Providing Worked Examples for Learning Multiple Principles

Lennart Schalk[1,a], Julian Roelle[2], Henrik Saalbach[3], Kirsten Berthold[4], Elsbeth Stern[5], &

Alexander Renkl[6]

Running title: WORKED EXAMPLES FOR MULTIPLE PRINCIPLES

[1]PH Schwyz

[2]Ruhr Universität Bochum

[3]Universität Leipzig

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/acp.3653

[4]Universität Bielefeld

[5]ETH Zürich

[6]Albert-Ludwigs-Universität Freiburg

[a]corresponding author: Lennart Schalk, PH Schwyz, Zaystrasse 42, 6410 Goldau, Switzerland, Email: lennart.schalk@phsz.ch

Keywords: worked examples, blocking, interleaving, comparison, contrast, self-explanations

The authors declare that there is no conflict of interest.

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements: Lennart Schalk and Julian Roelle contributed equally to this publication; both should be considered as primary authors.

Abstract

Worked examples support learning. However, if they introduce easy-to-confuse concepts or principles, specific ways of providing worked examples may influence their effectiveness. Multiple worked examples can be introduced blocked (i.e., several for the same principle) or interleaved (i.e., switching between principles), and can be presented sequentially or simultaneously. Crossing these two factors provides four ways of presenting worked examples: blocked/sequential, interleaved/sequential, blocked/simultaneous, and interleaved/simultaneous. In an experiment with university students ($N = 174$), we investigated how the two factors influence the acquisition of procedural and conceptual knowledge about different, but closely related (thus, easy-to-confuse) stochastic principles. Additionally, we assessed the ability of students to discriminate between principles with verification tasks. Simultaneous presentation benefitted procedural knowledge whereas interleaved presentation benefitted conceptual knowledge. No significant differences were found for verification tasks. The results suggest that it is worthwhile to adapt the presentation of the worked examples to the learning goals.

Providing Worked Examples for Learning Multiple Principles

Example-based learning is a well-known, widely used, and effective approach to introduce learners to new content (e.g., Renkl, 2014; van Gog & Rummel, 2010). This approach follows a typical instructional design. In the first step, learners receive instructional explanations explicating basic declarative knowledge regarding new principles and concepts. In the second step, learners study worked examples, which instantiate the to-be-learned principles or concepts in a specific context and which include a solution (e.g., Berthold & Renkl, 2009; Foster et al., 2018; Roelle et al., 2017; van Gog et al., 2008). For learning from worked examples to be effective (e.g., in comparison to problem solving), it is essential that learners focus on the structural features of the examples and relate them to the previously received instructional explanations. As not all learners engage in such *self-explanation activities* of their own accord (e.g., Chi et al., 1989; Renkl, 1997), combining worked examples with specific self-explanation prompts has been well established. These prompts should direct learners' attention to the relevant aspects, engage them in generating self-explanations, and help them process the examples with the focus intended by the instructor (e.g., Atkinson et al., 2003; Gerjets et al., 2008; Roelle & Renkl, 2020).

A combination of instructional explanations, worked examples, and self-explanation prompts has repeatedly been shown to be more effective than both unguided and guided problem solving (for overviews see, Renkl, 2014; Renkl & Eitel, 2019). Its implementation is relatively straightforward if learners study a single principle or concept. In this case, after having read the basic instructional explanations, learners receive multiple worked examples that instantiate the respective principle or concept. However, in educational practice, instructors often introduce several different but closely-related principles and concepts within one lesson. For example, physics teachers may present the three Newtonian laws of motion in a lesson (and then, of course, continue their instruction on these laws for some time), or mathematics teachers may present the stochastic principles of relevance of order and replacement in a mathematics lesson. If students have to learn several related, but different principles, sets of worked examples that support students in recognizing and understanding the important structural features of those principles are needed (e.g., Quilici & Mayer, 1996; Renkl, 2014). However, how these sets of worked examples should be provided to optimally support learning is not yet clear. Specifically, both the question of how the examples should be ordered and the question as to whether multiple examples should be provided simultaneously have received relatively scarce attention in research on worked examples.

Order of Examples: Blocked vs. Interleaved

One important design decision in providing sets of examples to introduce different principles relates to the *order* of the examples. More specifically, examples can be *blocked* or *interleaved*. Providing examples in a blocked order means that an educator would first present all examples for Principle_X, then all examples for Principle_Y, then all examples for Principle_Z, and so on. By contrast, in an interleaved order an educator would first present one worked example for Principle_X, then one for Principle_Y, then one for Principle_Z, then again one for Principle_X, one for Principle_Y, etc.

From a theoretical view, a major advantage of a blocked order is that it allows for *within-category comparisons* (e.g., Gerjets et al., 2008). That is, when processing several worked examples one after another, with all examples instantiating the same principle, this sequence might help learners to notice those common structural features or relations that are constitutive for the principle. This noticing sets the stage for schema abstraction (Gentner, 2010; Goldwater & Schalk, 2016; Quilici & Mayer, 1996), which enables solving novel tasks based on the same principle (i.e., knowledge transfer). A disadvantage, however, is that blocked examples scarcely engage learners in *between-category comparisons* (e.g., Gerjets et al., 2008), which would foster learners' ability to actively discriminate between principles. Not understanding the differences between principles is detrimental with respect to learning outcomes because learners may fail to understand when a principle does not apply. A further potential drawback of not having to decide which principle to choose could be that it might actually fool students because learning seems simple. The students know (or figure out quickly) which principle to apply for given problems and thus the experience of *fluency* increases (Rohrer et al., 2019) and the subjective cognitive load decreases. The literature on meta-comprehension suggests that the experience of fluency is used as a cue to judge the effectiveness of learning (Alter et al., 2007). However, fluency is not of a high diagnostic value for sustainable learning (e.g., Kornell et al., 2011). Hence, blocked orders of examples might trigger illusions of understanding, which lure learners into abandoning deep processing of the worked examples too early.

The outlined advantages and disadvantages of providing examples in a blocked order are reversed for providing worked examples in an interleaved order. Specifically, through enhancing between-category comparisons, an interleaved order of examples fosters learners' ability to discriminate between different principles in, for example, mathematics (for a meta-analysis, see Brunmair & Richter, 2019). In comparison to blocked examples, learners should experience less fluency and increased cognitive load when processing interleaved examples. Thus, given that learners must decide which principle to apply for each example, illusions of understanding may decrease. This notion is supported by the results of Paas and van Merriënboer (1994), who showed that learners studying highly variable worked examples (i.e., strong variation in surface features) compared to low variability worked examples reported higher cognitive load on a mental effort scale. A disadvantage, however, could be that in comparison to blocked examples learners receive less support in finding commonalities and similarities across worked examples that require application of the same principle, which could detrimentally affect schema abstraction and thus learning outcomes.

When the examples are provided sequentially—that is, when one example is presented after the other (a discussion of the potential role of simultaneity follows in the next section)—empirical findings indicate that blocked orders are inferior to interleaved orders in most contexts (Brunmair & Richter, 2019), although learners seem to prefer blocked orders. More specifically, in a study by Carvalho, Braithwaite, de Leeuw, Motz, and Goldstone (2016), students in an Introductory Psychology course could choose their own study sequence. They preferred blocked orders and performed better on the final exam compared to students who had to follow other sequences. A recent meta-analysis by Brunmair and Richter (2019) indicates that this advantage of blocking probably only holds for learning with expository texts or for word learning, and that interleaved order shows benefits for learning with visual materials and in mathematics education (see also Rohrer, 2012; Rohrer et al., 2019). Most of these results are based on empirical studies in which the examples are provided sequentially, one after another. Yet, when multiple principles and concepts are to be learned, the provision of several worked examples simultaneously is also a viable option.

