Special Issue Paper

Leonie Jasper and Insa Melle*

Scaffolding self-regulated problem solving: the influence of content-independent metacognitive prompts on students' general problem-solving skills

https://doi.org/10.1515/cti-2025-0012
Received January 31, 2025; accepted May 22, 2025; published online June 11, 2025

Abstract: The ability to solve problems is considered a key competence in today's society. However, solving domain-specific problems, such as those in chemistry, places high demands on students. Effective problem solving requires metacognitive strategies and their corresponding 'cold' executive functions, namely working memory and cognitive flexibility, which many students struggle with. To support students, we developed a web-based tool, *ChemApro*, designed to scaffold problem-solving processes by providing content-independent metacognitive prompts. The tool was used over several weeks in seven schools with $N = 153$ participating students ($M_{age} = 15.63$, $SD = 0.79$) in grades 10 and 11. Among other things, the study focuses on *ChemApro*'s effect on students' general problem-solving skills, and on how students perceive the tool in terms of its attractiveness and usability. In line with the study results, the use of *ChemApro* was descriptively associated with greater improvements in the treatment group's problem-solving skills compared to the corresponding baseline, particularly among those students with a lower cognitive level. However, the mixed ANOVA did not reveal significant interaction effects between group and time, although trends in the low cognitive level group approached significance. Additionally, students rated the tool's attractiveness and usability as moderate.

Keywords: self-regulated problem solving; metacognitive strategies; executive functions; scaffolding; ECRICE 2024

1 Introduction

Nowadays, problem solving is considered a key competency in both cross-domain and domain-specific contexts.[1,2] Therefore, problem solving is often referred to as one of the core skills of the 21st century.[3–7]

Solving (domain-specific) problems places high demands on students and can quickly lead them to feel overwhelmed.[8] This is partly due to the fact that effective self-regulated problem solving requires a high level of metacognitive strategies and their associated 'cold' executive functions.[8–16] In this context, various studies indicate that students often have an insufficient mastery of metacognitive strategies or are unaware of their use.[17,18] Deficits in metacognitive strategies are reflected in learners' inability to regulate their cognitive processes, including difficulties in applying planning, monitoring and evaluation strategies.[17] In addition, other studies show that many

Leonie Jasper and Insa Melle contributed equally to this work and share first authorship.

***Corresponding author: Insa Melle**, Department of Chemistry and Chemical Biology, TU Dortmund University, Otto-Hahn-Straße 6, 44227 Dortmund, Germany, E-mail: insa.melle@tu-dortmund.de. https://orcid.org/0000-0001-7112-456X
Leonie Jasper, Department of Chemistry and Chemical Biology, TU Dortmund University, Otto-Hahn-Straße 6, 44227 Dortmund, Germany. https://orcid.org/0009-0001-8362-6687

students struggle with their 'cold' executive functions, namely working memory and cognitive flexibility.[11,12,15,16] In the area of working memory, difficulties can be seen, for example, in students' inability to retain important information or memorize the steps needed to systematically solve complex tasks.[11,19] On the other hand, cognitive flexibility challenges arise when students persist with a previously successful strategy or solution rule, even when the task requirements have changed and the approach is no longer effective.[20] Consequently, such difficulties prevent these students from adequately self-regulating their problem-solving processes.

2 The tool *ChemApro*

It could be helpful to provide targeted support for students' self-regulated problem-solving processes, whereby students should proceed as independently and effectively as possible. One theory-based approach to provide such support is explicit scaffolding.[13,21,22] Within this approach, learners are provided with an external framework, that includes content-independent metacognitive prompts and thus addresses the process components during problem-solving activies.[13,21,22] To this end, we developed a web-based tool named *ChemApro* (short for: Chemistry Approach) as an external scaffold,[13] which, on the one hand, guides the students through the problem-solving process and, on the other hand, is adaptable to different problem types.

ChemApro is integrated into a website and comprises six phases of a general problem-solving strategy.[23] Additionally, a seventh phase was integrated to facilitate the consolidation and connection of new knowledge. Figure 1 shows an excerpt from the website and outlines the clickable steps that students go through with *ChemApro* in a self-regulated manner while solving a problem.

By clicking on each step (see Figure 1), learners can access flexibly designed metacognitive prompts, including guiding questions. These prompts are intended to stimulate and encourage the use of planning, monitoring and evaluating strategies. The implementation of the guiding questions for the 'Evaluate your work' step is illustrated below using another excerpt from the website (see Figure 2).

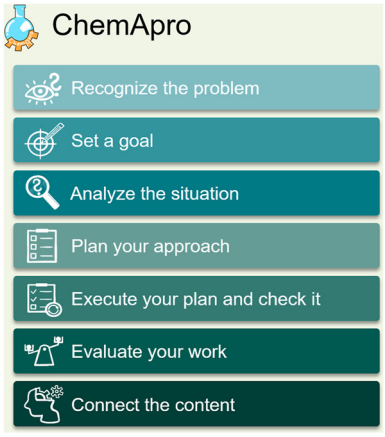

Figure 1: Seven phases of problem solving according to *ChemApro* (translated from German).

