Received: 12 March 2023 | Revised: 2 April 2024 | Accepted: 7 April 2024
DOI: 10.1111/jcal.12991

ARTICLE

Journal of Computer Assisted Learning WILEY

Video-based modeling examples and comparative self-explanation prompts for teaching a complex problem-solving strategy

Julius Moritz Meier[1,2] | Peter Hesse[3] | Stephan Abele[3] | Alexander Renkl[2] | Inga Glogger-Frey[1]

[1]Department of Psychology, University of Erfurt, Erfurt, Germany
[2]Department of Psychology, University of Freiburg, Freiburg, Germany
[3]Institute of Vocational Education and Vocational Didactics, Technische Universität Dresden, Dresden, Germany

Correspondence
Julius Moritz Meier, Department of Psychology, University of Erfurt, Nordhäuser Str. 63, 99089 Erfurt, Germany.
Email: julius.meier@uni-erfurt.de

Funding information
Federal Ministry of Education and Research, Germany; Federal Institute for Vocational Education and Training, Germany; research and transfer initiative 'ASCOT+ –Technology-based Assessment of Skills and Competences in Vocational Education and Training', Grant/Award Number: 21AP004C

Abstract

Background: In example-based learning, examples are often combined with generative activities, such as comparative self-explanations of example cases. Comparisons induce heavy demands on working memory, especially in complex domains. Hence, only stronger learners may benefit from comparative self-explanations. While static text-based examples can be compared easily, this is challenging for transient video-based modelling examples used in complex domains because simultaneous processing of two videos is not feasible.

Objectives: To allow for such comparisons, we combined video-based modelling examples with static representations (i.e., summarizing tables) of the observed optimal and a suboptimal solution of the problem-solving process. A comparative self-explanation prompt asked learners to compare the different solution approaches. Our study investigated the impact of video-based modelling examples versus independent problem-solving on cognitive load and problem-solving skill development. Moreover, we investigated the effects of comparative versus sequential self-explanation prompts, depending on learners' prior knowledge.

Methods: In an experiment, 118 automotive apprentices learned a car malfunction diagnosis strategy. Apprentices were divided into three groups: (1) modelling examples with comparative self-explanation prompts, (2) modelling examples with sequential prompts, and (3) no examples or prompts. Diagnostic knowledge and skills were assessed before and after the intervention. Cognitive load was measured retrospectively.

Results and conclusions: Despite no observed effects on cognitive load, modelling examples enhanced diagnostic knowledge and diagnostic skills with scaffolds, though not independent diagnostic skills without scaffolds. The need for more practice opportunities to foster independent diagnostic skills is assumed. Additionally, comparative prompts seem promising for learners with higher prior knowledge.

Takeaways: Video-based modelling examples were more beneficial for learning than practising to apply the diagnostic strategy. Static representations allow for comparisons of video examples and comparative prompts are promising for learners with higher prior knowledge (cf. expertise-reversal effect). Further research, especially on the effects on cognitive load, is necessary.

KEYWORDS
comparative self-explanation prompts, complex problem-solving, contrasting cases, diagnostic strategy, example-based learning, modelling examples

1 | INTRODUCTION

Example-based learning is a potent strategy for teaching problem-solving skills, particularly in ill-structured domains. Video-based modelling examples, often utilized in such cases (e.g., Meier et al., 2022; for a review see van Gog et al., 2019) reduce learning-irrelevant cognitive load compared to practicing instructed strategies. Consequently, more cognitive capacities are available for learning (Renkl, 2014; Sweller, 2006; van Gog et al., 2019). To benefit from the freed-up capacities, learners need to engage in learning-related activities, such as self-explanations, which can be prompted (Renkl & Eitel, 2019; Wylie & Chi, 2014). However, in situations where learning demands a substantial working memory capacity, additional instructional measures, such as self-explanation prompts, might risk cognitive overload and hinder learning (Große & Renkl, 2006; Sweller, 2006). Hence, this study firstly aimed to investigate the effects of video-based modelling examples and self-explanation prompts on cognitive load and learning outcomes for complex problems in an ill-structured domain. One form of self-explanation prompts asks learners to compare examples (Alfieri et al., 2013; Rittle-Johnson & Star, 2011). Comparing examples can be a powerful instructional strategy to have learners discover similarities and differences between example solutions and help them become sensitive to information that they might miss otherwise (Gentner et al., 2003; Richland et al., 2007; Schwartz & Bransford, 1998). Comparisons might be especially effective in complex and ill-structured domains, where good solutions can take many forms and are hard to differentiate from suboptimal solutions (Gentner et al., 2003; Renkl et al., 2009). For example, learning how to assess the quality of learning strategies can be supported by comparing good and bad examples of learning journal entries, that represent a learners' learning strategy (also referred to as contrasting cases; Glogger-Frey et al., 2015, 2022). However, comparison processes induce heavy demands on working memory (Holyoak, 2012). Hence, only stronger learners who can manage the increased demands can be expected to benefit from self-explanations including comparisons (i.e., *comparative self-explanation prompts*); weaker learners might be better off with studying and self-explaining examples sequentially (i.e., *sequential self-explanation prompts*; Rittle-Johnson et al., 2009). Consequently, the second aim of the present study was to compare the effects of *comparative* and *sequential self-explanation prompts* on cognitive load and learning outcomes for learners with different levels of prior knowledge when learning with video-based modelling examples in an ill-structured domain.

In addition, for effective comparison of examples, learners must be able to examine them simultaneously. This is possible with static and non-transient text-based worked examples, that can be studied side by side (e.g., Rittle-Johnson & Star, 2007). However, in the case of dynamic and transient video-based modelling examples, comparisons become more challenging. Since learners cannot simultaneously watch two video examples, a direct comparison of distinctive features becomes impractical. To address this limitation, a static and non-transient representation of the problem-solving process demonstrated in the video, such as a table-based summary highlighting key processes, could be provided. Instead of comparing two transient video examples learners could then compare two static representations. Consequently, in this study, we developed and tested *representation-based self-explanation prompts* to enable comparisons of problem-solving processes as demonstrated in video-based modelling examples.

1.1 | Example-based learning

When learning how to solve a problem, learning from examples is very effective and superior to practising how to solve problems (e.g., Renkl, 2014; Sweller, 2006; van Gog et al., 2019). This is true not only for well-structured domains, such as mathematics, where mostly text-based worked examples are used (e.g., Schalk et al., 2020) but also for ill-structured domains, where often video-based modelling examples are used (van Gog et al., 2019; van Gog & Rummel, 2010).

Usually, the effectiveness of examples is explained by *cognitive load theory* (CLT; Paas & van Gog, 2006; Sweller, 2006): CLT distinguishes three types of working memory load[1] (Sweller et al., 1998): First, *extraneous* cognitive load (ECL) is learning-irrelevant load, often induced by the (sub-optimal) design of learning materials. Second, *germane* cognitive load (GCL) results from learning-related activities. Third, *intrinsic* cognitive load (ICL) is determined by the learning material's complexity. More complex materials (i.e., materials with higher element interactivity) induce a higher ICL. However, the more prior knowledge learners have about a topic, the lower the ICL they experience; they are able to chunk different elements to one information unit (Sweller et al., 2011). Given the same task (i.e., same element interactivity) and the same learners (i.e., same prior knowledge), ICL is fixed. Therefore, to ensure that sufficient working memory resources are available for GCL, ECL must be minimized. Accordingly, most CLT guidelines for instructional design aim to reduce ECL (e.g., Mayer & Moreno, 2003).

When learning how to solve a problem, one possibility to reduce ECL (and potentially increase GCL) is to provide worked or modelling examples (Paas & van Gog, 2006; Sweller, 2006). When novices try to solve a problem, they often apply weak problem-solving strategies (Renkl, 2014). Applying such weak strategies is hardly conducive to learning and understanding a new strategy, so it can also be considered a learning-irrelevant activity inducing ECL (van Gog et al., 2019). Studying examples eliminates the need for weak strategies as learners can focus on the provided problem-solving steps, thereby reducing ECL. These freed cognitive capacities can then be used for learning (i.e., for GCL), which explains the frequently found beneficial effect on learning outcomes (Renkl, 2014; Sweller, 2006). However, to benefit from the reduced ECL, learners must actively use these freed-up capacities for learning (GCL), that is, they must engage in generative activities.

So far, research on video-based modelling examples has mostly focused on teaching quite simple problem-solving processes with rather short video-based examples. For example, Schmitz et al. (2017) successfully used short (i.e., 30 s–51 s) video-based examples to teach healthcare students a strategy for delivering bad news. Only few studies have investigated the effects of longer video-based modelling examples for teaching more complex problem-solving skills with some studies finding beneficial effects of modelling examples (e.g., Frerejean et al., 2018) and other studies finding no beneficial effects (e.g., Meier et al., 2023). Thus, further research in this regard is necessary to further investigate the boundary conditions of example-based learning (Sweller, 2023).

[1]Recently, Sweller et al. (2019) presented updates to the theory mainly with innovations to the concept of germane cognitive load: Most importantly, these updates suggest that intrinsic and germane load can be classified as one type of load, resulting in only two types of load that can be distinguished. However, we refer to the 1998 concept with three types of cognitive load as this concept is the basis for most of the research we refer to and because we had this original concept in mind when developing the learning materials and experimental design. Moreover, most cognitive load questionnaires assume a three-factor model (Krieglstein et al., 2022). Furthermore, a recent confirmatory factor analysis found stronger support for the three-factor model than for the two-factor model (Zavgorodniaia et al., 2020).

1.2 | Self-explanations

When learning from examples, learners need to engage in generative activities. An effective generative learning activity is *self-explaining* (e.g., Bisra et al., 2018; Rittle-Johnson et al., 2017). Self-explanations can occur spontaneously (Chi et al., 1989), but learners can also be *prompted to self-explain* (Renkl & Eitel, 2019). For instance, Hefter et al. (2014, 2015, 2018) conducted three studies examining the effects of a web-based training intervention comprising video examples and self-explanation prompts. The intervention aimed to promote both understanding and application of argumentative strategies. In all three studies, the self-explanation quality mediated the intervention's beneficial effect on the respective outcome measures (Hefter et al., 2014, 2015, 2018).

When learners explain content from examples to themselves, this can make them engage more deeply with the underlying principles of the example, as they basically try to make sense of the given learning materials (Wylie & Chi, 2014). Thus, self-explanations promote GCL (Renkl et al., 2009). However, very complex learning material could also induce such a high ICL that the working memory is pushed to its capacity limits. In such a case, modelling examples would still reduce ECL. However, this capacity gain may not be sufficient to deal with the increased GCL, which could produce a working memory overload and thus decrease learning (Große & Renkl, 2006; Sweller, 2006).