Simultaneity of Examples: Sequential vs. Simultaneous

When more than one worked example per principle is provided, instructors must decide—in addition to deciding the order of examples—whether they provide worked examples sequentially or simultaneously. When two or more examples are provided simultaneously, learners see several worked examples that instantiate the same principle (blocked order) or different principles (interleaved order) at once. For example, if three principles—X, Y, and Z—have to be learned in a blocked order, an educator would simultaneously present several worked examples, all instantiating principle X first, then several examples instantiating principle Y, and finally several examples instantiating principle Z. In an interleaved order, by contrast, simultaneous presentation would entail that instructors first present a set of worked examples simultaneously, with one of the examples instantiating principle X, another instantiating principle Y, and another instantiating principle Z. In the next step, she would again present three examples instantiating the three concepts, etc.

In comparison to sequential presentation, simultaneous presentation should entail the advantage that the abstraction of commonalities and similarities (blocked order) and the discrimination between different principles (interleaved order) is facilitated. Facilitation occurs when learners can compare and contrast the simultaneously provided examples without having to retrieve previous examples from memory. Retrieval demands have been shown to hinder knowledge construction activities, at least when the learning material is complex and learners have not yet thoroughly grasped the material (e.g., Roelle & Berthold, 2017; Roelle & Nückles, 2019). Simultaneous presentation does not require learners to retrieve or to hold (previously encoded) examples active in working memory to compare across examples. This off-loading of working memory should be reflected in lower cognitive load and higher perceived fluency during learning.

Empirical findings suggest that simultaneous presentation is more effective than sequential presentation. For instance, in terms of blocked orders, the meta-analysis of Alfieri, Nokes-Malach, and Schunn (2013) shows that fostering within-category comparison via simultaneous presentation of exemplars, cases, or worked examples is more beneficial than sequential presentation in various domains. In contrast, the evidence that simultaneous presentation additionally enhances the benefits of interleaved orders is mixed. We describe this evidence below so that we can directly derive the hypotheses for the present study from the existing empirical evidence.

Crossing Order and Simultaneity of Examples

Although both of the discussed factors (order: blocked vs. interleaved; simultaneity: sequential vs. simultaneous) and thus all four ways of providing examples (i.e., blocked/sequential, interleaved/sequential, blocked/simultaneous, interleaved/simultaneous) have been implemented in previous research, the evidence is relatively inconclusive. Most studies did not completely cross both factors and have used different dependent variables to monitor the process of learning and to assess learning outcomes. To our knowledge, only two prior studies completely crossed both factors (Hancock-Niemic et al., 2016; Sana et al., 2017). Moreover, only one of these provided worked examples. Specifically, Hancock-Niemic and colleagues (2016) provided students with worked examples to teach basic probability principles. Their interleaved/simultaneous presentation resulted in the best transfer performance; the other factorial combinations of providing examples did not differ from each other. In this study, however, since no specific self-explanation prompts were used, the potential of worked examples might not have been fully exploited. In another study by Sana and colleagues (2017, Exp. 3), undergraduates learned three different statistical concepts. They were asked to carefully read exemplars of the concepts. Afterwards, undergraduates classified novel exemplars. The interleaved conditions outperformed the blocked conditions. Simultaneity fostered learning for blocked but not for interleaved orders. However, the examples provided in this study consisted of descriptions of concrete situations in which one of the concepts could be applied without giving a solution. Hence, they did not qualify as worked examples and it is unclear whether the findings could be generalized to the provision of worked examples.

A few further studies compared subsets of the four factorial combinations of order and simultaneity to provide (worked) examples. Kang and Pashler (2012, Exp. 2) compared three conditions: blocked/sequential, interleaving/sequential, and interleaving/simultaneous. Learners processed paintings by different artists accompanied by the artist’s name but were not given specific prompts to self-explain in order to foster processing of the examples. In the posttest, the learners classified novel paintings. Interleaved/sequential and interleaved/simultaneous led to a better performance than blocked/sequential presentation; there was no effect of simultaneity in the interleaved groups. In a series of four studies using categorization tasks with feedback, Corral and colleagues (2018) showed that interleaved/simultaneous presentation led to better classification performance than blocked/simultaneous presentation across different kinds of stimuli. Corral and colleagues did not foster processing of the stimuli through the use of self-explanation prompts. In addition to these results from studies that compared several ways of providing examples, we mentioned above that meta-analyses and reviews indicate that blocked/simultaneous presentation is more beneficial than blocked/sequential presentation (Alfieri et al., 2013) for knowledge transfer (comprising procedural and conceptual knowledge), and that interleaved/sequential presentation is more beneficial than blocked/sequential presentation (Brunmair & Richter, 2019; Rohrer et al., 2019) for various learning outcomes in some contexts (e.g., mathematics). Further, several studies have indicated that interleaved/simultaneous presentation is superior to blocked/sequential presentation for constructing conceptual and procedural knowledge (e.g., Ziegler & Stern, 2014, 2016). The learning outcomes vary widely in these studies that compared several ways of providing examples. Some of the retention and transfer tasks required learners to classify novel examples, others to apply and execute procedures, or to demonstrate conceptual understanding. The process measures used to capture how learners experience the different ways of providing examples have rarely been assessed.

The Present Study

Given the limited evidence from experiments that simultaneously varied both factors of interest using worked examples, and the fact that the studies used outcome measures that are hardly comparable, we aimed to investigate the role of both factors on several outcome and process measures when learning with worked examples. As outcomes, we assessed procedural knowledge (i.e., being able to compute the solution to a problem), conceptual knowledge (i.e., being able to describe principles and reason qualitatively about them), as well as learners' classification accuracy: verifying a given solution procedure for a problem and verifying whether a given principle described a task correctly. As process measures, we assessed subjective fluency, cognitive load, and time needed to study the worked examples. Our hypotheses were as follows.

1. *Learning outcomes:* We made the same prediction for all learning outcomes (procedural knowledge, conceptual knowledge, and verification tasks). That is, we expected a main effect of order (blocked < interleaved) and a main effect of simultaneity (sequential <

simultaneous). Furthermore, we were interested whether, similar to Sana et al. (2017) and as suggested by the findings of Kang and Pashler (2012), we would find an interaction that would indicate that the benefits of simultaneity are more pronounced for the blocked than for the interleaved order.

2. *Fluency and cognitive load:* It is more challenging to derive precise hypotheses regarding the process measures of fluency and cognitive load because no prior study has assessed these measures across both factors. In view of our theoretical considerations, we expected a main effect of order (blocked > interleaved) and a main effect of simultaneity (sequential < simultaneous) for fluency. The directions of these hypotheses are reversed for cognitive load. That is, we also expected the main effects of order and simultaneity but in the opposite direction (i.e., blocked < interleaved, sequential > simultaneous). Although we did not have a specific hypothesis, for explorative purposes we furthermore investigated if there would be an interaction between order and simultaneity concerning these measures.

3. *Learning time:* Finally, for the time needed to study the worked examples, we expected a main effect of order with blocked orders leading to lower learning time than interleaved orders (see Rohrer et al., 2019) because learners do not have to switch between principles in blocked orders (blocked < interleaved). Furthermore, we expected that simultaneous presentation would yield faster learning time than sequential presentation (sequential > simultaneous) because providing worked examples simultaneously frees learners from retrieval demands. For explorative purposes, we were also interested if there would be an interaction between order and simultaneity.

Additionally, we checked to see whether randomization yielded comparable groups of learners regarding their pretest performance and whether all conditions resulted in learning gains (i.e., expecting improvements within all conditions). As a manipulation check, we analyzed which focus learners adopt in their self-explanations. That is, learners should focus on single examples when they are presented sequentially, but should focus on similarities or differences of examples if the presentation highlights these; that is, a focus on similarities should be present in the blocked/simultaneous condition and a focus on differences should be present in the interleaved/simultaneous condition.

