Fostering the learning of problem-solving skills: Prompting stepwise retrieval during worked examples.

Aafke Gruintjes, 6517234

Educational Sciences, University of Utrecht

Master thesis

First assessor: Margot van Wermeskerken

Second assessor: Jeroen Janssen

10-06-2024

Wordcount: 7891

Abstract

Problem-solving tasks are a common part of many school subjects, but acquiring problem-solving skills can be difficult. This study investigated the effectiveness of repeatedly prompting students to retrieve prior knowledge while studying worked examples. It was studied whether prompting stepwise retrieval during worked example study (retrieval condition) leads to higher performance on scaffolded and isomorphic problem-solving tasks compared to only studying worked examples (control condition). No effects of stepwise retrieval practice, positive or negative, were found on both the scaffolded and isomorphic tasks. This indicates that adding retrieval prompts to worked example study does not improve, nor hamper, the performance on problem-solving tasks, compared to regular worked example study. Moreover, this study was conducted with pre-vocational education students and findings suggest that our understanding of how these students learn may still be incomplete. This study is a first step in the investigation of combining worked examples with (stepwise) retrieval practice. Unexpectedly, it also provides insights for future research into the learning process of pre-vocational education students during worked example study. Multiple directions for future research are identified.

Keywords: problem-solving skills, retrieval practice, worked examples, prompting

Fostering the learning of problem-solving skills: Prompting stepwise retrieval during worked examples

Problem-solving tasks are a common part of many school subjects, especially in the STEM domain (science, technology, engineering, and mathematics) (Van Gog et al., 2019). During problem-solving tasks, students are presented with a (well-structured) problem and must follow a procedure that specifies the necessary actions to get from an initial state to a described goal state (Jonassen, 2000). Problem-solving tasks require the use of problem-solving skills. Problem-solving skills consist of both procedural knowledge, understanding what actions to perform and how to execute them, and conceptual knowledge, comprehending why to perform those actions (van Gog et al., 2019). To acquire problem-solving skills, unguided practice of problem-solving tasks is a widely adopted instructional approach in education (Yeo & Fasio, 2018). However, this instructional approach appears to be ineffective (Sweller et al., 2011) as it creates difficulty for novice learners, as they have not yet learned effective specific procedures to encounter the presented problem (Van Gog et al., 2019). Numerous studies have demonstrated that using worked examples, that demonstrate the solution steps applied to an example task, is more effective (resulting in enhanced learning outcomes) and more efficient (requiring less effort and/or time investment) for the acquisition of problem-solving skills compared to relying exclusively on unguided problem-solving practice (Paas, 1992; Sweller & Cooper, 1985; Sweller et al., 1998; Van Gog et al., 2011).

Where worked examples are known for effectively learning procedural knowledge (such as problem-solving skills), retrieval-based learning is known for effectively learning conceptual knowledge (Yeo & Fasio, 2018), and improving the long-term retention of studied information (Roediger & Butler, 2011). Retrieval practice, during which learners recall information from memory, has been extensively researched and the testing effect, caused by retrieval practice, has proven to be quite vigorous (Karpicke, 2017; Van Gog & Kester, 2012). However, when researching retrieval-based learning aiming at the development of problem-solving skills, the testing effect seems to lack applicability in that context (Adesope et al., 2017; Van Gog & Kester, 2012; Van Gog et al., 2015). Nevertheless, given the significant impact of both example-based learning and retrieval practice, it is suggested to further explore how example-based learning and retrieval practice can complement each other (Roelle, 2023). This study will therefore test whether it is effective to repeatedly stimulate students to retrieve prior knowledge while studying worked examples.

Instructional Methods

Since unguided problem-solving practice appears not to be effective when acquiring problem-solving skills (Sweller et al., 2011), research has been conducted into the use of different instructional methods. Two prominent instructional methods are example-based learning and retrieval-based learning (Adesope et al., 2017; Atkinson et al., 2000; Renkl, 2014; Rowland, 2014), each utilizing distinct cognitive processes for knowledge acquisition and retention (Karpicke, 2017; Van Gog et al., 2019).

Example-Based Learning

Extensive experimental research conducted in laboratory, educational, or professional training settings have demonstrated the effectiveness of example-based learning when acquiring problem-solving skills (Atkinson et al., 2000; Paas et al., 1998; Renkl, 2014, 2017; Sweller et al., 2011; van Gog & Rummel, 2010). Example-based learning can take various forms, including worked examples and modeling examples. Worked examples are learning tasks wherein an initial state, a desired goal state, and solution steps are explicitly presented. Additionally, it emphasizes the problem-solving process essential for achieving the goal (van Merriënboer & Kirschner, 2018). The effectiveness of worked examples is known as the *worked example effect* (Sweller & Cooper, 1985; Sweller et al., 1998). During modeling examples, students witness a live- or video demonstration of another person solving the problem (van Gog et al., 2019). Worked examples and modeling examples can be combined in a ‘hybrid’ form, in which written steps appear step-by-step during a video (Van Gog et al., 2019)[1]. Research shows that both worked examples and modeling examples are equally effective for acquiring problem-solving skills (Hoogerheide et al., 2014).

According to the Cognitive Load Theory, example-based learning is effective as it decreases cognitive load due to the presentation of the necessary problem-solving strategy (e.g., during worked

[1] In this study, we employ worked examples in a hybrid form (where the written steps appear step-by-step in a video).

examples). Learners retain more cognitive capacity to create cognitive schemas about how to solve the problem (Kalyuga et al., 2010; Sweller, 1988; Sweller et al., 1998). In contrast, when practicing unguided problem-solving novices often use ineffective problem-solving strategies, which results in high cognitive load and less cognitive capacity to create the necessary cognitive schemas (Van Gog et al., 2019).

Nevertheless, the benefits of worked example practice appear to be moderated by the retention interval. The retention interval hypothesis describes the observation that the worked example effect is consistently more noticeable in immediate problem-solving tests compared to delayed tests (Leahy et al., 2015; van Gog & Kester, 2012; van Gog et al., 2015; van Gog et al., 2006). Retrieval practice is often used for long-term retention of information, which therefore might support performance on delayed testing (Kapricke, 2017).

Retrieval Practice

Retrieval-based learning is effective for learning conceptual knowledge (Yeo & Fasio, 2018) and improving long-term retention of studied information (Roediger & Butler, 2011). Retrieval practice refers to learners actively retrieving information from memory and has various formats, such as quizzing or self-testing (Karpicke, 2017). The act of retrieving information activates pre-existing cognitive schemas and updates them with new information (Hinze et al., 2013). Retrieving information requires mental effort and is considered a desirable difficulty (Bjork & Bjork, 1992). The retrieval process and required mental effort strengthen the memory of the information (Dudai, 2012; Dudai et al., 2015). Ultimately it improves long-term retention more so than restudying (Roediger & Butler, 2011; Roediger & Karpicke, 2006a). This effect of improved long-term retention is also referred to as the *testing effect*. The testing effect has been extensively researched with educational materials such as word lists (Wheeler et al., 2003), and multimedia lessons (Johnson & Mayer, 2009), and has proven to be quite vigorous (Karpicke, 2017; Van Gog & Kester, 2012). A meta-analysis showed consistent and robust testing effects across studies, with moderate effect sizes (Adesope et al., 2017).

However, when researching learning tasks aiming at the development of problem-solving skills, the testing effect appears not to apply (Huang et al., 2023; Leahy et al., 2015; Van Gog &

Kester, 2012; Van Gog & Sweller, 2015; Van Gog et al., 2015). In the study by Van Gog and colleagues (2015), they compared worked example study and retrieval practice. In the retrieval condition, participants were asked to study a worked example and recall the worked example in the next round. In the worked example condition, participants did not retrieve the information but restudied worked examples. The retrieval condition did not outperform the worked example condition on the problem-solving tasks, which indicated that there was no testing effect. A possible explanation for this result could be that participants struggled with retrieval because they received no feedback on the retrieved information. However, Huang and colleagues (2023) found that providing feedback on the retrieved information did not improve learning problem-solving skills. In their study, participants received four worked example – problem pairs. Feedback was given (in the form of presenting the problem-solving steps) either after completing all four pairs or after the completion of two pairs. They found no effect of retrieval practice (with feedback) on learning problem-solving skills.

Note though, that in both Van Gog and colleagues (2015), and Huang and colleagues (2023), participants needed to retrieve the full worked example. This might be cognitively demanding, as the worked examples consist of several steps and successful retrieval of later steps depends on successful retrieval of earlier steps. Hence, to make the retrieval less demanding, a solution might be to have students engage in stepwise retrieval. Indeed, research in vocabulary learning showed that instructing students to engage in retrieval *before* presenting the answer options improved students' retrieval and subsequent learning (Van den Broek et al., 2023). This method of using a stepwise display of information to foster retrieval could also be applied to worked example practice, to scaffold the retrieval of the worked example. Moreover, stepwise display provides immediate feedback on the retrieved information, in contrast to the study of Huang and colleagues (2023), where feedback was given only after two or more retrieval actions.

The Present Study

The present study aimed to investigate the effectiveness of repeatedly stimulating students to retrieve prior knowledge while studying worked examples. By doing so, this study answered the question of whether the combination of worked example study and retrieval practice would yield an effective instructional method to enhance student learning of problem-solving skills (Roelle et al., 2023). To this end, pre-vocational secondary education students were assigned to a stepwise retrieval condition or a worked example (control) condition. The stepwise retrieval condition featured prompts during worked example study, which prompted students to retrieve the next step of the procedure. The control condition received worked examples without prompts for retrieval. The following research questions, sub-questions, and hypotheses were addressed:

> RQ: Does prompting stepwise retrieval during worked example study lead to higher performance on problem-solving tasks compared to only studying worked examples?
>
> > SQ1: Does prompting stepwise retrieval during worked example study lead to higher performance on scaffolded problem-solving tasks compared to only studying worked examples?
> >
> > SQ2: Does prompting stepwise retrieval during worked examples lead to higher performance on isomorphic problem-solving tasks compared to only studying worked examples?

Based on the results of Van den Broek and colleagues (2023), it is expected that the prompted retrieval condition will outperform the control condition in both the isomorphic and transfer tasks. Moreover, because scaffolded tasks rely less on retrieval (Van Gog & Kester, 2012), it is expected that the difference in performance, between the retrieval and control condition, on the scaffolded tasks will be smaller, yet still favor the retrieval condition. Furthermore, retrieval practice is allegedly more effortful than the restudy of worked examples (Bjork, 1994), potentially resulting in the retrieval condition experiencing higher invested mental effort than the control condition. To explore this, the invested mental effort of both conditions was measured.