Taken together, for an ill-structured domain, video-based modelling examples with self-explanation prompts can be expected to be more beneficial for learning than practising to apply a strategy – provided the learning material is not so complex as to cause cognitive overload. Consequently, the first aim of the present study was to investigate the effects of video-based modelling examples and self-explanation prompts on cognitive load and learning outcomes for complex learning material in an ill-structured domain.

1.3 | Comparative self-explanations

One potentially effective but also very challenging form of self-explanation prompts asks learners to compare examples (Alfieri et al., 2013; Rittle-Johnson & Star, 2011). According to the *example comparison principle* (Renkl, 2014), comparing several examples benefits the development of abstract schemata and allows learners to discover similarities and differences between these examples (Gentner et al., 2003; Richland et al., 2007; Schwartz & Bransford, 1998). Comparisons have been investigated in well-structured, algorithmic domains, such as algebra: In a study by Rittle-Johnson and Star (2007), text-based worked examples for algebraic equations were designed to showcase more and less efficient solution methods. Pairs of seventh-grade students received these worked examples either side-by-side with self-explanation prompts encouraging comparisons (i.e., comparison condition) or sequentially with prompts that did not encourage comparisons (i.e., control condition). Students in the comparison condition exhibited greater advancements in procedural knowledge and procedural flexibility (i.e., the ability to select and apply the correct problem-solving strategy depending on certain features of the problem to be solved). Furthermore, they demonstrated similar improvement in conceptual knowledge.

Comparisons are also beneficial for learning in ill-structured domains. For instance, Gentner et al. (2003) conducted three studies focusing on instructing undergraduate students in negotiation strategies using example pairs. Notably, in experiment 2, the instruction to compare examples for key differences and similarities was more beneficial for learning than studying the examples sequentially (Gentner et al., 2003). Another example is provided by the two studies by Glogger-Frey et al. (2015, 2022). To teach pre-service teachers how to assess students' elaboration strategies in learning journals, Glogger-Frey et al. (2015, 2022) provided participants with contrasting cases of learning journal extracts differing in the quality of elaborations. In both studies, the inclusion of contrasting cases within a worked example condition, where quality criteria for assessing elaboration were provided alongside the example cases, proved effective in facilitating learning.

When learners are encouraged by comparison prompts to assess how well a strategy is executed, they engage with the learning material in depth. Thereby, the development of a differentiated mental representation of the problem-solving strategy to be learned is promoted. From a CLT perspective, one could thus argue that the beneficial effects of comparisons are due to an increase of GCL (Sweller, 2006).

However, the effects of (different types of) comparisons on learning seem to depend on the learners' prior knowledge: Rittle-Johnson et al. (2009) tested seventh- and eighth-grade students' prior knowledge of algebra and then provided them with pairs of text-based worked examples of solved linear algebraic equations in three conditions: a method comparison condition (i.e., worked example pairs included the same equations but were solved with different methods), a problem comparison condition (i.e., worked example pairs included different equations that were solved with the same method), and a sequential condition without comparison. In terms of learning outcomes, students with little prior knowledge benefitted most in the problem comparison condition or the sequential condition. Students with higher prior knowledge benefitted the most when they compared methods (Rittle-Johnson et al., 2009). The authors argue that this finding is an example of an expertise-reversal effect (Kalyuga et al., 2003): The instructional approach that was most beneficial for novice learners with little prior knowledge was not beneficial for learners with higher prior knowledge and vice versa.

This effect can again be explained with the CLT: Comparison processes in general induce heavy demands on working memory (Holyoak, 2012). Thus, comparing examples side by side should induce a substantially higher ICL in comparison to sequentially studying examples one by one, as element interactivity is much higher. Consequently, especially for complex problems that are high in ICL, only learners with higher prior knowledge can be expected to benefit from comparisons, that is, to engage in GCL-related processes and to achieve better learning outcomes. For learners with less prior knowledge, comparisons with complex problems are likely to produce cognitive overload. For these learners, sequential study of examples might be more beneficial (e.g., Rittle-Johnson et al., 2009).

However, the assumption that comparisons or comparative prompts are appropriate for stronger learners has so far only been investigated in rather well-structured domains with text-based worked examples. To our knowledge, there are no studies investigating those effects for video-based modelling examples teaching a complex problem-solving strategy to learners with differing levels of prior knowledge. Consequently, the second aim of the present study was to compare the effects of *comparative and sequential self-explanation prompts* on cognitive load and learning outcomes for learners with different levels of prior knowledge when learning with video-based modelling examples in an ill-structured domain.

1.4 | Static representations for comparisons of video-based modelling examples

Example comparisons can be implemented relatively easily for text-based worked examples in well-structured domains. The static and non-transient format of these text-based examples allows comparisons of the examples' critical features side by side (see Glogger-Frey et al., 2015; Rittle-Johnson & Star, 2007). For problems in complex domains, often there is not only one correct solution method, and correct solutions can take very different forms. Especially in these domains, the systematic comparison of examples helps to distinguish good from suboptimal solutions (Gentner et al., 2003; Glogger-Frey et al., 2015, 2022; Renkl et al., 2009). However, example comparisons seem to be difficult to implement for video-based modelling examples. Comparing such video examples would require learners to either watch two videos at the same time or pause the videos repeatedly. Hence, it is not surprising, that (to our knowledge) there is no research published in peer-reviewed journals on comparing video examples.

Against this background, we propose that after watching (parts of) a video example, learners are provided with a static representation of the (so far) observed problem-solving process as the basis for comparing and explaining. Such a representation could be realized, for example, by a text-based or graphical summary (e.g., a table, a bullet-point summary, or a mindmap) of the problem-solving process so far, highlighting the critical steps or it might be an overview of a product created in the problem-solving process. Such a representation consisting of text and/or image is non-transient and thus better suited for comparison than a transient video. To allow for comparison, learners can also be provided with an additional representation of the current state of an alternative solution to the same problem. This could be a solution of lower quality (e.g., Glogger-Frey et al., 2015, 2022; Rittle-Johnson & Star, 2007). Hence, we developed and explored the use of *representation-based self-explanation prompts* that allow for comparisons of problem-solving processes as demonstrated in a video-based modelling example.

1.5 | Present study and research questions

The present study was conducted with automotive apprentices learning a complex problem-solving strategy, namely a strategy for diagnosing complex car malfunctions. Such complex malfunctions cannot be easily diagnosed with the help of the usually available computer-based expert system. Complex malfunctions are also characterized by the fact that they often share the same symptoms, although the causes can be fundamentally different (Abele, 2018; Abele & von Davier, 2019). In such cases, a systematic strategy is necessary. Applying the instructed diagnostic strategy includes filling in and executing a so-called diagnosis plan. This plan is a table in which the apprentices (or other technicians applying this strategy) list hypothesised causes, diagnostic measurement methods and their requirements, measurement results, as well as the conclusions drawn therefrom for the present malfunction. Thus, a filled-in diagnosis plan represents a summary of a diagnostic process. We developed modelling examples in a screencast format showing an expert applying the diagnostic strategy and filling out a diagnosis plan in a computer simulation (Gschwendtner et al., 2009; Meier et al., 2022, 2023). While studying the modelling examples, apprentices received self-explanation prompts that referred to the filled-in diagnosis plans that served as static representations of the problem-solving process. For each modelling example, we designed two versions of diagnosis plans: the expert plan from the modelling examples and a novice plan providing an overview of an alternative but worse application of the same strategy. In a first condition, apprentices received these two diagnosis plans side by side and answered *comparative prompts*, that is, they were asked to self-explain and compare how well the expert and the novice had filled out their diagnosis plans, respectively. In a second condition, apprentices answered *sequential prompts*, that is, apprentices self-explained first the expert plan and then the novice plan in the described manner, but without comparing them directly. In a third condition (*control*), which can be regarded as a learning by problem-solving condition, the apprentices did not receive any modelling examples and therefore no self-explanation prompts or worked-out diagnosis plans; instead, they tried to diagnose the car malfunctions in the computer simulation on their own.

We investigated the effects of modelling examples and the different self-explanation prompts on apprentices' diagnostic knowledge and skills and cognitive load. Diagnostic knowledge and skills were assessed before and after the intervention. Cognitive load was assessed once after the intervention. Building on cognitive load theory (e.g., Paas & van Gog, 2006) and the worked or modelling example effect (Renkl, 2014; van Gog et al., 2019), we investigated the following hypotheses:

- **Hypothesis 1.** *We expected apprentices learning with modelling examples to experience a lower extraneous and a higher germane cognitive load than apprentices trying to solve the respective problems on their own.*

- **Hypothesis 2.** *We expected a greater increase in diagnostic knowledge and skills from before to after the learning phase for apprentices learning with the modelling examples in comparison with apprentices attempting to solve the respective problems on their own.*

We assumed that, due to the higher element interactivity, the comparative self-explanation prompts would be more demanding than the sequential self-explanation prompts. In line with research on example comparison and the expertise reversal effect (e.g., Kalyuga et al., 2003; Renkl, 2014; Rittle-Johnson et al., 2009, 2017; Rittle-Johnson & Star, 2007), we hypothesised the following with respect to our different prompt types:

- **Hypothesis 3.** *We expected the more demanding comparative self-explanation prompts to induce a higher intrinsic cognitive load than the sequential self-explanation prompts.*

- **Hypothesis 4.** *We expected apprentices with low prior knowledge to perceive a higher germane cognitive load when learning with sequential self-explanation prompts than when learning with comparative self-explanation prompts. In contrast, apprentices with high prior knowledge were expected to experience a higher germane cognitive load when learning with comparative self-explanation prompts than when learning with sequential self-explanation prompts.*

- **Hypothesis 5.** *For participants with low prior knowledge, we expected a larger increase in diagnostic knowledge and skills when learning with sequential self-explanation prompts than when learning with comparative self-explanation prompts. In contrast, for participants with high prior knowledge, we expected a larger increase in diagnostic knowledge and skills when learning with comparative self-explanation prompts than when learning with sequential self-explanation prompts.*

Table 1 provides an overview of these five hypotheses.

2 | METHODS

2.1 | Participants and design

We conducted a computer-based experiment in two sessions with three experimental conditions at German vocational schools. Session 1 included the pretest. In session 2, the intervention and the posttest

TABLE 1 Overview of the hypotheses.

