PAPER

Check for updates

Should we scaffold it? Analysing the effect of task format and scaffolding on students' learning gain

David Kranz, [iD] [a] Paul P. Martin, [iD] [a] Michael Schween [iD] [b] and Nicole Graulich [iD] *[a]

An essential goal of science education is to support students in reasoning about the underlying mechanisms of observed phenomena, which requires well-designed instructional approaches. In organic chemistry, various approaches have been designed to support students' reasoning about mechanisms, including contrasting cases as a task format. Qualitative studies indicate that contrasting cases positively impact students' mechanistic reasoning since this task format encourages students to identify and analyse similarities and differences in chemical phenomena. Additionally, a prior mixed-methods study showed that scaffolded contrasting cases can advance undergraduate students' reasoning about mechanisms, but the effect varied depending on prior knowledge. Despite these valuable insights, research has not yet quantitatively analysed the effectiveness of scaffolded *versus* non-scaffolded contrasting cases, compared with single cases. This study quantitatively examines the effects of these instructional approaches on undergraduate organic chemistry students' learning gains, with a particular focus on the role of prior knowledge. Our findings suggest that non-scaffolded contrasting cases increase learning gains for students with low prior knowledge. Additionally, scaffolded contrasting cases support students with low prior knowledge in their open-ended reasoning about chemical mechanisms. Given these findings, organic chemistry instructors should consider contrasting cases as an alternative task format. However, instructors should introduce the scaffolding used in this study with practice sessions as it may otherwise increase cognitive load for students unaccustomed to its demands.

Introduction

In organic chemistry, students should be able to predict the selectivity of transformations, design new synthesis routes, and critically evaluate the plausibility of reactions—practices requiring mechanistic reasoning (Bhattacharyya and Bodner, 2005). Mechanistic reasoning is a scientific practice that goes beyond simply describing the outcomes of chemical reactions (Graulich, 2015; Dood and Watts, 2022). It involves deducing the underlying mechanisms behind observed phenomena to explain how and why changes occur at the molecular level (Russ *et al.*, 2008; Cooper *et al.*, 2016). However, students often struggle to reason about the mechanisms and causal processes underlying observed phenomena (Graulich, 2015; Ince, 2018; Dood and Watts, 2023). Furthermore, students may rely on different, sometimes unproductive strategies when working with similar representations across contexts (Braun and Graulich, 2024). Therefore, identifying instructional approaches that help students overcome these challenges is a key goal of chemistry education research.

Recent frameworks describing students' mechanistic reasoning in general chemistry (Sevian and Talanquer, 2014) and organic chemistry (*e.g.*, Caspari *et al.*, 2018; Bodé *et al.*, 2019; Deng and Flynn, 2021) highlight students' challenges in deriving implicit features from explicit representations (*e.g.*, molecules, electron arrows, charges, *etc.*). However, the ability to infer information from representations with a high level of abstractness allows students to find more plausible solutions (Weinrich and Sevian, 2017). For beginners in chemistry, it is challenging to determine which explicit features are relevant and how to derive the corresponding implicit properties (Graulich and Bhattacharyya, 2017). Students in chemistry, thus, tend to focus on single surface features or symbolic patterns rather than recognising implicit functional similarities, whether evaluating individual molecules or entire reactions (Domin *et al.*, 2008; Graulich and Bhattacharyya, 2017; Galloway *et al.*, 2018). However, considering implicit properties and underlying processes of a mechanism is essential for higher-level reasoning (Weinrich and Sevian, 2017) and leads to greater success when solving mechanistic problems that require knowledge transfer (Grove *et al.*, 2012). Since these student challenges have been documented in numerous studies, students require targeted instructional support (*e.g.*, Bachtiar *et al.*, 2022). Consequently, it is critical that evidence-based conclusions can be drawn about the effectiveness of such instructional support.

[a] *Institute of Chemistry Education, Justus-Liebig-University, Giessen, Heinrich-Buff-Ring 17, 35392 Giessen, Germany. E-mail: Nicole.Graulich@didaktik.chemie.uni-giessen.de*

[b] *Marburg, Faculty of Chemistry, Philips-University, Hans-Meerwein-Strasse 4, 35032 Marburg, Germany*

Contrasting cases *versus* single cases

Contrasting cases are a common approach across the natural sciences and mathematics for prompting students to recognise relevant features of a problem. In chemistry, contrasting cases are sets of chemical examples selected to highlight critical conceptual differences, such as how substituent effects influence reaction rates of mechanisms, so that students can purposefully compare and learn the underlying chemical concepts (Graulich and Schween, 2018). Alfieri *et al.* (2013) found in a meta-analysis that contrasting cases led to a significantly higher number of identified variables than single cases ($d = 0.60$, 95% CI [0.47, 0.72]). Contrasting cases were successfully used to promote mechanistic reasoning in organic chemistry (Caspari *et al.*, 2018; Bodé *et al.*, 2019; Caspari and Graulich, 2019; Rodemer *et al.*, 2020; Deng and Flynn, 2021; Watts *et al.*, 2021; Eckhard *et al.*, 2022; Kranz *et al.*, 2023; Haas *et al.*, 2024). Since contrasting cases have demonstrated positive effects on students' mechanistic reasoning in qualitative research, it is important to evaluate how effective this kind of instruction is when compared with tasks commonly used in organic chemistry, *e.g.*, predict-the-product tasks (*i.e.*, predominantly *single cases*) and whether contrasting cases lead to greater learning gains than traditional organic chemistry tasks.

Scaffolding

To support students in connecting features relevant for a problem, *scaffolding* (*i.e.*, a structured sequence of prompts) can be used to break down the reasoning process into several sub-steps. This supporting instruction has been widely used in science education (Lin *et al.*, 2012; Wilson and Devereux, 2014), among other applications, to guide students through the problem-solving process while working with contrasting cases in organic chemistry (Caspari and Graulich, 2019; Watts *et al.*, 2021; Kranz *et al.*, 2023). Scaffolding slows down the decision-making process (Caspari and Graulich, 2019). Furthermore, it provides students with the opportunity to activate necessary conceptual and procedural knowledge (Rittle-Johnson and Star, 2007; Rittle-Johnson and Star, 2009; Shemwell *et al.*, 2015; Chin *et al.*, 2016) and to reflect in depth on the concepts used (Graulich *et al.*, 2021). Scaffolding helps students generalise and transfer concepts to new examples (Lombrozo, 2006) and to identify more connections between features that are essential for successfully solving a task (Caspari and Graulich, 2019; Watts *et al.*, 2021). Since scaffolding provides step-by-step guidance in comparing the elements of a problem, such as chemical structures, it helps students derive and weigh implicit properties from explicit information (Caspari and Graulich, 2019). As a result, scaffolded contrasting cases may foster greater learning gains than non-scaffolded contrasting cases.

At present, there is no quantitative evidence for the effectiveness of contrasting cases and scaffolded contrasting cases compared to traditional tasks in the context of organic chemistry reaction mechanisms, nor is there quantitative evidence regarding the role of prior knowledge in moderating these effects. For future use of contrasting cases and scaffolding, it is, thus, critical to understand the distinct impact of each instructional approach relative to traditional tasks, such as single cases. Additionally, exploring how prior knowledge shapes the effectiveness of these approaches is crucial for identifying which students—based on their prior knowledge—gain the most from each approach.

Theoretical framework

Variation theory

Some fundamental principles underlying the contrasting cases task design are rooted in Variation Theory (Ling Lo, 2012). According to this theory, learning occurs when individuals become aware of previously unrecognised features of a phenomenon. Consequently, instructional approaches should guide students in analysing similarities and differences so that they can identify and weigh critical features of the phenomenon (Ling Lo and Marton, 2011). Varying certain elements makes them salient to the viewer, while other elements are kept invariant (Ling Lo and Marton, 2011; Bussey *et al.*, 2013). This enables students to notice critical features more quickly (Bussey *et al.*, 2013). In the context of science and mathematics education, the use of contrasting cases (Chin *et al.*, 2016; Graulich and Schween, 2018) appears, therefore, to be a promising approach for helping students identify more critical features, compared with learning from single cases (Alfieri *et al.*, 2013). Nonetheless, when instructors incorporate variations into exercises, these should not be applied at random—rather, the timing and purpose of these variations should be carefully considered (Ling Lo, 2012).

Guiding reasoning through scaffolding

The concept of scaffolding originates from Wood *et al.* (1976) and is closely linked to Vygotsky and Cole's (1978) theory of the zone of proximal development. This theory posits that assisted learning can offer greater potential for learning gains than unassisted learning because assistance can help students accomplish tasks they could not complete independently. From this perspective, the learning potential of a student answering a task regarding a reaction mechanism in organic chemistry without assistance is lower than that of a student who receives scaffolding for that task. Scaffolding can be used to direct students' problem-solving process by guiding them through a complex task (Pea, 2004). It aims to increase students' problem-solving skills (Wood *et al.*, 1976; Belland, 2011; Yuriev *et al.*, 2017).

As Reiser (2004) outlines, scaffolding supports problem-solving by both structuring and problematising tasks. These strategies not only facilitate engagement, but also benefit learners with low prior knowledge by helping them build on fragmented understanding. Students with high prior knowledge might either not profit from scaffolding or may be hindered by it since it can create redundancy, distraction, or lead to an expertise reversal effect (*i.e.*, a task design that is helpful for novices becomes ineffective for experts) (Kalyuga, 2007; Homer and Plass, 2010; Nückles *et al.*, 2010; Oksa *et al.*, 2010; Salden *et al.*, 2010).

Because scaffolding seeks to manage the mental demands placed on students, cognitive load theory offers a useful lens for

evaluating scaffold design. Cognitive load theory (Sweller, 1994) states that working memory is limited: individuals can process only a small amount of information at one time. Thus, this theory clarifies why some students may struggle with particular task designs more than other students. Cognitive load theory assumes that more prior knowledge leads to less intrinsic cognitive load (*i.e.*, the effort required to understand a given task) (Paas and Sweller, 2014). Similarly, sequencing tasks leads to a lower intrinsic cognitive load by reducing task complexity (Paas and Sweller, 2014). However, if the prompts are not helpful for the individual student, a scaffold with a complex structure may increase cognitive load, since the extraneous cognitive load (*i.e.*, the mental effort required to deal with the design of a task) also increases (Sweller, 2010). The design of a scaffold must, therefore, carefully consider to what extent the task should be structured into subtasks to reduce the intrinsic cognitive load, while keeping the structure simple enough to avoid increasing the extraneous cognitive load.

Study design and research questions

From the theoretical framework of this study, one can assume that different instructional approaches, including task format (*i.e.*, learning with single or contrasting cases) and structured guidance (*i.e.*, scaffolded contrasting cases), could have varying effects on student learning gains. To investigate this assumption, undergraduate organic chemistry students were randomly divided into three treatment groups:

(1) a group who worked on traditional organic chemistry tasks, predominately single cases (*sc*),

(2) a contrasting cases group without scaffolding (*cc*) and

(3) a group who received an additional scaffold with the contrasting cases (*ccsf*).

We hypothesised, based on prior findings, that working with scaffolded contrasting cases would lead to a greater effect on students' learning gains than non-scaffolded contrasting cases, while contrasting cases would produce a greater effect than single cases. This led to the first research question:

RQ1: How does the task format (*sc vs. cc vs. ccsf*) influence students' learning gain?

Building on previous findings (Kranz *et al.*, 2023), the effectiveness of scaffolded and non-scaffolded contrasting cases may depend on students' prior knowledge. The second research question, therefore, examines this relationship:

RQ2: To what extent does the learning gain in the treatment groups depend on students' prior knowledge?

To answer these questions, a pre-post intervention study was conducted.

Methods

Context of the study

Data collection took place between autumn 2021 and summer 2022 in four courses at three German universities, all of which had been taught the basics about nucleophilic substitution reactions prior to the study (*i.e.*, mechanisms, hyperconjugation, inductive effects and resonance effects). The data were collected on two consecutive days in each course. $N = 122$ students (33 chemistry student teachers, 62 chemistry bachelor students and 27 students of chemistry-related study programs, such as food chemistry or materials science) participated in the study. The students were between 18 and 40 years old, while 62 identified themselves as male, 59 as female, and one as non-binary. All, except four students, were German native speakers. We ensured that students were familiar with similar chemical content based on their course schedule and had not worked with contrasting cases or the scaffold before the data collection.

Learning material

Each student was randomly assigned to one of three treatment groups and received a set of tasks tailored to the respective group (problem-solving phase, see Fig. 1), which ended with a constructed-response task (CRT). All students had 40 minutes to complete the respective tasks. Because the CRTs needed different amounts of time across the groups, the *sc* and *cc* groups were provided with additional tasks—primarily traditional tasks from the organic chemistry curriculum—to ensure a comparable duration of the treatment. Overall, the learning materials were piloted using an online questionnaire in a study conducted across universities in Germany to provide evidence of valid and reliable data (Martin, 2022).

(1) *Single cases* (*sc*): Students in the *sc* group received several traditional organic chemistry tasks, for instance, predict-the-product and ranking tasks. Students had to explain the thermodynamic favourability of two single cases as CRTs (Fig. 1).

(2) *Contrasting cases* (*cc*): Students in the *cc* group worked with comparable tasks as the *sc* group and were prompted to explain the differing reaction rate of two contrasting cases as CRTs (Fig. 1).

(3) *Scaffolded contrasting cases* (*ccsf*): The *ccsf* group worked with the same contrasting cases as the *cc* group as CRTs (Fig. 1) but received additional subtasks provided in a scaffold grid (Fig. 2). The scaffold was adopted from Kranz *et al.* (2023) (Fig. 2), which was a modified version of the scaffold by Caspari and Graulich (2019). For this study, the scaffold provided a sequence of five scaffold prompts in total (Fig. 2) and was implemented in a digital format, which had the advantage that the grid expanded step-by-step after students' input. It is important to note that the prompts structured the problem-solving process only and provided no further conceptual information.