Method

During the data collection, it became apparent that the materials, instruments, and setup of the experiment did not fully align with the target group. Therefore, halfway through the data collection, the decision was made to adjust the materials and procedure, resulting in two waves of data collection.

Research Design

This study employed a between-subjects design with condition (retrieval versus control) as the between-subjects factor and post-test performance as the dependent variable. The post-test consists of scaffolded tasks and isomorphic tasks.

Participants

An a priori power analysis indicated a total sample size of 102 participants (51 in each condition) to achieve a power of 0.80 at a medium effect size of Cohen's $d = 0.5$ for an independent *t*-test.

Participants were recruited by approaching schools within the personal network. Participation was unpaid. 114 Dutch pre-vocational secondary education (*vmbo*) students from the same school participated in this study (wave 1: n = 54; wave 2: n = 60). 31 Participants needed to be excluded due to various reasons, including, failure to report their random ID (see below; wave 1: $n = 3$), no consent for data use (wave 1: $n = 7$), severe disruptions during data collection (wave 1: $n = 16$), and failure to finish in time (wave 2: $n = 5$). This left 28 participants in wave 1 and 55 participants in wave 2. In wave 1 the age ranged from 13 to 15 years ($M = 13.7$, $SD = .670$). 18 males (64.3 %), and 10 females (35.7 %). In wave 2 the age ranged from 13 to 16 years ($M = 13.8$, $SD = .696$). 28 males (50.9 %), 26 females (47.3 %), and 1 that identified as other than male or female (1.8 %). Participants were randomly assigned to either the prompted stepwise retrieval group (wave 1: $n = 17$; wave 2: $n = 30$) or the control group (wave 1: $n = 11$; wave 2: $n = 25$).

Ethical Approval

The study was approved by the Ethical Review Board of the Faculty of Social and Behavioral Sciences (FETC), Utrecht University, file number 24-0480. Two weeks before the experiment, passive parental consent was obtained. During wave 1, student consent about data usage was obtained at the start of the experiment. Since the materials were part of the curriculum, all students had to participate irrespective of consent. However, this procedure led to significant confusion among students and caused significant disruption. After consulting with the FETC, it was decided to omit the active consent of students in wave 2. A new protocol was approved, file number 24-0277.

All digital data was stored on Yoda, a data management service that helps to securely manage research data and is recommended by Utrecht University (Utrecht University, n.d.). The paper post-tests were scanned and stored along with the rest of the data. The paper versions will be destroyed four weeks after approval of the master thesis.

Materials and Instruments

The materials focused on learning to apply the balance method (*balans methode)* employed for solving mathematical equations.

Prior Knowledge Test

Because the material concerns curriculum content that has yet to be covered, it is assumed that students possess no prior knowledge about the subject, except for general mathematical knowledge required to comprehend the new content. To verify this assumption, the prior knowledge test comprised 2 questions about the basic understanding of the addition of variables within a formula (e.g., *please condense the following formula to the shortest possible version; 33d + 12d = g)*. To check whether the prior knowledge was accurate, the answers on the pre-test were categorized into the following categories: correct answer, wrong answer, and wrong answer due to mistakes with negative numbers.

During wave 1, confusion and frustration among the students caused by the pre-test were observed, despite clear explanations regarding the purpose of asking questions before the instruction (e.g., *to activate your prior knowledge, we now ask you to fill in 2 questions)*. To reduce disruptions and optimize the procedure, it was decided not to conduct a pre-test during wave 2.

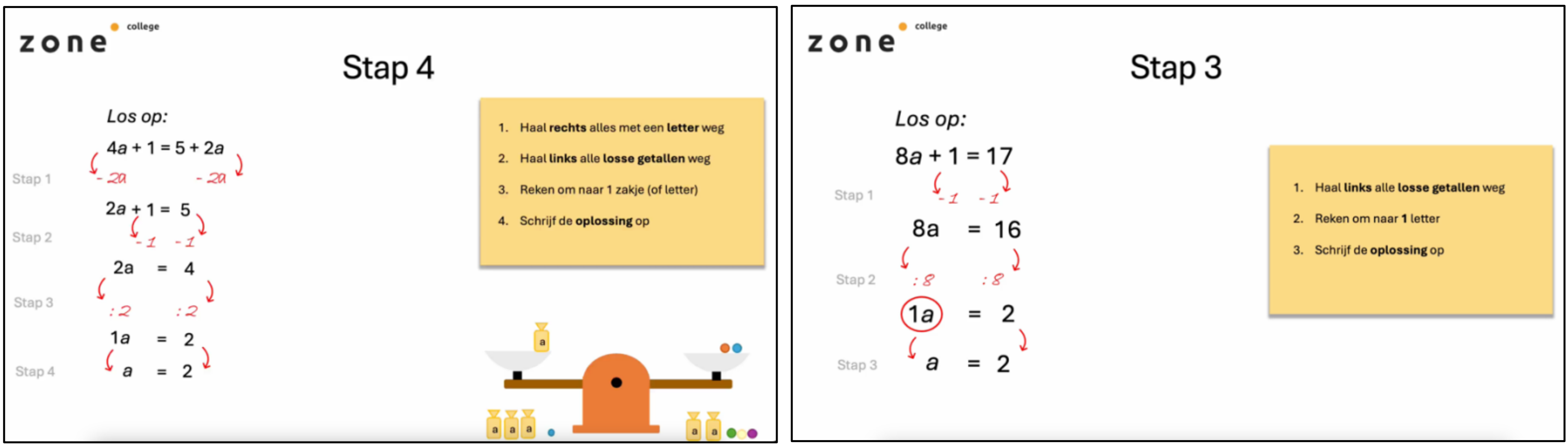
Worked Examples

For the acquisition of knowledge about the balance method, three worked examples were created based on the method *Modern Mathematics* (*Moderne Wiskunde*), in collaboration with a didactical content expert. A hybrid form of worked examples was utilized, wherein a video demonstrated step-by-step explanations, provided by a female voiceover. Because research shows that retrieval practice is effective when information is retrieved multiple times (Karpicke, 2017), three worked examples are used. For wave 1, the total video duration was 10.40 minutes (video 1: 6.02 minutes; video 2: 2.26 minutes; video 3: 2.12 minutes). For wave 2, the length of the videos was reduced as students in wave 1 indicated that the length of the videos was too long, and students showed a loss of concentration during the videos. Video duration, therefore, was reduced to 6.47 minutes (video 1: 3.24; video 2: 2.05; video 3: 1.47).

During the original worked examples, instruction was given about how to solve an equation within 4 steps. A preliminary check of the performance revealed that the tasks were overly challenging. To simplify the tasks, it was decided to reduce the number of steps in the worked examples from 4 to 3. Consequently, adjustments were made to the instruction worked example videos (e.g., the number of steps on the post-it and the removal of the scale). Figure 1 shows an impression of how the worked examples looked and how they differ between waves 1 and 2.

Figure 1

Screenshots of the Worked Example Videos (Left: Wave 1; Right: Wave 2)



Stap 4

Los op:

$4a + 1 = 5 + 2a$

Stap 1: $-2a \quad -2a$

$2a + 1 = 5$

Stap 2: $-1 \quad -1$

$2a = 4$

Stap 3: $:2 \quad :2$

$1a = 2$

Stap 4: $a = 2$

1. Haal **rechts** alles met een **letter** weg
2. Haal **links** alle **losse getallen** weg
3. Reken om naar 1 zakje (of letter)
4. Schrijf de **oplossing** op

Stap 3

Los op:

$8a + 1 = 17$

Stap 1: $-1 \quad -1$

$8a = 16$

Stap 2: $:8 \quad :8$

$\boxed{1a} = 2$

Stap 3: $a = 2$

1. Haal **links** alle **losse getallen** weg
2. Reken om naar **1** letter
3. Schrijf de **oplossing** op

Prompted Stepwise Retrieval

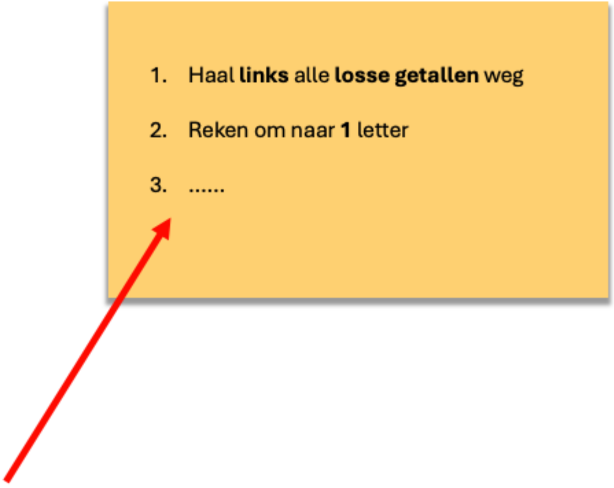
The prompted stepwise retrieval condition watched the same worked examples as the control condition. For the retrieval condition, there were pauses during the second and third worked examples. The pauses were right before each next step of the procedure became visible. A prompt appeared asking students '*What will be the next step?*', after which they wrote down their answer in Qualtrics. The prompts aimed to encourage students to retrieve the procedural steps of the balance method (e.g., *step 1: remove all individual numbers on the left part of the equation*). Prompts appeared in each worked example 4 times (a total of 8 prompts).

A preliminary check showed that students struggled with retrieval of the steps. It also was observed that students perceived the prompts as confusing. They often inquired whether they were required to solve a part of the equation. The didactical expert indicated that students find it challenging to recall the procedural steps and more often recall the executive step (the actual computation e.g., - 8 on both sides). To direct students more towards retrieving the procedural step, images of the section of the video where the procedural steps were mentioned were added. An example of this can be seen in Figure 2. The prompt was adjusted to '*Think about what the next step will be. Fill in what should be written on the dotted line'*.

To check whether the retrieved information in the retrieval condition was accurate, participants wrote down their retrieved answers during the pauses. The answers were categorized into the following categories: wrong answer, 'I don't know' or '?', identification of the executive step (e.g., -*8*), the new equation after executing the step (e.g., *1a* = *2*), the right procedural step (e.g., *convert to a single letter*), and the previous step.