Hypothesis	Learner characteristics	Factor: Learning with ...	Expected outcome
H1	All learners	Modelling examples	Lower ECL
			Higher GCL
		No modelling examples	Higher ECL
			Lower GCL
H2	All learners	Modelling examples	Higher increase in diagnostic knowledge and skills
		No modelling examples	Lower increase in diagnostic knowledge and skills
H3	All learners	Comparative prompts	Higher ICL
		Sequential prompts	Lower ICL
H4	Low prior knowledge learners	Comparative prompts	Lower GCL
		Sequential prompts	Higher GCL
	High prior knowledge learners	Comparative prompts	Higher GCL
		Sequential prompts	Lower GCL
H5	Low prior knowledge learners	Comparative prompts	Lower increase in diagnostic knowledge and skills
		Sequential prompts	Higher increase in diagnostic knowledge and skills
	High prior knowledge learners	Comparative prompts	Higher increase in diagnostic knowledge and skills
		sequential prompts	Lower increase in diagnostic knowledge and skills

took place. The sessions were conducted during school hours in the apprentices' classrooms. All material was presented in digital form.

We conducted two a-priori power analyses with Gpower 3.1 (Faul et al., 2007) to calculate the required sample sizes. We aimed for a power of 0.80. Based on previous studies on the worked example effect (e.g., Nievelstein et al., 2013; Schwonke et al., 2009; van Gog et al., 2011) and self-explanation prompts (e.g., Atkinson et al., 2003; Hilbert & Renkl, 2009), we expected medium effect sizes (e.g., Cohen's f > 0.25 or η^2 > 0.06; Cohen, 1988). For the analyses regarding H1, H3 and H4 (i.e., analyses of variance, ANOVAs), the required sample size was $N = 128$. For the analyses regarding H2 and H5 (i.e., repeated measures analyses of variance, RM-ANOVAs), the required sample size was $N = 34$. In total, 135 apprentices participated in session one and 131 participated in session two. However, only 118 apprentices participated in both sessions and can thus be included in the analyses. Thus, the sample size falls a little short of the sample size ($N = 128$) required for the ANOVAs, but is sufficient for the RM-ANOVAs. As data collection took place shortly before the summer holidays (6 weeks), a secondary data collection to achieve the required sample size was not feasible. A larger sample may have enabled us to demonstrate additional (smaller) effects. However, the effects we did discover can still be interpreted.

Apprentices were 20.08 years old ($SD = 2.04$), 114 were male, and 4 were female. For most apprentices ($n = 103$), German was their only first language, 13 reported another first language, and two apprentices reported that German was not their first language. Although these two non-native speakers might be at a disadvantage because of the German test and learning material, we did not exclude them from the analyses because they did not show extreme values (i.e. values more than 3 standard deviations above or below the mean) on any of the variables in either session. Regarding general school education, 13 apprentices had a university entrance qualification (*Abitur*), 96 apprentices had a secondary school leaving certificate (*Mittlere Reife*), and nine apprentices had a lower secondary school leaving certificate (*Hauptschulabschluss*). At the beginning of the first session, we assessed the apprentices' general prior knowledge about car diagnoses with a prior knowledge test. Then the apprentices completed the pretest on all repeated measures variables (i.e., diagnostic knowledge and skills). In the second session, the apprentices received the intervention, rated their cognitive load, and completed the posttest on the repeated measures variables. For the intervention, apprentices were assigned to one of three experimental conditions. In a first condition, apprentices learned with modelling examples and *comparative self-explanation prompts* ($n = 42$). In a second condition, apprentices received modelling examples and *sequential self-explanation prompts* ($n = 39$). In a third condition (*control*), apprentices received no modelling examples and, thus, no self-explanation prompts ($n = 37$). Instead, these apprentices tried to diagnose the malfunctions that were illustrated in the modelling examples themselves in the computer simulation.

2.2 | Intervention

2.2.1 | Diagnostic strategy

In collaboration with subject-matter experts and based on respective literature (e.g., Abele, 2014), we developed an intervention in which apprentices learned about a strategy to diagnose complex car malfunctions. This strategy comprised three steps: (1) When diagnosing car malfunctions, apprentices should first formulate hypotheses about possible causes for the present malfunctions. These hypotheses

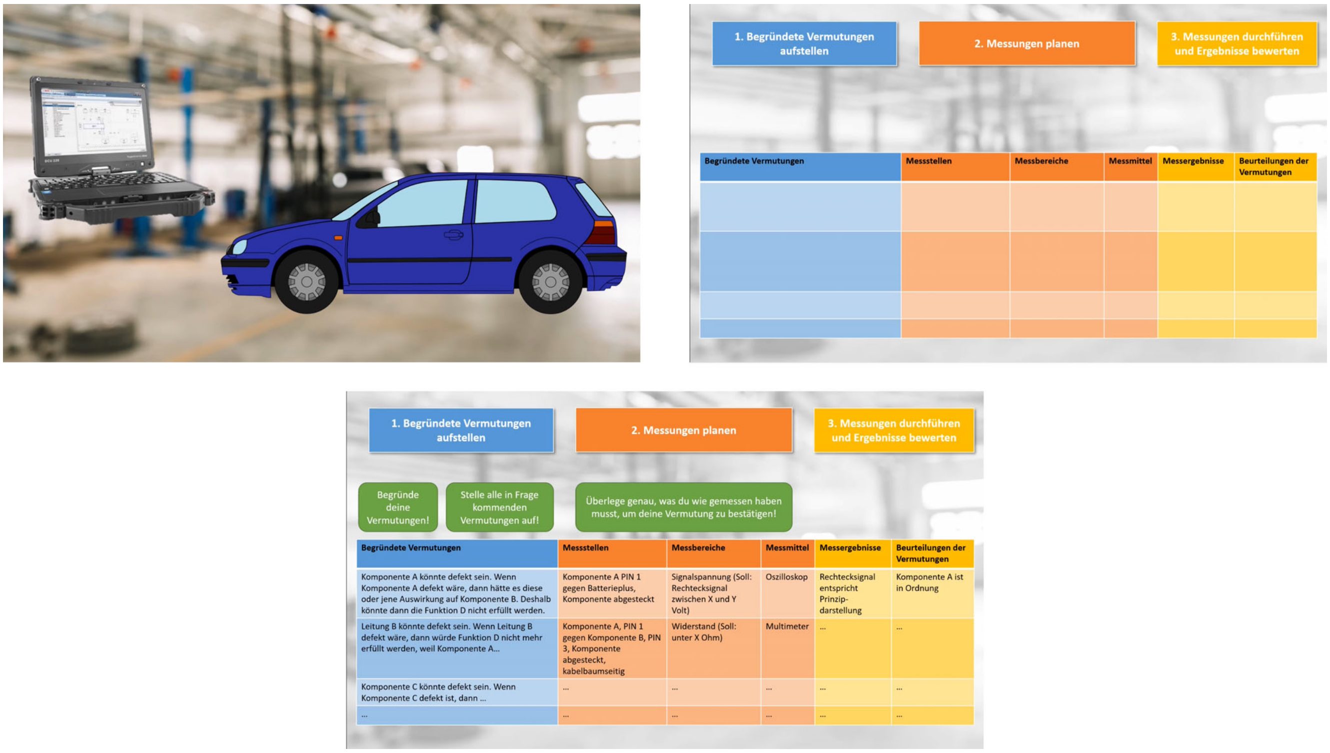

FIGURE 1 Screenshots of the instructional videos. The screenshots are from the instructional videos explaining the three diagnostic steps using the diagnosis plan. The three green boxes in the bottom screenshot contain the rules underlying the first two diagnostic steps.

should be reasoned, that is, based on the functional relationships of different relevant components in an automotive system. To formulate these reasoned hypotheses, apprentices learned about two underlying rules, namely the *reasoning rule* (i.e., 'formulate what function is probably impaired, what components are relevant to accomplishing that function, and how those components typically work together to accomplish the function'), and the *rule of completeness* (i.e., 'formulate all possible hypotheses and do not just rely on your first idea'). (2) The second diagnostic strategy step comprises the planning of (electro-technical) measurements to verify the hypotheses. The planning includes collecting information on (2a) measuring points, (2b) measuring range, and (2c) measuring equipment for each of the hypotheses. We emphasized the importance of these three points with the so-called *carefulness rule* (i.e., 'think carefully about what and how you must have measured to confirm your hypothesis'). (3) In the third and last diagnostic step the planned measurements are executed and the measurement results and with it the respective hypotheses are evaluated. Proceeding through these three steps of the diagnostic strategy was supported by a diagnosis plan. This diagnosis plan was a six-column table with the six columns corresponding to (1) reasoned hypotheses, (2a) measuring points, (2b) measuring ranges, (2c) measuring equipment, (3a) measuring results, and (3b) evaluations of the hypotheses. To teach apprentices this strategy as well as how to fill out the diagnosis plan, we developed an intervention consisting of instructional videos, two modelling examples, and three self-explanation prompts for each of the modelling examples. These learning materials are described below. Note that for the first modelling example, the instructional videos and the modelling example were presented in an *interleaved format*. This means that the instructional videos explaining the strategy initially and the first modelling example, which consisted of one video per step illustrating the application of the strategy, were shown in alternation. A detailed explanation and rationale for this format can be found in the Appendix A.

2.2.2 | Instructional videos

Six instructional videos (see Figure 1) briefly explained the three diagnostic steps with the three underlying rules and how to fill out the diagnosis plan along these steps (overall duration: 10:38 minutes). Participants from all three conditions received these instructional videos and thus learned about the diagnostic strategy.

2.2.3 | Modelling examples (first experimental variation)

The modelling examples showed an expert diagnosing a malfunction by applying the steps of the diagnostic strategy in the computer simulation while also filling out a diagnosis plan (see Figure 2). The expert verbalized his cognitive processes. Corresponding to the three diagnostic steps, both modelling examples consisted of three videos. The three videos of the first modelling example took 20:12 minutes, the second modelling example took 13:50 minutes.