Research instruments

Concept knowledge test. An online questionnaire (Graulich *et al.*, 2025) with an inventory of 13 items (after deletion of items that had a negative impact on reliability and very low inter-item correlation (Cronbach, 1951; Tavakol and Dennick, 2011; Sürücü and Maslaçi, 2020)) was used to assess students' organic chemistry concept knowledge before
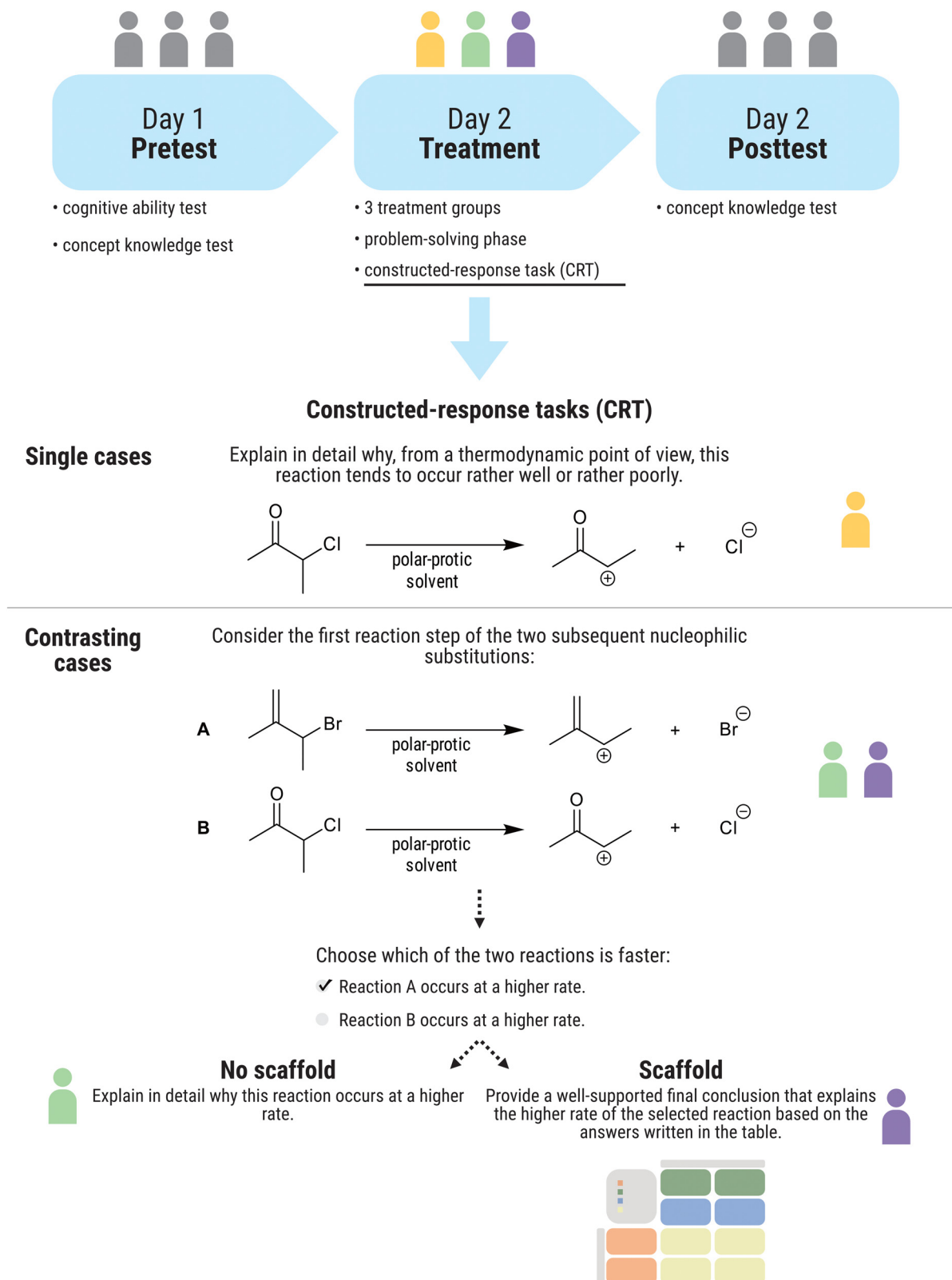
Fig. 1 Study design and illustration of the treatment groups. The second CRT was a contrasting case where 2-bromopropane and 2-bromo-2-methylpropane were contrasted.

and after the treatment (a sample of items is shown in Appendix 2). Items were selected so that they corresponded to the content of the organic chemistry lecture. The test included six multiple-choice items, two single-choice items, and one
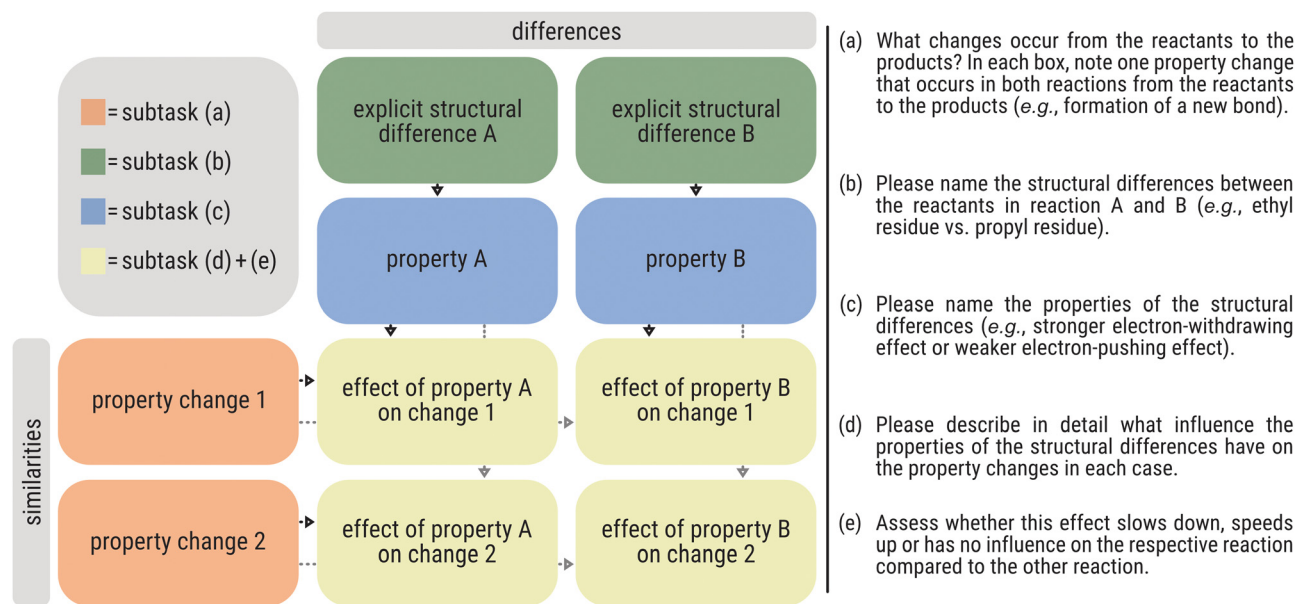
Fig. 2 Scaffold grid structured by subtasks.

(a) What changes occur from the reactants to the products? In each box, note one property change that occurs in both reactions from the reactants to the products (*e.g.*, formation of a new bond).

(b) Please name the structural differences between the reactants in reaction A and B (*e.g.*, ethyl residue vs. propyl residue).

(c) Please name the properties of the structural differences (*e.g.*, stronger electron-withdrawing effect or weaker electron-pushing effect).

(d) Please describe in detail what influence the properties of the structural differences have on the property changes in each case.

(e) Assess whether this effect slows down, speeds up or has no influence on the respective reaction compared to the other reaction.

ranking task as well as two closed-response and two open-ended contrasting cases with a total of 42 attainable points. The single- and multiple-choice items covered concepts, such as inductive and resonance effects (see Appendix 2). The ranking task required students to rank leaving groups according to their quality. In the two-tiered closed-response contrasting cases items, students first decided which reaction occurs faster and then selected an appropriate explanation from several options; points were only awarded if both responses were correct. The open-ended contrasting cases are shown in Appendix 3. The questionnaire is published by (Graulich *et al.*, 2025).

Internal consistency was $\alpha_{pre} = 0.81$ (95% CI [0.75; 0.86]) for the pretest and $\alpha_{post} = 0.80$ (95% CI [0.75; 0.85]) for the posttest, calculated with the R package *psych* (Revelle, 2022). These values represent good reliability (Cronbach, 1951; Tavakol and Dennick, 2011; Field *et al.*, 2012).

The aim of the questionnaire was to assess students' concept knowledge regarding several concepts necessary for solving the mechanistic problems used in the intervention (*e.g.*, inductive and resonance effects). The inventory had previously been piloted with $N = 72$ students from different German universities; based on classical test theory and item-response theory, the original 38 items were reduced to 18 (Martin, 2022), and finally 13 items with acceptable difficulty indices and a Kaiser–Meyer–Olkin value (Cureton and D'Agostino, 2013) above the threshold were retained for analysis.

Cognitive ability test. As part of the pretest, students completed a standardised cognitive ability test (Heller and Perleth, 2000) to assess their ability to visualise and mentally rotate objects, an important skill for identifying reaction sites in structural representations. The 8-minute test confirmed the randomised group assignment and has been used in similar research (Rodemer *et al.*, 2021). The students had 8 minutes to complete this test. Internal consistency was $\alpha_{cat} = 0.86$ (95% CI [0.82; 0.89]),

calculated with the R-library *psych* (Revelle, 2022) indicating a good reliability (Cronbach, 1951; Tavakol and Dennick, 2011; Field *et al.*, 2012).

Students' perceived task difficulty. After the CRTs, students rated task difficulty on a five-point Likert scale (1: *very easy*, 5: *very difficult*) as well as their confidence in solving the tasks (1: *very confident*, 5: *very unconfident*). These self-ratings served as an approximate indicator of cognitive load because difficulty estimates correlate with item complexity and confidence judgements relate to response time and overall mental effort (Bratfisch *et al.*, 1972; Chen and Chang, 2009; Gavas *et al.*, 2018; Hoch *et al.*, 2023).

Data collection

Students participated voluntarily in this study. The study was conducted on two consecutive days in each course. All students had the opportunity to win a voucher (*i.e.*, five €20 Amazon gift cards were randomly distributed among the students) for participating; nevertheless, the sessions were scheduled during regular lecture times. Approximately 80% of all enrolled students took part, mirroring the demographics of the overall cohorts. On the first day, a cognitive ability test and a pretest (*i.e.*, concept knowledge test) were administered. Demographic data were collected *via* a digital survey prior to the pretest. The cognitive ability test was administered according to Heller and Perleth (2000). Students then had 30 minutes to complete the concept knowledge test, which was presented in a randomised item order. To match datasets, students generated an eight-character code based on personal information; the data were anonymised so that no one except the respondent could link identities to datasets. Instructors were asked not to provide additional information about S_N1 mechanisms after the pretest. On the following day, students completed the tasks assigned to their treatment group, answered two constructed-response tasks

(CRTs), and then completed the posttest (*i.e.*, concept knowledge test) (Fig. 1). The intervention itself lasted approximately 40 minutes before students took the posttest. All instruments and the intervention were delivered digitally *via* iPads and laptops. The tasks were presented in German; answers were translated and reviewed by an English native speaker for publication (Appendix 1). This two-day schedule minimised carry-over effects between pre- and posttests and ensured identical conditions for all groups.

Data analysis

We qualitatively coded the open-ended responses from the concept knowledge test and the CRTs for subsequent analysis. In addition, we scored the pretest, posttest, and CRT responses for correctness. We also assessed students' perceived task difficulty and confidence. An overview of the analysis workflow is provided in Fig. 3.

The first author coded the entire dataset, and the second author independently coded a randomly selected 20% sample to calculate inter-rater reliability with the R package *DescTools* (Signorell *et al.*, 2022). After the first coding round, $\kappa = 0.73$ was obtained; following a discussion of discrepancies, agreement increased to $\kappa = 0.87$ (95% CI [0.82; 0.92]), indicating an *almost-perfect* level of agreement (Fleiss and Cohen, 1973; McHugh, 2012). Closed-response items on the pre- and posttest were evaluated using a binary rubric, awarding 1 point for each correct answer—except for the two-tiered items, where points were given only when both tiers were answered correctly.

Concept knowledge test. For scale construction, items were first separated into two sub-scales (*i.e.*, conceptual and procedural knowledge) but subsequently combined because the correlation between scales was $r = 0.57$, $p < 0.001$.

