains of chemistry can be modeled as a single overarching ability, but also as a multidimensional perspective that provides a richer account of students' strengths and weaknesses. Our study supports a layered view: assessing competence of basic chemistry ideas combines at least two distinct layers: representational competence (or fluency) and conceptual competence. This paper provides a renewed perspective on the interplay between the two layers of competence. We find that lower performance on the visual version of items reflects either students' limited representational competence and/or a fragmented understanding of the core concept. For example, in the NH_3 item, the abundance of visual details added information that seemed to overload students who knew the core concepts of reactants and products. But in other cases, with large performance differences, representational features can expose flawed conceptual reasoning, as in the mercury oxide item where the picture of the inflated balloon elicited a fragmented understanding of conservation of mass. Finally, clarity ratings help disentangle students' grasp of fundamental principles from their actual performance on visual assessment tasks, offering a more fine-grained lens on their reasoning. Overall, these findings suggest that in most cases, replacing verbal assessment items with visual ones exposes the fragmented nature of middle schoolers' chemistry knowledge.

Implications

For instructors, our findings demonstrate that visual items diagnose both reasoning about conceptual relations (*e.g.*, conservation or reactant–product distinctions) and acquaintance with the visuals themselves. Educators can leverage students' responses for formative purposes, but the version of items should be aligned with instructional goals. To probe the conceptual core, verbal items or simplified visuals are preferable. To foster representational competence, such as translating between symbolic equations and molecular depictions or interpreting graphs, items that require linking representations should be used, even if they increase difficulty, so that students engage explicitly with the representational layer. For example, when introducing chemical equations, formative assessment tasks can present several particulate diagrams and ask to explain what features of the visuals can help identify the reactants and products. Similarly, when working with graphs of mass or temperature changes, prompting students to verbalize what each axis represents can foster representational competence by making the representational layer itself an object of discussion, rather than a source of implicit difficulty.

In addition, responses to visual items should be examined for reliance on superficial details. When choices are driven by superficial (but familiar) pictorial cues, the item may not capture fragmented knowledge. In such cases, visuals should be replaced with verbal forms focusing on the core concept. Finally, clarity ratings can provide a lens for distinguishing conceptual understanding and representational demands, supporting a layered view of competence that combines conceptual understanding with representational fluency.

Limitations

This study has several limitations. First, the sample size, although reasonable, was limited, and the correlations reported may not generalize to broader populations. Second, the groups in phase 1 and phase 2 were not identical and the questionnaires were administered at different points in the school year: phase 1 toward the end of the academic year, and phase 2 in the middle of the following year. Third, further investigation of clarity ratings should compare them with other measures such as confidence and item difficulty judgements, in order to strengthen the validity evidence supporting the clarity rating as an assessment measure. Finally, the qualitative interviews provided valuable insights, yet the number of participants was small and included only female students from one school, which may limit the generalizability of the qualitative finding.

Conflicts of interest

There are no conflicts of interest to declare.

Data availability

The assessment instruments used in this article are included as part of the supplementary information (SI), and the full response datasets can be requested from the second author. Supplementary information contains the visual and verbal versions of the questionnaire, the interview protocol, and a table that shows the results from the teacher–practice survey in the groups of students that responded to the verbal and visual forms. See DOI: https://doi.org/10.1039/d5rp00372e.

References

Adadan E., (2013), Using multiple representations to promote grade 11 students' scientific understanding of the particle theory of matter, *Res. Sci. Educ.*, **43**, 1079–1105.

Adadan E., Irving K. E. and Trundle K. C., (2009), Impacts of multi-representational instruction on high school students' conceptual understandings of the particulate nature of matter, *Int. J. Sci. Educ.*, **31**(13), 1743–1775.

Barker V. and Millar R., (1999), Students' reasoning about chemical reactions: What changes occur during a context-based post-16 chemistry course? *Int. J. Sci. Educ.*, **21**(6), 645–665.

Brandriet A. R. and Bretz S. L., (2014), Measuring meta-ignorance through the lens of confidence: examining students' redox misconceptions about oxidation numbers, charge, and electron transfer, *Chem. Educ. Res. Pract.*, **15**(4), 729–746.

Bransford J. D. and Schwartz D. L., (1999), Chapter 3: Rethinking transfer: a simple proposal with multiple implications, *Rev. Res. Educ.*, **24**(1), 61–100.

Bruner J., Oliver R. R., and Greenfield P. M., (1966), *Studies in cognitive growth*, New York: Wiley.

Butcher K. R., (2006), Learning from text with diagrams: promoting mental model development and inference generation, *J. Educ. Psychol.*, **98**(1), 182.

Corbin J. M. and Strauss A., (1990), Grounded theory research: Procedures, canons, and evaluative criteria, *Qual. Sociol.*, **13**(1), 3–21.