Figure 2

Example of the Image Shown to Students during the Worked Example Retrieval (Wave 2)



1. Haal **links** alle **losse getallen** weg
2. Reken om naar **1** letter
3. ......

Mental Effort Rating

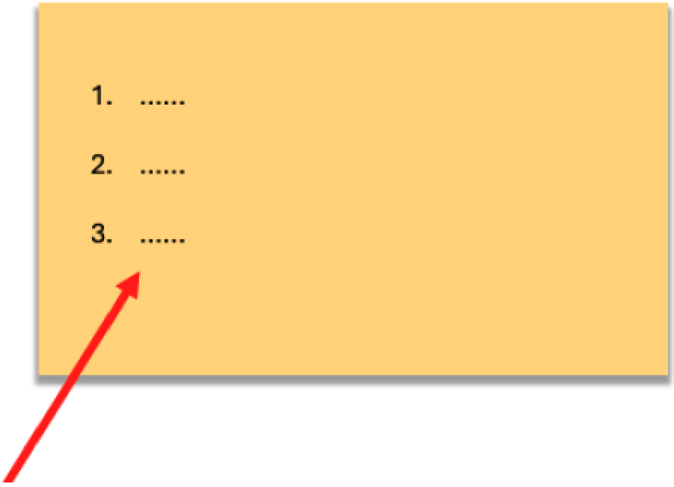
The invested mental effort was assessed using Paas's (1992) nine-point subjective rating scale, ranging from 1 (very, very low effort) to 9 (very, very high effort).

Retrieval Check

During wave 2, a retrieval check was conducted to investigate whether both conditions differed in retrieval performance (e.g., retrieval success). The check consisted of three questions and asked students to retrieve the three steps of the balance methods shown in the worked examples (e.g., *think back to the instruction. What was step 3 of the balance method? Write it down)*. The question was supported by an image of the post-it on which the procedural steps appeared during the instruction video to clarify to the students that we wanted them to name the procedural step instead of the executive step. Figure 3 shows an example of an image shown during the retrieval check.

Figure 3

Example of the Image Shown during the Retrieval Check Questions



Post-test

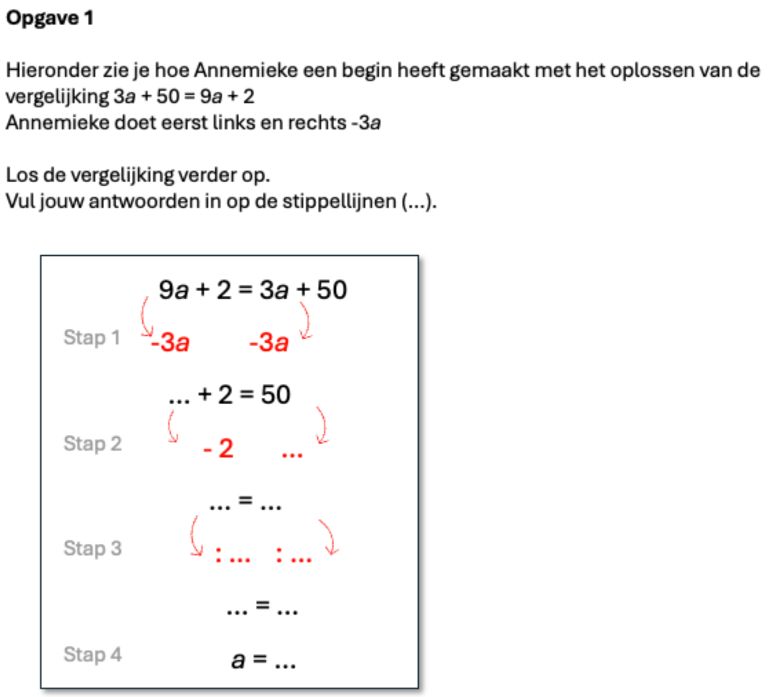
To align with the target group, paper-based post-tests were administered. Given the audience's lack of prior experience with computer-based mathematical tasks and to prevent potential increases in difficulty, tasks were presented in the traditional handwritten format familiar to them.

The post-test contained two scaffolded problem-solving tasks and three isomorphic tasks. In the scaffolded tasks used for wave 1, the first step was given, and students had to execute step 1 and retrieve and execute the other three steps. The three isomorphic tasks had a similar problem-solving procedure as the worked examples, but different surface features (e.g., the values and variables of the equation). The exact values of the example procedure no longer apply to the new task, and therefore participants must perform the necessary calculations themselves (Van Gog & Kester, 2012). Due to the restricted time available, and the characteristics of the target group, no transfer tasks were included in the experiment to avoid frustration and cognitive overload among the students. Internal consistency of the post-test was measured using Cronbach's alpha and was excellent for wave 1 ($\alpha = .96$) and high for wave 2 ($\alpha = .86$).

A preliminary check indicated that the level of the tasks was too high for the students and the high levels of frustration observed during the experiment supported this indication. The following adjustments were made for the materials of wave 2. For all tasks (scaffolded and isomorphic), equations were simplified to have no decimal or negative outcomes. Moreover, the level of scaffolding was increased on the scaffolded tasks. Previously, the scaffolded tasks presented step 1 and students had to execute step 1 and retrieve and execute (part of) steps 2 and 3. In the adjusted materials the highest scaffolded task presented steps 1 and 2 leaving students to execute step 2 and retrieve and execute step 3. Figure 4 shows the scaffolded tasks of wave 1 and Figure 5 shows the adjusted scaffolded tasks used for wave 2.

Figure 4

Scaffolded Tasks for Wave 1



Opgave 1

Hieronder zie je hoe Annemieke een begin heeft gemaakt met het oplossen van de vergelijking $3a + 50 = 9a + 2$
Annemieke doet eerst links en rechts $-3a$

Los de vergelijking verder op.
Vul jouw antwoorden in op de stippellijnen (...).

$9a + 2 = 3a + 50$

Stap 1 $-3a$ $-3a$

$... + 2 = 50$

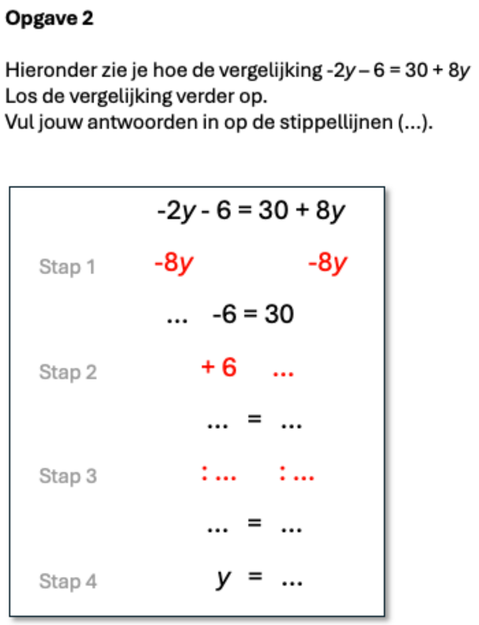
Stap 2 $- 2$ ...

... = ...

Stap 3 : ... : ...

... = ...

Stap 4 a = ...



Opgave 2

Hieronder zie je hoe de vergelijking $-2y – 6 = 30 + 8y$
Los de vergelijking verder op.
Vul jouw antwoorden in op de stippellijnen (...).

$-2y - 6 = 30 + 8y$

Stap 1 $-8y$ $-8y$

... $-6 = 30$

Stap 2 $+ 6$ ...

... = ...

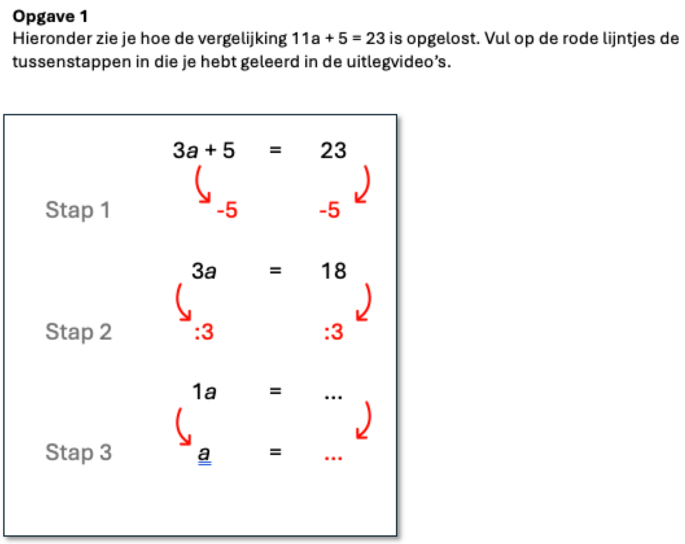
Stap 3 : ... : ...

... = ...

Stap 4 y = ...

Figure 5

Scaffolded Tasks for Wave 2 (More Heavily Scaffolded)



Opgave 1
Hieronder zie je hoe de vergelijking $11a + 5 = 23$ is opgelost. Vul op de rode lijntjes de tussenstappen in die je hebt geleerd in de uitlegvideo's.

$3a + 5$ = 23

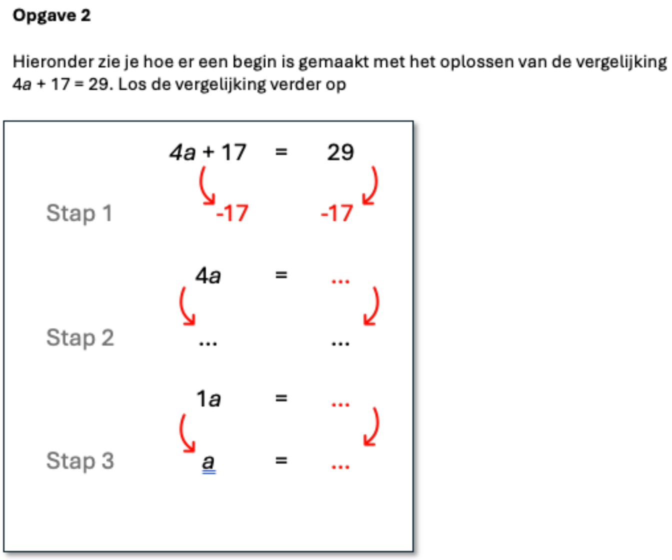
Stap 1 -5 -5

$3a$ = 18

Stap 2 :3 :3

$1a$ = ...

Stap 3 a = ...



Opgave 2

Hieronder zie je hoe er een begin is gemaakt met het oplossen van de vergelijking $4a + 17 = 29$. Los de vergelijking verder op

$4a + 17$ = 29

Stap 1 -17 -17

$4a$ = ...

Stap 2 ... ...

$1a$ = ...

Stap 3 a = ...

Procedure

Two weeks before the experiment parental consent was obtained, and students were informed about the experiment. The study was conducted during the regular classroom hours of the participants (50 min.). The tables and chairs in the classroom were arranged in an examination layout to prevent students from collaborating during the experiment.