Students' answers to the two open-ended questions of the pre- and posttest were qualitatively coded (Appendix 3) using an
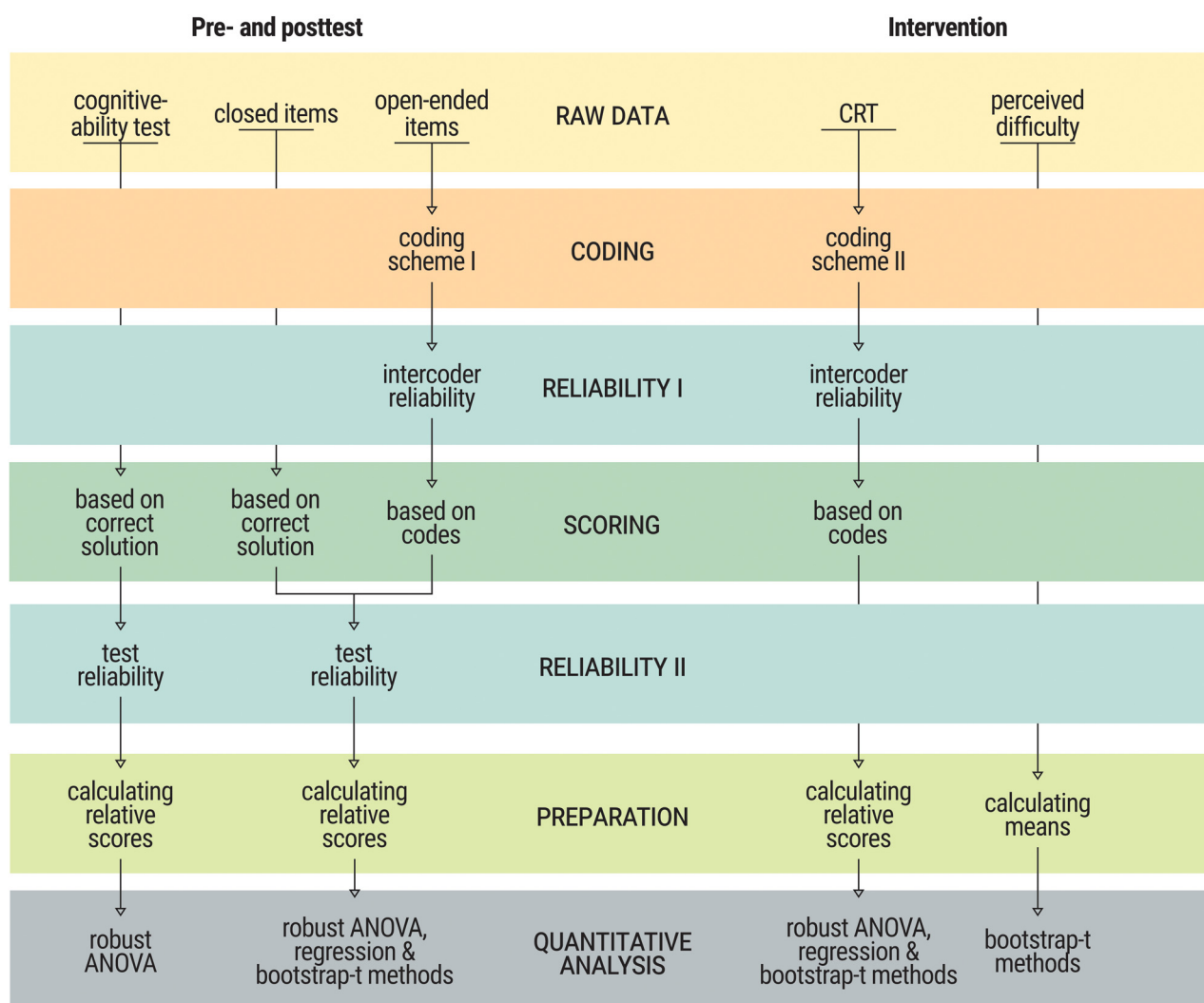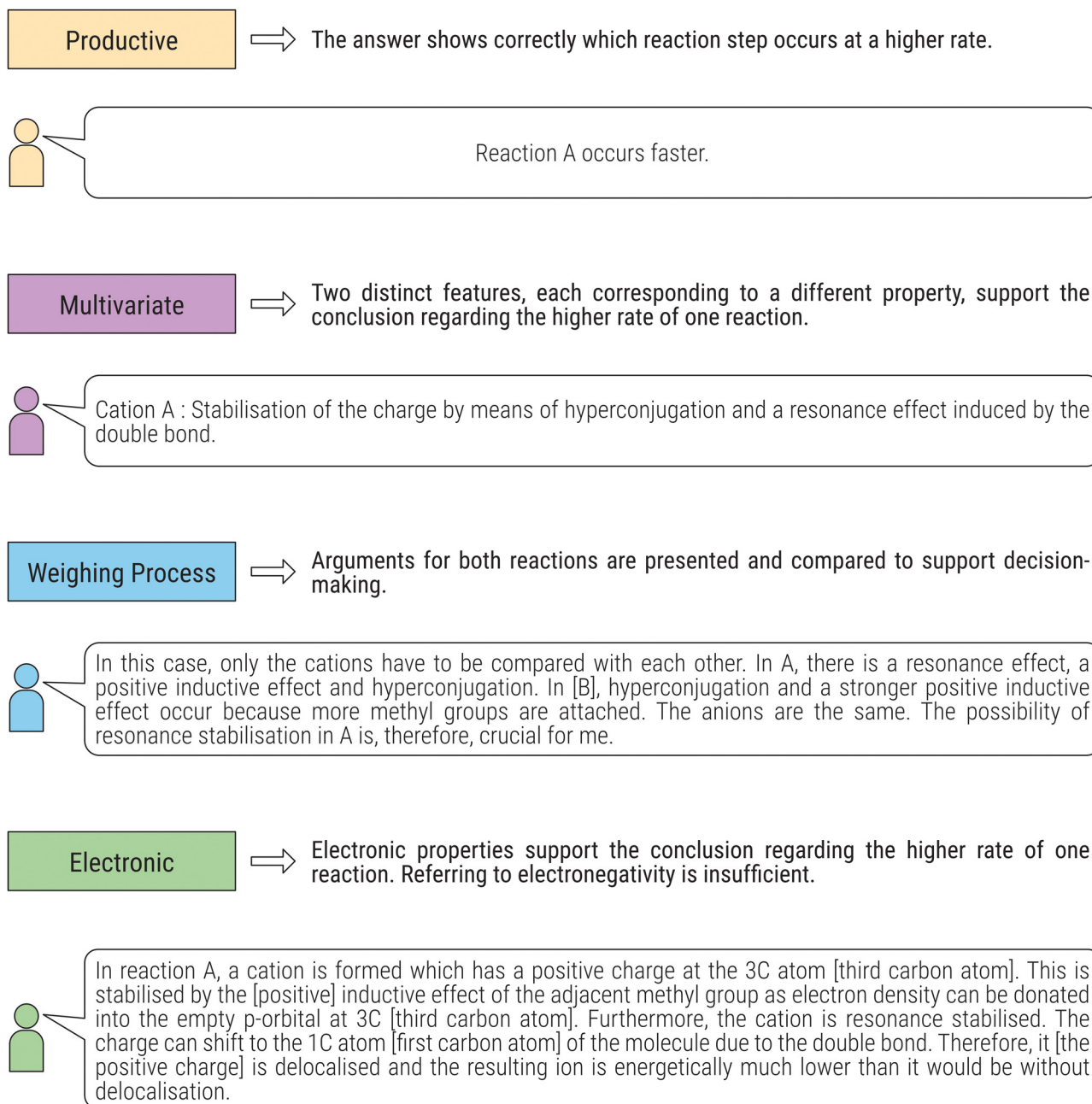

Fig. 3 Overview of the analysis steps from raw data input to quantitative analysis. Scores of students with missing data were omitted, resulting in 109 student datasets for analysis (*sc*: 38 students, *cc*: 33 students, *ccsf*: 38 students). Students' answers to the two open-ended questions of the pre- and posttest were coded qualitatively (Appendix 3) based on an inductive coding scheme with four different binary codes (1: present or 0: not present, max. 4 points) (Fig. 4).

inductive coding scheme with four binary categories (1 = present or 0 = absent, maximum 4 points) (Fig. 4). The validity of this coding scheme was corroborated in a related study employing machine learning techniques (Martin *et al.*, 2024), which are increasingly used to analyse students' mechanistic reasoning (Martin and Graulich, 2023).

Qualitative analysis of students' answers in the CRTs. To assess how students performed in the CRTs across treatment groups, their responses were qualitatively coded using the *modes of reasoning* by Sevian and Talanquer (2014) (Fig. 5), as

it appropriately reflects reasoning in both single and contrasting cases. Chemical correctness was *not* coded because we were primarily interested in students' reasoning and how it was influenced by task format and scaffolding. In-depth reasoning may be logically sound even when it does not employ canonically correct ideas. In contrast, a correct answer can be based on no or merely descriptive reasoning. Therefore, we focused our CRT analysis on the modes of reasoning. Similarly, Weinrich and Talanquer (2016) applied the modes of reasoning without assessing answer correctness (Weinrich and Talanquer, 2016).

Productive ⟹ The answer shows correctly which reaction step occurs at a higher rate.

> Reaction A occurs faster.

Multivariate ⟹ Two distinct features, each corresponding to a different property, support the conclusion regarding the higher rate of one reaction.

> Cation A : Stabilisation of the charge by means of hyperconjugation and a resonance effect induced by the double bond.

Weighing Process ⟹ Arguments for both reactions are presented and compared to support decision-making.

> In this case, only the cations have to be compared with each other. In A, there is a resonance effect, a positive inductive effect and hyperconjugation. In [B], hyperconjugation and a stronger positive inductive effect occur because more methyl groups are attached. The anions are the same. The possibility of resonance stabilisation in A is, therefore, crucial for me.

Electronic ⟹ Electronic properties support the conclusion regarding the higher rate of one reaction. Referring to electronegativity is insufficient.

> In reaction A, a cation is formed which has a positive charge at the 3C atom [third carbon atom]. This is stabilised by the [positive] inductive effect of the adjacent methyl group as electron density can be donated into the empty p-orbital at 3C [third carbon atom]. Furthermore, the cation is resonance stabilised. The charge can shift to the 1C atom [first carbon atom] of the molecule due to the double bond. Therefore, it [the positive charge] is delocalised and the resulting ion is energetically much lower than it would be without delocalisation.

Scoring: present = 1 P., not present = 0 P.

Fig. 4   Coding scheme for the open-ended contrasting cases in the concept knowledge test. The examples refer to item 1 in Appendix 3.
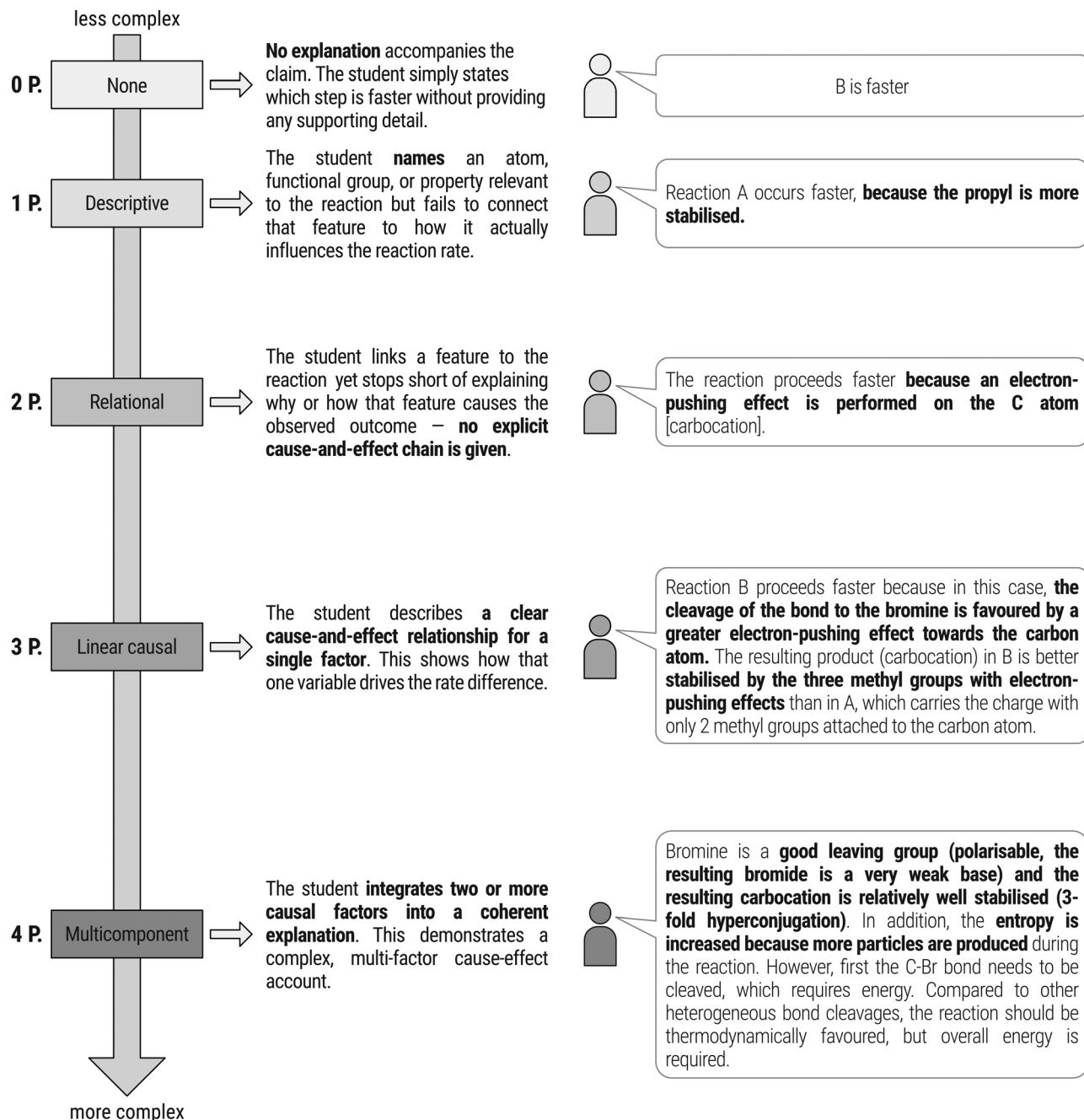
Fig. 5   Coding scheme for the CRTs with code descriptions and student examples. The student examples refer to the two tasks in the CRTs. One of the tasks is shown in Fig. 1. The second task was a contrasting case where 2-bromopropane and 2-bromo-2-methylpropane were contrasted.