We tested our hypotheses using learning materials about stochastic principles. These materials had been developed for previous studies which investigated learning with worked examples (e.g., Berthold & Renkl, 2009; Deiglmayr & Schalk, 2015); we adapted them slightly to suit the needs of the present study. The stochastic principles are so-called mathematical urn models (described in more detail below). While these are taught in secondary schools, the previous study by Deiglmayr and Schalk (2015) indicated that undergraduates are not very familiar with these models (if they are not studying mathematics or computer science); thus, these contents also provide a suitable learning task for undergraduates (see also Berthold et al., 2009).

Method

Participants

The study was approved by the institutional ethics committee. All participants gave written informed consent that their data may be analyzed. We aimed to test at least 40 participants per condition based on studies by Sana and colleagues (2017), who had a maximum of 35 participants per condition in their Exp. 3, and Hancock-Niemic and colleagues (2016), who had a maximum of 30 participants per condition. In total, participants comprised 174 undergraduates studying in Zurich, Switzerland, with a mean age of 22.1 years (74 female, 96 male, 4 did not specify) majoring in various subjects. This sample size allows for the detection of small to medium effect sizes ($f > .2$; $d > .4$) with a power of $1-\beta = 80\%$. We excluded undergraduates majoring in mathematics and closely related subjects such as computer science and physics. Participants were randomly assigned to one of the four sequences (blocked/sequential: 44 undergraduates, interleaved/sequential: 44, blocked/simultaneous: 44, interleaved/simultaneous: 42). All participants were native German speakers and were paid 50 Swiss francs. They were tested in the Decision Science Laboratory at ETH Zurich (https://www.descil.ethz.ch/), where participants can work undisturbed in small cubicles equipped with a computer, a 19" screen, and a keyboard.

Procedure

The experiment lasted two hours. All materials were delivered via a computer-based learning and assessment environment. Participants started with the pretest. Afterwards, they read a short general instructional explanation about urn models and stochastics (identical for all conditions) and then processed the worked examples and responded to the self-explanation prompts. The learners could process the examples as long as they wanted but could not go back to an example once they had clicked a button to proceed to the next page of the learning environment. After four worked examples, participants answered questions about fluency and cognitive load. These assessments were repeated three more times, always after four worked examples. For the analyses below, the scores were aggregated across the four assessments. Additionally, we recorded the time needed to work through all the learning materials. A tenminute break followed, after which participants completed the posttest. Finally, participants were debriefed and paid.

Materials

Learning materials. Participants learned about mathematical urn models, which are used to calculate probabilities of combinations of random events. Basic examples include repeated dice throwing or drawing objects from a vessel such as an urn (hence the metaphorical name "urn models"). Urn models can be applied in four distinct cases (we refer to the cases as "principles" in the following). The four principles capture situations in which the order of events/objects is either relevant or irrelevant and in which events/objects are taken out of the urn and are then either returned to the urn or kept outside of it (i.e., with/without replacement). That is, the four stochastic principles are order relevant/with replacement (P1), order relevant/without replacement (P2), order irrelevant/with replacement (P3), and order irrelevant/without replacement (P4).

To create the worked examples, we embedded the four principles (P1-P4) in four different contexts (C1-C4). Context simply denotes different cover stories, such as the probability of blindly picking a specific cookie out of a box containing four different kinds of cookies. By combining principles with contexts, we embedded each principle in every context, obtaining 16 worked examples (see Figure 1 for an example in which all four principles are embedded in one context, and Figure 2 for how the examples were presented). In the *blocked/sequential* condition, we provided the 16 worked examples blocked by principle individually (i.e., first, the examples for P1 embedded in all contexts, then the examples for P2 in all contexts, etc., sequentially). In the *interleaved/sequential* condition, we again provided the worked examples individually, but the worked examples for the different principles were interleaved (i.e., first, the examples for P1-P4 embedded in C1, then the examples for P1-P4 in C2, etc.). In the *blocked/simultaneous* condition, we always presented four worked examples simultaneously (i.e., on the same screen), blocked by principle (i.e., first, all examples for P1 embedded in all four contexts, then all examples for P2 in all contexts, etc.). In the *interleaved/simultaneous* condition, we again presented four worked examples simultaneously, but interleaved the principles (i.e., first the P1-P4 examples embedded in C1, then the P1-P4 examples in C2, etc.).

Self-explanation prompts accompanied the worked examples (see Figure 2). Participants wrote 16 self-explanations in all conditions. To scaffold processing of the worked examples in accordance with the four ways of providing worked examples, the prompts differed. That is, we consider the prompts to be intrinsic parts of the design of the different ways of example presentation. Specifically, in the sequential conditions, participants always saw only one worked example at a time; therefore, they were prompted to "Describe the solution of the worked example. What do you notice?". In the simultaneous conditions, participants saw four worked examples simultaneously on one screen; thus, they also saw four self-explanation prompts per screen. The first self-explanation prompt in the simultaneous conditions was: "Describe the solution of the first and second worked examples. Compare them!". The second prompt was: "Describe the solution of the third worked example and compare it with solutions for the first and second examples." The third prompt was: "Describe the solution of the fourth worked example and compare it to the other three." The fourth prompt was: "Compare all four solutions of these worked examples."

In total, participants produced 2784 self-explanations (174 participants x 16 self-explanations). Their content was used only to assess whether the four ways of providing worked examples do indeed lead to different foci. Therefore, each self-explanation was categorized as belonging to one of the following categories: focus on a single worked example, focus on the similarities of two or more examples, focus on the differences between two or more examples, focus on the similarities and differences of two or more examples, and off topic. Two coders blind to the hypotheses independently categorized a subset of 640 self-explanations. Their interrater reliability was Cohen's Kappa = .80; disagreements were solved by discussing them. Given this reliability, one of the coders categorized all remaining self-explanations.

Process measures gathered during the learning phase. To assess fluency, we used the Flow Short Scale by Engeser and Rheinberg (2008) comprising 10 items (e.g., "My thoughts run fluidly and smoothly."). Participants rated these items on a scale from 1 (not at all) to 7 (very much). Cronbach's Alpha was larger than .81 for all four assessments.

To assess cognitive load, we posed a widely-used single mental effort item (i.e., "How much effort did you invest to understand the last four worked examples?") taken from Paas (1992). Participants responded on a scale from 1 (very little effort) to 9 (very high effort).[1]

Pretest and posttest. Based on tests used in prior research (e.g., Berthold & Renkl, 2009; Deiglmayr & Schalk, 2015), we designed a pretest and four outcome assessments (posttests). In all assessments, all tasks were weighted equally and scored with one point if correct and with 0

[1] In our study, we additionally used a cognitive load questionnaire developed by Berthold and Renkl (2009). However, the results of this questionnaire are hard to interpret, since the items relate to several different types of cognitive load. Therefore, we do not present the results of this questionnaire. In addition, there were neither statistically significant main effects nor interaction (we would, of course, make the results available to interested researchers).

points if incorrect (see Table 1 for an example task for each test). The pretest comprised 12 tasks (max. 12 points). Four tasks assessed a basic skill necessary to understand stochastic principles and compute solutions to stochastic problems—that is, the skill to multiply fractions. Four tasks assessed the skill required to determine the probability of a single event, which can also be considered a basic skill to understand more complex stochastic problems. Four tasks assessed the skill used to compute the joint probability of two events. Solving and understanding those kinds of two-step problems is the skill that is taught in the learning materials. These two-step problems were repeated in the posttest for procedural understanding, allowing us to run an implementation check to test whether students benefitted from the learning materials in all conditions.

As learning outcomes, we assessed procedural and conceptual knowledge as well as the ability to verify given solutions and principles. The procedural knowledge assessment comprised 12 tasks which required participants to compute a solution to a word problem (max. 12 points). Four tasks were identical to the pretest (see description above). Four tasks were two-step word problems that were embedded in contexts that were the same as those used in the learning materials but were based on novel numbers. Four tasks were two-step word problems embedded in novel contexts.