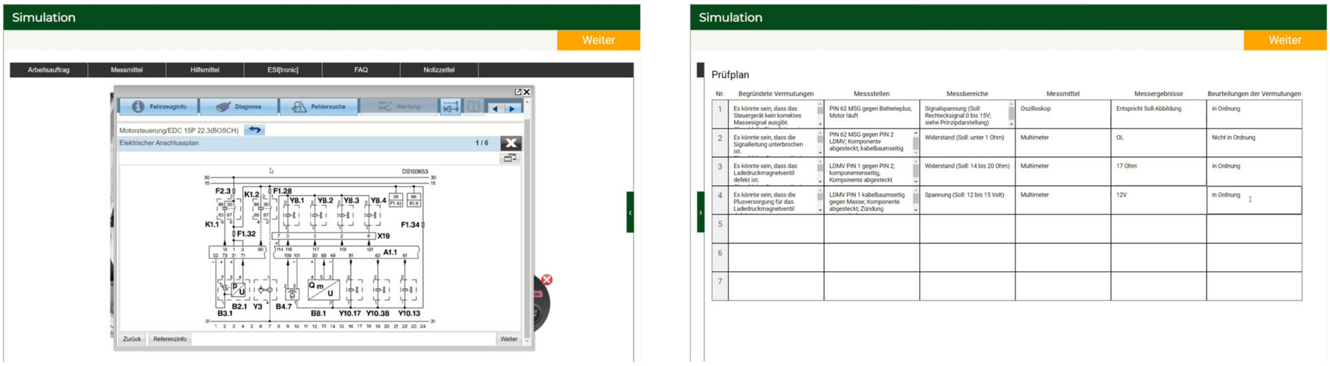

FIGURE 2 Screenshots of the first modelling example. The left screenshot shows how the expert uses the computer-based expert system to open an electrical circuit diagram. These diagrams illustrate the interrelationships between electrotechnical components and are thus an important resource for formulating hypotheses. The right screenshot shows how the expert fills in the diagnosis plan.

The modelling examples constituted the first experimental variation as – dependent on the experimental condition – apprentices either learned with modelling examples or tried to solve the respective problem on their own, that is, they tried to diagnose the malfunction on their own.

2.2.4 | Self-explanation prompts (second experimental variation)

Three self-explanation prompts were given after diagnostic steps 1 and 2 in the modelling examples that asked learners to explain how well the three underlying rules in these diagnostic steps, namely the reasoning rule, the rule of completeness, and the carefulness rule were applied in the example. The prompts had an open-book format, that is, the respective rule was displayed at the top of the page (Hiller et al., 2020). Besides the rule, apprentices were provided with (a relevant section of) the diagnosis plan as it had been filled out by the expert in the modelling examples (i.e., expert solution). The apprentices also received a novice solution of the same diagnostic step for the same problem, namely they were provided with (a section of) a diagnosis plan as it had been filled out by a less experienced hobby mechanic.

The format of the self-explanation prompts constituted the second experimental variation: In the *comparative self-explanation prompt condition*, the apprentices received the expert solution and the novice solution at the same time side by side and were instructed to compare the solutions, to look for similarities and differences, and to explain how differently well the expert and the hobby mechanic applied the respecting rule. After each prompt, apprentices were provided with a solution: In a written text it was explained and demonstrated that, for example regarding the rule of completeness, the expert had formulated all possible hypotheses while the hobby mechanic's diagnosis plan was not complete. In the *sequential self-explanation prompt condition*, the apprentices received the expert solution and the novice solution successively. For both the expert and the novice solution the apprentices were asked to explain how well the expert or the novice had applied the respecting rule. After providing an answer, apprentices received the corresponding solution. Note that apprentices in the control condition did not receive modelling examples and thus also no prompts. Instead, these apprentices tried to diagnose the malfunctions that were illustrated in the modelling examples themselves in the computer simulation.

2.3 | Testing materials

We used different tests to investigate the effects of modelling examples and comparative versus sequential self-explanation prompts: Only in the pretest in session 1, we assessed apprentices' *prior knowledge*. Both in the pretest and posttest, various tests were administered to measure the apprentices' development in *diagnostic knowledge* and *diagnostic skills*. In the posttest only, we assessed the participants' *cognitive load* during learning. These tests are described below. In most tests, closed and open question items were used. Closed items were scored automatically. For open question items, the first author and a subject matter expert (i.e., the second author) developed a coding scheme. Then, a student assistant and the first author scored 25% of all answers and adjusted the coding schemes until achieving an interrater reliability of Cohen's $\kappa > 0.6$. Then the student assistant independently scored the remaining answers. For some items, coding required very detailed automotive diagnostic expertise and no sufficient reliability could be established in the codings of the student assistant and the first author. In these cases, the first author coded the answers. For items where this applies, this is noted separately in the detailed description below.

2.3.1 | Prior knowledge

As a measure of general prior knowledge, we assessed the apprentices' *diagnosis-relevant reception competence*. This competence includes the ability to read different diagnosis-relevant documents, such as electrical circuit diagrams, and is thus required for successful diagnoses of automotive malfunctions. For this test, we used a selection of five out of 24 items from the diagnosis-relevant reception competence (DRC) test by Norwig et al. (2021), as in our previous study (Meier et al., 2022, 2023). To prevent floor and ceiling effects, we selected items for their midrange solution range (ranging from 32% to 71% in Norwig et al., 2021) and with the highest item-total correlation (>0.43 for all 5 items in Norwig et al., 2021). Apprentices could achieve up to five points on this DRC test. DRC test scores were used to test for moderating effects of prior knowledge (i.e., hypotheses H4 and H5).

2.3.2 | Diagnostic knowledge and skills

We applied one test to measure the apprentices' diagnostic knowledge and two tests to measure their diagnostic skills. These tests were applied both in sessions 1 and 2.

Diagnostic knowledge

To measure apprentices' diagnostic knowledge, two questions were asked in the *strategy description test*: First, apprentices described by which steps they would proceed in a diagnosis when there is only little assistance from a computer-based expert system (i.e., complex diagnosis). Apprentices could achieve six points for this question. Interrater reliability between the student assistant and the first author was acceptable (session 1: Cohen's $\kappa = 0.864$; session 2: Cohen's $\kappa = 0.689$). Second, apprentices described what would go through their minds when reading the error memory of a car and thinking about why the component/subsystem named in the error memory entry might be malfunctioning. Apprentices could achieve three points for this question. This second question required extensive knowledge of electrotechnical car systems and was thus coded by the first author. Taken together, the maximum achievable score for the strategy description test was nine points.

Diagnostic skills

Two tests were administered to measure apprentices diagnostic skills: First, the *strategy completion test* assessed scaffolded diagnostic skills. Apprentices were successively provided with three diagnostic scenarios – one scenario for each of the diagnostic steps. For each scenario, apprentices answered different open and closed questions to describe or carry out (parts of) the diagnostic strategy and thereby complete the three diagnostic steps. For example, in the scenario regarding the second step, after reading the respective diagnostic scenario, apprentices studied a circuit diagram and described an appropriate measurement, thereby completing the second diagnostic step, that is, planning measurements. All open questions in the strategy completion test were scored by the first author and not by the student assistant. Apprentices could achieve up to 47 points on this strategy completion test.

Secondly, to test independent diagnostic skills, in the *strategy application test* participants performed two diagnoses in the computer simulation both in sessions 1 and 2. For these independent diagnoses, apprentices first read a description of the malfunction and then diagnosed it. Eventually, apprentices were asked to describe the malfunction and how it could be repaired. Apprentices had 30 minutes to complete one diagnosis. The maximum score for each diagnosis was four points, resulting in a maximum score of eight points for the strategy application test. Interrater agreement was acceptable (first diagnosis, session 1: Cohen's $\kappa = 0.625$; second diagnosis, session 1: Cohen's $\kappa = 0.756$; first diagnosis, session 2: Cohen's $\kappa = 0.657$; second diagnosis, session 2: Cohen's $\kappa = 0.681$).

Cognitive load

After the intervention, we assessed the apprentices' *intrinsic* (two items), *germane* (two items), and *extraneous cognitive load* (three items) while learning on a seven-point Likert-scale (Klepsch et al., 2017; Klepsch & Seufert, 2020, 2021). Reliability was acceptable (intrinsic load: Cronbach's $\alpha = 0.62$; germane load: Cronbach's $\alpha = 0.64$; extraneous load: Cronbach's $\alpha = 0.61$).

2.3.3 | Procedure

The procedures in sessions 1 and 2 are displayed in Table 2. All material was presented on computers in digital form in a page-based learning environment. Once participants left a page, they could not go back.

3 | RESULTS

Exploratory analyses revealed participants with scores on dependent variables in the pretest or posttest that were more than 3 standard deviations below or above the grand mean. These participants were removed as outliers ($n = 6$). Exploratory analyses also showed a large variance in the quality of responses to the self-explanation prompts, as some apprentices gave meaningless answers (e.g., only single letters). Apprentices who answered less than 80% of the prompts meaningfully were consequently excluded from further analyses ($n = 14$). The final sample thus consisted of $N = 99$ apprentices with $n = 38$ apprentices learning with modelling examples and comparative self-explanation prompts, $n = 27$ apprentices learning with examples and sequential self-explanation prompts, and $n = 34$ apprentices in the control condition.

3.1 | Effects of modelling examples (H1 and H2)

To analyse the effects of modelling examples versus independent problem-solving on cognitive load, we conducted an analysis of variance (ANOVA) with *example condition* (i.e., modelling examples yes vs. no) as between-subjects variable. To analyse the effects on diagnostic knowledge and skills, we performed a repeated measures

TABLE 2 Procedures in sessions 1 and 2.

Phase	Content	Planned duration in min	Actual duration in min
Session 1			
Phase 1	Introduction to study and computer simulation, demographics	35	31
	Assessment of motivation[a]	5	4
	Strategy description test	10	4
Break		15	22
Phase 2	Strategy application test: First diagnosis in simulation	30	28
	Strategy application test: Second diagnosis in simulation	30	22
	Strategy completion test	20	25
Break		15	23
Phase 3	Diagnosis-relevant reception competence test	10	7
	Expertise of car technology test[b]	50	50
TOTAL SESSION 1		**220**	**216**
Session 2			
Phase 1	Refresher on computer simulation	5	4
	Instructional videos and modelling example 1 in interleaved format	55	44
	Modelling example 2	30	24
	Cognitive load rating	5	1
Break		15	27
Phase 2	Assessment of motivation[a]	5	2
	Strategy description test	10	3
	Strategy application test: First diagnosis in simulation	30	20
	Strategy application test: Second diagnosis in simulation	30	17
Break		15	28
Phase 3	Strategy completion test	20	14
TOTAL SESSION 2		**220**	**184**

[a]In our previous study (Meier et al., 2022, 2023) we assessed the apprentices' motivation (i.e., self-efficacy, interest, perception of challenge, and incompetence fear). In the present study, we assessed the apprentices' motivation with the same items to ensure that neither the modeling examples nor the prompts had negative effects on the apprentices' motivation. We found no evidence of such a negative effect. As we did not have any hypotheses regarding the effects of conditions on the apprentices' motivation, we do not report results for these variables.
[b]This expertise test on different automotive systems was not related to research questions investigated in the present paper and is thus not presented or analysed here.

analysis of variance (RM-ANOVA) with *example condition* as between-subjects variable and *timepoint* (pretest vs. posttest) as within-subjects variable. Table 3 displays descriptive data. Table 4 shows the results of the test of statistical significance. A significance level of 0.05 applies to all analyses. As effect size we used $\eta^2_{partial}$ with 0.01, 0.06, and 0.14 corresponding to a small, medium, and large effect, respectively (Lakens, 2013).