The data were coded by the first author, and the second author independently coded a randomly selected 20% subsample. Inter-rater reliability reached $\kappa = 0.91$ (95% CI [0.80; >0.99]), indicating an *almost-perfect* level of agreement (Fleiss and Cohen, 1973; McHugh, 2012). For further analysis, the sum of both CRT scores was divided by the maximum achievable score (8 points). This normalisation yielded a proportional score between 0 and 1, facilitating comparisons across groups.

Approximation of cognitive load. To approximate cognitive load in the CRTs, we calculated each student's mean confidence rating and perceived difficulty rating. Hoch *et al.* (2023) showed

that difficulty estimates are a good indicator for actual item complexity, while difficulty and confidence rating better predict task success than mental effort ratings. Moreover, there is a relationship between difficulty ratings and item response time, and preliminary indications of a relationship between confidence ratings and item response time (Hoch *et al.*, 2023). Bratfisch *et al.* (1972) established a linear relation between perceived task difficulty performance on intelligence test items, arguing that perceived difficulty also reflects '[...] a person's feelings, attitudes, motivation, *etc.* [...]' (Bratfisch *et al.*, 1972). Chen and Chang (2009) expanded these findings and identified a significant

positive correlation between perceived difficulty and cognitive load. Gavas *et al.* (2018) likewise reported that metacognitive confidence levels correlate with overall cognitive load. On this basis, the combined difficulty-confidence metric was adopted as a pragmatic proxy for intrinsic and extraneous cognitive load in the present study.

Intervention effects: statistical methods used. Descriptive statistics were computed using the R package *pastecs* (Grosjean and Ibanez, 2018). To test normality, Shapiro-Wilk tests (Shapiro and Wilk, 1965) were performed. Not all group scores were normally distributed. Homogeneity of variance was examined *via* Levene's test (Levene, 1961) in the R package *car* (Fox and Weisberg, 2019); variances were similar across groups. Because of the non-normal distributions, we relied on robust procedures from *WRS2* (Mair and Wilcox, 2020) and additional functions supplied by R. R. Wilcox (Wilcox, 2023) for all between- and within-group comparisons (Wilcox, 2021). Field and Wilcox (2017), Wilcox (2021) and Field *et al.* (2012) provide the analytic guidelines we followed.

To analyse the impact of the three treatments on students' posttest and CRT score, we performed a robust heteroscedastic one-way analysis of variance (ANOVA) with bootstrapped confidence intervals (n_{boot} = 2000) for trimmed means (tr = 0.2). To compare pairwise between the groups a bootstrap-t method for trimmed means was used. To compare learning gains from pre- to posttest we used robust repeated-measures ANOVA with trimmed means for within-subjects comparisons. To determine which group had which learning gain a bootstrap-t test for marginal trimmed means was used for *post hoc* pre- *versus* post-comparisons.

To estimate the influence of the treatments while controlling for the pretest score we fitted a robust linear regression using *robustbase* (Maechler *et al.*, 2022). The grouping factor was dummy-coded, with *sc* as the reference. We specified our model for the posttest score using $posttest \sim pretest + cc + ccsf + (cc \times pretest) + (ccsf \times pretest)$ (R^2 = 0.72). An analogous model (R^2 = 0.44) predicted CRT scores. Pretest, posttest and CRT scores were centred (M = 0) and scaled (s = 0.5); dummy variables were centred at zero to allow effect size comparison (Gelman, 2008).

To explore prior knowledge effects, we created *low* and *high prior knowledge* subgroups using a mean split of the pretest score (*e.g.*, McNamara and Kintsch, 1996; Salmerón *et al.*, 2006; Abramovich *et al.*, 2013). This dichotomy was chosen to retain adequate sample sizes; finer splits would have produced groups too small for robust testing. Group comparisons within low and high prior knowledge were evaluated with the bootstrap-t test for trimmed means. Data preprocessing and transformation were performed using *reshape2*, *tidyr* and *dplyr* (Wickham, 2007; Wickham *et al.*, 2019; Wickham and Girlich, 2022; Wickham *et al.*, 2022). Graphs were produced with *ggplot2* (Wickham, 2016) and *chemometrics* (Filzmoser and Varmuza, 2017).

The significance criterion was $\alpha < 0.05$. Post-hoc p-values were Bonferroni-adjusted (Bonferroni, 1936). We used *AKP* as a robust effect size (Algina *et al.*, 2005) that works well with trimmed means, making it appropriate when using

bootstrap-t tests or non-normal distributions. The thresholds are analogous to Cohens' d (Mair and Wilcox, 2020; Wilcox, 2021), $d \geq 0.2 \triangleq$ small effect; $d \geq 0.5 \triangleq$ medium effect and $d \geq 0.8 \triangleq$ large effect (Cohen, 1988). ξ can be interpreted like r (Cohen, 1992; Mair and Wilcox, 2020).

Results

Group composition

To ensure baseline equivalence between treatment groups, we compared pretest scores and cognitive ability test scores across groups. No significant differences could be found for the pretest scores, $F(2, 43.73)$ = 2.01, p = 0.147, ξ = 0.27, 95% CI [0.07; 0.47] and the cognitive ability test scores between the groups, $F(2, 43)$ = 1.21, p = 0.307, ξ = 0.23, 95% CI [0.05; 0.44]. These findings indicate that the treatment groups did not differ meaningfully at baseline.

RQ1: Influence of treatment groups on learning gain

Effects of the treatment groups on students' posttest scores. A one-way between-subjects ANOVA revealed no significant group differences in posttest scores, $F(2, 43.94)$ = 0.47, p = 0.630, ξ = 0.19, 95% CI [0.03; 0.39]. Thus, students had similar posttest scores across all three groups.

A robust repeated measures ANOVA revealed a statistically significant main effect, $F(1, 66)$ = 22.56, p < 0.001, regarding the learning gains from pre- to posttest across all groups, albeit with a small effect size (Fig. 6). However, the three treatment groups exhibited differing learning gains. Only the *cc* and *ccsf* groups demonstrated significant learning gains from pretest to posttest, while the *sc* group showed no significant learning gain. Notably, the *cc* group achieved a medium-sized learning gain (medium effect, *AKP* = 0.65), which was larger than the small effect observed in the *ccsf* group (small effect, *AKP* = 0.46). These findings reveal that while both contrasting case conditions yielded significant learning gains, the gain observed in the *cc* group was greater.

**Effects of the treatment groups on students' modes of reasoning in the constructed response tasks (CRTs).** The three groups showed no significant difference in CRT scores, $F(2, 42.51)$ = 1.86, p = 0.168, ξ = 0.26, 95% CI [0.07; 0.47]. Fig. 7 reveals only minor variation in the distribution of students' application of the modes of reasoning, *e.g.*, more descriptive instances in the *cc* group; more multicomponent causal ones in the *ccsf* group. Thus, task format and scaffolding did not translate into overall differences in CRT score.

RQ2: Dependence of learning gain on students' prior knowledge

Influence of the treatment group on the posttest score depending on the pretest score. The robust regression with prior knowledge as a covariate revealed significant main effects of the pretest score (standardised(std.) $\beta_{pretest}$ = 0.88, p < 0.001) and *cc* group membership (std. β_{cc} = 0.14, p = 0.036) on the posttest score (Fig. 8). The strong effect of the pretest score

Pre- and posttest score by group
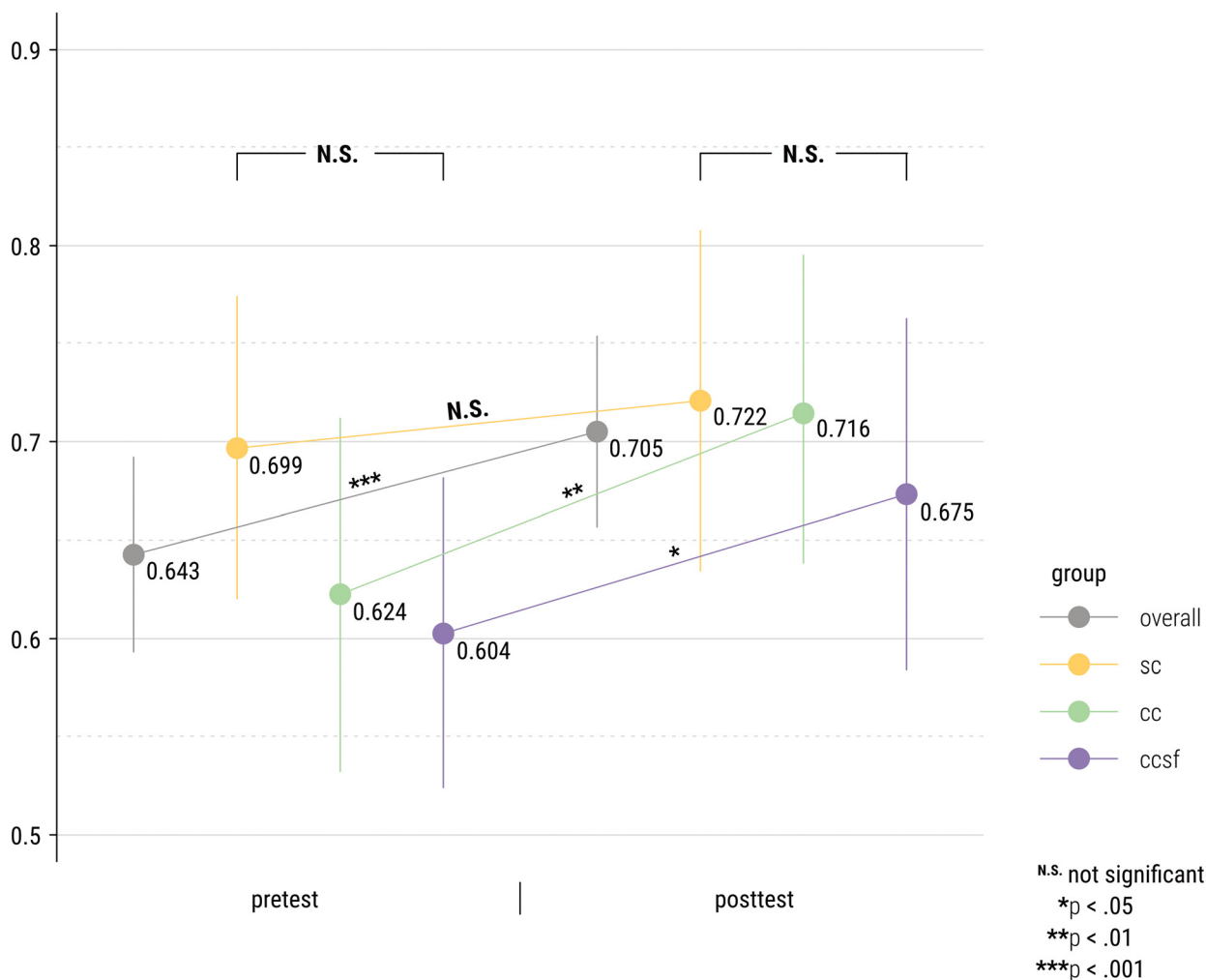
test results (pre-post comparisons)

Group	μ_{post}	μ_{pre}	p	$\mu_{post}-\mu_{pre}$	AKP
overall	0.705	0.643	**<.001**	0.063, 95% CI [0.038; 0.088]	**0.46, 95% CI [0.25; 0.70]**
sc	0.722	0.699	.934	0.024, 95% CI [-0.016; 0.064]	0.23, 95% CI [-0.10; 0.64]
cc	0.716	0.624	**.002**	0.092, 95% CI [0.043; 0.140]	**0.65, 95% CI [0.22; 1.21]**
ccsf	0.675	0.604	**.022**	0.071, 95% CI [0.027; 0.115]	**0.46, 95% CI [0.17; 0.88]**

Fig. 6 Results of robust tests of students' learning gain from pre- to posttest. Numbers near the data points in the plot represent trimmed means for the respective group and test. N.S., *, **, and *** represent whether the respective test was significant. The table below shows the corresponding results for each group. The μ represents the trimmed means, while $\mu-\mu$ represents a trimmed mean difference with the corresponding 95% confidence interval. *AKP* represents the effect size as introduced in the data analysis part of the methods with the corresponding 95% confidence interval. Significant p-values and effect sizes are printed bold. All p-values were adjusted with Bonferroni correction.

indicates that students with higher prior knowledge tended to perform substantially better on the posttest. Also, *cc* group membership leads to significantly higher posttest scores than *sc* group membership with a medium effect, after accounting for prior knowledge.