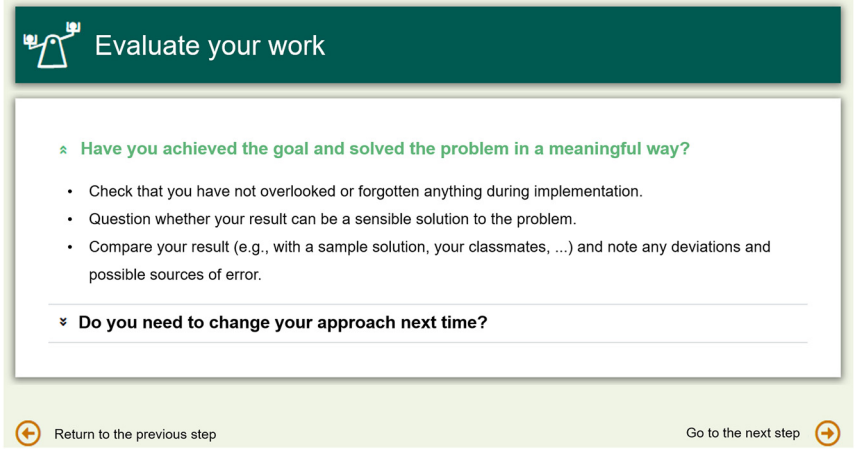

Figure 2: Guiding questions and action-oriented instructions for implementing the problem-solving step 'Evaluate your work' (translated from German).

However, given heterogeneity in regular chemistry classes, specific action-oriented instructions have also been provided for those students who require additional assistance in executing the single steps. These instructions are embedded via drop-down function so that students can retrieve them if necessary (see Figure 2).

3 Research questions and design

The aim of this study is to examine the effect of *ChemApro* on students' general problem-solving skills. Given the fact that problem-solving skills are related to intelligence,[24] it is also useful to conduct a separate analysis of the effects of *ChemApro* on problem-solving skills based on different levels of students' cognitive ability. This is important because these abilities may influence how effectively students benefit from the developed tool. In addition to the potential effects described above, this study will also evaluate the attractiveness and usability of *ChemApro*. In summary, the following research questions will be addressed in this paper:

Q1 To what extent does the use of *ChemApro* affect students' general problem-solving skills?
Q2 To what extent is the development of students' problem-solving skills affected by their cognitive level?
Q3 How do students evaluate *ChemApro* with regard to…?
   a. attractiveness
   b. usability

In order to answer these research questions, the present study is designed as an explanatory, quasi-experimental field study, that is conducted in a pre-post-design with two non-randomized, naturally occurring parallel school classes. To provide an insight into the research design and the course of the study, Figure 3 shows how the intervention was structured.

Within the scope of the intervention we integrated *ChemApro* into regular chemistry classes over a period of about ten weeks (see Figure 3). During this period, the treatment group (TG) was expected to use *ChemApro* approximately three times, while the baseline group (BG) received problem-oriented lessons on the same content but without using the tool. The integration of *ChemApro* into TG's chemistry classes was carried out autonomously by the teachers, who embedded the tool into their independently planned, problem-oriented teaching units. It should be noted that the tasks were not standardized across all classes, but varied according to the respective instructional designs of the teachers, although these were identical in the treatment group and in the baseline of each school. Before and after the intervention, we collected and analyzed various variables and, in some cases, assessed their development during the intervention period, e.g., students' use of metacognitive strategies or, as presented in this paper, students' general problem-solving skills. In addition, process-related data were collected from the treatment group. This included, first, the learners' documentation, which provided insights into learners' processing quality of each *ChemApro* step, and second, log file data, which was evaluated to examine the learners' usage behavior of *ChemApro* (see Figure 3).

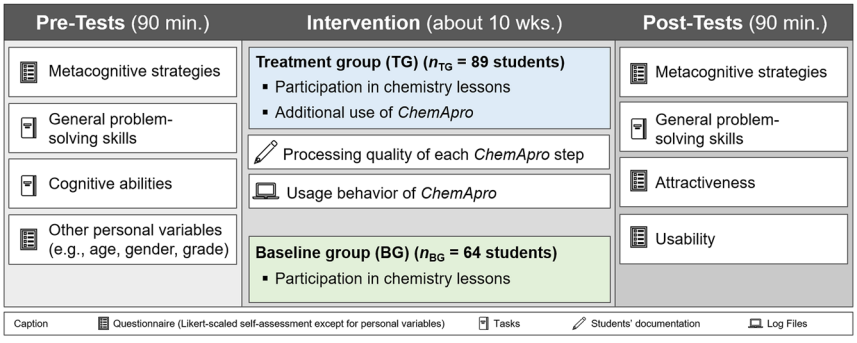

Figure 3: Quasi-experimental pre-post research design of the study.