The conceptual knowledge assessment comprised 10 tasks (max. 10 points). Four tasks were in a single choice format that required participants to select a verbal description (out of four descriptions) that correctly described the solution to a given problem. Six tasks were open questions taken from and analyzed with the procedure described in Berthold and Renkl (2009).

Finally, we used two verification assessments with a total of 16 tasks (max. 8 points per assessment). In the eight solution procedure verification tasks, a computational solution to a word problem was provided, and participants had to decide whether this solution was correct or incorrect (half of the solutions were correct). In the eight principle verification tasks, the word problems were accompanied by a verbal description of one of the four principles (e.g., order relevant/with replacement). Participants had to decide whether this description correctly or incorrectly fitted the word problem (half of the descriptions were correct).

Results

Table 2 shows the mean scores and standard deviations for all conditions on all outcome and process measures. An α-level of .05 was used for all tests.

Randomization and Manipulation Checks

First, we tested whether randomization yielded comparable groups of learners regarding their pretest performance. A 2x2-ANOVA yielded no statistically significant main effects for *simultaneity* (simultaneous vs. sequential), $F(1, 170) = 1.76$, $p = .186$, $\eta_p^2 = .01$, or *order* (blocked vs. interleaved), $F(1, 170) = 0.36$, $p = .547$, $\eta_p^2 = .00$, or an interaction, $F(1, 170) = 0.00$, $p = .978$, $\eta_p^2 = .00$. That is, the groups did not differ statistically in their pretest performance. Additionally, we tested for knowledge gains in the learning phase. A comparison of the performance on the items that were parallel between the pretest and the posttest yielded a large, statistically significant effect for all four conditions, all *F*s > 33.08, all *p*s < .001, all $\eta_p^2 > .38$. Hence, all participants gained knowledge substantially in the learning phase.

Second, as manipulation checks, we analyzed learners' self-explanations. We compared whether learners focused their self-explanations on single examples only, or whether they described similarities, differences, or similarities and differences across two or more examples.

In the first step of this analysis, we ran a 2x2x4 mixed ANOVA to determine whether there were differences in the number of the four types of self-explanations and whether these differences depended on the factors order and simultaneity. There was a statistically significant three-way interaction, $F(3, 510) = 67.81$, $p < .001$, $\eta_p^2 = .28$. Furthermore, there were statistically significant interactions between order and type of self-explanation, between simultaneity and type of self-explanation, $F(3, 510) = 86.95$, $p < .001$, $\eta_p^2 = .33$, and $F(3, 510) = 369.02$, $p < .001$, $\eta_p^2 = .68$, respectively, and a main effect of type of self-explanation, $F(3, 510) = 528.31$, $p < .001$, $\eta_p^2 = .75$. Hence, the distribution of the four types of self-explanation depended on both the order and simultaneity of the examples.

To evaluate these group differences regarding the four types of self-explanation, we conducted separate 2x2 ANOVAs for the four types of self-explanation (with a Bonferroni-corrected $\alpha = .012$). For *self-explanations focusing on single examples*, we found a statistically significant main effect of simultaneity, $F(1, 170) = 756.77$, $p < .001$, $\eta_p^2 = .81$, a statistically significant main effect of order, $F(1, 170) = 34.51$, $p < .001$, $\eta_p^2 = .16$, and a statistically significant interaction, $F(1, 170) = 16.92$, $p < .001$, $\eta_p^2 = .09$. Interpreted together, learners in sequential conditions generated many more self-explanations focusing on single examples than learners in simultaneous conditions; and while the two sequential conditions are fairly similar, there is a larger difference in the simultaneous conditions (see Table 2). That is, the blocked/simultaneous condition resulted in more self-explanations focusing on single examples than did the interleaved/simultaneous condition.

For *self-explanations focusing on differences*, we found statistically significant main effects of simultaneity, $F(1, 170) = 68.13$, $p < .001$, $\eta_p^2 = .28$, order, $F(1, 170) = 63.77$, $p < .001$, $\eta_p^2 = .27$, and an interaction, $F(1, 170) = 54.16$, $p > .001$, $\eta_p^2 = .24$. This pattern indicates that learners in the interleaved/simultaneous condition were the only ones who focused on differences, a focus that was almost non-existent in the other three conditions (see Table 2).

For *self-explanations focusing on similarities*, we found statistically significant main effects of simultaneity, $F(1, 170) = 277.59$, $p < .001$, $\eta_p^2 = .62$, order, $F(1, 170) = 221.54$, $p < .001$, $\eta_p^2 = .56$, and an interaction, $F(1, 170) = 181.56$, $p < .001$, $\eta_p^2 = .51$. This pattern indicates that learners in the blocked/simultaneous condition were the only ones who focused on similarities. This focus was almost non-existent for the other three conditions (see Table 2).

For *self-explanations focusing on both similarities and differences*, we found a statistically significant effect of simultaneity, $F(1, 170) = 65.37$, $p < .001$, $\eta_p^2 = .27$, as well as a marginally significant effect of order, $F(1, 170) = 5.97$, $p = .016$, $\eta_p^2 = .03$, and a marginally significant interaction, $F(1, 170) = 6.35$, $p = .013$, $\eta_p^2 = .03$. That is, almost no self-explanations focusing on similarities and differences were generated in the sequential conditions, while learners in the simultaneous conditions described similarities and differences, which they did (slightly) more frequently in the interleaved/simultaneous condition than in the blocked/simultaneous condition (see Table 2). Jointly, these analyses of the foci of learners' self-explanations clearly indicate that the experimental variation worked as intended.

Learning Outcomes

The pretest score was significantly correlated with all learning outcomes, $.282 < r < .515$, all ps < .001. Therefore, to reduce error variance, we included the pretest scores as a covariate in all outcome analyses. The assumption of homogeneity of regression slopes was not violated in any analysis. We formulated the identical hypothesis for all four learning outcomes. Therefore, all outcome analyses use the 2x2-ANCOVA statistical model.

We found the following pattern with the procedural knowledge assessment: the ANCOVA showed a statistically significant effect of the pretest, $F(1, 169) = 54.20$, $p < .001$, $\eta_p^2 = .24$. Furthermore, we found a statistically significant main effect of simultaneity, $F(1, 169) = 7.29$, $p = .008$, $\eta_p^2 = .04$. As expected, learners in the simultaneous conditions performed better than learners in the sequential conditions. Unexpectedly, there was neither a statistically significant main effect of order, $F(1, 169) = 1.05$, $p = .307$, $\eta_p^2 = .01$, nor an interaction, $F(1, 169) = 0.02$, $p = .874$, $\eta_p^2 = .00$.

The ANCOVA for the conceptual knowledge assessment showed a statistically significant effect of the pretest, $F(1, 169) = 58.58$, $p < .001$, $\eta_p^2 = .25$. As for the procedural knowledge assessment, there was no interaction, $F(1, 169) = 0.16$, $p = .685$, $\eta_p^2 = .01$. In contrast to the results for procedural knowledge, there was no statistically significant main effect of simultaneity, $F(1, 169) = 2.27$, $p = .134$, $\eta_p^2 = .01$, but we now found the expected main effect of order, $F(1, 169) = 4.35$, $p = .038$, $\eta_p^2 = .02$. Learners in the interleaved conditions performed better than learners in the blocked conditions on the conceptual knowledge test.

With the two kinds of verification tasks, we unexpectedly found no statistically significant differences. That is, for the tasks that required learners to verify provided solution procedures, we found no statistically significant main effect of simultaneity, $F(1, 169) = 0.49$, $p = .483$, $\eta_p^2 = .00$, order, $F(1, 169) = 1.96$, $p = .163$, $\eta_p^2 = .01$, or interaction, $F(1, 169) = 1.26$, $p = .263$, $\eta_p^2 = .01$. Only the covariate was a statistically significant predictor of learners' performance, $F(1, 169) = 26.57$, $p < .001$, $\eta_p^2 = .13$. For the verification tasks that required learners to verify provided principles, the covariate was again the only statistically significant predictor of learners' performance, $F(1, 169) = 14.81$, $p < .001$, $\eta_p^2 = .08$. There were no statistically significant main effects of simultaneity, $F(1, 169) = 0.38$, $p = .536$, $\eta_p^2 = .00$, of order, $F(1, 169) = 1.42$, $p = .234$, $\eta_p^2 = .01$, nor an interaction, $F(1, 169) = 1.02$, $p = .314$, $\eta_p^2 = .01$.