Importantly, a significant negative interaction between pretest score and *cc* group membership (std. $\beta_{\text{pretest}\times\text{cc}} = -0.31$, $p = 0.006$) suggests that the effectiveness of the contrasting cases in the *cc* group depended on students' prior knowledge (for

visualisation, see Fig. 8). Follow-up pairwise comparison showed that low prior knowledge students in the *cc* group outperformed their counterparts with low prior knowledge in the *sc* group ($F_t = 2.032$, $p = 0.037$, $\hat{\psi} = 0.144$, 95% CI [0.007; 0.281], *AKP* = 0.78, 95% CI [0.06; 1.86]). No difference was found among students with high prior knowledge between the *cc* and *sc* groups ($p = 0.945$). This suggests that the contrasting cases were particularly beneficial for students with lower prior knowledge.
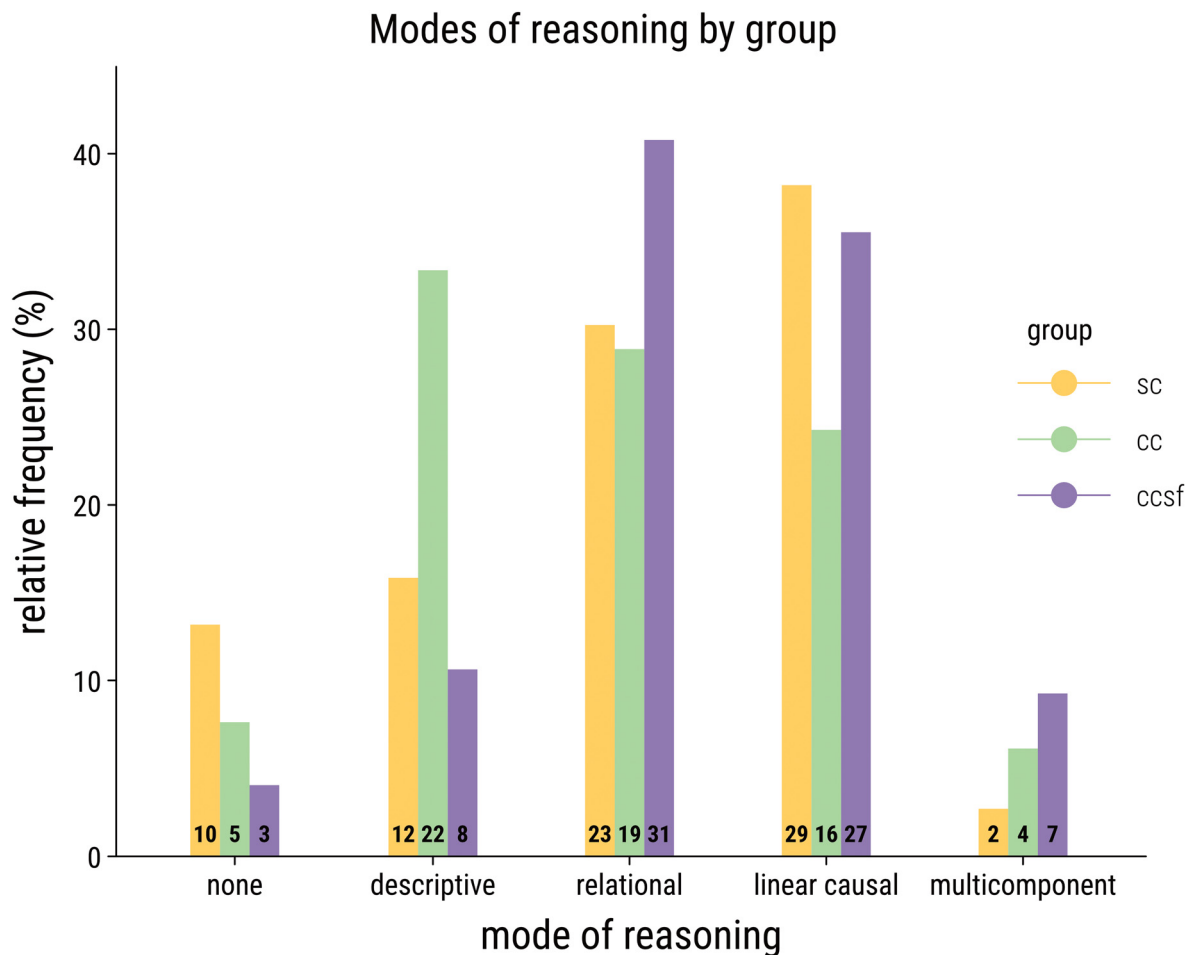
Modes of reasoning by group

Fig. 7 Distribution of the modes of reasoning by groups. The numbers on the bars show the absolute numbers of the respective modes in the respective groups. Since two CRTs were answered the numbers of answers in the diagram ($n_{\text{answers sc}} = 76$, $n_{\text{answers cc}} = 66$, $n_{\text{answers ccsf}} = 76$) are twice the number of students in each group.

Influence of the treatment group on students' CRT score depending on the pretest score. Using a similar robust regression for the CRT scores revealed no significant interaction effects between pretest score and *cc* group or pretest score and *ccsf* group (std. $\beta_{\text{cc} \times \text{pretest}} = -0.17$, $p = 0.235$ and std. $\beta_{\text{ccsf} \times \text{pretest}} = -0.35$, $p = 0.053$). The analysis revealed medium-sized significant main effects of the pretest score (std. $\beta_{\text{pretest}} = 0.61$, $p < 0.001$) and *ccsf* group membership (std. $\beta_{\text{ccsf}} = 0.30$, $p = 0.002$) on students' use of the modes of reasoning (Fig. 9). The effect of the pretest score indicates that students with higher prior knowledge generally reached higher modes of reasoning. In addition, the *ccsf* group showed significantly higher modes of reasoning than the *sc* group, after accounting for prior knowledge.

Discussion

RQ1: Influence of treatment groups on learning gain

Based on prior research on contrasting cases in organic chemistry, we expected both the *cc* and *ccsf* groups to outperform the *sc* group because contrasting cases help students notice implicit structural features by comparing alternatives (Caspari and Graulich, 2019; Kranz *et al.*, 2023). Additionally, given that scaffolding is intended to help students connect conceptual knowledge and reason mechanistically (Caspari *et al.*, 2018; Graulich and Caspari, 2021; Watts *et al.*, 2021; Kranz *et al.*, 2023), we had anticipated that the *ccsf* group would outperform both the *cc* and *sc* groups. However, these expectations were not supported by the posttest scores or the CRT results. Neither posttest scores nor CRT scores differed significantly between the groups.

A plausible explanation for this finding is the strong influence of prior knowledge on students' posttest (Fig. 8) and CRT scores (Fig. 9). Students with high prior knowledge achieved similar posttest scores across all groups (*sc*: $M = 0.800$; *cc*: $M = 0.808$; *ccsf*: $M = 0.830$), whereas those with low prior knowledge exhibited substantial differences (*sc*: $M = 0.490$; *cc*: $M = 0.607$; *ccsf*: $M = 0.568$). Students with high prior knowledge also showed similar CRT scores across groups (*sc*: $M = 0.610$, *cc*: $M = 0.607$, *ccsf*: $M = 0.658$), while students with low prior knowledge varied substantially in their CRT scores (*sc*: $M = 0.298$, *cc*: $M = 0.368$, *ccsf*: $M = 0.543$). These findings suggest that students with high prior knowledge may benefit less from scaffolded support or contrasting cases, as they are already able to recognise, link, and apply relevant concepts (Kalyuga, 2007).

Interaction effects of groups and pretest score regarding the posttest score
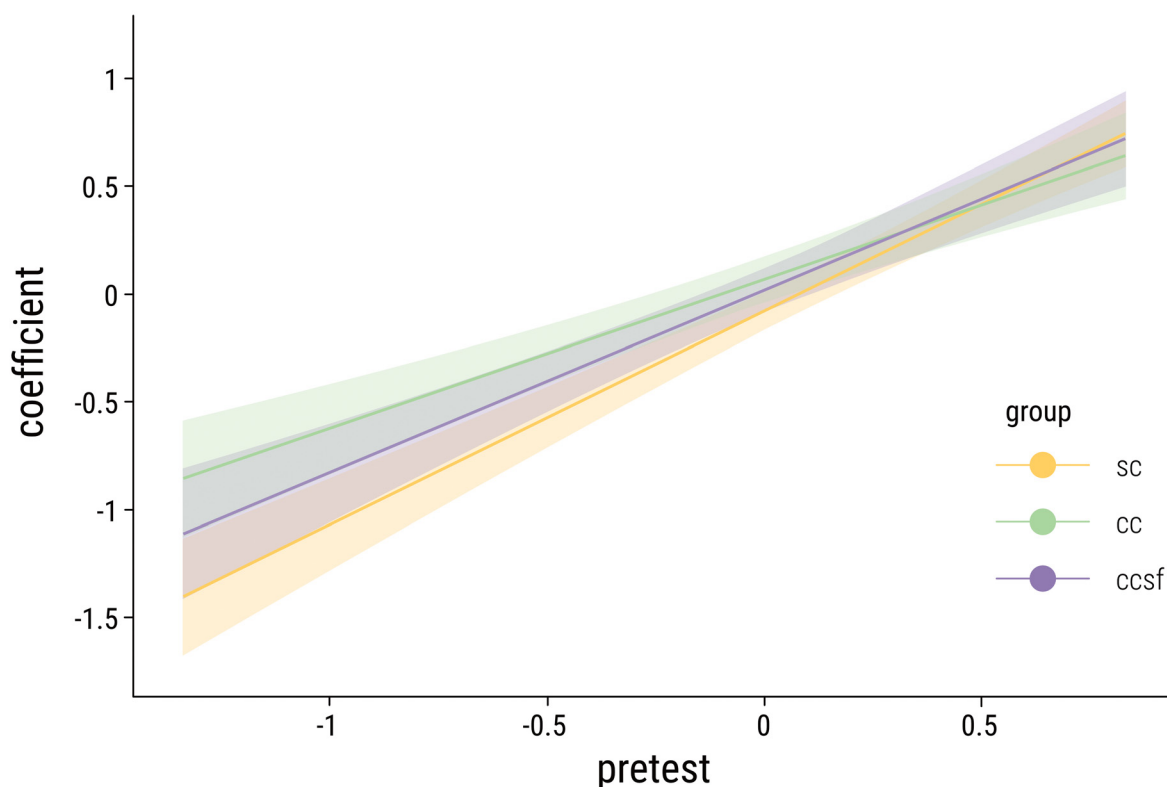
model parameters

Variable	Std. β	SE	t	p
pretest	.88	.05	16.90	**< .001**
cc	.14	.07	2.12	**.036**
ccsf	.07	.06	1.10	.275
pretest:cc	-.31	.11	-2.79	**.006**
pretest:ccsf	-.06	.13	-0.51	.615

Fig. 8 Interaction effects between the treatment groups and the pretest score on the posttest score and model parameters of the robust linear model for the posttest score. The x-axis shows the standardised pretest score, while the y-axis shows the resulting posttest score. The lines in the plot represent the interaction effects. The steepness of the slopes shows how large the association between pretest score and posttest score is in each group.

This interpretation aligns with our regression results, which indicate that the scaffold in the *ccsf* condition primarily benefited students with low prior knowledge on both posttest and CRT scores, while students with high prior knowledge showed no significant effects. Due to the similar scores among students with high prior knowledge across treatments, statistical analyses did not yield significant differences. However, task design appears to play a critical role for students with low prior knowledge.

When examining students' learning gains, rather than their posttest scores, the *cc* and *ccsf* groups showed significant learning gains from pretest to posttest compared to the *sc* group (Fig. 6). Particularly, the *cc* group exhibited a larger learning gain (Δ = 0.092, *AKP* = 0.65) compared to the *ccsf* group (Δ = 0.071, *AKP* = 0.46)—an unexpected finding given the assumed support by the scaffold.

Cognitive load theory may account for this unexpected finding (Sweller *et al.*, 2019). Students in the *ccsf* group reported higher perceived difficulty and lower confidence (F_t = 1.91, p = 0.046, $\hat{\psi}$ = 0.34, 95% CI [0.01; 0.67], *AKP* = 0.47, 95% CI [0.005; 0.946]). This suggests increased extraneous cognitive load, potentially attenuating their learning gain. Hence, overly complex scaffolds can increase extraneous cognitive load, especially for students unfamiliar with them (Sweller *et al.*, 2019). In contrast, non-scaffolded contrasting cases may leave more cognitive resources available for
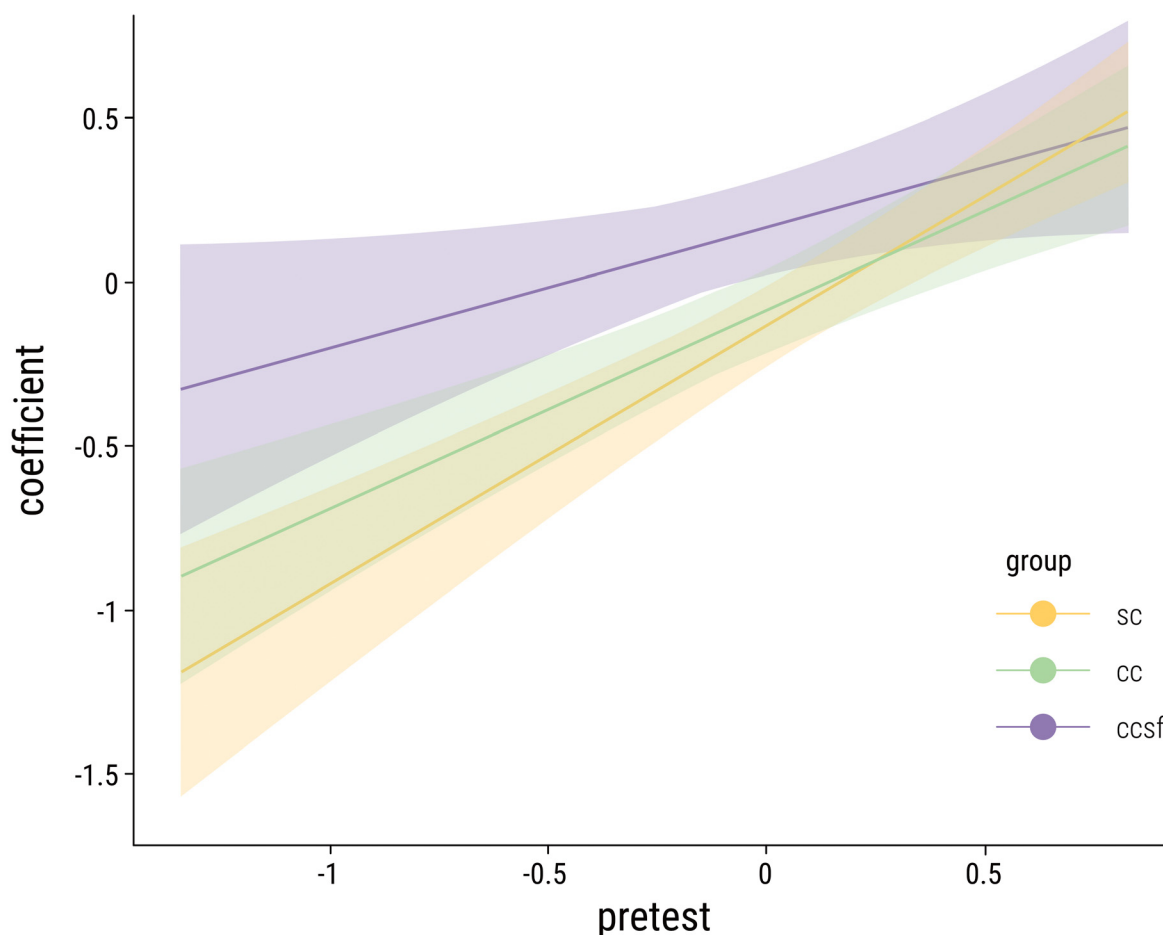
Fig. 9 Interaction effects between the treatment groups and the pretest score on the CRT score and model parameters of the robust linear model for the CRT score. The *x*-axis shows the standardised pretest score, while the *y*-axis shows the resulting posttest score. The lines in the plot represent the interaction effects. The steepness of the slopes shows how large the association between pretest score and CRT score is in each group.

applying concept knowledge, possibly explaining the medium-sized gain observed in the *cc* group.