The decision of choosing this kind of research design was based on findings from meta-studies such as that by Dignath and Büttner,[25] which demonstrated that interventions to promote self-regulation are more effective when implemented over an extended period. The advantages in this case are twofold: Firstly, it allows students to learn in their familiar environment. Secondly, it increases the external validity of the study. However, to implement such a long study duration, it was indispensable to work with naturally existing groups without any randomization.

4 Methods

The sample of the main study initially consisted of a total of $N = 211$ students ($M_{age} = 15.73$, $SD = 0.868$) from grades 10 and 11 of two different school types in Germany (*Gymnasium*/grammar school and *Gesamtschule*/comprehensive school) with a total of 7 participating teachers. In Germany, the *Gymnasium* and the higher grades of the *Gesamtschule* are comparable. All students have the *Abitur* (high school diploma) as their goal. Both the *Gymnasium* and the *Gesamtschule* students were in their first year of high school. In *Gymnasien*, this was grade 10 and in *Gesamtschulen*, it was grade 11. The same curricula apply to both schools. In each school, there was a treatment group and a baseline group of the same year, both taught by the same teacher. The teachers decided at random which class should work with *ChemApro* and which should not.

However, the analysis of the log file data revealed that two of the participating teachers did not make *ChemApro* fully available to their students as instructed. This refers in particular to the fact that students had no opportunity to use the tool in its web-based format and its whole functionality. When these two teachers were asked why they deviated from the prescribed use in their lessons, they explained that school-related technical difficulties prevented them from utilizing the tool to its full extent. Therefore, the classes that did not have full access to *ChemApro* and the corresponding parallel classes were excluded from the analysis. Consequently, the results for research questions Q1–Q3 are presented only for the students who had the opportunity to use *ChemApro* fully (TG) and the corresponding parallel classes (BG). This reduced sample thus amounts to $N = 153$ students ($M_{age} = 15.63$, $SD = 0.79$), of which are $n_{Gymnasium} = 115$ students from *Gymnasium* and $n_{Gesamtschule} = 38$ students from the *Gesamtschule*, with a total of 5 participating teachers. Both the TG and the BG include students from both types of school and are made up as follows: $n_{TG} = 89$ students ($n_{Gymnasium} = 60$, $n_{Gesamtschule} = 29$; $M_{age} = 15.73$, $SD = 0.78$) and $n_{BG} = 64$ students ($n_{Gymnasium} = 55$, $n_{Gesamtschule} = 9$; $M_{age} = 15.48$, $SD = 0.80$).

To ensure the comparability of the two groups, *t*-tests were used to test differences in the mean values of various person-related prerequisites. The non-parametric Mann-Whitney-*U*-test was used if the normal distribution assumption was violated or if there was a lack of homogeneity of variance. For categorical characteristics such as gender, the chi-square test was used. We found that the TG and the BG did not differ significantly in central personal prerequisites such as cognitive ability ($t(151) = 0.651$, $p = 0.516$, $d = 0.107$), gender ($\chi^2(3) = 0.647$, $p = 0.886$, $\varphi = 0.052$) or in problem-solving skills ($t(151) = -0.027$, $p = 0.979$, $d = 0.004$), and metacognitive strategy use ($t(138) = -0.670$, $p = 0.504$, $d = 0.115$) at the 'pre' measurement point. However, they differ significantly in terms of age with a small effect size ($t(151) = 1.907$, $p = 0.058$, $d = 0.313$; $U = 2{,}257.00$, $Z = -2.413$, $p = 0.016$, $\varphi = 0.195$), although on a descriptive level they differ on average by only 0.25 (3 months), which is not considered a meaningful difference. This is probably due to the fact that the TG contains a larger proportion of students from the *Gesamtschule* compared to the BG. This is important because the students from the *Gesamtschule* are one grade above and also older than those students from the *Gymnasium*.

The data used for answering Q1 and Q2, and thus to investigate the potential impact of *ChemApro* on students' general problem-solving skills from pre to post, were collected at both measurement points using an established testing instrument. This test is based on selected and subsequently translated items that were originally published by the OECD[26] and were used as a part of the 2003 PISA survey to measure the problem-solving skills of 15-year-old students. According to the OECD,[26] we define general problem-solving competence as the ability to solve problems in real-life situations that go beyond the specific contexts of individual school subjects. The test was carried out in a paper-pencil format with a time limit of 45 min and contains a total of 14 items, including 6 closed and 8 open-ended items.[26] An example of a closed item is the 'Cinema Outing' item, which is presented in single choice