Process Measures

For all process measures, we computed the same statistical model—a 2x2 ANOVA—to check our hypotheses. For fluency, the ANOVA did not show the expected main effects of simultaneity, $F(1, 170) = 2.28$, $p = .132$, $\eta_p^2 = .01$, and order, $F(1, 170) = 0.05$, $p = .810$, $\eta_p^2 = .00$. However, there was a statistically significant interaction between the two factors, $F(1, 170) = 5.70$, $p = .018$, $\eta_p^2 = .03$. This interaction emerged because there was no statistically significant difference between the two blocked conditions (blocked/sequential vs. blocked/simultaneous, $p = .556$, $\eta_p^2 = .00$); in the interleaved conditions, however, learners in the interleaved/simultaneous condition reported a statistically significant lower fluency than learners in the interleaved/sequential condition ($p = .005$, $\eta_p^2 = .09$).

For cognitive load, we also did not find the expected main effects of simultaneity, $F(1, 170) = 0.46$, $p = .495$, $\eta_p^2 = .00$, and order, $F(1, 170) = 0.02$, $p = .868$, $\eta_p^2 = .00$, but a statistically significant interaction, $F(1, 170) = 5.79$, $p = .017$, $\eta_p^2 = .03$. This interaction emerged because there was no statistically significant difference between the blocked conditions (blocked/sequential vs. blocked/simultaneous, $p = .200$, $\eta_p^2 = .02$); in the interleaved conditions, however, learners in the interleaved/simultaneous condition reported investing more effort than learners in the interleaved/sequential condition ($p = .042$, $\eta_p^2 = .05$).

Finally, for the time needed to study the worked examples, the ANOVA indicated no statistically significant interaction, $F(1, 170) = 1.26$, $p = .262$, $\eta_p^2 = .01$. Nor did we find the expected main effect of order, $F(1, 170) = 0.84$, $p = .359$, $\eta_p^2 = .00$. The main effect of simultaneity was statistically significant, $F(1, 170) = 13.63$, $p < .001$, $\eta_p^2 = .07$, but the direction of the effect did not fit our hypothesis. Contrary to our expectation, learning in the sequential conditions took less time than learning in the simultaneous conditions.

Discussion

We investigated outcomes and processes for undergraduates learning about four easy-to-confuse stochastic principles with different ways of providing worked examples. Specifically, we manipulated two factors, simultaneity and order, to provide worked examples. Our results only partly fit our expectations for the outcome and process measures that we derived from prior empirical work.

As outcomes, we assessed procedural and conceptual knowledge as well as performance on two kinds of verification tasks. We analyzed the focus of the self-explanations as a manipulation check, and, as process measures, we assessed fluency, cognitive load, and time needed to study the worked examples. For procedural knowledge, we found the expected evidence that providing worked examples simultaneously was more beneficial than providing them sequentially, but neither the expected benefit nor the expected interaction for an interleaved over a blocked order. In contrast, for conceptual knowledge, we found the expected benefit that providing worked examples in an interleaved order was more beneficial than providing them in a blocked order, but we found neither the expected benefit nor the expected interaction for providing worked examples simultaneously over providing them sequentially. We found no statistically significant differences for either of the two verification tasks. For fluency and cognitive load, we did not find the expected main effect(s), but an interaction. That is, there was a difference only in the two conditions in which worked examples were provided interleaved: learners in the interleaved/simultaneous condition experienced lower fluency and higher cognitive load than learners in the interleaved/sequential condition. Finally, for time needed to study the worked examples, we expected—and found—a main effect of simultaneity; this effect, however, was not in the direction we expected. Learners in the two conditions in which worked examples were provided simultaneously spent more time on the learning materials than learners in the sequential conditions. In short, no combination of the two factors in providing worked examples clearly stands out, either with regard to learning processes or concerning outcomes.

We thus failed to replicate specific effects that have received quite a lot of empirical support in previous studies. We did not find statistically clear-cut advantages for providing worked examples in a blocked/simultaneous presentation in comparison to a blocked/sequential presentation, a well-documented effect (Alfieri et al., 2013), nor did we find clear-cut advantages to providing worked examples in an interleaved/sequential presentation in comparison to a blocked/sequential presentation, also a well-documented effect (Rohrer et al., 2019). Given that our research design has sufficient power to detect medium-sized differences between sequences and the analyses of the self-explanation indicating that the conditions shifted the focus of the learners, we can provide tentative explanations for this replication failure. First, all of our conditions implement strong instructional aids using worked examples with specific self-explanation prompts, which by themselves already positively influence learning (Renkl, 2014). It is possible that these strong aids mitigated the effects of the two factors, order and simultaneity. Second, we used several outcome and process measures taken from the research literature. To the best of our knowledge, no such broad assessment has previously been attempted. Our results indicate differential advantages: Procedural knowledge benefitted from providing worked examples simultaneously, conceptual knowledge benefitted from providing them interleaved. Since most prior studies used single scales when investigating the different ways of providing worked examples, these differential advantages may not have been detected. Thus, there is a third conceivable explanation related to this aspect: none of the sequences may be perfect from an applied perspective. The last two tentative explanations deserve further consideration.

A theoretical approach from cognitive psychology may provide an integrating explanation of the basic learning processes triggered by the four different ways of providing examples. Structure-mapping theory (Gentner, 1983, 2010) postulates that knowledge acquisition comprises four processes: schema abstraction, difference detection, rerepresentation, and inference projection. Given that our stimuli were complex worked examples coupled with self-explanation prompts, we cannot precisely identify and track these processes individually. Nevertheless, the four ways that we investigated resulted in different foci in the self-explanations; thus, they may also differ in how strongly they trigger the single structure-mapping processes and affect procedural and conceptual knowledge development differently. Consider as examples the foci present in the blocked/simultaneous and the interleaved/simultaneous conditions. Learners described similarities across examples in the blocked/simultaneous condition more often than in the other conditions, which may indicate stronger triggering of the process of abstraction. In the interleaved/simultaneous condition, in contrast, learners described differences between examples more often than in the other conditions, which may indicate stronger triggering of the process of difference detection. We deem the exploration of this possibility to be an exciting future direction for basic research. From an applied perspective, however, it is obvious that learners need to abstract schemata for single principles, to distinguish between principles by diagnostic features, to change their conceptions by rerepresenting their knowledge, and to generalize their knowledge via inference projection. To become mathematically literate in school settings and beyond, learners need to develop and integrate both knowledge types. It is not an either-or question. Thus, investigating additional ways of providing worked examples which support the construction of procedural and conceptual knowledge equally will be highly relevant for applied research.

Limitations

We admit to some shortcomings in our study. First, we did not equate the learning time across conditions. We made this decision because it seemed more ecologically valid to give the students the time they needed and not force them to adhere to a strict schedule (see also Rohrer et al., 2019). Moreover, we were interested in how long it takes to process an equal number of worked examples depending on the different ways of providing them. Stricter time limitations might have produced different results, but this would be a different research question.

Second, we gathered various outcome and process measures, a broad assessment that would have allowed us to compute many statistical models (e.g., various mediation models). We refrained from doing that. As there is no prior research to build on, all of these models would have been explorative. Moreover, we used an experimental design to contrast the four sequences on the learning outcomes, and our sample size was determined accordingly. More complex models would require larger sample sizes to achieve acceptable power. However, we will provide the data for researchers who are interested in running a specific exploratory model.