Daniel K. L., Bucklin C. J., Austin Leone E. and Idema J., (2018), Towards a definition of representational competence, *Towards a framework for representational competence in science education*, pp. 3–11.

Ding L. and Beichner R., (2009), Approaches to data analysis of multiple-choice questions, *Phys. Rev. Spec. Top.--Accel. Beams*, **5**(2), 020103.

Doran R. L., (1980), *Basic Measurement and Evaluation of Science Instruction*, National Science Teachers Association, 1742 Connecticut Ave., NW, Washington, DC, 2009.

Eilam B. and Gilbert J. K., (2014), The significance of visual representations in the teaching of science, in Gilbert J. K. (ed.), *Science teachers' use of visual representations*, Routledge, pp. 3–28.

Follmer D. J. and Clariana R., (2022), Predictors of adults' metacognitive monitoring ability: the roles of task and item characteristics, *J. Exp. Educ.*, **90**(3), 570–592.

Garcia-Mila M., Marti E., Gilabert S. and Castells M., (2014), Fifth through Eighth grade students' difficulties in constructing bar graphs: data organization, data aggregation, and integration of a second variable, *Math. Thinking Learn.*, **16**(3), 201–233.

Gkitzia V., Salta K. and Tzougraki C., (2020), Students' competence in translating between different types of chemical representations, *Chem. Educ. Res. Pract.*, **21**(1), 307–330.

Goodman L. A. and Kruskal W. H., (1954), Measures of association for cross classifications, *J. Am. Stat. Assoc.*, **49**(268), 732–764.

Hadenfeldt J. C., Neumann K., Bernholt S., Liu X. and Parchmann I., (2016), Students' progression in understanding the matter concept, *J. Res. Sci. Teach.*, **53**(5), 683–708.

Hoch E., Sidi Y., Ackerman R., Hoogerheide V. and Scheiter K., (2023), Comparing mental effort, difficulty, and confidence appraisals in problem-solving: a metacognitive perspective. *Educ. Psychol. Rev.*, **35**(2), 61.

Johnson P., (1998), Children's understanding of changes of state involving the gas state, Part 1: boiling water and the particle theory, *Int. J. Sci. Educ.*, **20**(5), 567–583.

Johnstone A. H., (1991), Why is science difficult to learn? Things are seldom what they seem, *J. Comput. Assist. Learn.*, **7**(2), 75–90.

Joo H., Park J. and Kim D., (2021), Visual representation fidelity and self-explanation prompts in multi-representational adaptive learning, *J. Comput-Assist. Learn.*, **37**(4), 1091–1106.

Kohl P. B. and Finkelstein N. D., (2005), Student representational competence and self-assessment when solving physics problems, *Phys. Rev. Spec. Top.--Accel. Beams*, **1**(1), 010104.

Kozma R., Chin E., Russell J. and Marx N., (2000), The roles of representations and tools in the chemistry laboratory and

their implications for chemistry learning, *J. Learn. Sci.*, **9**(2), 105–143.

Kozma R. and Russell J., (2005), Students becoming chemists: developing representational competence. *Visualization Sci. Educ.*, **1**, 121–146.

Kruger J. and Dunning D., (1999), Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments, *J. Personality Soc. Psychol.*, **77**(6), 1121–1134.

Küchemann S., Malone S., Edelsbrunner P., Lichtenberger A., Stern E., Schumacher R., and Kuhn J., (2021), Inventory for the assessment of representational competence of vector fields, *Phys. Rev. Phys. Educ. Res.*, **17**(2), 020126.

Langbeheim E., (2015), Reinterpretation of students' ideas when reasoning about particle model illustrations, *Chem. Educ. Res. Pract.*, **16**(3), 697–700.

Langbeheim E., Akaygun S., Adadan E., Hlatshwayo M. and Ramnarain U., (2023), Relating pictorial and verbal forms of assessments of the particle model of matter in two communities of students, *Int. J. Sci. Math. Educ.*, **21**(8), 2185–2201.

Langbeheim E., Ben-Eliyahu E., Adadan E., Akaygun S. and Ramnarain U. D., (2022), Intersecting visual and verbal representations and levels of reasoning in the structure of matter learning progression, *Chem. Educ. Res. Pract.*, **23**(4), 969–979.

Langbeheim E. and Levy S. T., (2018), Feeling the forces within materials: bringing inter-molecular bonding to the fore using embodied modelling, *Int. J. Sci. Educ.*, **40**(13), 1567–1586.

Levy D., (2013), How dynamic visualization technology can support molecular reasoning, *J. Sci. Educ. Technol.*, **22**(5), 702–717.

Lin S.-Y. and Singh C., (2013), Using isomorphic problems to learn introductory physics, *Am. J. Phys.*, **81**(7), 597–606.

Lin Y. I., Son J. Y. and Rudd J. A., (2016), Asymmetric translation between multiple representations in chemistry, *Int. J. Sci. Educ.*, **38**(4), 644–662.