format.[26] This item involves the planning of a visit to the cinema by three young people who want to see a predefined movie during their vacation, taking into account individual time constraints, age-related requirements and the cinema schedule.[26] Accordingly, the student must use all the given information to choose the day on which the three can go to the cinema.[26] An example of an open-ended item is the 'Transit System' item.[26] This item presents a subway map with three lines, indicating a starting point and a destination. Students have to determine the most efficient route in terms of cost and travel time, taking into account the duration of transfers and the fare structure based on the number of stations traveled through.[26] For the evaluation, the closed items were rated on a scale from 0 to 1 (0 = *no credit* and 1 = *full credit*), while the open-ended items were rated from 0 to 2 (0 = *no credit*, 1 = *partial credit*, and 2 = *full credit*) or from 0 to 3 (0 = *no credit*, 1 = *partial credit* (I), 2 = *partial credit* (II), and 3 = *full credit*) using a coding scheme which was also developed by the OECD.[26] For each rated item, the scores awarded were subsequently assigned, which is also derived by the OECD[26] on the basis of their degree of difficulty and were adopted without modification for this study. In this test, students could achieve a maximum score of 7,979 points. The results are reported as the percentage points (%) of the total score achieved.[26] Accordingly, the presented results range from 0 % (= 0.00) to 100 % (= 1.00) of the maximum score. Overall, the test showed an acceptable internal consistency (Cronbach's α = 0.710). Furthermore, the reliability of the coding scheme was ensured by double coding, which yielded satisfactory results for the open-ended (ICC_{unjust} = 0.990) and closed questions (Cohen's κ = 1.000).

A shortened version of the validated CFT 20-R by Weiß and Weiß[27] was used to take into account the students' different cognitive abilities as one of the core personal prerequisites (Q2). The CFT 20-R consists of two test parts, each divided into four subtests.[27] For the purposes of the study, we conducted only the first part of the test with its four subtests in a paper-pencil format with a given test time of about 35 min, where students were given a test booklet and an answer sheet on which they could write down their answers.[27] However, this test omits verbal descriptions of the items and uses only pictorial representations, which can be answered in a single-choice format with five possible answers.[27] From the point of view of heterogeneity, this test offers the advantage that no language tasks are used, thus ensuring that language barriers are minimized.[27] The data was then analyzed and interpreted using the age-specific norms defined in the test manual by Weiß and Weiß,[27] whereby the number of correctly solved tasks was transformed into a corresponding value as an indicator of cognitive abilities. Based on these scores, students were divided into quartiles and thus into four cognitive levels (CL): low, lower-middle, upper-middle, and high CL.

Two questionnaires were administered at post time (Q3) to investigate how students rate *ChemApro* in terms of both its attractiveness[28] (adapted) and its usability[29] (translated into German). Each questionnaire contained a total of ten closed items asking the students to rate the attractiveness of the tool on a 6-point Likert scale from 1 (*not attractive*) to 6 (*attractive*) and its usability on a 5-point Likert scale from 1 (*low usability*) to 5 (*high usability*). Like the other tests used in this study, these two tests were also carried out in a paper-pencil format. The attractiveness questionnaire demonstrated good internal consistency (Cronbach's α = 0.855), while the internal consistency of the usability test turned out to be moderate, but also sufficient (Cronbach's α = 0.760). To answer research questions Q3, a reduced sample size must be assumed, as not all students completed the tests.

5 Results

Initially, descriptive data are considered in order to determine the development in students' problem-solving skills from pre to post (Q1). For this purpose, the mean scores achieved per group and their mean differences were calculated. These results are presented in Table 1. With regard to the mean scores at time pre, the data show that all students have already started with moderate general problem-solving skills. From pre to post the students in the TG improve their skills by 9.8 % (= 0.098), whereas the according BG improve only by 6.6 % (= 0.066) (see Table 1).

Afterwards, additional statistical calculations were conducted using paired-samples *t*-tests for each group to analyze whether there is a statistically significant improvement in problem-solving skills per group (see Table 1). The results indicate that the TG consistently demonstrated a significant improvement in problem-solving skills

Table 1: Development of the students' general problem-solving skills (pre-post) using *t*-test for dependent samples. As the data on general problem-solving skills are not always normally distributed at each measurement point, non-parametric Wilcoxon-tests were used to confirm the results of the *t*-tests if necessary.

Sample	*n*	Time	*M*	Mean difference pre-post	*t*-test effect size (*d*)	Wilcoxon test effect size (φ)
TG	89	Pre	0.602	0.098**	0.75	0.63
		Post	0.700			
BG	64	Pre	0.603	0.066**	0.38	–
		Post	0.669			

Significance level: *$p < 0.05$; **$p < 0.01$, ***$p < 0.001$.

from pre to post, with a moderate effect size. The BG also showed a significant increase in their problem-solving skills over time, however, the effect size is noticeably smaller compared to the TG.