RQ2: Dependence of learning gain on students' prior knowledge

Building on prior evidence that students with low prior knowledge benefit most from scaffolding (Kranz *et al.*, 2023), we investigated to what extent learning gains depend on students'

prior knowledge. Among other findings, a robust linear regression examining the effect of treatment group on students' posttest scores as a function of their pretest scores (Fig. 8) revealed a significant negative interaction between membership in the *cc* group and prior knowledge (std. $\beta_{cc \times pretest} = -0.31$, $p = 0.006$). Practically, this indicates that contrasting cases reduced the differences in posttest scores between

students with low and high prior knowledge. Specifically, students with low prior knowledge in the *cc* group outperformed their peers in the *sc* group ($F_t = 2.032$, $p = 0.037$, $\hat{\psi} = 0.144$, 95% CI [0.007; 0.281], $AKP = 0.78$, 95% CI [0.06; 1.86]). In contrast, no such difference was found among students with high prior knowledge ($p = 0.945$). Contrasting cases highlight explicit structural differences between molecules, which can support students in reasoning about these differences (Watts *et al.*, 2021). Therefore, a contrasting cases task may help students with low prior knowledge to identify more structural differences than they would in a single-case task—and to incorporate these differences into their reasoning while working on the concept knowledge test.

Similar effects were not observed in the *ccsf* group when analysing their posttest scores. It is, thus, possible that the stepwise instructions of the unfamiliar scaffold distracted students from focusing on critical features of the reaction mechanisms, or that the increased perceived task difficulty contributed to this null effect (Moos and Pitton, 2014).

For the CRT scores, the regression (Fig. 9) showed no significant interaction effects (std. $\beta_{cc\times pretest} = -0.17$, $p = 0.235$ and std. $\beta_{ccsf\times pretest} = -0.35$, $p = 0.053$). However, among students with low prior knowledge, the *ccsf* group achieved the highest CRT score (0.543), compared to the *cc* (0.368) and *sc* groups (0.298). This difference in CRT scores between the *ccsf* and *sc* group has a large effect size ($F_t = 2.171$, $p = 0.048$, $\hat{\psi} = 0.233$, 95% CI [0.002; 0.465], $AKP = 0.90$, 95% CI [0.20; 2.42]). In contrast, no significant difference was found among students with high prior knowledge ($p = 0.165$). The CRT responses of students with low prior knowledge in the *ccsf* group showed more linear causal and multicomponent causal modes of reasoning (Fig. 10). The subtask prompts of the scaffold may have facilitated more elaborate reasoning. Being instructed to explicitly identify and connect the different features of the shown reactions may have led students with low prior knowledge to integrate multiple chemical concepts while working on the tasks (Weinrich and Talanquer, 2016; Kranz *et al.*, 2023), achieving more elaborate reasoning in the CRTs than students
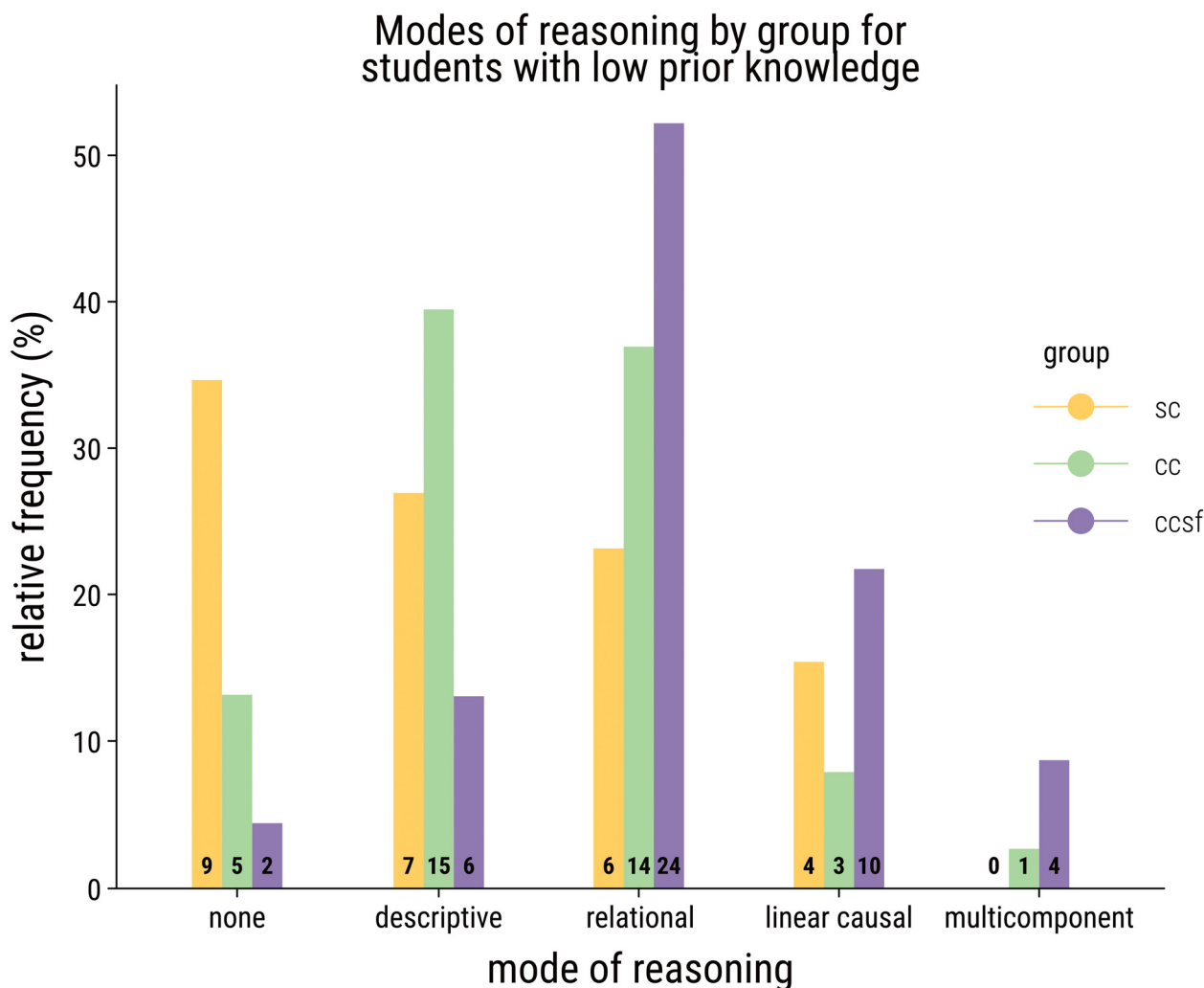

Fig. 10 Distribution of the qualitatively coded modes of reasoning based on students' answers (students with pretest scores below the mean). The numbers on the bars show the absolute numbers of the respective modes in the respective groups. Since two CRTs were answered the numbers of answers in the diagram ($n_{answers\,sc\,low} = 26$, $n_{answers\,cc\,low} = 38$, $n_{answers\,ccsf\,low} = 46$) are twice the number of students in each group.

with low prior knowledge in the *sc* group. These findings are consistent with findings of a previous study by Kranz *et al.* (2023), which documents that students with low prior knowledge also significantly benefited from working with the scaffolded contrasting case, showing greater score gains than their peers with high prior knowledge, whose scores remained stable. Haas *et al.* (2024) could also show that scaffolded contrasting cases contribute to more in-depth reasoning. In a qualitative study, Caspari and Graulich (2019) provide an explanation for this finding. They observed in their study that students working with scaffolded contrasting cases were in 80% of the cases able to identify more than one implicit influence on the reaction rate compared to non-scaffolded contrasting cases—a prerequisite for higher modes of reasoning. However, this more elaborate reasoning did not translate into better posttest scores as mentioned above.

The absence of a main effect for the *cc* group in the CRT scores is notable, especially given the posttest differences. We had expected that students with low prior knowledge in the *cc* group would outperform those in the *sc* group on the CRTs, which was not observed. These results suggest that scaffolded prompts enable students to engage in more elaborate mechanistic reasoning compared to non-scaffolded contrasting cases.

Overall, these findings suggest for students with low prior knowledge in the CRTs that:

(1) non-scaffolded contrasting cases seem to foster more descriptive and relational modes of reasoning compared to students in the *sc* group.

(2) scaffolded contrasting cases seem to foster even higher modes of reasoning, such as linear causal or multicomponent causal, compared to students in the *cc* group, although they reported a higher perceived difficulty (perceived difficulty: $F_t = 1.91$, $p = 0.046$).

Students with high prior knowledge seem to be able to solve tasks effectively regardless of the task format. Consistent with the expertise-reversal effect, additional prompts may become redundant or distracting for these students (Kalyuga *et al.*, 2003; Kalyuga, 2007).

Conclusions

This study compared the effect of the task format, single (*sc*) or contrasting cases (*cc*), and the type of support, scaffolded contrasting cases (*ccsf*), on students' learning gains. Both the *cc* and *ccsf* groups demonstrated significant learning gains from pretest to posttest, with the *cc* group showing a more substantial gain (medium effect size) than the *ccsf* group (small effect size). This difference may be attributable to the added complexity of the scaffolded task, which likely contributed to higher perceived difficulty and lower confidence among students in the *ccsf* group. However, neither the *cc* nor the *ccsf* group significantly outperformed the *sc* group in terms of overall posttest scores.

In the second part of the analysis, students' prior knowledge was considered as a covariate. A robust linear regression revealed a significant interaction between *cc* group membership and pretest scores on posttest scores. Pairwise comparisons demonstrate that contrasting cases are particularly beneficial for students with low prior knowledge. Although, we did not observe a significant interaction between pretest scores and *ccsf* group membership on the CRT scores, pairwise comparisons show a significant advantage of the *ccsf* group over the *sc* group among students with low prior knowledge when evaluating the modes of reasoning. This suggests that scaffolded contrasting cases can effectively support students with low prior knowledge. This aligns with qualitative research that documents that scaffolded contrasting cases enhance causal reasoning (Haas *et al.*, 2024).

Limitations

The analysis was conducted with 109 datasets, which is still a relatively small sample by quantitative standards. To validate the results, a follow-up study should include a larger number of students.

When analysing the data, some group parameters were not normally distributed. A larger number of students would likely have produced normal distributions. Nevertheless, we selected robust test procedures (Field *et al.*, 2012; Field and Wilcox, 2017; Wilcox, 2021) as the most appropriate way to obtain interpretable results.

Statistical tests showed no significant prior knowledge differences between the three treatment groups. Qualitatively, however, small disparities in pretest scores were noticeable and should not be overlooked. Descriptively, students in the *sc* group had higher average pretest scores, which may have affected their learning gain. Future studies could pretest prior knowledge first and then allocate students to homogeneous groups.

There is still debate if perceived task difficulty and confidence are proxies for cognitive load compared with established scales (Paas *et al.*, 2003). As justified in the Methods, we treated them only as rough estimates.

The intervention covered a short time span and, therefore, cannot fully demonstrate the extent of the learning gains that would occur if the task formats were used for a more extended period. Longitudinal studies are needed to confirm and extend these findings for contrasting cases and scaffolded contrasting cases.

For administrative reasons, we collected the demographic data at the beginning of the pretest. Doing so may inadvertently have primed participants' social identities and induced stereotype threat, potentially influencing their subsequent engagement with the intervention.

Implications for research

Our work offers several implications for future research. First, a scaffold may have different effects on procedural and conceptual knowledge (Anderson *et al.*, 2001), which is why developing an instrument that more precisely distinguishes these variables

would allow clearer statements about the specific impact of contrasting cases and scaffolding. Especially given the findings that students with low prior knowledge had higher modes of reasoning in the CRTs, but this was not reflected in higher learning gains. Accordingly, future research needs to more clearly differentiate between students' conceptual and procedural knowledge in undergraduate organic chemistry.

As an alternative to the scaffold used here, future work should examine adaptive support that adapts to students' concept knowledge and mechanistic reasoning in real time, *e.g.*, by directed prompting or providing the missing conceptual pieces (Lieber *et al.*, 2022a, 2022b). Such adaptive scaffolds could vary in complexity according to prior knowledge. To develop such adaptive scaffolds, additional research is needed to clarify how the reasoning process varies across different levels of prior knowledge and how scaffolds could be tailored accordingly. To date, effects of fading, such as gradually removing support has not been explored with this scaffold type (Puntambekar and Hubscher, 2005; McNeill *et al.*, 2006; Lin *et al.*, 2012).