Meltzer D. E., (2005), Relation between students' problem-solving performance and representational format, *Am. J. Phys.*, **73**(5), 463–478.

Nieminen P., Savinainen A. and Viiri J., (2010), Force concept inventory-based multiple-choice test for investigating students' representational consistency, *Phys. Rev. Spec. Top.--Accel. Beams*, **6**(2), 020109.

Nussbaum J. and Novick S., (1982), Alternative frameworks, conceptual conflict, and accommodation: toward a principled teaching strategy, *Instruct. Sci.*, **11**(3),183–200.

Osborne R. J. and Cosgrove M. M., (1983), Children's conceptions of the changes of state of water, *J. Res. Sci. Teach.*, **20**(9), 825–838.

Özmen H. and Ayas A., (2003), Students' difficulties in understanding of the conservation of matter in open and closed-system chemical reactions, *Chem. Educ. Res. Pract.*, **4**(3), 279–290.

Ozuru Y., Kurby C. A. and McNamara D. S., (2012), The effect of metacomprehension judgment task on comprehension monitoring and metacognitive accuracy, *Metacogn. Learn.*, **7**, 113–131.

Piaget J. and Inhelder B., (1974), *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*, Basic Books.

Planinic M., Boone W. J., Krsnik R. and Beilfuss M. L., (2006), Exploring alternative conceptions from Newtonian dynamics and simple DC circuits: links between item difficulty and item confidence, *J. Res. Sci. Teach.*, **43**(2), 150–171.

Potgieter M., Malatje E., Gaigher E. and Venter E., (2010), Confidence versus performance as an indicator of the presence of alternative conceptions and inadequate problem-solving skills in mechanics, *Int. J. Sci. Educ.*, **32**(11), 1407–1429.

Ralph V. R. and Lewis S. E., (2020), Impact of representations in assessments on student performance and equity, *J. Chem. Educ.*, **97**(3), 603–615.

Rau M. A., (2015), Enhancing undergraduate chemistry learning by helping students make connections among multiple graphical representations, *Chem. Educ. Res. Pract.*, **16**(3), 654–669.

Rau M. A., (2017), Conditions for the effectiveness of multiple visual representations in enhancing STEM learning, *Educ. Psychol. Rev.*, **29**, 717–761.

Rau M. A., Michaelis J. E. and Fay N., (2015), Connection making between multiple graphical representations: a multi-methods approach for domain-specific grounding of an intelligent tutoring system for chemistry, *Comput. Educ.*, **82**, 460–485.

Sass S., Wittwer J., Senkbeil M. and Köller O., (2012), Pictures in test items: effects on response time and response correctness, *Appl. Cogn. Psychol.*, **26**(1), 70–81.

Simon H. A. and Hayes J. R., (1976), The understanding process: problem isomorphs, *Cogn. Psychol.*, **8**(2), 165–190.

Stankov L., Lee J., Luo W. and Hogan D., (2012), Confidence: a better predictor of academic achievement than self-efficacy, self-concept, and anxiety? *Learn. Individual Differences*, **22**(6), 747–758.

Stavy R., (1990), Children's conception of changes in the state of matter: From liquid (or solid) to gas, *J. Res. Sci. Teach.*, **27**(3), 247–266.

Susac A., Planinic M., Bubic A., Jelicic K. and Palmovic M., (2023), Effect of representation format on conceptual question performance and eye-tracking measures, *Phys. Rev. Phys. Educ. Res.*, **19**(2), 020114.

Sweller J., Ayres P. and Kalyuga S., (2011), Intrinsic and extraneous cognitive load, in Sweller J., Ayres P. and Kalyuga S. (ed.), *Cognitive load*, Springer, pp. 1–24.

Talanquer V., (2022), The complexity of reasoning about and with chemical representations, *JACS Au*, **2** (12), 2658–2669.

Tiffin-Richards S. P., Lenhart J. and Marx P., (2022), When do examinees change their initial answers? The influence of task instruction, response confidence, and subjective task difficulty. *Learn. Instruct.*, **82**, 101654.

Tonyali B., Ropohl M. and Schwanewedel J., (2023), What makes representations good representations for science education? A teacher-oriented summary of significant

findings and a practical guideline for the transfer into teaching, *Chem. Teacher Int.*, **5**(4), 413–425.

Ufer S. and Neumann K., (2018), Measuring competencies, *International handbook of the learning sciences*, Routledge, pp. 433–443.

Vojíř K. and Rusek M., (2022), Of teachers and textbooks: lower secondary teachers' perceived importance and use of chemistry textbook components, *Chem. Educ. Res. Pract.*, **23**(4), 786–798.

Vosniadou S., (2013), Conceptual change in learning and instruction: the framework theory approach, *International handbook of research on conceptual change*. Routledge., pp. 11–30.