Regarding cognitive load, which was measured only once after the learning phase, the ANOVA indicated no effects of modelling examples as compared to independent problem-solving on participants' intrinsic, germane, or extraneous cognitive load (H1). Regarding learning outcome variables measured both in pre- and post-test (H2), the RM-ANOVA indicated two-way interaction effects of timepoint and example condition on the participants' scores in the strategy description test (small to medium effect) and the strategy completion test (small to medium effect). Figures 3 and 4 illustrate these interactions. For the strategy description test and the strategy completion test, apprentices who learned with modelling examples had a higher increase in test scores from the pretest to the posttest than apprentices in the control group who did not learn with the modelling examples.

Besides these interaction effects, we found three main effects of the timepoint on dependent variables that were not affected by interaction effects, namely on the participants' strategy application test score (large effect), their interest in diagnoses (medium effect), and their perception of diagnoses as a challenge (medium to large effect). Descriptive data in Table 3 indicate that the participants' scores in the strategy application test increased from the pretest to the post-test. Participants' interest and their perception of diagnoses as a challenge decreased from the pretest to the post-test.

TABLE 3 Descriptive data of dependent variables for the control condition (i.e., no modeling examples) and modeling examples condition.

	No modelling examples ($n = 34$)				Modelling examples ($n = 65$)			
	Pretest		Posttest		Pretest		Posttest	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Intrinsic load[a]	–	–	4.54	1.14	–	–	4.15	1.40
Germane load[a]	–	–	5.01	1.29	–	–	5.28	1.41
Extraneous load[a]	–	–	3.33	0.99	–	–	3.24	1.25
Strategy description test score[b]	0.97	0.87	1.41	1.58	0.86	0.98	1.95	1.61
Strategy completion test score[c]	14.78	5.31	16.65	6.41	16.58	4.95	21.09	7.22
Strategy application test score[d]	0.85	1.42	2.29	1.90	0.78	1.28	1.58	1.58

[a]7-point Likert-scale ranging from 1 = *absolutely not true* to 7 = *absolutely true*.
[b]0–9 points.
[c]0–47 points.
[d]0–8 points.

TABLE 4 Main and interaction effects of the example condition and timepoint on dependent variables.

				Statistical test results			
Analysis	Hypothesis	Independent variable(s)	Dependent variables	Df	F	*p*	$\eta^2_{partial}$
ANOVA		Example condition	Intrinsic load	1, 97	1.961	0.165	0.020
	H1		Germane load	1, 97	0.139	0.710	0.001
	H1		Extraneous load	1, 97	0.817	0.368	0.008
RM-ANOVA		Timepoint	Strategy description test score	1, 97	22.350	**<0.001**	0.187
			Strategy completion test score	1, 97	28.460	**<0.001**	0.227
			Strategy application test score	1, 97	41.389	**<0.001**	0.299
	H2	Timepoint* Example condition	Strategy description test score	1, 97	4.030	**0.047**	0.040
	H2		Strategy completion test score	1, 97	4.880	**0.030**	0.048
			Strategy application test score	1, 97	3.388	0.069	0.034

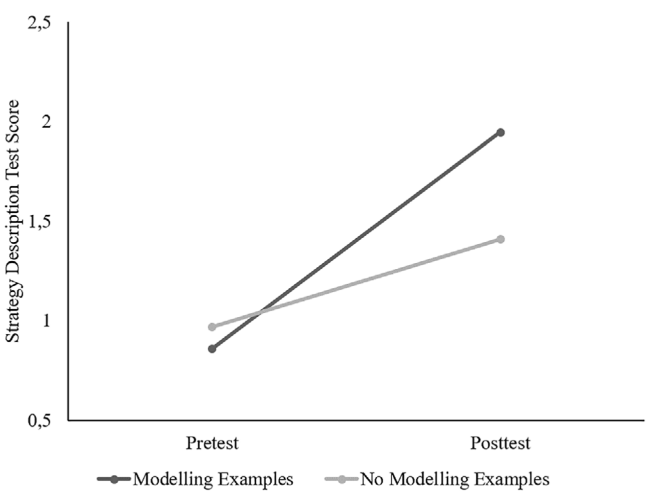

FIGURE 3 Interaction effect of timepoint and example condition on strategy description test score.

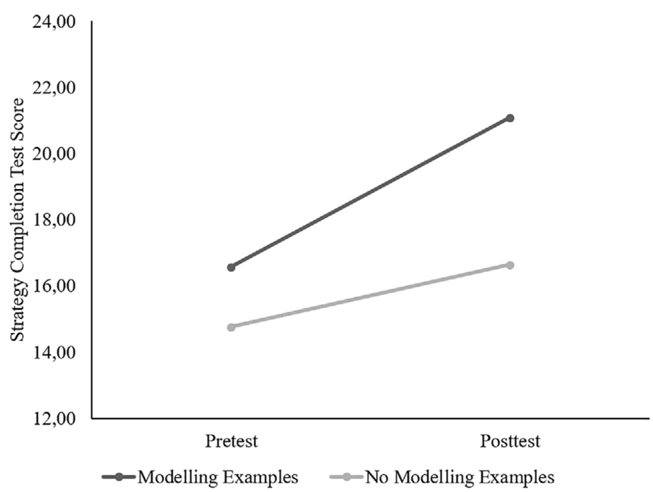

FIGURE 4 Interaction effect of timepoint and example condition on strategy completion test score.

Taken together, hypothesis H1 is not supported and must therefore be rejected, as participants in the modelling example condition did not perceive a lower extraneous and higher germane cognitive load. Hypothesis H2 is partially supported, as participants in the modelling example condition showed a greater increase in diagnostic knowledge (i.e., strategy description test) and scaffolded diagnostic

TABLE 5 Descriptive data of dependent variables and the moderator variable for the sequential and comparative self-explanation prompt condition.

	Sequential SE-prompts (*n* = 27)				Comparative SE-prompts (*n* = 38)			
	Pretest		Posttest		Pretest		Posttest	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
DRC test score[a]	3.44	0.97	–	–	3.55	1.13	–	–
Intrinsic load[b]	–	–	4.26	1.62	–	–	4.08	1.23
Germane load[b]	–	–	5.17	1.39	–	–	5.36	1.43
Extraneous load[b]	–	–	3.44	1.38	–	–	3.10	1.15
Strategy description test score[c]	0.78	0.97	1.44	1.37	0.92	1.00	2.32	1.68
Strategy completion test score[d]	15.67	5.41	19.80	7.26	17.24	4.56	22.01	7.14
Strategy application test score[e]	0.81	1.33	1.63	1.64	0.76	1.26	1.55	1.55

[a]0–5 points; this test was a prior knowledge measure and used as a moderation variable.
[b]7-point Likert-scale ranging from 1 = *absolutely not true* to 7 = *absolutely true*.
[c]0–9 points.
[d]0–47 points.
[e]0–8 points.

TABLE 6 Main and interaction effects of the prompt condition, the moderation variables, and timepoint on dependent variables.

				Statistical test results			
Analysis	Research question	Independent variable(s)	Dependent variables	Df	F	*p*	$\eta^2_{partial}$
ANOVA	H3	Prompt condition	Intrinsic load	1, 61	0.358	0.552	0.006
			Germane load	1, 61	0.252	0.617	0.004
			Extraneous load	1, 61	1.078	0.303	0.017
		Prompt condition* DRC test score	Intrinsic load	1, 61	0.717	0.492	0.023
	H4		Germane load	1, 61	0.611	0.546	0.020
			Extraneous load	1, 61	0.672	0.515	0.022
RM-ANOVA		Timepoint	Strategy description test score	1, 61	26.244	**<0.001**	0.301
			Strategy completion test score	1, 61	40.710	**<0.001**	0.400
			Strategy application test score	1, 61	20,876	**<0.001**	0.255
		Timepoint*Prompt condition	Strategy description test score	1, 61	2.739	0.103	0.043
			Strategy completion test score	1, 61	0.114	0.737	0.002
			Strategy application test score	1, 61	0.004	0.947	0.000
	H5	Timepoint*Prompt condition*DRC test score	Strategy description test score	1, 61	4.171	**0.045**	0.064
	H5		Strategy completion test score	1, 61	1.296	0.259	0.021
	H5		Strategy application test score	1, 61	0.422	0.518	0.007

skills (i.e., strategy completion test) but not in independent diagnostic skills (i.e., strategy application test).

3.2 | Effects of comparative versus sequential self-explanation prompts (H3, H4 and H5)

To analyse the effects of comparative versus sequential self-explanation prompts on cognitive load in the post-test, we conducted an ANOVA with *prompt condition* (i.e., comparative vs. sequential prompts) as between-subjects variable. To analyse the effects on diagnostic knowledge and skills, we performed an RM-ANOVA with *prompt condition* as between-subjects variable and *timepoint* (pretest vs. post-test) as within-subjects variable. To test for moderating effects of prior knowledge, we included the *diagnosis-relevant reception competence (DRC) test score* as additional continuous factors for both analyses. This factor was grand mean-centred for these analyses (Schneider et al., 2015). Table 5 shows descriptive data. Table 6 provides the results of the tests on statistical significance.