The statistical examination of whether the two groups, TG and BG, differ in terms of their growth in general problem-solving skills was carried out using a two-factor analysis of variance with repeated measures (mixed ANOVA) (see Table 2). The results show that there is neither a significant effect of the between-subject factor group nor a significant interaction effect group * time. Thus, although the problem-solving skills of each group increase significantly over time (within-subject factor time), the TG differ not significantly from the BG.

Regarding research question Q2, the descriptive data for each group were analyzed according to the four cognitive level (CL) quartiles (see Table 3). Additionally, *t*-tests were conducted separated by these four levels in order to examine the development in problem-solving skills for each CL in the TG and BG from pre to post. These results are also shown in Table 3.

As it can be seen in Table 3, the data reveal that across all cognitive levels, the TG consistently show significant improvements from pre to post with moderate to large effect sizes. The BG can also improve their problem-solving skills significantly in three out of four CLs (lower-middle, upper-middle and high) over time. But it is noteworthy that students from the TG with a low cognitive level improve significantly by 14.7 % (= 0.147) from pre to post on the problem-solving skills test, with a large effect size, whereas the corresponding BG improve by only 4.0 % (= 0.04), which was not statistically significant.

In addition, a mixed ANOVA was conducted to examine whether there was a difference between the growth in problem-solving skills of TG und BG, taking into account their cognitive level (see Table 4).

As the data in Table 4 indicate, no significant interaction for the between-subject factor group * time can be found for any of the four CLs. Although the descriptive data and the results of the *t*-test may suggest differences between TG und BG in the group with CL low over time (see Table 3), this is not confirmed by the mixed ANOVA (see Table 4). However, for the group with CL low the results approach significance and show a moderate effect size (see Table 4).

In terms of addressing research questions Q3a and Q3b, the descriptive statistics results, presented in Table 5, indicate that the TG rated the attractiveness and the usability of *ChemApro* at a moderate level.

Table 2: Analysis of the increase in general problem-solving skills (pre-post) in relation to group differences between TG and BG using a mixed ANOVA. The dependent variable cannot be assumed to be normally distributed in all subsamples. Since the mixed ANOVA is considered to be relatively robust against the violation of the normal distribution, it is used anyway.[30] The presence of variance homogeneity has been demonstrated by a non-significant Levene test. It is not necessary to check the sphericity, as only two measurement points are included in the calculation.

Factor	Main effect		Effect size (η^2_{part})
	F ratio	*p*	
Group	0.22	0.637	0.001
Time	43.79***	<0.001	0.225
Group * time	1.65	0.201	0.011

Significance level: *$p < 0.05$; **$p < 0.01$, ***$p < 0.001$.

Table 3: Development of the students' problem-solving skills (pre-post) according to the four CL using *t*-test for dependent samples. As the data on problem-solving skills are not always normally distributed at each measurement point, non-parametric Wilcoxon-tests were used to confirm the results of the *t*-tests if necessary.

CL	Sample	*n*	Time	*M*	Mean difference pre-post	*t*-test effect size (*d*)	Wilcoxon test effect size (φ)
Low	TG	17	Pre	0.421	0.147***	1.22	–
			Post	0.568			
	BG	18	Pre	0.433	0.040	0.16	–
			Post	0.473			
Lower-middle	TG	21	Pre	0.561	0.067*	0.50	–
			Post	0.628			
	BG	11	Pre	0.609	0.092[a]	0.62	–
			Post	0.701			
Upper-middle	TG	22	Pre	0.641	0.106**	0.70	–
			Post	0.747			
	BG	16	Pre	0.699	0.066	0.45	–
			Post	0.765			
High	TG	29	Pre	0.707	0.085**	0.75	0.62
			Post	0.793			
	BG	19	Pre	0.678	0.076*	0.54	0.51
			Post	0.754			

Significance level: *$p < 0.05$; **$p < 0.01$, ***$p < 0.001$. CL = cognitive level quartiles. [a]It may seem surprising at first that the difference is not significant, even though the effect size is quite large. However, $p = 0.067$ (with a small *n*) is close to significance.

Table 4: Analysis of the increase in problem-solving skills (pre-post) in relation to group differences between TG and BG, taking into account their cognitive level using a mixed ANOVA. The dependent variable cannot be assumed to be normally distributed in all subsamples. Since the mixed ANOVA is considered to be relatively robust against the violation of the normal distribution, it is used anyway.[30] The presence of variance homogeneity has been demonstrated by a non-significant Levene test. It is not necessary to check the sphericity, as only two measurement points are included in the calculation.