Third, our process measures are rather coarse-grained. Using more fine-grained process measures might help to gain a better understanding of how differential benefits emerge. For example, the use of eye-tracking may help scrutinize whether and how learners integrate various worked examples in the simultaneous conditions.

Fourth, the four ways of providing worked examples differ not only in simultaneity and order of worked examples, but also in the self-explanation prompts. This design feature of our study is not atypical. For example, Alfieri and colleagues (2013) describe in their meta-analysis that comparison typically starts with a prompt to compare, a prompt that is typically absent in sequential sequences. Nevertheless, the differences between prompts can be regarded as confounding simultaneity and prompts. It would be highly interesting if future studies were to use the prompts for sequential sequences that are typically applied in simultaneous sequences.

Conclusion

In our study, we combined several instructional techniques. From a basic science perspective, this might seem problematic as it becomes impossible to precisely tease apart influencing factors. From an applied perspective, however, our approach to combine several proven instructional strategies to teach basic stochastic principles corresponds well with complex, real-life classroom situations. We assume that our materials could be implemented in classrooms without adaptation, but additional research is needed to determine the sequence in which our materials should be presented. Nevertheless, when integrating our results with prior research on the effectiveness of the different sequences, we arrive at a Solomonic decision: different ways of providing worked examples have different advantages. The primary learning goal—whether it is building up procedural knowledge or conceptual knowledge—will determine the choice one makes.

References

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

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *134*(4), 569–576. http://dx.doi.org/10.1037/0096-3445.136.4.569

Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solve problems: Combining fading with prompting fosters learning. *Journal of Educational Psychology*, *95*, 774–783.

Berthold, K., Eysink, T. H. S., & Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science*, *37*, 345–363.

Berthold, K., & Renkl, A. (2009). Instructional aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology*, *101*(1), 70–87. http://dx.doi.org/10.1037/a0013247

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

Carvalho, P. F., Braithwaite, D. W., De Leeuw, J. R., Motz, B. A., & Goldstone, R. L. (2016). An in vivo study of self-regulated study sequencing in introductory psychology courses. *PLoS ONE*, *11*(3), e0152115. https://doi.org/10.1371/journal.pone.0152115

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*, 145–182. https://doi.org/10.1016/0364-0213(89)90002-5

Corral, D., Kurtz, K. J., & Jones, M. (2018). Learning relational concepts from within- versus between category comparisons. *Journal of Experimental Psychology: General*, *147*(11), 1571–1596. http://dx.doi.org/10.1037/xge0000517

Deiglmayr, A., & Schalk, L. (2015). Weak versus strong knowledge interdependence: A comparison of two rationales for distributing information among learners in collaborative learning settings. *Learning and Instruction*, *40*, 69–78. https://doi.org/10.1016/j.learninstruc.2015.08.003

Engeser, S., & Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion*, *32*(3), 158–172. https://doi.org/10.1007/s11031-008-9102-4

Foster, N. L., Rawson, K. A., & Dunlosky, J. (2018). Self-regulated learning of principle-based concepts: Do students prefer worked examples, faded examples, or problem solving? *Learning and Instruction*, *55*, 124–138. https://doi.org/10.1016/j.learninstruc.2017.10.002

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170. https://doi.org/10.1016/S0364-0213(83)80009-3

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*, 752–775. https://doi.org/10.1111/j.1551-6709.2010.01114.x

Gerjets, P., Scheiter, K., & Schuh, J. (2008). Information comparisons in example-based hypermedia environments: Supporting learners with processing prompts and an interactive comparison tool. *Educational Technology Research and Development*, *56*, 73–92. https://doi.org/10.1007/s11423-007-9068-z

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*(7), 729–757. http://dx.doi.org/10.1037/bul0000043

Hancock-Niemic, M. A., Lin, L., Atkinson, R. K., Renkl, A., & Wittwer, J. (2016). Example-based learning: Exploring the use of matricesand problem variability. *Educational Technology Research and Development*, *64*(1), 115–136. https://doi.org/10.1007/s11423-015-9403-8

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*, 97–103. https://doi.org/10.1002/acp.1801

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The Ease-of-Processing Heuristic and the Stability Bias: Dissociating Memory, Memory Beliefs, and Memory Judgments. *Psychological Science*, *22*(6), 787–794. https://doi.org/10.1177/0956797611407929

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

Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*(1), 122–133.

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, *88*(1), 144–161.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, *21*(1), 1–29. https://doi.org/10.1016/S0364-0213(99)80017-2

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. Rawson (Eds.), Cambridge Handbook of Cognition and Education* (pp. 528–549). Cambridge University Press. https://doi.org/10.1017/9781108235631.022

Roelle, J., & Berthold, K. (2017). Effects of incorporating retrieval into learning tasks: The complexity of the tasks matters. *Learning and Instruction*, *49*, 142–156. https://doi.org/10.1016/j.learninstruc.2017.01.008

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

Roelle, J., & Nückles, M. (2019). Generative learning versus retrieval practice in learning from text: The cohesion and elaboration of the text matters. *Journal of Educational Psychology*, *111*(8), 1341–1361. https://doi.org/10.1037/edu0000345

Roelle, J., & Renkl, A. (2020). Does an option to review instructional explanations enhance example-based learning? It depends on learners' academic self-concept. *Journal of Educational Psychology*, *112*(1), 131–147. https://doi.org/10.1037/edu0000365

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review*, *24*, 355–367. https://doi.org/10.1007/s10648-012-9201-3

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2019). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, *advance online publication*. http://dx.doi.org/10.1037/edu0000367

Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*, *109*(1), 84–98. http://dx.doi.org/10.1037/edu0000119

van Gog, T., Paas, F., & van Merriënboer, J. J. G. (2008). Effects of studying sequences of process-oriented andproduct-oriented worked examples on troubleshootingtransfer 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*, 155–174. https://doi.org/10.1007/s10648-010-9134-7

Ziegler, E., & Stern, E. (2014). Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. *Learning and Instruction*, *33*, 131–146. https://doi.org/10.1016/j.learninstruc.2014.04.006

Ziegler, E., & Stern, E. (2016). Consistent advantages of contrasted comparisons: Algebra learning under direct instruction. *Learning and Instruction*, *41*, 41–51.

Table 1

Example Tasks for All Outcome Assessments

Pretest	
Multiplication of fractions	1/4 * 3/7 =
Probability of single events	What is the probability of getting a 6 when throwing a dice once?
Two-step problem	You and your friend are taking part in a two-day mountain bike course. On both days, the instructor will bring 7 bike helmets, each of a different color (blue, green, orange, silver, brown, red, and yellow). The helmets are distributed randomly and returned to the instructor at the end of the day. On both days you will receive a helmet first and your friend second. What is the probability that you will get the red helmet on the first day of the course and your friend the yellow one?
Posttest – Procedural knowledge	
Two-step problem (context similar to learning materials)	The five ski jumpers—Adam, Urs, Beat, Christoph, and Daniel—often competed against each other on the old Engelberg ski jump. Everyone came first, second, third, fourth, or fifth with the same frequency. So, the five jumpers are equally good, while who jumps furthest depends on random factors (e.g., wind conditions). Now a new ski jump has been built in Engelberg and Adam, Urs, Beat, Christoph, and Daniel will be the first to try it out. There are two rounds. What is the probability that Beat will jump furthest in both rounds?
Two-step problem (novel context)	Four friends—Henrik, Michael, Peter, and Roland—are very interested in music. Every week their music teacher brings them four different CDs, one CD with pop, one with rock, one with techno, and one with classical music, and randomly distributes them among the friends. This week Peter and Michael are late, so the teacher gives first Henrik and then Roland a CD. What is the probability that Henrik will get the classical and Roland the techno CD?
Posttest – Conceptual knowledge	
Single-choice task	For the distribution of the parts for a school theater play, a black box is filled with notes on which the parts are described. The leads are those of the princess and the prince. One after the other, each child draws a piece of paper and keeps it. So far, half of the children have drawn a piece of paper, but the pieces of paper with the leads are still in the box. What can you say about the next child's chance to play one of the leads? 1. This child's chance to play a lead is greater than the first child's chance to play a lead. 2. This child's chance to play a lead is less than the first child's chance to play a lead. 3. This child's chance to play a lead is as great as the chance of the first child to play a lead. 4. There is a relationship between this child's chance to play a lead and the first child's chance to play a lead, but that depends on how many different parts there are.
Open question	Five people with their guinea pigs take part in a labyrinth competition. Whichever guinea pig finds the exit first is the winner. All guinea pigs have the same chance to win. If you want to predict the winner and the second one, the chance that you are right is 1/5 * 1/4. If you want to predict the first two without telling who is first and who is second, the chance that you are right is 2*(1/5 * 1/4). Why is the second calculation multiplied by 2, but not the first?