At the start of the experiment, students were orally briefed on the research objectives, and it was communicated to the students that participation or outcomes of this experiment would have no academic consequences (e.g., impact on their grades). Moreover, students were told that asking questions during the experiment about the mathematical content was not allowed. Following this, students were able to ask questions about the experiment and clarify any misconceptions. Subsequently, students independently accessed Qualtrics on their laptops. Upon accessing Qualtrics, the verbally conveyed information was reiterated in written form. After which informed consent was taken. Subsequently, participants filled in demographic information and a prior knowledge test, after which they entered the worked example phase. Students were randomly assigned to one of the two conditions. For all worked examples, written instructions were provided to the students, stating that they were not allowed to pause the videos or watch them at an increased speed and were only allowed to watch the videos once. Qualtrics was set up so that the 'next page' button would only appear once the entire video had been viewed. All conditions viewed the first worked example. The control condition then viewed the two remaining worked examples after which they entered the invested mental effort rating. The retrieval practice condition started with the adjusted worked examples after the first worked example. Each time a retrieval prompt was given during the worked examples, participants needed to write down their retrieved answers in the text box on the screen. After the two worked examples with stepwise retrieval, they also started the invested mental effort rating. After the mental effort rating, all students progressed to the post-test phase. They were provided with a random ID number via Qualtrics, which they recorded on their paper answer sheet, enabling later linkage between the Qualtrics data and responses. The post-test was completed on paper. Despite instructions to the contrary, many students asked questions during the post-test. When it was indicated that questions could not be answered, high levels of frustration often ensued. To address this, the researcher instead adopted a coaching approach, providing the following instruction in response to a question about the content '*Let's calmly go over the questions together (researcher reads the question aloud). What do you remember about the videos? Write that down, you can do it!'*. This often resulted in students completing the assignment independently. At the end of the experiment, participants were thanked for their participation.

As described, it became apparent during the data collection that the materials, instruments, and setup of the experiment did not fully align with the target group. Therefore, the following adjustments were made in the procedure for wave 2. The informed consent and prior knowledge test were discontinued, and a retrieval check for all conditions was added (after the mental effort rating). Figures 6 and 7 show a schematic overview of the procedure for waves 1 and 2. Red indicates that an element was only applicable during wave 1 (and removed during wave 2), and blue indicates an element that was only applicable during wave 2.

Figure 6

Schematic Overview of the Procedure Wave 1

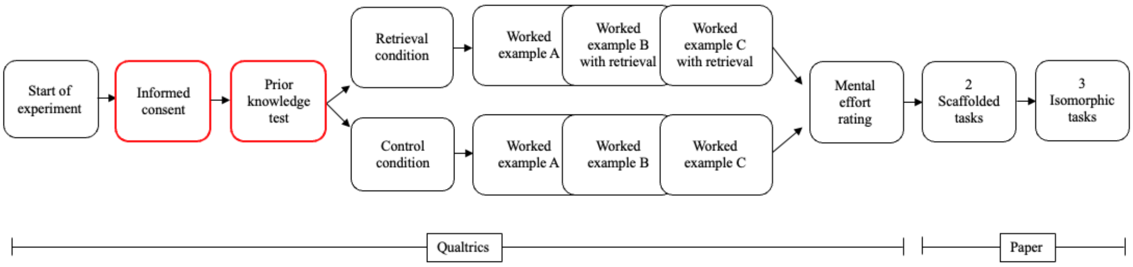
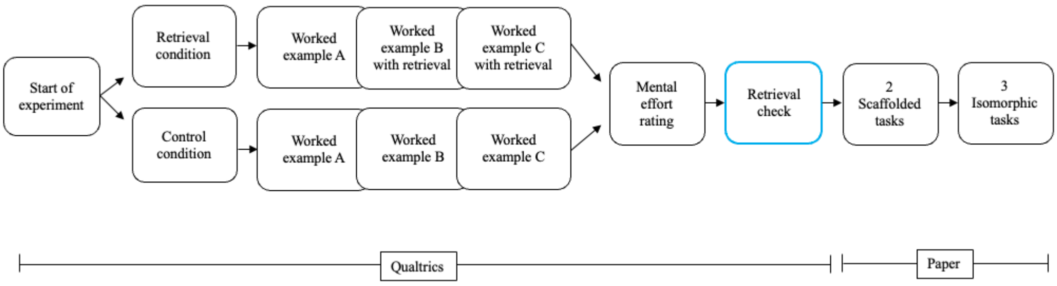


Figure 7

Schematic Overview of the Procedure Wave 2



Data Analysis

Scoring Post-test

Problem-solving accuracy on the post-test tasks was scored by assigning 1 point for each correctly retrieved and executed step in the problem-solving process. Depending on the nature of the task (e.g., scaffolded, or isomorphic) a maximum of 4 points per task was assigned. For the 2 scaffolded tasks and the 3 isomorphic tasks, a sum score for each participant was calculated resulting in a scaffolded score and an isomorphic score. To ensure the reliability of the scores, two independent raters (the first author and another master's student) scored 20% of the participants and Cohen's kappa was calculated, resulting in a moderate agreement for wave 1: $\kappa = .447$, $p = .004$, and a substantial agreement for wave 2: $\kappa = .686$, $p <.001$.

Data Analysis

SPSS Statistics 29 will be used to conduct all analyses. In case of normally distributed data, independent *t*-tests will be conducted with condition as independent variable and post-test performance as dependent variable. In case of violation of assumptions, non-parametric Mann-Whitney U tests will be used instead.

Results

Table 1 shows the descriptive statistics of the scaffolded and isomorphic items per condition for waves 1 and 2. The variables 'test performance' on the scaffolded items and isomorphic items were not normally distributed for both conditions and both waves as evidenced by Shapiro-Wilk tests (all *p*'s ≤ .007). Hence the non-normal distribution, the Mann-Whitney U test was used as a nonparametric alternative. As an effect size for the Mann-Whitney U Test *r* was used (Clark-Carter, 2009). For all analyses, a significance level of 0.05 was used.

Table 1

Mean, SD Median, and Range of the Post-test Scores) on the Scaffolded and Isomorphic Tasks per Condition for Wave 1 and 2.

	Control				Retrieval			
	M	*SD*	*Med*	Range	*M*	*SD*	*Med*	*Range*
Wave 1								
Pretest (0-2)	1.0	0.7	1.5	0.0 - 1.5	0.5	0.7	0.0	0.0 - 1.5
Scaffolded items (0-8)	2.0	2.5	1.0	0 - 7	1.8	2.6	0.0	0 - 7
Isomorphic items (0-12)	3.4	4.2	1.0	0 - 10	1.8	3.8	0.0	0 - 12
Mental effort (1-9)	5.5	2.1	6.0	2 - 9	6.8	2.6	8.0	1 - 9
Wave 2								
Scaffolded items (0-5)	4.6	0.8	5.0	2 - 5	4.5	1.3	5.0	0 - 5
Isomorphic items (0-9)	6.0	2.9	7.0	0 - 9	5.9	2.2	6.0	0 - 8
Mental effort (1-9)	4.9	2.0	5.0	1 – 9	5.2	2.1	5.0	1 - 9

Note. Wave 1: Control condition: (n = 11); Retrieval condition (n = 17). Wave 2: Control condition *(n* = 25); Retrieval Condition (*n* = 30).

SQ1: Performance on Scaffolded Items

A Mann-Whitney U test indicated that there were no significant differences between the performance on the scaffolded items for the control condition and retrieval condition in wave 1 (*U* = 83.50, *z* = -.50, *n* = 28, *p* = .618, *r* = .09). This test was repeated for wave 2, and again, no significant differences were found between the performance on the scaffolded items for the control condition and retrieval condition (*U* = 365.50, *z* = -.24, *n* = 55, *p* = .811, *r* = .24). These findings suggest that adding retrieval prompts to worked example study does not improve, nor hamper, the performance on scaffolded problem-solving tasks compared to regular worked example study.

SQ2: Performance on Isomorphic Items

A Mann-Whitney U test indicated that there were no significant differences between the performance on the isomorphic items for the control condition and retrieval condition in wave 1 (*U* = 69.50, *z* = -1.28, *n* = 28, *p* = .200, *r* = .03). This test was repeated for wave 2, and again, no significant differences were found between the performance on the isomorphic items for the control condition and retrieval condition (*U* = 319.50, *z* = -.96, *n* = 55, *p* = .338, *r* = .13). These findings suggest that adding retrieval prompts to worked example study does not improve, nor hamper, the performance on isomorphic problem-solving tasks compared to regular worked example study.

Mental Effort

An independent samples *t*-test was used to investigate differences in the average mental effort reported by the participants of wave 1 between the control condition and the retrieval condition. The *t*-test did not reveal a significant difference, *t*(26) = -1.39, *p* = .176. This test was repeated for wave 2, and again, revealed no significant difference in average mental effort between the control condition and the retrieval condition *t*(53) = -.45, *p* = .654. These findings suggest that adding retrieval prompts to worked examples study does not increase nor decrease the mental effort during worked example study.

Additional Explorative Analyses

Pre-test

During wave 1 a pre-test was conducted. Table 2 shows the frequencies of the answer options per condition. The frequencies show that on Q1 50% of the answers were right, and 50% were wrong. During Q2, no students answered the question correctly (right answer 0%). Moreover, 39.3% of all incorrect answers to Q2 resulted from students making errors with negative numbers.

To test whether both conditions in wave 1 differed significantly in pre-knowledge an independent samples *t*-test was used. The *t*-test was statistically non-significant *t*(26) = 1.62, *p* = .827.

this indicates that there were no significant differences in pre-knowledge between the conditions in wave 1.

Table 2

Frequencies of Answer Options on Q1 and Q2 Pre-test Wave 1

	Control (*n* = 11)		Retrieval (*n* = 17)		Total (*n* = 28)	
	Count	% within condition	Count	% within condition	Count	% within condition
Q1						
Wrong answer	4	36.4	10	58.8	14	50.0
Right answer	7	63.6	7	41.2	14	50.0
Q2						
Wrong answer	4	36.4	13	76.5	17	60.7
Wrong answer due to + - mistake	7	63.6	4	23.5	11	39.3
Right answer	0	0	0	0	0	0

Retrieval Performance

To evaluate the quality of the retrieval, the retrieval answers from the retrieval condition were categorized. Table 3 shows the frequencies of each category for all 6 retrieval questions.