Finally, mixed-methods studies linking process data (*e.g.*, log files) with outcome measures could reveal how and when students actually engage with scaffold prompts. Such analyses could help uncover the mechanisms through which scaffolding supports learning and identify the conditions under which it is most effective.

Implications for teaching

This study demonstrates that non-scaffolded contrasting cases lead to higher learning gains than traditionally used single cases for students with low prior knowledge in organic chemistry. Consequently, contrasting cases can be introduced early in a course to familiarise students with discipline-specific reasoning, or whenever new mechanisms are taught.

Although both the *cc* and *ccsf* groups improved from pretest to posttest, scaffolded contrasting cases did not automatically produce larger gains than contrasting cases alone. Nevertheless, qualitative research documents that scaffolded contrasting cases enhance causal reasoning (Haas *et al.*, 2024). Our CRT findings—higher modes of reasoning among students with low prior knowledge in the *ccsf* group—support this benefit. Instructional strategies, such as providing worked-out examples in class, could support students' understanding of how to approach contrasting cases or use scaffolds, and thus minimise negative effects on cognitive load. Using the scaffold regularly and from the outset may prevent cognitive overload (Sweller *et al.*, 2019) and build routines for activating key concepts. At the same time, scaffolds should remain optional for advanced students who may not need them (Kalyuga, 2007). Teachers might adopt a *scaffold-on-demand* approach: provide the grid only when students request help or show signs of impasse, thereby balancing support and autonomy.

Ethical statement

Although IRB approval is not required at German Universities, the data collection was based on ethical guidelines (German Research Foundation (DFG), 2022) and gave students the opportunity to opt-out anytime. Students were informed about their rights regarding their data and about the usage of the data (European Union, 2016) both in a written and verbal format.

Conflicts of interest

There are no conflicts of interest to declare.

Data availability

The data are not publicly available as participants of this study did not consent for their data to be shared publicly.

Appendices

Appendix 1   Translations

Translations of students' German original quotes into English.

Original (German)	Translation (English)
Die Reaktion A läuft schneller ab.	Reaction A occurs faster.
Kation A: Stabilisierung der Ladung möglich durch eine Hyperkonjugation und einen +M Effekt ausgelöst durch die Doppelbindung.	Cation A: Stabilisation of the charge by means of hyperconjugation and a resonance effect induced by the double bond.
Hier muss man nur die Kationen miteinander vergleichen. Bei A wirkt ein +M-Effekt, +I-Effekt und Hyperkonjugation. Bei [B] tritt Hyperkonjugation und ein stärkerer +I-Effekt auf, da mehr Methylgruppen anhängen. Die Anionen sind gleich. Die Möglichkeit einer mesomeren Ladungsstabilisierung bei A ist daher für mich ausschlaggebend.	In this case, only the cations have to be compared with each other. In A, there is a resonance effect, a positive inductive effect and hyperconjugation. In [B], hyperconjugation and a stronger positive inductive effect occur because more methyl groups are attached. The anions are the same. The possibility of resonance stabilisation in A is, therefore, crucial for me.
Bei Reaktion A entsteht ein Kation, welches am 3C-Atom eine positive Ladung trägt. Diese wird durch den Induktiven Effekt der anliegenden Methylgruppe stabilisiert, indem Elektronendichte in das leere p-Orbital an 3C doniert werden kann. Weiterhin ist das Kation	In reaction A, a cation is formed which has a positive charge at the 3C atom [third carbon atom]. This is stabilised by the [positive] inductive effect of the adjacent methyl group as electron density can be donated into the empty p-orbital at 3C [third carbon atom]. Furthermore, the

Table (*continued*)

Original (German)	Translation (English)
mesomeriestabilisiert. Die Ladung kann durch Überklappen der Doppelbindung auf das 1C-Atom des Moleküls umverlagert werden. Dadurch ist sie delokalisiert und das entstandene Ion liegt energetisch wesentlich niedriger, als es ohne Delokalisierung der Fall wäre.	cation is resonance stabilised. The charge can shift to the 1C atom [first carbon atom] of the molecule due to the double bond. Therefore, it [the positive charge] is delocalised and the resulting ion is energetically much lower than it would be without delocalisation.
B ist schneller.	B is faster.
Reaktion A läuft schneller ab, da der Propylrest besser stabilisiert ist.	Reaction A occurs faster, because the propyl is more stabilised.
Die Reaktion läuft schneller ab, da eine elektronenschiebende Wirkung auf das C-Atom ausgeübt wird.	The reaction proceeds faster because an electron-pushing effect is performed on the C atom [carbocation].
Reaktion B verläuft schneller, da hier die Bindungsspaltung des Broms durch mehr elektronenschiebenden Effekt an das C [Kohlenstoffatom] begünstigt ist. Das entstehende Produkt (Carbenium-Ion) bei B ist durch die 3 Methylgruppen mit elektronennschiebenden Effekten besser stabilisiert als bei A mit nur 2 Methylgruppen am C [Kohlenstoffatom], welches die Ladung trägt.	Reaction B proceeds faster because in this case, the cleavage of the bond to the bromine is favoured by a greater electron-pushing effect towards the carbon atom. The resulting product (carbocation) in B is better stabilised by the three methyl groups with electron-pushing effects than in A, which carries the charge with only 2 methyl groups attached to the carbon atom.
Brom ist eine gute Abgangsgruppe (polarisierbar, das entstehende Bromid ist eine sehr schwache Base), das entstehende Carbeniumion ist relativ gut stabilisiert (3fache Hyperkonjugation). Außerdem wird die Entropie vergrößert, da im Zuge der Reaktion mehr Teilchen entstehen. Trotzdem muss erst einmal die C–Br – Bindung gespalten werden, was Energie benötigt. Relativ zu anderen heterogenen Bindungsspaltungen sollte die Reaktion thermodynamisch begünstigt sein, insgesamt benötigt man jedoch Energie.	Bromine is a good leaving group (polarisable, the resulting bromide is a very weak base) and the resulting carbocation is relatively well stabilised (3-fold hyperconjugation). In addition, the entropy is increased because more particles are produced during the reaction. However, first the C–Br bond needs to be cleaved, which requires energy. Compared to other heterogeneous bond cleavages, the reaction should be thermodynamically favoured, but overall energy is required.

Appendix 2 Exemplary items of the concept knowledge test

Appendix 3 Open-ended contrasting case tasks
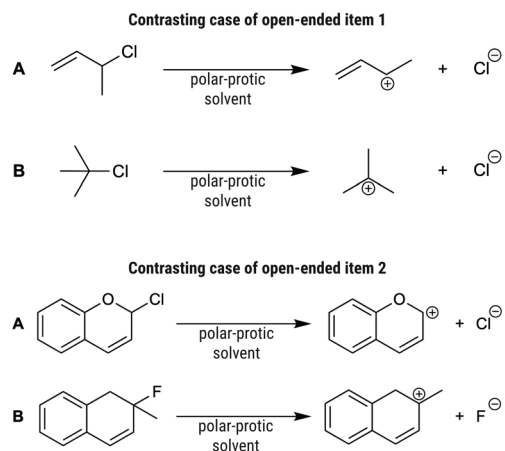
for his productive discussions and support. This work was supported by the German Research Foundation DFG (Deutsche Forschungsgemeinschaft) under grant number: 446349713.

Acknowledgements

References

Abramovich S., Schunn C. and Higashi R. M., (2013), Are badges useful in education? it depends upon the type of badge and expertise of learner, *Educ. Tech. Res.*, **61**, 217–232.

Alfieri L., Nokes-Malach T. J. and Schunn C. D., (2013), Learning through case comparisons: a meta-analytic review, *Educ. Psychol.*, **48**(2), 87–113.

Algina J., Keselman H. and Penfield R. D., (2005), An alternative to Cohen's standardized mean difference effect size: a robust parameter and confidence interval in the two independent groups case, *Psychol. Methods*, **10**(3), 317–328.

Anderson L. W., Krathwohl D. R., Airasian P. W., Cruikshank K. A., Mayer R. E., Pintrich P. R. and Wittrock M. C., (2001), *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*, White Plains, NY: Longman.

Bachtiar R. W., Meulenbroeks R. F. and van Joolingen W. R., (2022), Mechanistic reasoning in science education: a literature review, *EURASIA J. Math., Sci Tech. Ed.*, **18**(11), em2178.

Belland B. R., (2011), Distributed cognition as a lens to understand the effects of scaffolds: the role of transfer of responsibility, *Educ. Psychol. Rev.*, **23**(4), 577–600.

Bhattacharyya G. and Bodner G. M., (2005), "It gets me to the product": How students propose organic mechanisms, *J. Chem. Educ.*, **82**(9), 1402.

Bodé N. E., Deng J. M. and Flynn A. B., (2019), Getting past the rules and to the WHY: causal mechanistic arguments when judging the plausibility of organic reaction mechanisms, *J. Chem. Educ.*, **96**(6), 1068–1082.

Bonferroni C. E., (1936), Teoria statistica delle classi e calcolo delle probabilita [Statistical theory of classification and probability calculus], *Pubblicazioni del R. istituto superiore di scienze economiche e commericiali di firenze*, **8**, 3–62.

Bratfisch O., Borg G. and Dornic S., (1972), Perceived item-difficulty in three tests of intellectual performance capacity, *Rep. Inst. Appl. Psychol. Univ. Sweden*, **29**, 1–17.

Braun I. and Graulich N., (2024), Exploring diversity: student's (un-) productive use of resonance in organic chemistry tasks through the lens of the coordination class theory, *Chem. Educ. Res. Pract.*, **25**(3), 643–671.

Bussey T. J., Orgill M. and Crippen K. J., (2013), Variation theory: a theory of learning and a useful theoretical framework for chemical education research, *Chem. Educ. Res. Pract.*, **14**(1), 9–22.

Caspari I. and Graulich N., (2019), Scaffolding the structure of organic chemistry students' multivariate comparative mechanistic reasoning, *Int. J. Phys. Chem. Educ.*, **11**(2), 31–43.

Caspari I., Kranz D. and Graulich N., (2018), Resolving the complexity of organic chemistry students' reasoning through the lens of a mechanistic framework, *Chem. Educ. Res. Pract.*, **19**(4), 1117–1141.

Chen I. and Chang C.-C., (2009), Cognitive load theory: an empirical study of anxiety and task performance in language learning, *Electron. J. Res. Educ. Psychol.*, **7**(2), 729–746.

Chin D. B., Chi M. and Schwartz D. L., (2016), A comparison of two methods of active learning in physics: inventing a general solution versus compare and contrast, *Instr. Sci.*, **44**(2), 177–195.

Cohen J., (1988), *Statistical power analysis for the behavioral sciences*, New York: Academic Press.

Cohen J., (1992), A power primer, *Psychol. Bull.*, **112**(1), 155–159.

Cooper M. M., Kouyourndjian H. and Underwood S. M., (2016), Investigating students' reasoning about acid–base reactions, *J. Chem. Educ.*, **93**(10), 1703–1712.

Cronbach L. J., (1951), Coefficient alpha and the internal structure of tests, *Psychometrika*, **16**(3), 297–334.

Cureton E. E. and D'Agostino R. B., (2013), *Factor analysis: An applied approach*, New York: Psychology press.

Deng J. M. and Flynn A. B., (2021), Reasoning, granularity, and comparisons in students' arguments on two organic chemistry items, *Chem. Educ. Res. Pract.*, **22**(3), 749–771.

Domin D. S., Al-Masum M. and Mensah J., (2008), Students' categorizations of organic compounds, *Chem. Educ. Res. Pract.*, **9**(2), 114–121.

Dood A. J. and Watts F. M., (2022), Mechanistic reasoning in organic chemistry: a scoping review of how students describe and explain mechanisms in the chemistry education research literature, *J. Chem. Educ.*, **99**(8), 2864–2876.

Dood A. J. and Watts F. M., (2023), Students' strategies, struggles, and successes with mechanism problem solving in organic chemistry: a scoping review of the research literature, *J. Chem. Educ.*, **100**(1), 53–68.

Eckhard J., Rodemer M., Bernholt S. and Graulich N., (2022), What do university students truly learn when watching tutorial videos in organic chemistry? An exploratory study focusing on mechanistic reasoning, *J. Chem. Educ.*, **99**(6), 2231–2244.

European Union, (2016), Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), *Off. J. Eur. Union*, **L119**, 1–88.

Field A. P., Miles J. and Field Z., (2012), *Discovering statistics using R*, London: Sage.

Field A. P. and Wilcox R. R., (2017), Robust statistical methods: a primer for clinical psychology and experimental psychopathology researchers, *Behav. Res. Ther.*, **98**, 19–38.

Filzmoser P. and Varmuza K., (2017), *chemometrics: Multivariate statistical analysis in chemometrics (version: 1.4.2)*, **https://cran.r-project.org/src/contrib/Archive/chemometrics/chemometrics_1.4.2.tar.gz**.

Fleiss J. L. and Cohen J., (1973), The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, *Educ. Psychol. Measurement*, **33**(3), 613–619.

Fox J. and Weisberg S., (2019), *An {R} companion to applied regression*, Thousand Oaks, CA, USA: Sage.

Galloway K. R., Leung M. W. and Flynn A. B., (2018), A comparison of how undergraduates, graduate students, and professors organize organic chemistry reactions, *J. Chem. Educ.*, **95**(3), 355–365.

Gavas R. D., Tripathy S. R., Chatterjee D. and Sinha A., (2018), Cognitive load and metacognitive confidence extraction from pupillary response, *Cogn. Syst. Res.*, **52**, 325–334.