CL	Factor	Main effect		Effect size (η^2_{part})
		F ratio	*p*	
Low	Group	0.812	0.374	0.024
	Time	8.040**	0.008	0.196
	Group * time	2.627	0.115	0.074
Lower-middle	Group	1.034	0.317	0.033
	Time	9.455**	0.004	0.240
	Group * time	0.224	0.639	0.007
Upper-middle	Group	0.578	0.452	0.016
	Time	12.237**	0.001	0.254
	Group * time	0.663	0.421	0.018
High	Group	0.360	0.552	0.008
	Time	18.996***	< 0.001	0.292
	Group * time	0.063	0.803	0.001

Significance level: *$p < 0.05$; **$p < 0.01$, ***$p < 0.001$. CL = cognitive level quartiles.

Table 5: Descriptive statistics for the attractiveness (1 = *not attractive* to 6 = *attractive*) and usability test (1 = *low usability* to 5 = *high usability*) for the TG at the post measurement point.

Variable	*n*	*M*	*SD*
Attractiveness	87	3.08	0.87
Usability	88	2.96	0.65

6 Discussion

Initially, data show a significant improvement in problem-solving skills across both groups over time, as indicated by the significant within-subject factor time in the mixed ANOVA. However, even if problem-solving skills increase over time, for example, due to the influence of lessons, the TG consistently demonstrates larger effect sizes compared to the BG regarding the *t*-test results. The mixed ANOVA revealed only a small, non-significant interaction between group and time. At this stage, it must therefore be concluded that in this study the use of content-independent metacognitive prompts did not lead to significant improvements in students' general problem-solving skills. In order to strengthen the potential impact of *ChemApro*, a more frequent and continuous use in the classroom seems sensible. In addition, a replication of the study with a larger sample could help to increase the statistical power and prove possible effects more reliably.[30]

With regard to the increase in general problem-solving skills, taking into account the cognitive level, the *t*-test results show that students with a low cognitive level (CL) in the TG showed substantial significant improvements (14.7 %) from pre to post with a large effect size, while the corresponding BG's results for the same subgroup were smaller and not statistically significant (4.0 %). Despite this finding, no significant interaction effects (group * time) for the individual CLs can be determined in the mixed ANOVA. Consequently, the emerging difference at the low cognitive level cannot be statistically proven. The reason may be the small subsample sizes, particularly within certain cognitive levels, which likely reduced the statistical power of the mixed ANOVA and limited its ability to detect significant effects.[30] Future studies with larger sample sizes and more balanced group distributions are needed to validate these findings.

Furthermore, the evaluation of *ChemApro*'s attractiveness and usability indicates moderate levels. This result aligns with findings from the literature that highlight the challenges associated with implementing external scaffolds. Studies such as Vo, Sarkar, White and Yuriev[21] indicate that learners often do not perceive external scaffolds as an integral part of the problem-solving process. Instead, they tend to view their use as an additional task, which reduces their willingness to engage with the scaffold.[21] The consequence of this can be that students perceive the tool as less appealing and therefore lead them to rate it lower in terms of attractiveness and usability.

For future research, as already mentioned, an increasing sample size is a critical step to examine the effect of the tool. The relatively small sample sizes, especially in certain subgroups of CL, limit the generalizability of the findings, so that a larger sample size would enable more robust conclusions to be drawn through greater statistical power. Another limitation of this study is its design, as it does not fully control for the influence of external factors, such as teacher behavior, which might affect how students interact with the tool. An alternative study design, such as an experimental study, in which the content is predetermined and implemented by a researcher, could be considered to control for potential confounding variables. In this way, factors such as the influence of the teacher or the lesson content itself can be controlled in order to ensure that the observed effects can be more reliably attributed to the scaffold itself. At the same time, implementing a design that accounts for these factors would strengthen the internal validity of the study and provide more accurate conclusions regarding the effectiveness of the tool. On the other hand, if researchers implement the intervention, there is a risk that the treatment will be carried out with a notably higher quality in both the TG and the BG, which may limit the comparability of the results. Additionally, it would also be important to investigate the specific features of *ChemApro* that contribute most to skill development, as well as to assess its efficacy across diverse educational contexts. Finally, the combination of qualitative methods, such as student interviews or focus groups, in which students of varying cognitive abilities are observed in detail, and quantitative methods, as chosen in this work, could lead to a deeper understanding of how students engage with and benefit from the tool.

In order to consider additional interaction between the improvement of problem-solving skills, the use of metacognitive prompts and self-regulation, in a next step we will include the students' use of metacognitive strategies as a second factor in the calculation. Metacognitive strategies were collected at the pre and post measurement points using a Likert-scaled self-assessment questionnaire. To gain a deeper understanding of how

students interact with *ChemApro*, we will analyze process-related data in a next step. This will include an analysis of students' documentation, which will be graded on the quality of their completion of individual steps in their self-directed problem-solving process. In addition, we will conduct a detailed examination of the log file data to see how the students used the tool.