Table 3

Frequency and Percentages of Answers of Retrieval Condition on Retrieval Questions

	Wave 1		Wave 2	
	Total frequency	%	Total frequency	%
Wrong answer	58	42.65	65	36.11
‘I don’t know’ or ‘?’	43	31.62	19	10.56
Executive step	7	5.15	37	20.56
New equation	12	8.82	4	2.22
Right answer	15	11.03	49	27.22
Previous step	1	0.74	6	3.33
Total	136	100	180	100

Post-test Retrieval

During wave 2 a post-test retrieval check was conducted. To evaluate the retrieval performance of both conditions on the post-test, the answers were categorized. Table 4 shows the frequencies of each answer category for all 3 retrieval questions split for conditions.

Table 4

Frequency and Percentages of Answers on Post-test Retrieval Check

	Control condition		Retrieval condition	
	Total frequency	%	Total frequency	%
Wrong answer	33	44.00	65	36.11
‘I don’t know’ or ‘?’	12	31.62	19	10.56
Executive step	3	5.15	37	20.56
New equation	0	8.82	4	2.22
Right answer	15	11.03	49	27.22
Previous step	27	0.74	6	3.33
Total	75	100	180	100

Discussion

This study investigated the effectiveness of repeatedly prompting students to retrieve prior knowledge while studying worked examples. It was studied whether prompting stepwise retrieval during worked example study leads to higher performance on scaffolded and isomorphic problem-solving tasks compared to only studying worked examples.

SQ1: Differences in Performance on Scaffolded Items

The results of this study indicate that there is no significant difference in problem-solving accuracy between the different conditions on the scaffolded items for both waves. This indicates that prompting stepwise retrieval during worked example study does not improve (nor hamper) the performance on scaffolded problem-solving tasks compared to regular worked example study.

SQ2: Differences in Performance on Isomorphic Items

Results also indicate that there is no significant difference in problem-solving accuracy between the different conditions on the isomorphic items for both waves. This indicates that prompting stepwise retrieval during worked example study does not improve (nor hamper) the performance on isomorphic problem-solving tasks compared to regular worked example study.

RQ: Differences in Performance

It was expected that the retrieval condition would outperform the control condition on both the scaffolded and isomorphic items, which would mean that retrieval practice enhances the learning of problem-solving skills during worked example study. By segmenting the retrieval into steps instead of retrieving the entire worked example, we aimed to make the retrieval process less demanding. Although retrieval without feedback can be effective (Roediger & Butler, 2011), based on Van den Broek and colleagues (2023), we expected that the immediate feedback students would receive on their attempt to retrieve the procedural step would enhance performance. Nevertheless, as mentioned above, no effects of stepwise retrieval practice, positive or negative, were found. This indicates that adding retrieval prompts to worked example study does not improve nor hamper the performance on problem-solving tasks compared to when no prompts are included (regular worked example study).

Several possible explanations could give insight into the absence of effect. On the scaffolded tasks, the scaffolding reduces the intrinsic load of the task by showing part of the worked example; students no longer must retrieve and execute all steps but only one or two. Therefore, scaffolded tasks rely less on retrieval. This could explain why the retrieval condition did not outperform the control condition on the scaffolded tasks. In contrast to scaffolded tasks, isomorphic tasks require more retrieval (Van Gog & Kester, 2012). However, the isomorphic tasks followed the scaffolded tasks. It may be that the scaffolded task (which shows part of the procedure) functioned as a retrieval practice. This might have helped the control condition to refresh their memory of the procedure due to this additional retrieval opportunity. Indeed, it was observed during the experiment that (some) students used the scaffolded task as a guide for the isomorphic tasks.

Moreover, an influential factor during retrieval practice is the retention interval (Agarwal et al., 2017). The time between initial retrieval practice and the final test impacts retention. Restudying improves short-term retention, whereas retrieval practice improves long-term retention (Karpicke & Roediger, 2007; Roediger & Karpicke, 2006a). During this experiment, students were tested immediately after studying the worked examples. Therefore, it might be that the control condition (who only restudied the worked examples) profited more from the restudy than the retrieval condition benefited from the retrieval practice. Future research could explore how delayed testing (and therefore more emphasis on long-term retention) would impact the performance of both conditions.

Furthermore, the retrieval success during this experiment was low. Retrieval practice appears to be most effective when retrieval success is high (Karpicke, 2017). However, results show that only 11.03% of the retrieval answers in wave 1, and 27.22% in wave 2, were successful. Most retrieval answers were incorrect (wave 1: 42.65%; wave 2: 36.11%). Additionally, a large amount of retrieval opportunities did not take place as students did not make a retrieval attempt (wave 1: 31.62%; wave 2: 10.56%). Therefore, the lack of significant difference in problem-solving accuracy between both conditions on the tasks might be due to the low retrieval success rate.

Another alternative explanation for why there was no beneficial effect found of stepwise retrieval when learning from worked examples is that there was a mismatch between what learners had to retrieve during worked example study (i.e., retrieval of step) and what learners had to do during the post-test (i.e., apply the steps). Solving isomorphic problem tasks requires more than just recall; they also involve answer construction (Van Gog & Kester, 2012). It might be that the retrieval practice supported only the retrieval of the procedure but not the answer construction, resulting in the retrieval condition benefiting less from the retrieval practice. Moreover, to achieve effective answer construction, students should focus on the procedure during the worked examples so that they can build a cognitive schema, rather than on the exact values, as these do not apply to the isomorphic test problems (Van Gog & Kester, 2012). It appears that some students in the retrieval condition struggled to focus on the procedure and instead focused on the exact values in the worked examples (i.e., when asked to retrieve the steps during the post-retrieval test 20.56% of wave 2 indicated the executive steps instead of the procedural steps). Future research on how pre-vocational education students approach learning from worked examples is therefore suggested.

Mental Effort

In both waves, there were no significant differences in mental effort between the conditions. This is surprising since retrieval practice (retrieval condition) is assumed to require more mental effort than just restudying (control condition) (Bjork, 1994). Possibly, students did not invest much effort when having to retrieve information in the stepwise retrieval condition, which also corresponds to their low retrieval success. Another alternative explanation for not finding any differences in mental effort between the control condition and stepwise retrieval condition might be that mental effort was rated after completion of all worked examples instead of after each worked example. Indeed, research has indicated that when multiple worked examples are provided and mental effort is measured after the last example, the mental rating is largely influenced by the final worked example (Schmeck et al., 2015). Future research therefore might investigate whether, and if so how, mental effort ratings change when measuring in between worked examples.

Moreover, it might be that students struggled to discriminate between mental effort and difficulty. Mental effort and difficulty appraisal are closely related (Hoch et al., 2023) and it might be that students could not discriminate between these two based on the one-scale question of Paas (1992) (see Klepsch et al., 2017). Therefore, future research could use a multiple-item scale to measure mental effort to get a clear idea of mental effort during worked example study with retrieval practice.

Retrieval Performance

The retrieval success of the retrieval condition in both waves was measured. Midway through the data collection, adjustments were made because the procedure and tasks did not fully align with the target group. Modifications were made to the worked examples (e.g., shortened duration), post-test tasks (e.g., reduction in number of steps), and the pre-test was removed (a detailed description of the modifications can be found in the method section). It appears the adjustments to the materials from wave 1 to wave 2 may have had positive effects on retrieval success in wave 2. During wave 2, students gave the wrong answer less often (wave 1: 42.65%; wave 2: 36.11%) and indicated less often that they could not retrieve (wave 1: 31.62%; wave 2: 10.56%). Moreover, the percentage of correct answers was higher during wave 2 (wave 1: 11.03%; wave 2: 27.22%). It could be that the shortened duration of the videos allowed students to longer maintain their focus. It could also be that the reduction in the number of steps in the worked examples resulted in less intrinsic load and the student's working memory could process the procedure better. To gain more insight into which video durations and number of procedural steps are effective for this student population, further research is necessary.

Post-test Retrieval

During wave 2, a post-test was conducted to evaluate the retrieval performance on both conditions. When asked to retrieve the steps of the procedure, the retrieval condition was more successful (see Table 4). These findings seem to suggest that even though there were no differences in performance on the post-test tasks (performing the to-be-retrieved procedure), the retrieval condition was better able to identify the procedural steps of the procedure. As mentioned above, completion of an isomorphic problem-solving task requires more than just the retrieval of procedural steps, it also requires answer construction. These findings can function as a starting point for future research to investigate whether, and if so how, retrieval practice could support the answer construction process. For instance, retrieval practice might encompass not only the procedural steps but also the executive steps.

Practical Implications for Research in Pre-vocational Education

Pre-vocational education is a distinct pathway within Dutch secondary education. It is considered the least advanced of the three tracks and aims to prepare students for vocational education. The current findings showed that some students did not retrieve the procedural step (e.g., *remove all individual numbers on the left part of the equation*) but the executive step (e.g., - *8 on both sides)* during the retrieval practice (wave 2: 20,56% - Table 3). This is not surprising as pre-vocational education students are often highly practical and application-oriented (Van der Neut et al., 2005). Moreover, it was observed during the experiment that students struggled with understanding the importance of the procedure, as they often did not realize that the same procedural steps could be applied to different examples. Students frequently approached a new example as entirely new and regularly asked for clarification on which procedure to use for this new example, not realizing that the explained procedure could be applied to *all* examples. This aligns with research showing that pre-vocational education students struggle with identifying, selecting, and processing (important) information during learning (Van der Neut et al., 2005).

Moreover, frustration was frequently observed during the experiment (particularly during wave 1). The frustration may have arisen due to the difficulty level being too high or the videos being too long (exceeding their concentration span). Frustration can lead to negative motivation among pre-vocational education students (Van Steensel et al., 2013), which might hinder their learning. Research shows that pre-vocational education students need substantial support with (self-)regulation, motivation, and concentration during learning (Hamstra & Van de Ende, 2006; Van der Neut et al., 2005). Frustration was observed, for instance, during the pre-test in wave 1. Students expressed that they were unhappy about being ‘tested’ before receiving any instruction (even though it was communicated that the pre-test only served to gauge prior knowledge). It might be that students, due to lack of prior instruction, perceived their resources as insufficient for the pre-test tasks, leading to frustration as an expression of a potential threat state. The biopsychological model of challenge and threat examines how individuals respond to tasks (Blascovich, 2008a). Before undertaking a task, individuals assess the demands of the task (demand evaluation) and whether they possess the necessary resources to effectively meet these demands (resource evaluation). A threat state occurs when they perceive their resources as inadequate (Seery, 2011). Moreover, research shows that many pre-vocational education students have limited success experiences in educational contexts (Hamstra & Van de Ende, 2006). Future research into how pre-vocational education students evaluate their resources and whether their history of limited educational success might influence their evaluation (negatively) is recommended. This could contribute to a better understanding of how pre-vocational education students can be encouraged to enter a challenge state rather than a potential threat state while engaging in learning tasks.