Gelman A., (2008), Scaling regression inputs by dividing by two standard deviations, *Stat. Med.*, **27**(15), 2865–2873.

German Research Foundation (DFG), (2022), Guidelines for safeguarding good research practice, *Code of Conduct.*, DOI: **10.5281/zenodo.6472827**.

Graulich N., (2015), The tip of the iceberg in organic chemistry classes: how do students deal with the invisible? *Chem. Educ. Res. Pract.*, **16**(1), 9–21.

Graulich N. and Bhattacharyya G., (2017), Investigating students' similarity judgments in organic chemistry, *Chem. Educ. Res. Pract.*, **18**(4), 774–784.

Graulich N. and Caspari I., (2021), Designing a scaffold for mechanistic reasoning in organic chemistry, *Chem. Teach. Int.*, **3**(1), 19–30.

Graulich N., Kranz D. and Martin P. P., (2025), *Concept knowledge test Organic Chemistry*, DOI: **10.17605/OSF.IO/U8NPC**.

Graulich N., Langner A., Vo K. and Yuriev E., (2021), Scaffolding metacognition and resource activation during problem solving: a continuum perspective, in Tsaparlis G. (ed.), *Problems and problem solving in chemistry education: Analysing data, looking for patterns and making deductions*, London: Royal Society of Chemistry, pp. 38–67.

Graulich N. and Schween M., (2018), Concept-oriented task design: Making purposeful case comparisons in organic chemistry, *J. Chem. Educ.*, **95**(3), 376–383.

Grosjean P. and Ibanez F., (2018), *pastecs: Package for Analysis of Space-Time Ecological Series (version: 1.3.21)*, **https://cran.r-project.org/src/contrib/Archive/pastecs/pastecs_1.3.21.tar.gz**.

Grove N. P., Cooper M. M. and Cox E. L., (2012), Does mechanistic thinking improve student success in organic chemistry? *J. Chem. Educ.*, **89**(7), 850–853.

Haas D. B., Watts F. M., Dood A. J. and Shultz G. V., (2024), Analysis of organic chemistry students' developing reasoning elicited by a scaffolded case comparison activity, *Chem. Educ. Res. Pract.*, **25**(3), 742–759.

Heller K. A. and Perleth C., (2000), *KFT 4–12+ R kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision [CAT 4–12+ R cognitive-ability test for 4th to 12th grade]*, Beltz Test.

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.

Homer B. D. and Plass J. L., (2010), Expertise reversal for iconic representations in science visualizations, *Instr. Sci.*, **38**(3), 259–276.

Ince E., (2018), An overview of problem solving studies in physics education, *J. Educ. Learn.*, **7**(4), 191–200.

Kalyuga S., (2007), Expertise reversal effect and its implications for learner-tailored instruction, *Educ. Psychol. Rev.*, **19**(4), 509–539.

Kalyuga S., Ayres P., Chandler P. and Sweller J., (2003), The expertise reversal effect, *Educ. Psychol.*, **38**(1), 23–31.

Kranz D., Schween M. and Graulich N., (2023), Patterns of reasoning–exploring the interplay of students' work with a scaffold and their conceptual knowledge in organic chemistry, *Chem. Educ. Res. Pract.*, **24**(2), 453–477.

Levene H., (1961), Robust tests for equality of variances, in *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, ed. Olkin I., Ghurye S. G., Hoeffding W., Madow W. G. and Mann H. B., Stanford, CA: Stanford University Press, pp. 278–292.

Lieber L. S., Ibraj K., Caspari-Gnann I. and Graulich N., (2022a), Closing the gap of organic chemistry students' performance with an adaptive scaffold for argumentation patterns, *Chem. Educ. Res. Pract.*, **23**(4), 811–828.

Lieber L. S., Ibraj K., Caspari-Gnann I. and Graulich N., (2022b), Students' individual needs matter: a training to adaptively address students' argumentation skills in organic chemistry, *J. Chem. Educ.*, **99**(7), 2754–2761.

Lin T.-C., Hsu Y.-S., Lin S.-S., Changlai M.-L., Yang K.-Y. and Lai T.-L., (2012), A review of empirical evidence on scaffolding for science education, *Int. J. Sci. Math. Educ.*, **10**(2), 437–455.

Ling Lo M., (2012), *Variation theory and the improvement of teaching and learning*, Göteborg: Acta Universitatis Gothoburgensis.

Ling Lo M. and Marton F., (2011), Towards a science of the art of teaching: Using variation theory as a guiding principle of pedagogical design, *Int. J. Lesson Learn. Stud.*, **1**(1), 7–22.

Lombrozo T., (2006), The structure and function of explanations, *Trends Cogn. Sci.*, **10**(10), 464–470.

Maechler M., Rousseeuw P., Croux C., Todorov V., Ruckstuhl A., Salibian-Barrera M., Verbeke T., Koller M., Conceicao E. L. T. and Palma M. A. D., (2022), *robustbase: Basic Robust Statistics (version: 0.95-0)*, **https://cran.r-project.org/src/contrib/Archive/robustbase/robustbase_0.95-0.tar.gz**.

Mair P. and Wilcox R. R., (2020), Robust statistical methods in R using the WRS2 package, *Behav. Res. Methods*, **52**(2), 464–488.

Martin P. P., (2022), *Entwicklung und Validierung eines Diagnoseinstruments zu konzeptuellem und prozeduralem Wissen am Beispiel nukleophiler Substitutionsreaktionen in der Organischen Chemie [Development and validation of a diagnostic tool for conceptual and procedural knowledge using the example of nucleophilic substitution reactions in organic chemistry]*, unpublished scientific thesis, Justus-Liebig-University.

Martin P. P. and Graulich N., (2023), When a machine detects student reasoning: a review of machine learning-based formative assessment of mechanistic reasoning, *Chem. Educ. Res. Pract.*, **24**(2), 407–427.

Martin P. P., Kranz D. and Graulich N., (2024), Revealing Rubric Relations: Investigating the Interdependence of a Research-Informed and a Machine Learning-Based Rubric in Assessing Student Reasoning in Chemistry, *Int. J. Artif. Intell. Educ.*, 1–39.

McHugh M. L., (2012), Interrater reliability: the kappa statistic, *Biochem. Med.*, **22**(3), 276.

McNamara D. S. and Kintsch W., (1996), Learning from texts: effects of prior knowledge and text coherence, *Discourse Process.*, **22**(3), 247–288.

McNeill K. L., Lizotte D. J., Krajcik J. and Marx R. W., (2006), Supporting students' construction of scientific explanations by fading scaffolds in instructional materials, *J. Learn. Sci.*, **15**(2), 153–191.

Moos D. C. and Pitton D., (2014), Student teacher challenges: using the cognitive load theory as an explanatory lens, *Teach. Educ.*, **25**(2), 127–141.

Nückles M., Hübner S., Dümer S. and Renkl A., (2010), Expertise reversal effects in writing-to-learn, *Instr. Sci.*, **38**(3), 237–258.

Oksa A., Kalyuga S. and Chandler P., (2010), Expertise reversal effect in using explanatory notes for readers of Shakespearean text, *Instr. Sci.*, **38**(3), 217–236.

Paas F. and Sweller J., (2014), Implications of cognitive load theory for multimedia learning, in *The Cambridge handbook of multimedia learning*, ed. Mayer R. E., New York: Cambridge University Press, **vol. 2**, pp. 27–42.

Paas F., Tuovinen J. E., Tabbers H. and Van Gerven P. W. M., (2003), Cognitive load measurement as a means to advance cognitive load theory, *Educ. Psychol.*, **38**(1), 63–71.

Pea R. D., (2004), The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity, *J. Learn. Sci.*, **13**(3), 423–451.

Puntambekar S. and Hubscher R., (2005), Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educ. Psychol.*, **40**(1), 1–12.

Reiser B. J., (2004), Scaffolding complex learning: the mechanisms of structuring and problematizing student work, *J. Learn. Sci.*, **13**(3), 273–304.

Revelle W., (2022), *psych: Procedures for Personality and Psychological Research (version: 2.2.9)*, **https://cran.r-project.org/src/contrib/Archive/psych/psych_2.2.9.tar.gz**.

Rittle-Johnson B. and Star J. R., (2007), Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations, *J. Educ. Psychol.*, **99**(3), 561–574.

Rittle-Johnson B. and Star J. R., (2009), Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving, *J. Educ. Psychol.*, **101**(3), 529.

Rodemer M., Eckhard J., Graulich N. and Bernholt S., (2020), Decoding case comparisons in organic chemistry: eye-tracking students' visual behavior, *J. Chem. Educ.*, **97**(10), 3530–3539.

Rodemer M., Eckhard J., Graulich N. and Bernholt S., (2021), Connecting explanations to representations: benefits of highlighting techniques in tutorial videos on students' learning in organic chemistry, *Int. J. Sci. Educ.*, **43**(17), 2707–2728.

Russ R. S., Scherr R. E., Hammer D. and Mikeska J., (2008), Recognizing mechanistic reasoning in student scientific inquiry: a framework for discourse analysis developed from philosophy of science, *Sci. Educ.*, **92**(3), 499–525.

Salden R. J., Aleven V., Schwonke R. and Renkl A., (2010), The expertise reversal effect and worked examples in tutored problem solving, *Instr. Sci.*, **38**(3), 289–307.

Salmerón L., Kintsch W. and Caas J. J., (2006), Reading strategies and prior knowledge in learning from hypertext, *Mem. Cogn.*, **34**(5), 1157–1171.

Sevian H. and Talanquer V., (2014), Rethinking chemistry: a learning progression on chemical thinking, *Chem. Educ. Res. Pract.*, **15**(1), 10–23.

Shapiro S. S. and Wilk M. B., (1965), An analysis of variance test for normality (complete samples), *Biometrika*, **52**(3), 591–611.

Shemwell J. T., Chase C. C. and Schwartz D. L., (2015), Seeking the general explanation: a test of inductive activities for learning and transfer, *J. Res. Sci. Teach.*, **52**(1), 58–83.

Signorell A., Aho K., Alfons A., Anderegg N., Aragon T., Arachchige C., Arppe A., Baddeley A., Barton K., Bolker B., Borchers H. W., Caeiro F., Champely S., Chessel D., Chhay L., Cooper N., Cummins C., Dewey M., Doran H. C. and *et al.*, (2022), *DescTools: Tools for descriptive statistics. (version: 0.99.47)*, **https://cran.r-project.org/src/contrib/Archive/DescTools/DescTools_0.99.47.tar.gz**.

Sürücü L. and Maslaçi A., (2020), Validity and reliability in quantitative research, *Bus. Manag. Stud. Inter. J.*, **8**(3), 2694–2726.

Sweller J., (1994), Cognitive load theory, learning difficulty, and instructunal design, *Learn. Instr.*, **4**(4), 295–312.

Sweller J., (2010), Element interactivity and intrinsic, extraneous, and germane cognitive load, *Educ. Psychol. Rev.*, **22**, 123–138.

Sweller J., van Merriënboer J. J. and Paas F., (2019), Cognitive architecture and instructional design: 20 years later, *Educ. Psychol. Rev.*, **31**, 261–292.

Tavakol M. and Dennick R., (2011), Making sense of Cronbach's alpha, *Int. J. Med. Educ.*, **2**, 53.

Vygotsky L. S. and Cole M., (1978), *Mind in society: Development of higher psychological processes*, Cambridge, MA: Harvard University Press.

Watts F. M., Zaimi I., Kranz D., Graulich N. and Shultz G. V., (2021), Investigating students' reasoning over time for case comparisons of acyl transfer reaction mechanisms, *Chem. Educ. Res. Pract.*, **22**(2), 364–381.

Weinrich M. L. and Sevian H., (2017), Capturing students' abstraction while solving organic reaction mechanism problems across a semester, *Chem. Educ. Res. Pract.*, **18**(1), 169–190.

Weinrich M. L. and Talanquer V., (2016), Mapping students' modes of reasoning when thinking about chemical reactions used to make a desired product, *Chem. Educ. Res. Pract.*, **17**(2), 394–406.

Wickham H., (2007), Reshaping data with the reshape package, *J. Stat. Softw.*, **21**(12), 1–20.

Wickham H., (2016), *ggplot2: Elegant graphics for data analysis*, New-York: Springer-Verlag.

Wickham H., Averick M., Bryan J., Chang W., McGowan L. D. A., François R., Grolemund G., Hayes A., Henry L. and Hester J., (2019), Welcome to the tidyverse, *J. Open Source Softw.*, **4**(43), 1686.

Wickham H., François R., Henry L. and Müller K., (2022), *dplyr: A grammar of data manipulation (version: 1.0.10)*, **https://cran.r-project.org/src/contrib/Archive/dplyr/dplyr_1.0.10.tar.gz**.

Wickham H. and Girlich M., (2022), *tidyr: Tidy messy data (version: 1.2.1)*, **https://cran.r-project.org/src/contrib/Archive/tidyr/tidyr_1.2.1.tar.gz**.

Wilcox R. R., (2021), *Introduction to robust estimation and hypothesis testing*, Cambridge, MA: Academic press.

Wilcox R. R., (2023), *Rallfun-v41.txt (version: 2023.04.15)*, **https://osf.io/spvzc**.

Wilson K. and Devereux L., (2014), Scaffolding theory: high challenge, high support in academic language and learning (ALL) contexts, *J. Acad. Lang. Learn.*, **8**, 91–100.

Wood D., Bruner J. S. and Ross G., (1976), The role of tutoring in problem solving, *J. Child Psychol. Psychiatry*, **17**(2), 89–100.

Yuriev E., Naidu S., Schembri L. S. and Short J. L., (2017), Scaffolding the development of problem-solving skills in chemistry: guiding novice students out of dead ends and false starts, *Chem. Educ. Res. Pract.*, **18**(3), 486–504.