7 Conclusions

In this study, we investigated the impact of a web-based metacognitive scaffold, *ChemApro*, by examining its effect on the improvement of students' general problem-solving skills in chemistry classes. In this context, the results of the descriptive statistics and the *t*-tests indicate a potential benefit of *ChemApro*, even though the mixed ANOVA revealed no significant interaction effects - and thus no statistically significant difference between the group with scaffolding and the baseline group. Furthermore, we investigated the role of cognitive level in mediating the effectiveness of the tool. Taking into account descriptive data as well as the results of the *t*-tests, we found that students with a lower cognitive level tended to benefit the most from working with *ChemApro*. Despite this finding, this second mixed ANOVA also failed to detect any significant interaction effects in terms of group comparison by individual cognitive levels. However, even though for the students with CL 'low' the results are close to significance and show a moderate effect size, it must be concluded that there is no statistical difference between treatment group and baseline group in consideration of cognitive level at measurement point 'post'. Additionally, this study gains insights into how students perceive external scaffolds, specifically in terms of their attractiveness and usability.

Acknowledgments: First of all, we would like to thank the teachers and students who participated in our study. We would also like to thank Rejan Mohamed for illustrating the icons implemented in *ChemApro* and Christoph Pietzarka for his support in designing the website. We would also like to thank the staff of the Statistical Consulting and Analysis Centre (SBAZ) at the TU Dortmund University, who provided methodology and data analysis support for this project.
Research ethics: Not applicable.
Informed consent: The parents of the students provided written informed consent for their children to participate in this study. They were given important information about the project, data protection, data security and further processing of the data. Participation in the study was voluntary for the students. There were no disadvantages for non-participation. Teachers and heads of schools also provided their written informed consent. They were also given important information about the project, data protection, data security, and further processing of the data. Participation in the study was voluntary for the teachers and heads of schools.
Author contributions: LJ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, IM: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft. All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: The authors acknowledge the use of OpenAI's generative AI technologies ChatGPT 4o and DeepL and DeepL Write for some translations.
Conflict of interest: The authors state no conflict of interest.
Research funding: None declared.
Data availability: The datasets presented in this article are not readily available because in the context of the written consent form, we assured the students that their data will only be stored on devices belonging to our university. Hence, we are not permitted to disclose the data publicly. Requests to access the datasets should be directed to IM.