Furthermore, during wave 1, students indicated that the example of the scale (part of the first worked example) was not helpful (it was distracting and considered childish) (Figure 1 shows an example of the scale). The example of the scale closely followed the mathematical teaching method used by the school and therefore is also present in the students’ textbooks. Small-scale research in language education indicates that students find it annoying and hindering when method developers include examples or short texts that are not essential to the learning material (Toorenaar & Rijlaarsdam, 2007). This may have also been the case with the scale example. Information that is not essential can be perceived as extraneous load and research indicates that extraneous load hinders learning (Sweller et al., 2019).

In conclusion, there is an urgent need to do more research within this population, to better understand their learning process and how we can improve support in this regard. The above-mentioned could serve as a starting point for further research into how the learning of this population is affected by their application-oriented nature, how affective elements (e.g., frustration) affect their learning process, and which aspects of instructional material design hinder or help them.

Limitations and Strengths of this Study

Although these findings advance our understanding of the fostering of problem-solving skills through prompting stepwise retrieval, it is important to note several potential limitations. As mentioned above, it is possible that the scaffolded tasks served as a reminder, potentially masking performance differences. Consequently, this may have resulted in an impure retrieval measurement. Given the characteristics of the target group and their low frustration tolerance, it was not feasible to eliminate the scaffolded tasks to prevent this issue. In future research, scaffolded tasks could be handled differently (e.g., removed or counterbalanced). Furthermore, since the post-test was on paper, students could look back at the scaffolded items to retrieve the procedure, potentially hindering pure comparison between the control and retrieval conditions. However, administrating the post-test on paper aligned with the target group, who were not accustomed to math tasks on the computer. Moreover, the research setting differs from the actual educational setting. While students in educational environments often practice a procedure with the explanation available (and with the possibility to ask for guidance from the teacher), in this study, students were required to perform the procedure immediately without access to the explanation or guidance.

Due to unanticipated procedural challenges during this experiment, the sample is divided into two waves and is smaller than expected. Therefore, future research into prompting stepwise retrieval during worked example study is recommended with a larger sample. Moreover, it is recommended to conduct future research on stepwise retrieval during worked example study with upper general education (*havo*) and pre-university education (*vwo*) students. This could give insight into if there are potential differences in the effectiveness of prompted stepwise retrieval during worked example study, between pre-vocational education and general/pre-university education students.

An important strength of this study is the target group involved. The pre-vocational secondary education track constitutes a significant portion of our educational population; half of our third-year secondary education students, 94.000 pre-vocational secondary education students in total (Ministerie van Volksgezondheid, Welzijn en Sport, 2021). This study takes a step towards understanding the use of prompted retrieval practice during worked example study, as well as gaining insights into conducting educational research within this population.

Conclusion

The purpose of the current study was to determine whether the learning of problem-solving skills could be fostered by prompting stepwise retrieval practice during worked example study. This study has identified no significant impact of prompted stepwise retrieval practice during worked example study on the acquisition of problem-solving skills.

Due to the large amount of evidence on behalf of the effectiveness of retrieval practice, future research on the combination of retrieval practice and worked examples is important. Research in language education indicates that stepwise retrieval can indeed bring benefits (Van den Broek et al., 2023). This study is a first step in exploring stepwise retrieval in the context of worked example study and provides several starting points for future research into combining stepwise retrieval and worked example study. Interestingly, and perhaps more importantly and surprisingly, this study has shown that pre-vocational secondary education students are a highly interesting target group, with much to be discovered about how these students learn. Therefore, it would be a promising future perspective to involve this group in educational research with renewed interest and curiosity.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659-701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger III, H. L. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory*, *25*(6), 764-771. https://doi.org/10.1080/09658211.2016.1220579

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*(2), 181–214. https://doi.org/10.3102/00346543070002181

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: knowing about knowing* (pp. 185-205). MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In: A. F. Healy, S. M. Kosslyn, R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (vol. 2, pp. 35-67). Erlbaum.

Blascovich, J. (2008a). Challenge and threat. In A. J. Elliot (Ed.), *Handbook of approach and avoidance motivation* (pp. 431-455). Psychology Press.

Clark-Carter, D. (2009). *Quantitative psychological research: The complete handbook.* Psychology Press.

Dudai, Y. (2012). The restless engram: Consolidations never end. *Annual Review of Neuroscience*, *35*, 227-247. https://doi.org/10.1146/annurev-neuro-062111-150500

Dudai, Y., Karni, A., & Born, J. (2015). The consolidation and transformation of memory. *Neuron*, *88*(1), 20-32. https://doi.org/10.1016/j.neuron.2015.09.004

Hamstra, D. G., & Van den Ende, J. (2006). De vmbo-leerling. Onderwijspedagogische en ontwikkelingspsychologische theorieën [The prevocational student. Educational pedagogical and psychological theories]. CPS, Christian Pedagogical Study Centre. https://collegedebrink.weebly.com/uploads/8/1/7/5/81756180/rapport_de_vmbo_leerling_onderwijspedagogische.pdf

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language*, *69*(2), 151-164. https://doi.org.10.1016/j.jml.2013.03.002

Hoch, E., Sidi, Y., Ackerman, R., Hoogerheide, V., & Scheiter, K. (2023). Comparing mental effort, difficulty, and confidence appraisals in problem-solving: A metacognitive perspective. *Educational Psychology Review, 35*(2), 35-61. https://doi.org/10.1007/s10648-023-09779-5

Hoogerheide, V., Loyens, S. M., & Van Gog, T. (2014). Comparing the effects of worked examples and modeling examples on learning. *Computers in Human Behavior, 41*, 80-91. https://doi.org/10.1016/j.chb.2014.09.013

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*, 621–629. https://doi.org/10.1037/a0015183

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational technology research and development, 48*(4), 63-85. https://doi.org/10.1007/BF02300500

Kalyuga, S., Renkl, A., & Paas, F. (2010). Facilitating flexible problem solving: A cognitive load perspective. *Educational Psychology Review, 22,* 175–186. http://doi.org/10.1007/s10648-010-9132-9

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. T. Wixted (Ed.), *Cognitive Psychology of Memory* (Vol. 2, pp. 487–514). Academic Press. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of experimental psychology: learning, memory, and cognition*, *33*(4), 704. https://doi.org/10.1037/0278- 7393.33.4.704

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

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review, 27,* 291–304. http://doi.org/10.1007/s10648-015-9296-4

Ministerie van Volksgezondheid, Welzijn en Sport. (2021). *Jongeren in het voortgezet onderwijs.* Jeugdmonitor. Geraadpleegd op 26 mei 2024, van https://longreads.cbs.nl/jeugdmonitor-2021/school/#:~:text=Ongeveer%20de%20helft%20(50%2C2,leerlingen%20daalt%20al%20enige%20jaren.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skills in statistics: A cognitive load approach. *Journal of Educational Psychology*, *84*, 429-434. https://doi.org/10.1037/00220663.84.4.429

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

Renkl, A. (2017). Instruction based on examples. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction*. (2nd ed., pp. 286-309). Routledge.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, *15*(1), 20-27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roelle, J., Endres, T., Abel, R., Obergassel, N., Nückles, M., & Renkl, A. (2023). Happy together ? On the relationship between research on retrieval practice and generative learning using the case of follow-up learning tasks. *Educational Psychology Review, 35*(4), 102. https://doi.org/10.1007/s10648-023-09810-9

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140,* 1432-1463. https://doi.org/10.1037/a0037559

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,* 93-114. https://doi.org/10.1007/s11251-014-9328-3

Seery, M. D. (2011). Challenge or threat? Cardiovascular indexes of resilience and vulnerability to potential stress in humans. *Neuroscience & Biobehavioral Reviews, 35*(7), 1603-1610. https://doi.org/10.1016/j.neubiorev.2011.03.003

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

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2,* 59–89. https://doi.org/10.1207/s1532690xci0201_3

Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251-296. https://doi.org/10.1023/A:1022193728205

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

Toorenaar, A., & Rijlaarsdam, G. (2007). 'Een spiekbriefje leren maken?!' Sommige schoolboeken zijn gewoon te moeilijk. *Levende Talen Magazine, 94*(2), 17-19. https://lt-tijdschriften.nl/ojs/index.php/ltm/article/view/302/295

Utrecht University. (n.d.). *Yoda - a research data management service*. uu.nl. Retrieved June 10, 2024, from https://www.uu.nl/en/research/yoda

Van den Broek, G. E., Gerritsen, S. L., Oomen, I. R. J., Velthoven, E., van Boxtel, F. H. J., Kester, L., & van Gog, T. (2023). Optimizing multiple-choice questions for retrieval practice: Delayed display of answer alternatives enhances vocabulary learning. *Journal of Educational Psychology, 115*(8), 1087-1109. https://doi.org/10.1037/edu0000810

Van den Steensel, R., Oostendam, R., & Van Gelderen, A. (2013). Vermijding en frustratie: Het belang van negatieve motivaties voor de leesvaardigheid van vmbo'ers [Avoidance and frustration: The importance of negative motivation for the reading comprehension of prevocational students]. *De aarzalende lezer over de streep; recente wetenschappelijke inzichten,* 105-122. https://www.researchgate.net/publication/265965234_Vermijding_en_frustratie_Het_belang_van_negatieve_motivaties_voor_de_leesvaardigheid_van_vmbo'ers

Van der Neut, I., Teurlings, C., & Kools, Q. (2005). Inspelen op leergedrag van vmbo-leerlingen [Responding tot he learning behavior of prevocational secondary education students]. *Tilburg: IVA Beleidsonderzoek En Advies, 9*, 10-11.

https://collegedebrink.weebly.com/uploads/8/1/7/5/81756180/inspelen_op_leergedrag_van_vmbo-leerlingen.pdf

Van Gog, t., & Kester, L. (2012). A test of the testing effect: Acquiring problem-solving skills from worked examples. *Cognitive Science, 36*(8), 1532-1541. https://doi.org/10.1111/cogs.12002

Van Gog, T., Kester, L., Dirkx, K., Hoogerheid, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educational Psychology Review, 27*, 265–289 https://doi.org/10.1007/s10648-015-9297-3

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*, 212-218. https://doi.org/10.1016/j.cedpsych.2010.10.004

Van Gog, T., Paas, F. G. W. C., & Van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction, 16,* 154–164. http://doi.org/10 .1016/j.learninstruc.2006.02.003

Van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*, *22*, 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. Rawson (Eds.), *The Cambridge handbook of cognition and education* (pp. 183-208). Cambridge University Press. https://doi.org/10.1017/9781108235631.009

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27,* 247–264. http://doi.org/10.1007/s10648-015-9310-x

Van Merriënboer, J. J. G., & Kirschner, P. A. (2018). *Ten Steps to Complex Learning. A Systematic Approach to Four-Component Instructional Design* (3rd ed.). Taylor & Francis.