Posttest – Verification assessment	
Verify solution procedure	You and your friend are taking part in a two-day mountain bike course. On both days, the instructor will bring 6 bike helmets, all in different colors (green, blue, orange, silver, brown, and red). The helmets are distributed randomly and returned to the instructor at the end of the day. On both days you will receive a helmet first and your friend second. How likely are you and your friend to get the red and green helmets on the first day of the course (it doesn't matter who gets which color)? To solve the problem, you have to calculate the following: 2* (1/6 * 1/5). Is this solution correct?
Verify principle	You and your friend are taking part in a two-day mountain bike course. On both days, the instructor will bring 6 bike helmets, all in different colors (green, blue, orange, silver, brown, and red). The helmets are distributed randomly and returned to the instructor at the end of the day. On both days you will receive a helmet first and your friend second. What is the probability that you will get the green helmet on the first day of the course and your friend the blue one? This task corresponds to the principle "order relevant/without replacement". Is this principle correct?

Note. See Methods section for detailed descriptions of the different assessments. Tasks have been translated from the original German tests.

Table 2

Means and Standard Deviations of All Learning Outcome and Process Measures Across All Four Ways of Providing Worked Examples

	blocked/sequential	interleaved/sequential	blocked/simultaneous	interleaved/simultaneous
Pretest	.80 (.15)	.81 (.13)	.77 (.14)	.78 (.15)
Self-explanations: Single	14.88 (1.91)	15.54 (1.10)	3.04 (3.15)	6.78 (3.11)
Self-explanations: Differences	0.09 (0.42)	0.25 (0.96)	0.31 (0.80)	4.21 (3.12)
Self-explanations: Similarities	0.63 (1.29)	0.15 (.42)	10.84 (3.99)	1.23 (1.41)
Self-explanations: Differences+Similarities	0.22 (0.15)	0.00 (0.00)	1.65 (3.16)	3.11 (2.25)
Posttest: Procedural	.62 (.23)	.67 (.21)	.69 (.21)	.72 (.23)
Posttest: Conceptual	.74 (.20)	.82 (.17)	.69 (.22)	.75 (.17)
Verification: Solution	.86 (.18)	.92 (.11)	.89 (.10)	.90 (.14)
Verification: Principle	.79 (.20)	.86 (.17)	.82 (.17)	.83 (.20)
Fluency	4.66 (0.97)	4.95 (0.80)	4.78 (0.92)	4.42 (0.80)
Cognitive load	3.24 (1.31)	2.72 (1.53)	2.84 (1.56)	3.44 (1.65)
Example processing time (in min)	23.44 (7.62)	23.13 (7.35)	27.23 (11.68)	30.25 (11.50)

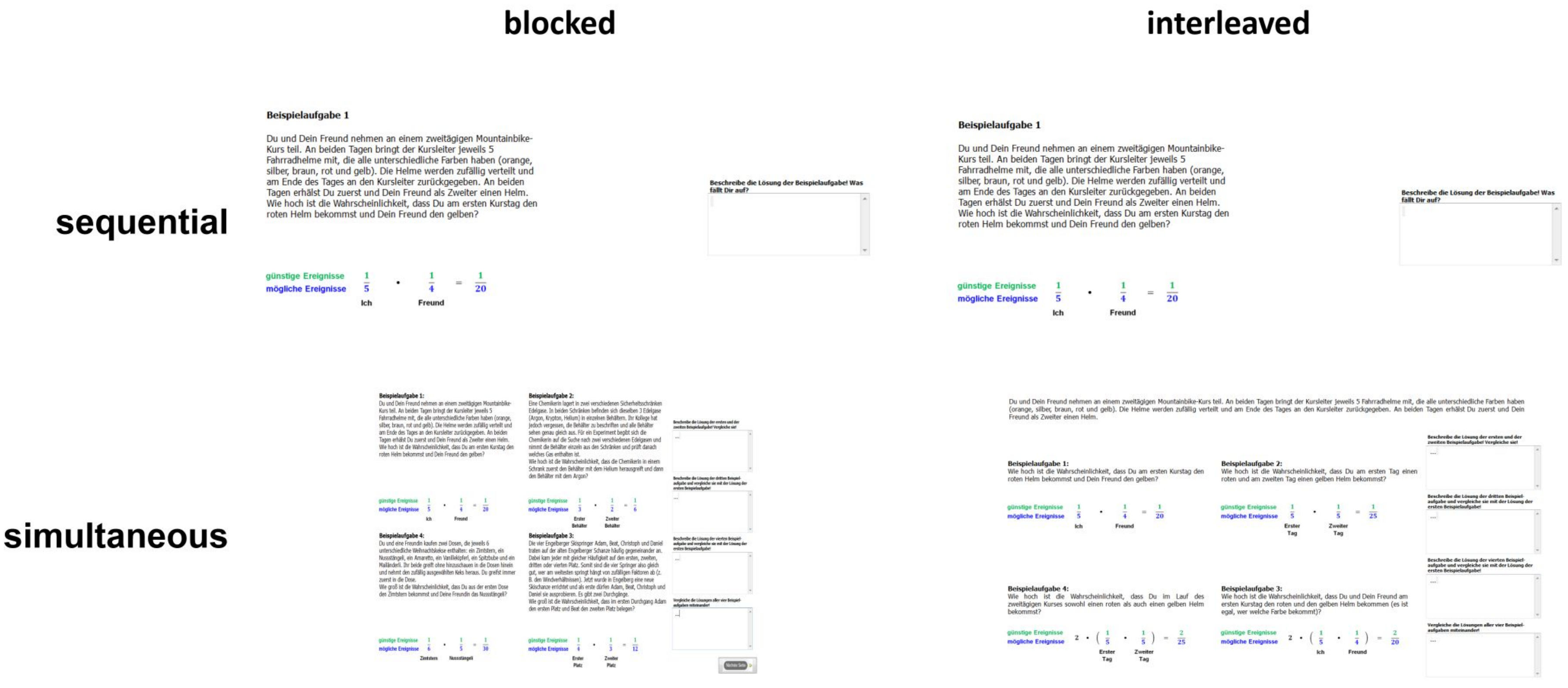
Figure captions

Figure 1. Conceptual representation of four of the worked examples used in the present study. The four stochastic principles (see Learning Materials section for details) are represented in one context. Worked examples are translated from the original German materials.

Figure 2. Representation of the two factors, order (blocked/interleaved) and simultaneity (simultaneous/sequential). In all four conditions, 16 worked examples and self-explanation prompts were presented in an integrated format (i.e., example + solution + prompt). In the sequential conditions, only one worked example was presented on a single screen; in the simultaneous conditions, four worked examples and self-explanation prompts were presented together on a single screen. See text for further information.