References

1. Csapó, B.; Funke, J., Eds. *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*; OECD Publishing: Paris, 2017.
2. Funke, J. Problemlösen [Problem Solving]. In *Denken - Urteilen, Entscheiden, Problemlösen*; Betsch, T.; Funke, J.; Plessner, H., Eds.; Allgemeine Psychologie für Bachelor. Springer: Berlin, 2011; pp 137–202.
3. Ahonen, A. K.; Kinnunen, P. How Do Students Value the Importance of Twenty-First Century Skills? *Scand. J. Educ. Res.* **2015**, *59* (4), 395–412.
4. Bialik, M.; Fadel, C.; Nilsson, P.; Trilling, B.; Groff, J. *Skills for the 21st Century: What Should Students Learn?*; Center for Curriculum Redesign: Boston, MA, 2015. https://curriculumredesign.org/wp-content/uploads/CCR-Skills_FINAL_June2015.pdf (accessed 2015-31-01).
5. Foster, N.; Piacentini, M., Eds. *Innovating Assessments to Measure and Support Complex Skills*; OECD Publishing: Paris, 2023.
6. González-Pérez, L. I.; Ramírez-Montoya, M. S. Components of Education 4.0 in 21st Century Skills Frameworks: Systematic Review. *Sustainability (Basel)* **2022**, *14* (3), 1493.
7. Kennedy, T. J.; Sundberg, C. W. 21st Century Skills. In *Science Education in Theory and Practice: An Introductory Guide to Learning Theory*. In *Springer Texts in Education*; Akpan, B.; Kennedy, T. J., Eds.; Springer International Publishing: Switzerland, 2020; pp 479–496.
8. Zumbach, J.; Ortler, C.; Deibl, I.; Moser, S. Using Prompts to Scaffold Metacognition in Case-Based Problem Solving within the Domain of Attribution Theory. *J. Probl. Based Learn.* **2020**, *7* (1), 21–31.
9. Baars, M.; Leopold, C.; Paas, F. Self-explaining Steps in Problem-Solving Tasks to Improve Self-Regulation in Secondary Education. *J. Educ. Psychol.* **2018**, *110* (4), 578–595.
10. Cancer, A.; Iannello, P.; Salvi, C.; Antonietti, A. Executive Functioning and Divergent Thinking Predict Creative Problem-Solving in Young Adults and Elderlies. *Psychol. Res.* **2023**, *87* (2), 388–396.
11. Diamond, A. Executive Functions. *Annu. Rev. Psychol.* **2013**, *64*, 135–168.
12. Diamond, A.; Ling, D. S. Conclusions about Interventions, Programs, and Approaches for Improving Executive Functions that Appear Justified and Those that, Despite Much Hype, Do Not. *Dev. Cogn. Neurosci.* **2016**, *18*, 34–48.
13. Graulich, N.; Langner, A.; Vo, K.; Yuriev, E. Scaffolding Metacognition and Resource Activation during Problem Solving: A Continuum Perspective. In *Problems and Problem Solving in Chemistry Education: Analysing Data, Looking for Patterns and Making Deductions*; Tsaparlis, G., Ed.; Royal Society of Chemistry: Cambridge, 2021; pp 38–67.
14. Marulis, L. M.; Nelson, L. J. Metacognitive Processes and Associations to Executive Function and Motivation during a Problem-Solving Task in 3–5 Year Olds. *Metacogn. Learn.* **2021**, *16* (1), 207–231.
15. Ropovik, I. Do Executive Functions Predict the Ability to Learn Problem-Solving Principles? *Intelligence (Norwood)* **2014**, *44*, 64–74.
16. Schäfer, J.; Reuter, T.; Leuchter, M.; Karbach, J. Executive Functions and Problem-Solving-The Contribution of Inhibition, Working Memory, and Cognitive Flexibility to Science Problem-Solving Performance in Elementary School Students. *J. Exp. Child Psychol.* **2024**, *244*, 105962.
17. Ohtani, K.; Hisasaka, T. Beyond Intelligence: A Meta-analytic Review of the Relationship among Metacognition, Intelligence, and Academic Performance. *Metacogn. Learn.* **2018**, *13* (2), 179–212.
18. Winne, P. H.; Azevedo, R. Metacognition. *The Cambridge Handbook of the Learning Sciences. In Cambridge Handbooks in Psychology*; Sawyer, R. K., Ed., 2nd ed.; Cambridge University Press: Cambridge, 2014; pp 63–87.
19. Alloway, T. P.; Alloway, R. G. Investigating the Predictive Roles of Working Memory and IQ in Academic Attainment. *J. Exp. Child. Psychol.* **2010**, *106* (1), 20–29.
20. Strobach, T. *Kognitive Psychologie [Cognitive Psychology]*; Kohlhammer: Stuttgart, 2020.
21. Vo, K.; Sarkar, M.; White, P. J.; Yuriev, E. Problem Solving in Chemistry Supported by Metacognitive Scaffolding: Teaching Associates' Perspectives and Practices. *Chem. Educ. Res. Pract.* **2022**, *23* (2), 436–451.
22. Yuriev, E.; Naidu, S.; Schembri, L. S.; Short, J. L. Scaffolding the Development of Problem-Solving Skills in Chemistry: Guiding Novice Students Out of Dead Ends and False Starts. *Chem. Educ. Res. Pract.* **2017**, *18* (3), 486–504.
23. Funke, J.; Fischer, A.; Holt, D. V. Competencies for Complexity: Problem Solving in the Twenty-First Century. In *Assessment and Teaching of 21st century skills: Research and Applications*, 2018. Care, E.; Griffin, P.; Wilson, M., Eds.; Educational Assessment in an Information Age; Springer International Publishing: Switzerland, 2018; pp 41–53.
24. Wenke, D.; Frensch, P. A.; Funke, J. Complex Problem Solving and Intelligence: Empirical Relation and Causal Direction. In *Cognition and Intelligence: Identifying the Mechanisms of the Mind*; Sternberg, R. J.; Pretz, J. E., Eds.; Cambridge University Press: Cambridge, 2005; pp 160–187.
25. Dignath, C.; Büttner, G. Components of Fostering Self-Regulated Learning among Students. A Meta-Analysis on Intervention Studies at Primary and Secondary School Level. *Metacogn. Learn.* **2008**, *3* (3), 231–264.
26. OECD *Problem Solving for Tomorrow's World: First Measures of Cross-curricular Competencies from PISA 2003*; OECD Publishing: Paris, 2004.
27. Weiß, R. H.; Weiß, B. *CFT 20-R: Grundintelligenztest Skala [CFT 20-R: Basic Intelligence Test Scale]* (2nd Revision); Hogrefe Verlag GmbH & Co. KG: Gottingen, 2006.
28. Tepner, M.; Roeder, B.; Melle, I. Effektivität des Gruppenpuzzles im Chemieunterricht der Sekundarstufe I [Effectiveness of jigsaw-Classroom in Lower Secondary Classes of Chemistry]. *Z. Didakt. Nat. Wiss.* **2009**, *15*, 7–29.
29. Brooke, J. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation in Industry*; Jordan, P. W.; Thomas, B.; Weerdmeester, B. A.; McElland, I. I., Eds.; Taylor and Francis: London, 1996; pp 207–212.
30. Bühner, M.; Ziegler, M. *Statistik für Psychologen und Sozialwissenschaftler [Statistics for Psychologists and Social Scienticts]*, 2nd ed.; Pearson Deutschland GmbH: Munich, 2017. https://elibrary.pearson.de/book/99.150005/9783863268091.