Wheeler, M., Ewers. M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trails. *Memory, 11*(6), 571-580. https://doi.org/10.1080/09658210244000414

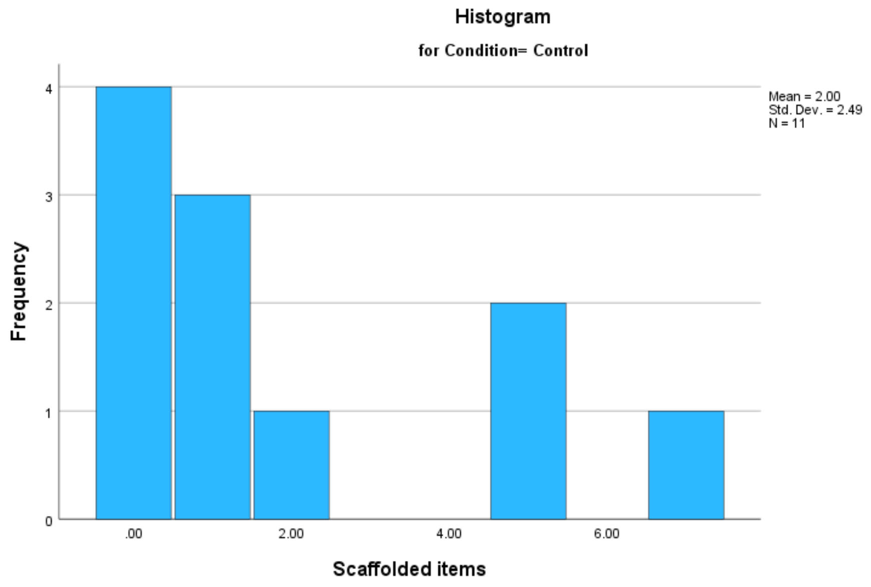
Yeo, D. J., Fazio, L. K. (2019). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *Journal of Educational Psychology*, *111*(1), 73-90. https://doi.org/10.1037/edu0000268