Regarding cognitive load, which was measured only once in the post-test, the ANOVA indicated no effects of comparative versus

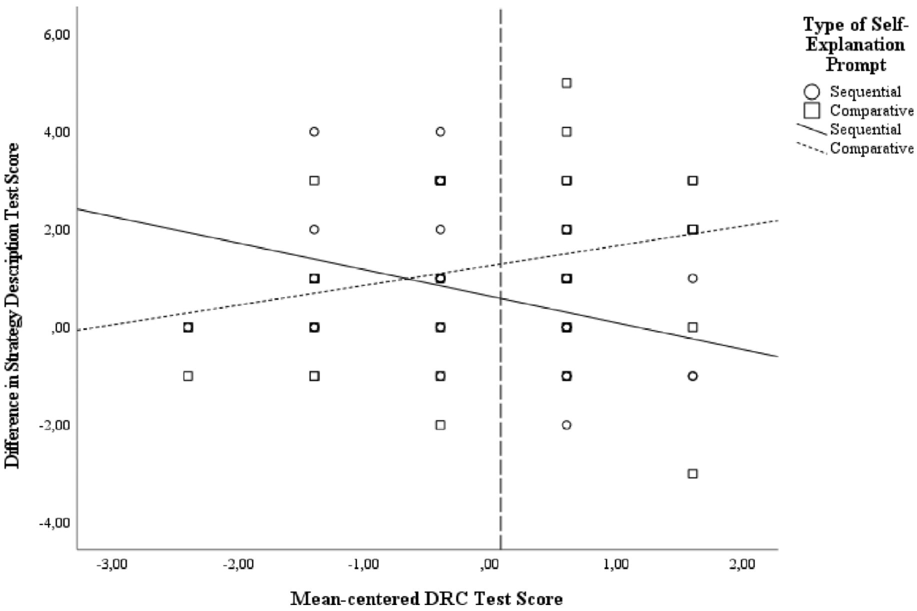

FIGURE 5 Scatter plot of grand mean-centred DRC test scores against the difference in strategy description test scores for both prompt conditions. The depicted effect is only significant for mean-centred DRC test scores >0.08, that is, right of the vertical longer dashed line.

sequential self-explanation prompts (i.e., neither main effects nor interaction effects with the DRC test score) on participants' intrinsic, germane, or extraneous cognitive load (H3 and H4). Regarding learning outcome variables measured in the pretest and post-test (H5), the RM-ANOVA indicated one significant three-way interaction of timepoint, prompt condition and mean-centred DRC test score on strategy description test scores (medium effect). Figure 5 indicates that participants with low DRC test scores benefitted (in terms of a higher increase in strategy description test scores) from sequential self-explanation prompts while participants with higher DRC test scores rather benefitted from comparative self-explanation prompts. To further explore this interaction, the Johnson-Neyman procedure (Hayes & Matthes, 2009; Montoya, 2019) was applied by using the SPSS-macro PROCESS by Hayes (2022). We tested where in the distribution of mean-centred DRC test scores the condition (i.e., comparative versus sequential prompts) had a statistically significant effect on the difference of strategy description test scores, calculated as post-test score minus pretest score. We found that the interaction effect was significant for learners with mean-centred DRC test scores larger than 0.08, which is essentially the half of participants with higher prior knowledge, $t(74) = 1.99$, $p = 0.05$.

No two-way interactions of timepoint and prompt condition were found. We found significant main effects of timepoint on the strategy description test score (large effect), strategy completion test score (large effect), strategy application test score (large effect), interest (large effect), and perception of challenge (medium to large effect). These effects correspond to the effects we had already observed when comparing the two example conditions.

Taken together, hypotheses H3 and H4 need to be rejected: Apprentices learning with comparative self-explanation prompts did not experience a higher intrinsic cognitive load than apprentices learning with sequential self-explanation prompts (H3). Moreover, apprentices' prior knowledge, as measured with the DRC test, did not moderate the effects of the different self-explanation prompts on germane cognitive load (H4). Hypothesis H5 is partially supported: regarding the strategy description test, apprentices with higher prior knowledge benefitted more from comparative self-explanation prompts.

4 | DISCUSSION

The present paper aimed to investigate the effects of video-based modelling examples versus independent problem-solving on cognitive load and learning outcomes when learning a complex problem-solving strategy. Furthermore, we compared the effects of comparative and sequential self-explanation prompts in combination with static representations of problem-solving processes for learners with different levels of prior knowledge. In the following paragraphs and in Table 7, we first give an overview of the investigated hypotheses and whether they could be confirmed.

We expected apprentices learning with modelling examples to experience a lower extraneous (ECL) and a higher germane cognitive load (GCL; H1) and subsequently to have a higher increase in diagnostic knowledge and skills (H2). Contrary to H1, we detected no effects of modelling examples on ECL or GCL. H2 could be confirmed for diagnostic knowledge and for scaffolded diagnostic skills, but not for independent diagnoses in the computer simulation. Moreover, contrary to H3, participants did not report higher intrinsic cognitive load (ICL) when learning with comparative prompts than when learning with sequential self-explanation prompts. Furthermore, H4 also had to be rejected: Low prior knowledge apprentices did not experience a higher GCL when learning with sequential self-explanation prompts, and higher prior knowledge apprentices did not experience a higher GCL when learning with comparative self-explanation prompts. Finally, we expected low prior knowledge apprentices to benefit more

TABLE 7 Overview and assessment of the hypotheses.

Hypothesis	Learner characteristics	Factor: Learning with …	Expected outcome	Assessment of hypothesis
H1	All learners	Modelling examples	Lower ECL	Rejected
			Higher GCL	
		No modelling examples	Higher ECL	
			Lower GCL	
H2	All learners	Modelling examples	Higher increase in diagnostic knowledge and skills	Confirmed for diagnostic knowledge and scaffolded diagnostic skills
		No modelling examples	Lower increase in diagnostic knowledge and skills	
H3	All learners	Comparative prompts	Higher ICL	Rejected
		Sequential prompts	Lower ICL	
H4	Low prior knowledge learners	Comparative prompts	Lower GCL	Rejected
		Sequential prompts	Higher GCL	
	High prior knowledge learners	Comparative prompts	Higher GCL	
		Sequential prompts	Lower GCL	
H5	Low prior knowledge learners	Comparative prompts	Lower increase in diagnostic knowledge and skills	Confirmed for high prior knowledge learners and diagnostic knowledge
		Sequential prompts	Higher increase in diagnostic knowledge and skills	
	High prior knowledge learners	Comparative prompts	Higher increase in diagnostic knowledge and skills	
		Sequential prompts	Lower increase in diagnostic knowledge and skills	

from learning with modelling examples along with sequential self-explanation prompts, and higher prior knowledge apprentices to benefit more in terms of learning outcomes when learning with comparative self-explanation prompts (H5). For the strategy description test asking for declarative knowledge about the diagnostic strategy, the hypothesised effect could be partially confirmed: participants with higher prior knowledge rather benefitted from comparative self-explanation prompts (H5).

In the remainder of the discussion, we first address our non-findings regarding cognitive load and discuss the challenges of consistent and standardized assessments of cognitive load, followed by a discussion of the effects of modelling examples on learning outcomes (second section), and a discussion of the self-explanation prompts (third section) Finally, we will present limitations of the study and implications for future research.

4.1 | Non-effects on cognitive load and challenges when assessing it

Contrary to H1, H3 and H4, we detected no effects of modelling examples on ECL or GCL. This finding challenges the prevalent assumption that worked examples reduce ECL and increase GCL as they make the use of ineffective problem-solving strategies unnecessary and free up working memory capacities for learning (Renkl, 2014; Sweller, 2006). Notably, many studies supporting this notion relied on a single measure of cognitive load, such as the mental effort scale by Paas (1992) used by Spanjers et al. (2012) and Van Gog et al. (2008). Other publications solely focused on learning outcomes without directly examining cognitive load but still argued with CLT (e.g., Rourke & Sweller, 2009). One could therefore argue that the assumption of working examples reducing ECL is much less well covered by evidence than is widely supposed. Therefore, we suggest that researchers investigating the effects of worked or modelling examples on learning should also investigate the effects on cognitive load to close this gap (Sweller, 2017). Moreover, it would be important for researchers to agree on comparable instruments to measure cognitive load to allow for overarching conclusions. An exception and one first step in this direction is study 3 by Klepsch and Seufert (2020), who developed the cognitive load instruments we also used in the present study. In their study, participants learned to solve complex math problems: In a learning phase, participants received such a math problem as well as the correct final result. In an example condition, participants additionally received the worked-out solution (i.e., a text-based worked example), while participants in the control or problem-solving condition had to work out the solution path themselves. Participants in the example condition reported a lower ECL and had better learning outcomes. Although we did use the same instruments to measure cognitive load, we did not find any effects of the modelling examples on ECL. One explanation for this might be that reliability was considerably worse in our study (i.e., Cronbach's α between 0.61 and 0.64, compared to McDonald's ω between 0.74 and 0.83 as reported by Klepsch & Seufert, 2020) Moreover, notable differences exist between the two studies: the duration of the learning phase (10 minutes in Klepsch and Seufert vs. over 60 minutes in our study) and the complexity of the learning material (simple text-based worked examples in mathematics vs. complex video-based modelling examples in an ill-defined domain combined with self-explanation prompts). It seems reasonable to assume that the duration of a learning phase as well as the timing of the cognitive load assessment affects the outcomes. For example, in two experiments Schmeck et al. (2015) found that learners who worked on six rather short problem-solving tasks of different complexity rated their mental effort and perceived task difficulty after all problems higher than the average of the six ratings made during problem solving. Similarly, in a series of experiments van Gog et al. (2012) found that a single mental effort rating following a series of tasks resulted in a higher mental effort than the average of ratings given after each individual task (Experiment 1 and 2), especially for complex tasks (Experiment 3). In order to be able to accurately assess cognitive load when learning or working on long and complex problem-solving processes such as the diagnosis process presented in this paper, we recommend repeated measurements of cognitive load.

4.2 | Utilization deficiency as possible reason for missing effects on independent diagnoses

In line with the literature on example-based learning (e.g., Renkl, 2014; Sweller, 2006; van Gog et al., 2019), H2 could be partially confirmed: Apprentices learning with modelling examples had a higher increase in diagnostic knowledge and they were also more proficient in applying the strategy when they worked on scaffolded tasks. However, these learning gains did not transfer to independent diagnoses in the computer simulation. One possible explanation might be that apprentices experienced a utilization deficiency (Hübner et al., 2010; Miller, 1994). This term describes that using a newly learned strategy, which is not yet automated, requires so much cognitive capacity that only little capacity remains to process the new problem that is to be solved (i.e., the malfunction in the simulation to be diagnosed). Consequently, the application of the strategy to the novel context fails. Such a utilization deficiency can be overcome by practice, that is, apprentices would have needed more time and opportunity to practice the diagnostic strategy (possibly with additional support) before performing diagnoses on their own.