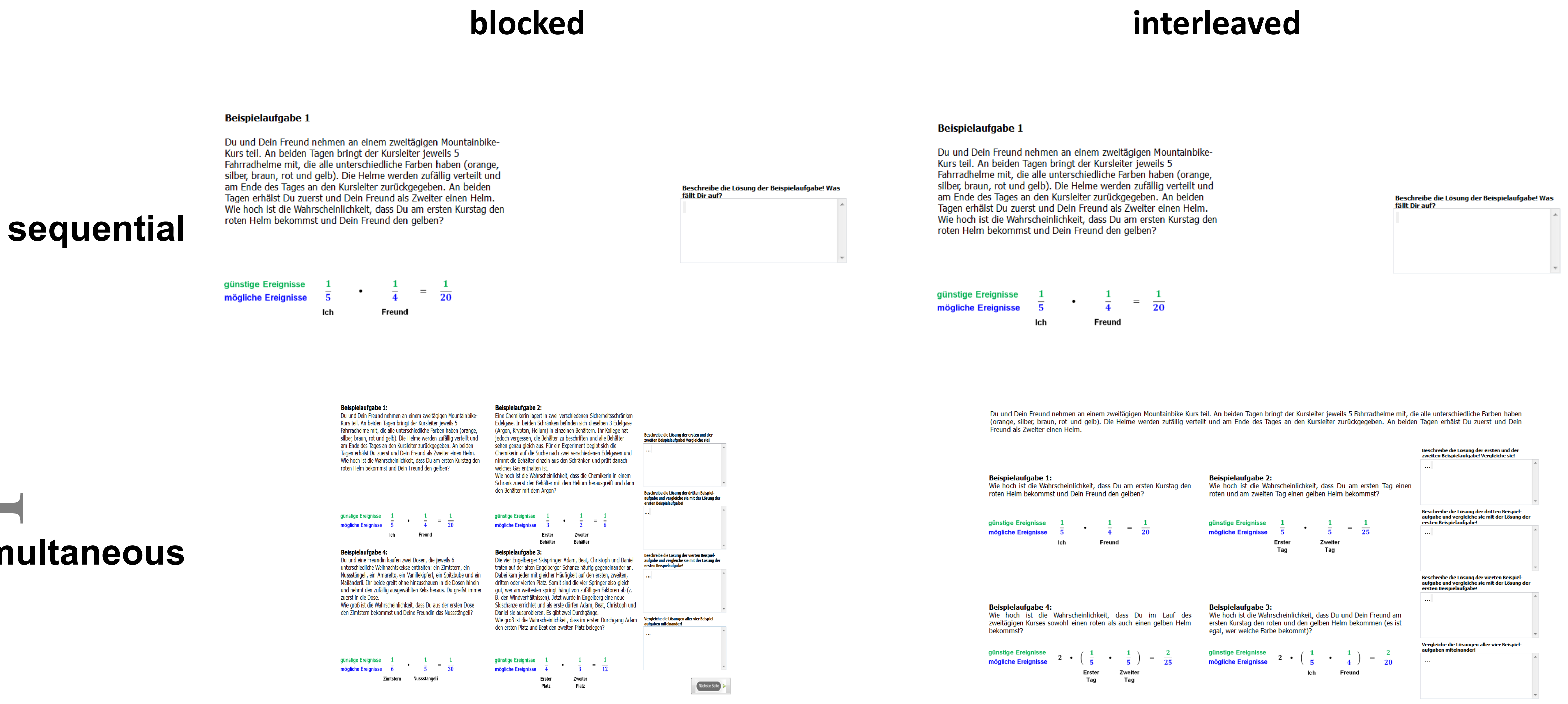
You and your friend are taking part in a two-day mountain bike course. On both days, the instructor will bring 5 bike helmets, each with different colors (orange, silver, brown, red and yellow). The helmets are distributed randomly and returned to the instructor at the end of the day. On both days you will receive a helmet first and your friend second.		

	without replacement	**with replacement**
order relevant	What is the probability that you will get the red helmet on the first day of the course and your friend the yellow one? 1/5 * 1/4 = 1/20	What is the probability that you will get a red helmet on the first day and a yellow helmet on the second day? 1/5 * 1/5 = 1/25
order irrelevant	What is the probability that you and your friend will get the red and yellow helmet on the first day of the course (it doesn't matter who gets which color)? 2* (1/5 * 1/4) = 2/20	What is the probability that you will get both a red and a yellow helmet during the two-day course? 2 * (1/5 * 1/5) = 2/25


blocked
interleaved
sequential
simultaneous
Beispielaufgabe 1
Du und Dein Freund nehmen an einem zweitägigen Mountainbike-Kurs teil. An beiden Tagen bringt der Kursleiter jeweils 5 Fahrradhelme mit, die alle unterschiedliche Farben haben (orange, silber, braun, rot und gelb). Die Helme werden zufällig verteilt und am Ende des Tages an den Kursleiter zurückgegeben. An beiden Tagen erhälst Du zuerst und Dein Freund als Zweiter einen Helm. Wie hoch ist die Wahrscheinlichkeit, dass Du am ersten Kurstag den roten Helm bekommst und Dein Freund den gelben?
Beschreibe die Lösung der Beispielaufgabe! Was fällt Dir auf?
günstige Ereignisse
mögliche Ereignisse
Ich
Freund

You and your friend are taking part in a two-day mountain bike course. On both days, the instructor will bring 5 bike helmets, each with different colors (orange, silver, brown, red and yellow). The helmets are distributed randomly and returned to the instructor at the end of the day. On both days you will receive a helmet first and your friend second.

	without replacement	**with replacement**
order relevant	What is the probability that you will get the red helmet on the first day of the course and your friend the yellow one? 1/5 * 1/4 = 1/20	What is the probability that you will get a red helmet on the first day and a yellow helmet on the second day? 1/5 * 1/5 = 1/25
order irrelevant	What is the probability that you and your friend will get the red and yellow helmet on the first day of the course (it doesn't matter who gets which color)? 2* (1/5 * 1/4) = 2/20	What is the probability that you will get both a red and a yellow helmet during the two-day course? 2 * (1/5 * 1/5) = 2/25


blocked
interleaved
sequential
simultaneous
Beispielaufgabe 1
Du und Dein Freund nehmen an einem zweitägigen Mountainbike-Kurs teil. An beiden Tagen bringt der Kursleiter jeweils 5 Fahrradhelme mit, die alle unterschiedliche Farben haben (orange, silber, braun, rot und gelb). Die Helme werden zufällig verteilt und am Ende des Tages an den Kursleiter zurückgegeben. An beiden Tagen erhälst Du zuerst und Dein Freund als Zweiter einen Helm. Wie hoch ist die Wahrscheinlichkeit, dass Du am ersten Kurstag den roten Helm bekommst und Dein Freund den gelben?
Beschreibe die Lösung der Beispielaufgabe! Was fällt Dir auf?
günstige Ereignisse
mögliche Ereignisse
Ich
Freund
Beispielaufgabe 2:
Wie hoch ist die Wahrscheinlichkeit, dass Du am ersten Tag einen roten und am zweiten Tag einen gelben Helm bekommst?
Beispielaufgabe 3:
Wie hoch ist die Wahrscheinlichkeit, dass Du und Dein Freund am ersten Kurstag den roten und den gelben Helm bekommen (es ist egal, wer welche Farbe bekommt)?
Beispielaufgabe 4:
Wie hoch ist die Wahrscheinlichkeit, dass Du im Lauf des zweitägigen Kurses sowohl einen roten als auch einen gelben Helm bekommst?
Beschreibe die Lösung der ersten und der zweiten Beispielaufgabe! Vergleiche sie!
Beschreibe die Lösung der dritten Beispielaufgabe und vergleiche sie mit der Lösung der ersten Beispielaufgabe!
Beschreibe die Lösung der vierten Beispielaufgabe und vergleiche sie mit der Lösung der ersten Beispielaufgabe!
Vergleiche die Lösungen aller vier Beispielaufgaben miteinander!
Erster Tag
Zweiter Tag