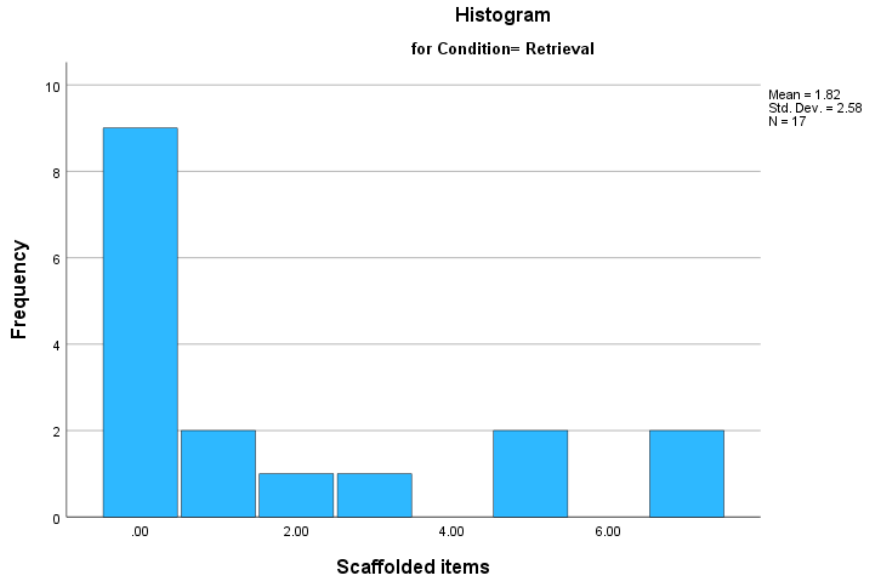
Appendix A

Histograms used for visual inspection normality assumption

Wave 1 – Scaffolded Items

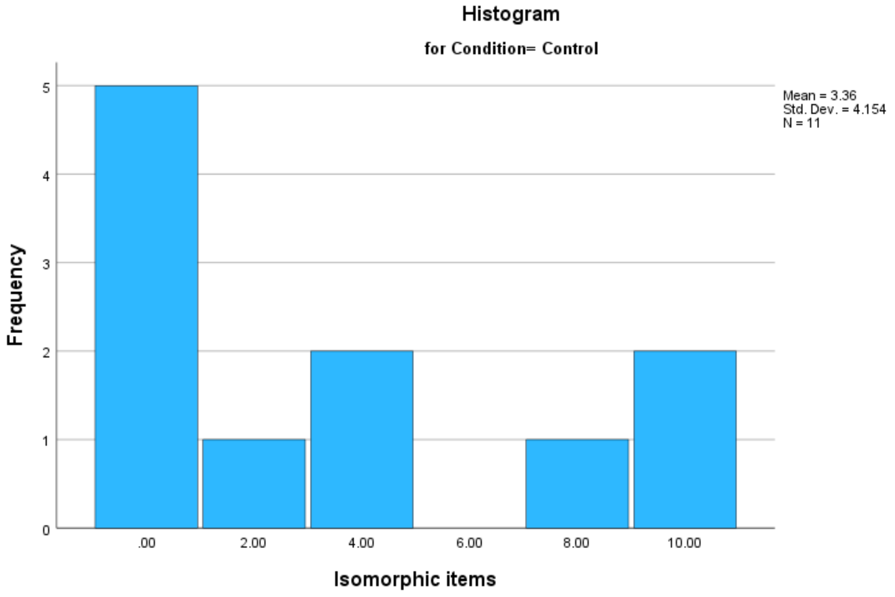
Histograms

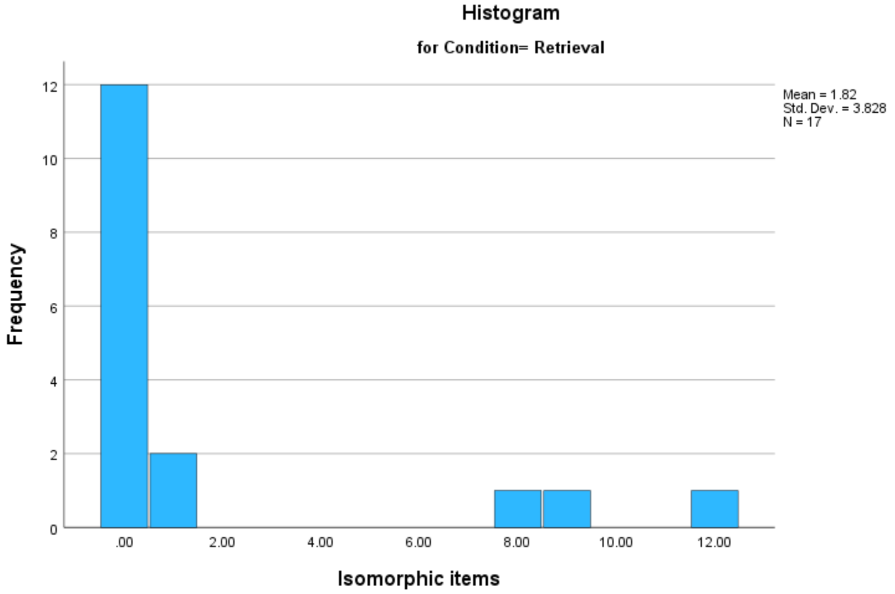




Wave 1 – Isomorphic Items

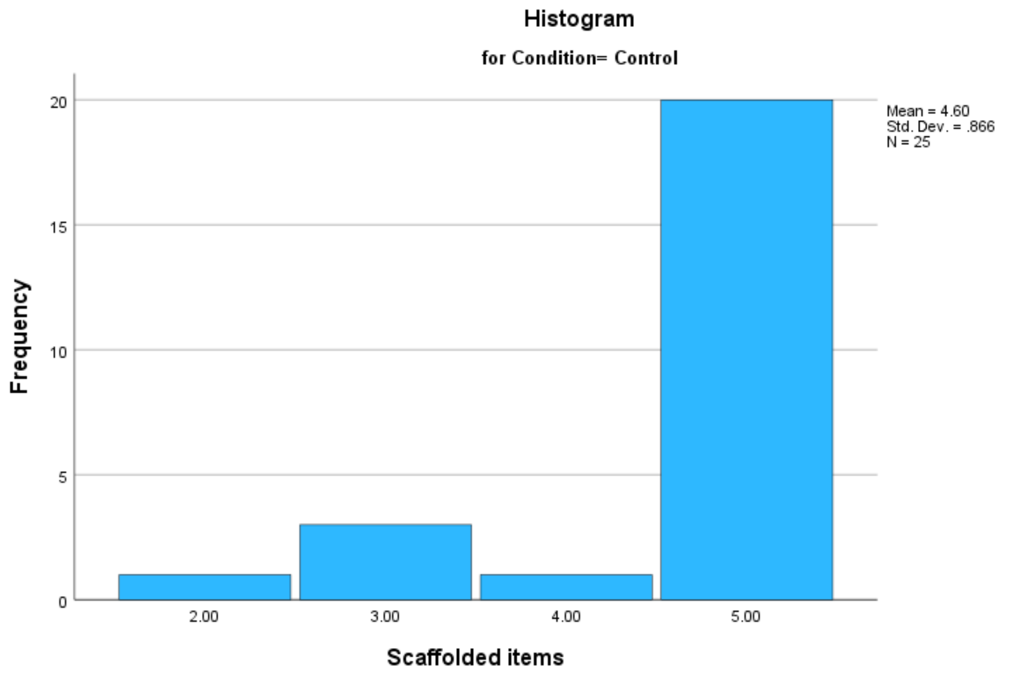
Histograms

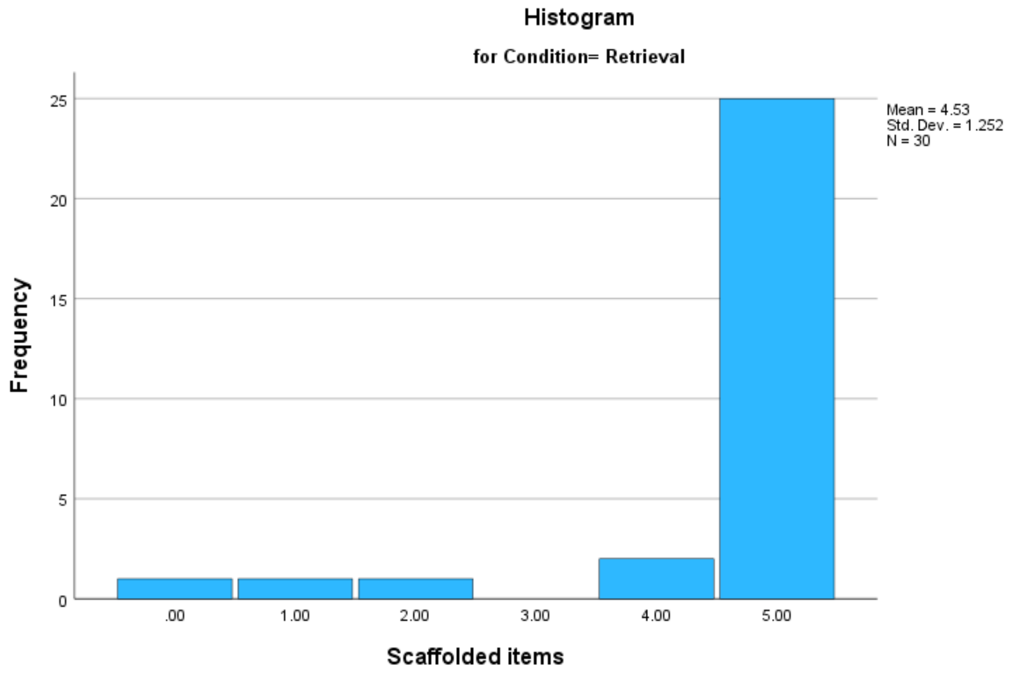




Wave 2- Scaffolded Items

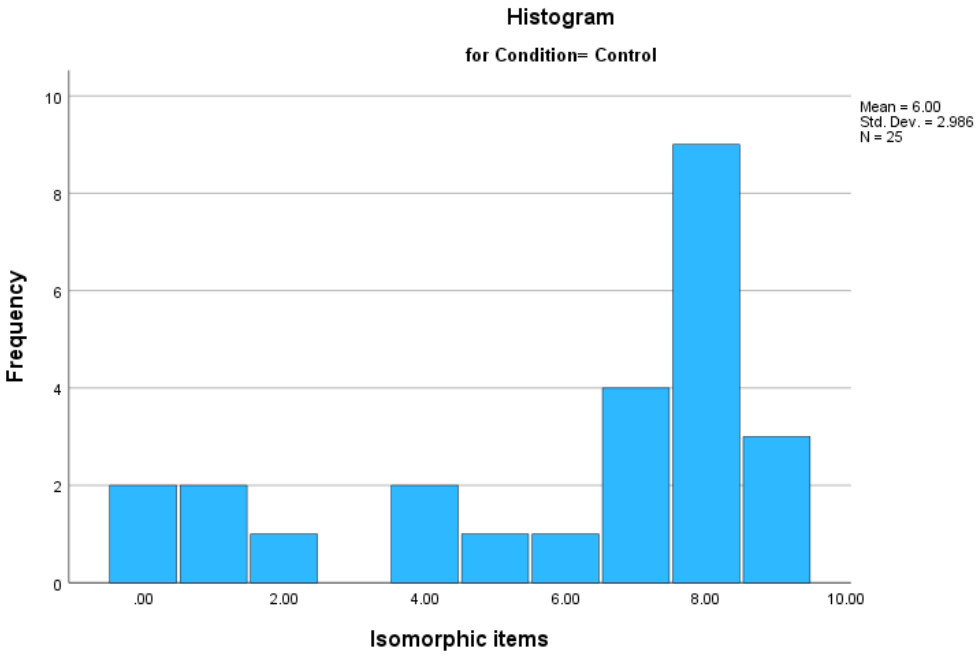
Histograms

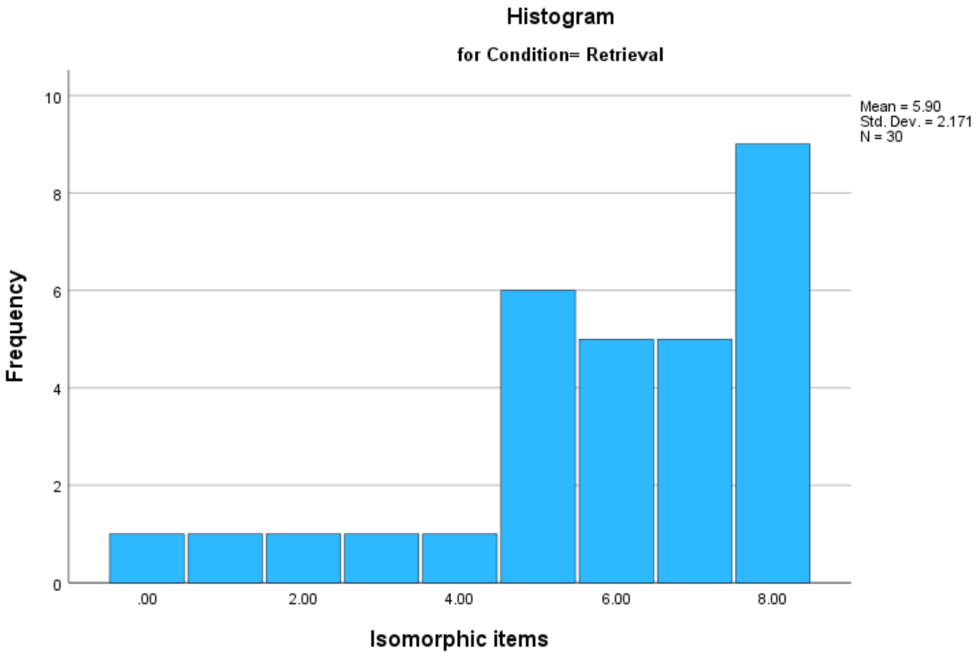




Wave 2 – Isomorphic Items

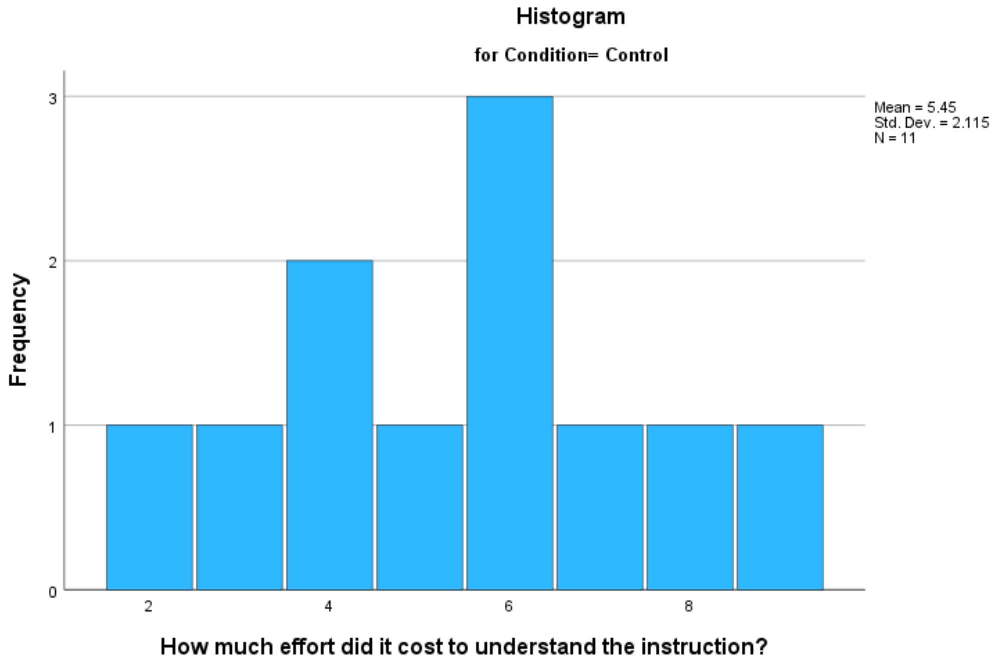
Histograms

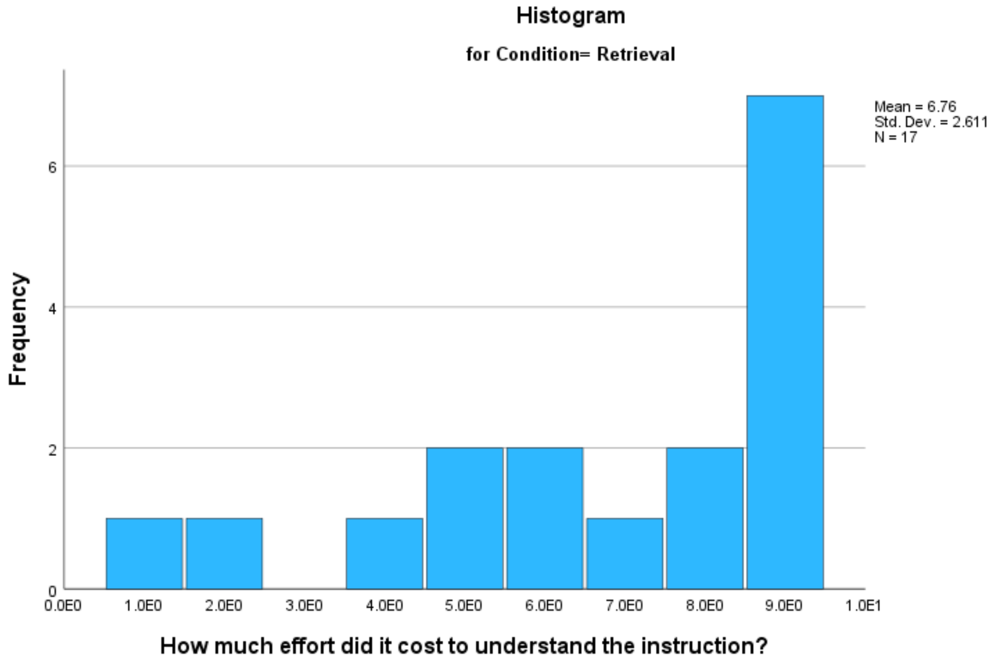




Wave 1 – Mental effort

Histograms


Histogram
for Condition= Control
Frequency
How much effort did it cost to understand the instruction?
Mean = 5.45
Std. Dev. = 2.115
N = 11


Histogram
for Condition= Retrieval
Frequency
0.0E0
1.0E0
2.0E0
3.0E0
4.0E0
5.0E0
6.0E0
7.0E0
8.0E0
9.0E0
1.0E1
How much effort did it cost to understand the instruction?
Mean = 6.76
Std. Dev. = 2.611
N = 17