4.3 | Differential effects of self-explanation prompts depending on prior knowledge

Also H5 could be partially confirmed: participants with higher prior knowledge rather benefitted from comparative self-explanation prompts. This finding is consistent with Rittle-Johnson et al. (2009): In their study, students learned to solve algebraic equations. Students with low prior algebra knowledge benefitted most from sequential example study or comparing problem types (a rather less complex comparison), rather than solution methods comparison (a rather complex comparison). Conversely, students with prior algebra knowledge learned more from comparing solution methods (Rittle-Johnson et al., 2009). Both these and our findings align with the expertise-reversal effect, suggesting different instructional approaches based on learners' prior knowledge levels (Kalyuga et al., 2003).

4.4 | Strengths and limitations

The study makes an important contribution to improving vocational education for automotive apprentices: Although the diagnosis of car malfunctions is a crucial aspect of the work of car mechatronics technicians (Spöttl et al., 2011), only about 15% of the apprentices master strategies for the diagnosis of complex malfunctions at the end of their apprenticeship (Abele & von Davier, 2019). Over all conditions, there were large increases in scores on the diagnostic knowledge tests as well as in the two independent diagnoses in the simulation. Accordingly, the developed intervention might be helpful to teach more apprentices a diagnosis for complex car malfunctions. Future research could investigate whether these positive effects persist over a longer period of time and whether they could also be confirmed in diagnoses on real vehicles.

Moreover, with the present study, we make a first proposal on how the positive effect of comparing examples can also be used for video-based modelling examples by combining them with static representations of the problem-solving process. Future studies on modelling examples could follow this direction and could investigate whether findings regarding the effects of comparisons of text-based worked examples also hold true for such representation-based comparisons of video-based modelling examples.

One limitation of the study concerns the procedure in the control group. While the instructional videos and the first modelling example were presented in an interleaved format for the modelling example conditions, such an interleaved format was not feasible for the control group (see Appendix A). For the apprentices in the control group, it would have meant first learning about only the first step of the strategy in the instructional videos. Then, analogous to the modelling examples, they would have had to carry out only the first step of the diagnosis in the simulation, that is, formulating hypotheses. Then they would have learned about the second step of the diagnostic strategy in the instructional videos and would have again only carried out the second step in the simulation themselves. Such a switch between instructional videos and the simulation would have required the simulation to save and reload intermediate states of the diagnostic process. This was technically not feasible. Hence, apprentices in the control group did first watch all six instructional videos on the diagnostic strategy and then tried to solve the problem in the computer simulation.

Another limitation could be the diverse composition of our sample. Note that to undergo apprenticeship training as an automotive mechatronic, no higher school qualification is required in Germany. This situation resulted in participants with different school qualifications, potentially introducing bias into the experiment. To validate the study results, a conceptual replication or an extension study (Zwaan et al., 2018) with a more homogeneous sample would be advisable. However, such an extension study with a more homogeneous sample would have lower ecological validity as the different levels of school qualification of our apprentices is to be expected in the reality of vocational classrooms.

5 | CONCLUSION

The present study investigated the effects of video-based modelling examples and various self-explanation prompts on automotive apprentices' diagnostic knowledge, skills, and cognitive load. For this study, we developed a video-based intervention to teach a diagnostic strategy. The intervention in general had beneficial effects on both diagnostic knowledge and skills and could therefore be implemented in the vocational education of automotive mechatronics apprentices, ideally with modelling examples. Unexpectedly, our findings revealed no effects of modelling examples on cognitive load. We argue that future research focusing on the effects of modelling long for complex problem-solving processes should use repeated measurements of cognitive load. While we observed that modelling examples can improve diagnostic knowledge and scaffolded diagnostic skills more strongly than independent problem solving, this effect may not carry through to diagnostic skills without further practice. Finally, our findings provide further evidence for the expertise reversal effect: participants with higher prior knowledge benefitted more from comparative prompts.

In summary, video-based modelling examples can be more beneficial for learning than practising to apply the diagnostic strategy. Static representations of key aspects of a problem-solving process can be used for comparisons of video examples. Comparative prompts are promising for learners with higher prior knowledge (cf. expertise-reversal effect). Further research, especially on the effects on cognitive load measured multiple times throughout a complex problem-solving process, is necessary.

AUTHOR CONTRIBUTIONS

Julius Moritz Meier: Conceptualization; data curation; formal analysis; investigation; methodology; writing – original draft. **Peter Hesse:** Conceptualization; investigation; project administration. **Stephan Abele:** Conceptualization; funding acquisition; methodology; project administration; resources; supervision; writing – review and editing. **Alexander Renkl:** Conceptualization; methodology; supervision; writing – review and editing. **Inga Glogger-Frey:** Conceptualization; funding acquisition; methodology; project administration; resources; supervision; writing – review and editing.

ACKNOWLEDGEMENTS

This work was supported by the Federal Ministry of Education and Research, Germany, and the Federal Institute for Vocational Education and Training, Germany, within the project 'DigiDIn-Kfz – Digitale Diagnostik und Intervention im Kfz-Wesen' [Digital diagnostics and intervention in the automotive sector], which is a subproject of the research and transfer initiative 'ASCOT+ –Technology-based Assessment of Skills and Competences in Vocational Education and Training' (grant number 21AP004C). Open Access funding enabled and organized by Projekt DEAL.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/jcal.12991.

DATA AVAILABILITY STATEMENT

The data supporting this study's findings are available from the corresponding author upon reasonable request.

ORCID

Julius Moritz Meier https://orcid.org/0000-0003-2671-4995
Inga Glogger-Frey https://orcid.org/0000-0002-1409-2116

REFERENCES

Abele, S. (2014). *Modellierung und Entwicklung berufsfachlicher Kompetenz in der gewerblich-technischen Ausbildung [modelling and development of vocational competence in vocational-technical education and training]*. Steiner.

Abele, S. (2018). Diagnostic problem-solving process in professional contexts: Theory and empirical investigation in the context of car mechatronics using computer-generated log-files. *Vocations and Learning*, *11*(1), 133–159. https://doi.org/10.1007/s12186-017-9183-x

Abele, S., & von Davier, M. (2019). CDMs in vocational education: Assessment and usage of diagnostic problem-solving strategies in car mechatronics. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Springer. https://doi.org/10.1007/978-3-030-05584-4_22

Alfieri, L., Nokes-Malach, T. J., & Schunn, C. D. (2013). Learning through case comparisons: A meta-analytic review. *Educational Psychologist*, *48*(2), 87–113. https://doi.org/10.1080/00461520.2013.775712

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, *95*(4), 774–783. https://doi.org/10.1037/0022-0663.95.4.774

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, *30*(3), 703–725. https://doi.org/10.1007/s10648-018-9434-x

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). L. Erlbaum Associates.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Frerejean, J., van Strien, J. L. H., Kirschner, P. A., & Brand-Gruwel, S. (2018). Effects of a modelling example for teaching information problem solving skills. *Journal of Computer Assisted Learning*, *34*(6), 688–700. https://doi.org/10.1111/jcal.12276

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, *95*(2), 393–408. https://doi.org/10.1037/0022-0663.95.2.393

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., & Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learning and Instruction*, *39*, 72–87. https://doi.org/10.1016/j.learninstruc.2015.05.001

Glogger-Frey, I., Treier, A.-K., & Renkl, A. (2022). How preparation-for-learning with a worked versus an open inventing problem affect subsequent learning processes in pre-service teachers. *Instructional Science*, *50*(3), 451–473. https://doi.org/10.1007/s11251-022-09577-6

Große, C. S., & Renkl, A. (2006). Effects of multiple solution methods in mathematics learning. *Learning and Instruction*, *16*(2), 122–138. https://doi.org/10.1016/j.learninstruc.2006.02.001

Gschwendtner, T., Abele, S., & Nickolaus, R. (2009). Computersimulierte Arbeitsproben: Eine Validierungsstudie am Beispiel der Fehlerdiagnoseleistungen von Kfz-Mechatronikern [can troubleshooting skills of car mechatronic technicians validly be assessed using computer-based simulations of real work sample?]. *Zeitschrift Für Berufs- Und Wirtschaftspädagogik*, *105*(4), 557–578.

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). The Guilford Press.

Hayes, A. F., & Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods*, *41*(3), 924–936. https://doi.org/10.3758/BRM.41.3.924

Hefter, M. H., Berthold, K., Renkl, A., Riess, W., Schmid, S., & Fries, S. (2014). Effects of a training intervention to foster argumentation skills while processing conflicting scientific positions. *Instructional Science*, *42*(6), 929–947. https://doi.org/10.1007/s11251-014-9320-y

Hefter, M. H., Renkl, A., Riess, W., Schmid, S., Fries, S., & Berthold, K. (2015). Effects of a training intervention to foster precursors of evaluativist epistemological understanding and intellectual values. *Learning and Instruction*, *39*, 11–22. https://doi.org/10.1016/j.learninstruc.2015.05.002

Hefter, M. H., Renkl, A., Riess, W., Schmid, S., Fries, S., & Berthold, K. (2018). Training interventions to Foster skill and will of argumentative thinking. *The Journal of Experimental Education*, *86*(3), 325–343. https://doi.org/10.1080/00220973.2017.1363689

Hilbert, T. S., & Renkl, A. (2009). Learning how to use a computer-based concept-mapping tool: Self-explaining examples helps. *Computers in Human Behavior*, *25*(2), 267–274. https://doi.org/10.1016/j.chb.2008.12.006

Hiller, S., Rumann, S., Berthold, K., & Roelle, J. (2020). Example-based learning: Should learners receive closed-book or open-book self-explanation prompts? *Instructional Science*, *48*(6), 623–649. https://doi.org/10.1007/s11251-020-09523-4

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (1st ed., pp. 234–259). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0013

Hübner, S., Nückles, M., & Renkl, A. (2010). Writing learning journals: Instructional support to overcome learning-strategy deficits. *Learning and Instruction*, *20*(1), 18–29. https://doi.org/10.1016/j.learninstruc.2008.12.001

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, *8*, 1997. https://doi.org/10.3389/fpsyg.2017.01997

Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science*, *48*(1), 45–77. https://doi.org/10.1007/s11251-020-09502-9

Klepsch, M., & Seufert, T. (2021). Making an effort versus experiencing load. *Frontiers in Education*, *6*, 645284. https://doi.org/10.3389/feduc.2021.645284

Krieglstein, F., Beege, M., Rey, G. D., Ginns, P., Krell, M., & Schneider, S. (2022). A systematic meta-analysis of the reliability and validity of subjective cognitive load questionnaires in experimental multimedia learning research. *Educational Psychology Review*, *34*(4), 2485–2541. https://doi.org/10.1007/s10648-022-09683-4

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Meier, J., Hesse, P., Abele, S., Renkl, A., & Glogger-Frey, I. (2023). Better self-explaining backwards or forwards? Prompting self-explanation in video-based modelling examples for learning a diagnostic strategy. *Instructional Science*. https://doi.org/10.1007/s11251-023-09651-7

Meier, J., Spliethoff, L., Hesse, P., Abele, S., Renkl, A., & Glogger-Frey, I. (2022). Promoting car mechatronics apprentices' diagnostic strategy with modeling examples: Development and evaluation of a simulation-based learning environment. *Studies in Educational Evaluation*, *72*, 101117. https://doi.org/10.1016/j.stueduc.2021.101117

Miller, P. H. (1994). Individual differences in children's strategic behaviors: Utilization deficiencies. *Learning and Individual Differences*, *6*(3), 285–307. https://doi.org/10.1016/1041-6080(94)90019-1

Montoya, A. K. (2019). Moderation analysis in two-instance repeated measures designs: Probing methods and multiple moderator models. *Behavior Research Methods*, *51*(1), 61–82. https://doi.org/10.3758/s13428-018-1088-6

Nievelstein, F., van Gog, T., van Dijck, G., & Boshuizen, H. P. A. (2013). The worked example and expertise reversal effect in less structured tasks: Learning to reason about legal cases. *Contemporary Educational Psychology*, *38*(2), 118–125. https://doi.org/10.1016/j.cedpsych.2012.12.004

Norwig, K., Güzel, E., Hartmann, S., & Gschwendtner, T. (2021). Tools to tap into the content of human minds – Think-Aloud-Interviews und Cognitive Labs als zentrale Bausteine zur Identifikation von Barrieren in Fehlerdiagnoseprozessen bei Auszubildenden des Kfz-Handwerks und zur Entwicklung addressantespezifischer Lehr–/Lernarrangements ("Tools to tap into the content of human minds" – The use of Think-Aloud and Cognitive Laboratory Interviews for Identifying Cognitive Barriers of Car Mechatronics Apprentices Diagnosing Electronic Car Systems and for Developing Educational Interventions Tailored to the Apprentices' Needs.). *Zeitschrift Für Berufs- Und Wirtschaftspädagogik*, *117*, 658–693. https://doi.org/10.25162/zbw-2021-0025

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Paas, F., & van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction*, *16*(2), 87–91. https://doi.org/10.1016/j.learninstruc.2006.02.004

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science*, *38*(1), 1–37. https://doi.org/10.1111/cogs.12086

Renkl, A., & Eitel, A. (2019). Self-explaining: Learning about principles and their application. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (1st ed., pp. 528–549). Cambridge University Press. https://doi.org/10.1017/9781108235631.022

Renkl, A., Hilbert, T., & Schworm, S. (2009). Example-based learning in heuristic domains: A cognitive load theory account. *Educational Psychology Review*, *21*(1), 67–78. https://doi.org/10.1007/s10648-008-9093-4

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science*, *316*(5828), 1128–1129. https://doi.org/10.1126/science.1142103

Rittle-Johnson, B., Loehr, A. M., & Durkin, K. (2017). Promoting self-explanation to improve mathematics learning: A meta-analysis and instructional design principles. *ZDM*, *49*(4), 599–611. https://doi.org/10.1007/s11858-017-0834-z

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve equations. *Journal of Educational Psychology*, *99*(3), 561–574. https://doi.org/10.1037/0022-0663.99.3.561

Rittle-Johnson, B., & Star, J. R. (2011). The power of comparison in learning and instruction: Learning outcomes supported by different types of comparisons. In *Psychology of learning and motivation* (Vol. 55, pp. 199–225). Elsevier. https://doi.org/10.1016/B978-0-12-387691-1.00007-7

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2009). The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge of equation solving. *Journal of Educational Psychology*, *101*(4), 836–852. https://doi.org/10.1037/a0016026

Roelle, J., Hiller, S., Berthold, K., & Rumann, S. (2017). Example-based learning: The benefits of prompting organization before providing examples. *Learning and Instruction*, *49*, 1–12. https://doi.org/10.1016/j.learninstruc.2016.11.012

Rourke, A., & Sweller, J. (2009). The worked-example effect using ill-defined problems: Learning to recognise designers' styles. *Learning and Instruction*, *19*(2), 185–199. https://doi.org/10.1016/j.learninstruc.2008.03.006

Schalk, L., Roelle, J., Saalbach, H., Berthold, K., Stern, E., & Renkl, A. (2020). Providing worked examples for learning multiple principles. *Applied Cognitive Psychology*, *34*(4), 813–824. https://doi.org/10.1002/acp.3653

Schmeck, A., Opfermann, M., van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, *43*(1), 93–114. https://doi.org/10.1007/s11251-014-9328-3

Schmitz, F. M., Schnabel, K. P., Stricker, D., Fischer, M. R., & Guttormsen, S. (2017). Learning communication from erroneous video-based examples: A double-blind randomised controlled trial. *Patient Education and Counseling*, *100*(6), 1203–1212. https://doi.org/10.1016/j.pec.2017.01.016

Schneider, B. A., Avivi-Reich, M., & Mozuraitis, M. (2015). A cautionary note on the use of the analysis of covariance (ANCOVA) in classification designs with and without within-subject factors. *Frontiers in Psychology*, *6*, 474. https://doi.org/10.3389/fpsyg.2015.00474

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, *16*(4), 475–522.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior*, *25*(2), 258–266. https://doi.org/10.1016/j.chb.2008.12.011

Spanjers, I. A. E., Gog, T., & Merriënboer, J. J. G. (2012). Segmentation of worked examples: Effects on cognitive load and learning: Segmentation of worked examples. *Applied Cognitive Psychology*, *26*(3), 352–358. https://doi.org/10.1002/acp.1832

Spöttl, G., Becker, M., & Musekamp, F. (2011). Anforderungen an Kfz-Mechatroniker und Implikationen für die Kompetenzerfassung [requirements for automotive mechatronics technicians and implications for the assessment of competencies]. In R. Nickolaus (Ed.), *Lehr-Lernforschung in der gewerblich-technischen Berufsbildung (Zeitschrift für Berufs- und Wirtschaftspädagogik: Sonderbd. 25)* (pp. 37–53). Steiner.

Sweller, J. (2006). The worked example effect and human cognition. *Learning and Instruction*, *16*(2), 165–169. https://doi.org/10.1016/j.learninstruc.2006.02.005

Sweller, J. (2017). The role of independent measures of load in cognitive load theory. In R. Zheng (Ed.), *Cognitive load measurement and application: A theoretical framework for meaningful research and practice*. Routledge.

Sweller, J. (2023). The development of cognitive load theory: Replication crises and incorporation of other theories can Lead to theory expansion. *Educational Psychology Review*, *35*(4), 95. https://doi.org/10.1007/s10648-023-09817-2

Sweller, J., Ayres, P. L., & Kalyuga, S. (2011). Cognitive load theory. *Springer*. https://doi.org/10.1007/978-1-4419-8126-4

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology*, *36*(3), 212–218. https://doi.org/10.1016/j.cedpsych.2010.10.004

van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures: Measuring mental effort. *Applied Cognitive Psychology*, *26*(6), 833–839. https://doi.org/10.1002/acp.2883

van Gog, T., Paas, F., & van Merriënboer, J. J. G. (2008). Effects of studying sequences of process-oriented and product-oriented worked examples on troubleshooting transfer efficiency. *Learning and Instruction*, *18*(3), 211–222. https://doi.org/10.1016/j.learninstruc.2007.03.003

van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*, *22*(2), 155–174. https://doi.org/10.1007/s10648-010-9134-7

van Gog, T., Rummel, N., & Renkl, A. (2019). Learning how to solve problems by studying examples. In J. Dunlosky & K. A. Rawson (Eds.), *The Cambridge handbook of cognition and education* (1st ed., pp. 183–208). Cambridge University Press. https://doi.org/10.1017/9781108235631.009

Wylie, R., & Chi, M. T. H. (2014). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 413–432). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.021

Zavgorodniaia, A., Duran, R., Hellas, A., Seppala, O., & Sorva, J. (2020). Measuring the cognitive load of learning to program: A replication study. *United Kingdom & Ireland Computing Education Research Conference*, 3–9. https://doi.org/10.1145/3416465.3416468

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, E120. https://doi.org/10.1017/S0140525X17001972

How to cite this article: Meier, J. M., Hesse, P., Abele, S., Renkl, A., & Glogger-Frey, I. (2024). Video-based modeling examples and comparative self-explanation prompts for teaching a complex problem-solving strategy. *Journal of Computer Assisted Learning*, *40*(4), 1852–1870. https://doi.org/10.1111/jcal.12991

APPENDIX A: INTERLEAVED PRESENTATION FORMAT OF INSTRUCTIONAL VIDEOS AND MODELLING EXAMPLES

For the first modelling example, instructional videos and the modelling example were presented in an *interleaved format*: After a general introduction to the diagnostic strategy, only the first diagnostic step was explained in the instructional videos. This video was followed by the first part of the first modelling example which showed the expert only performing the first step in the simulation. After the corresponding self-explanation prompts, apprentices watched the instructional video explaining the second diagnostic step followed by the second part of the modelling example and so on. We chose this interleaved format for the following reason: In a previous study on the same learning topic, we found no beneficial effects of the modelling examples on participants' learning outcomes (Meier et al., 2023). In this study, the first learning phase took 35 minutes and comprised five instructional videos that already included partial examples (16:33 minutes) and four practice tasks that presumably supported knowledge organization well (Roelle et al., 2017). This extensive instruction may have provided all apprentices with enough knowledge of the respective strategy rendering the modelling examples useless. In the present study, to omit the partial examples within the instructional videos, we decided on the interleaved format in learning phase one, in which only parts of the diagnostic strategy are instructed and then immediately illustrated by parts of the subsequent modelling examples. For technical reasons, this interleaved format was not viable for the control group. Apprentices in this group did first watch all six instructional videos on the diagnostic strategy and then tried to solve the problem in the simulation.

When learning with the second modelling example, apprentices had already been introduced to the diagnostic strategy. Hence, here apprentices only watched the three videos of the second modelling examples and worked on the three self-explanation prompts, but we did not provide apprentices with additional instructional videos.