Educ Psychol Rev (2016) 28:649–672
DOI 10.1007/s10648-016-9377-z

REPLICATION

# How Effective is Example Generation for Learning Declarative Concepts?

Katherine A. Rawson[1] · John Dunlosky[1]

Published online: 24 June 2016

**Abstract** Declarative concepts (i.e., key terms and corresponding definitions for abstract concepts) represent foundational knowledge that students learn in many content domains. Thus, investigating techniques to enhance concept learning is of critical importance. Various theoretical accounts support the expectation that example generation will serve this purpose, but few studies have examined the efficacy of this technique. We conducted three experiments involving 487 undergraduates to investigate the effects of example generation on concept learning and examined factors that may moderate its effectiveness. Students read a short text that introduced eight concepts. Some students were then prompted to generate concrete examples of each concept followed by definition restudy, whereas others only restudied definitions for the same amount of time. Two days later, students completed final tests involving example generation and definition cued recall. Meta-analytic outcomes indicated that example generation yields moderate improvements in learning of declarative concepts, relative to restudy only. Each experiment also included additional groups to investigate potential moderators. Example generation tended to be more effective with spaced versus massed restudy. Despite strong correlations between the quality of examples generated during practice and final test performance, experimental manipulations that improved example quality did not improve learning. In sum, the current work establishes that example generation enhances concept learning and provides an important foundation for further investigating factors that moderate its benefits to learning.

**Keywords** Example generation · Declarative concepts · Concept learning · Restudy

A common learning objective in many courses is for students to learn *declarative concepts*, which are key terms and corresponding definitions for abstract concepts (Rawson et al. 2015). By definition, a declarative concept has some level of abstraction that represents a type-token

✉ Katherine A. Rawson
krawson1@kent.edu

[1] Department of Psychological Sciences, Kent State University, P.O. Box 5190, Kent, OH 44242-0001, USA

relationship with specific situations or events to which that concept may be applied. Indeed, one reason that declarative concepts feature prominently in many courses is that they are important for understanding and improving outcomes in real-world contexts. For example, the concept of *confirmation bias* (i.e., people's tendency to seek out or attend to information that confirms a belief and to discount or ignore counterevidence) can be instantiated in contexts as wide ranging as financial investing, medical decision-making, scientific reasoning, politics, law, and so on. Given that declarative concepts represent a core component of foundational knowledge that students are expected to learn in many domains, exploring effective techniques to promote understanding and application of declarative concepts is of critical importance.

One learning technique that may support the acquisition of conceptual knowledge is *example generation*, in which students generate specific concrete examples of the abstract declarative concepts. Exploring example generation is important for practical purposes, given that it involves a relatively easy-to-use technique that could be broadly applicable to supporting learning for a wide range of learners and topics. Additionally, example generation may support various kinds of learning outcomes, from memory and understanding of concept meanings to transfer for application in novel contexts and on other transfer dimensions (Barnett and Ceci 2002). Finally, students report using example generation as a learning technique while they are studying (Gurung 2005; Gurung et al. 2010; Weinstein et al. 2013). Despite the practical relevance of example generation as a potential means for enhancing concept learning, few studies have examined the effectiveness of this technique or moderators of its effects.

Accordingly, the purpose of the current research was to investigate the effects of example generation on concept learning, as well as factors that may enhance or constrain the effectiveness of this technique. Below, we first summarize available evidence from prior research exploring the influence of example generation on concept learning. We then report three experiments that systematically examine the extent to which example generation enhances learning of declarative concepts.

## Empirical Evidence for the Effectiveness of Example Generation

In the seminal study of example generation, Hamilton (1989) presented undergraduates with an expository text on principles of operant conditioning that included three introductory paragraphs and then four target paragraphs, one for each of four target key concepts (positive and negative reinforcement, positive and negative punishment). Each paragraph defined the concept, provided four examples of the concept, and explained how each example illustrated the concept. After each paragraph, students answered adjunct multiple-choice questions tapping definition recognition and concept application. Participants in the example generation group were also then asked to write down two personal examples of each concept. All participants completed final tests immediately after the learning phase. The example generation group outperformed the control group on a problem-solving test in which various classroom scenarios were described and students were prompted to apply the concepts to address disruptive student behavior. However, the example generation and control groups did not differ on other final test measures, including recall of the concept definitions, recall of the provided examples, or classification of new examples. Hamilton (1999) used the same basic procedure and did not find any advantage of example generation over the no-generation control group on any measure. In subsequent studies (Hamilton 1990, 1997, 1999, 2004), performance on some final test measures was lower for example generation versus other learning techniques, including studying additional provided examples or comparing and contrasting concepts.

In sum, outcomes of Hamilton's studies do not support the expectation that example generation would enhance learning of declarative concepts. However, potential methodological limitations of these studies should be noted. For example, all studies involved a limited number and range of concepts (the same four concepts from operant conditioning), and some of the studies involved small samples of participants. Furthermore, final tests were always administered immediately after learning, and research on other learning techniques have shown that effects that are not apparent on immediate tests often emerge on delayed tests (e.g., the effects of spacing and self-testing are often not apparent and can even show a reversal on immediate tests, whereas sizeable advantages of spacing over massing and self-testing over restudy consistently emerge on delayed tests; e.g., Chan 2009; Kornell et al. 2011; Rawson 2012; Rawson and Kintsch 2005; Roediger and Karpicke 2006). Perhaps of greatest concern, the initial instructional materials provided an extensive set of examples for each concept, leaving open the possibility that students reproduced the examples provided in the base materials rather than attempting to generate their own examples.

Only two other published studies have examined the effects of example generation on learning of declarative concepts. During a 3-week unit on energy that included 26 key concepts (e.g., convection, conduction), Gorrell et al. (1991) had fifth-grade students complete a homework assignment in which they were instructed to generate an example for half of the concepts. On final multiple-choice tests administered at the end of the unit and again 6 months later, pretest-posttest improvement was greater for example generation items than for control items. However, these outcomes should be interpreted with some caution, because parents were permitted to help students with their homework assignment (i.e., some examples may functionally have been provided rather than generated). Finally, Dornisch et al. (2011) had undergraduates in an educational psychology class read a 3000-word text with prompts inserted at various points. Each prompt asked the student to generate an example of a concept, provided the student with an additional example of a concept, or asked the student to answer an elaborative interrogation question (explaining why a fact or situation might be true). No performance differences were observed on the final tests (including matching, multiple-choice, open-ended, and factual recall questions) that were administered immediately after learning and 1 week later. However, the dosage of practice may have been insufficient to produce effects, given that the lengthy 3000-word text only included 13 prompts (approximately one after every two paragraphs).

In sum, few studies have examined the effectiveness of example generation, outcomes are mixed, and interpretive difficulties arise due to methodological limitations. Given the practical appeal of example generation as a learning technique, further research is warranted.

## Overview of the Current Experiments

We conducted three experiments to systematically examine the extent to which example generation enhances learning of declarative concepts. Students first read a short textbook excerpt that introduced eight key concepts and then studied each of the key concepts in isolation. In the example generation group, the concepts were then presented one at a time for students to generate a concrete example that illustrated that concept, followed by an opportunity to restudy the definition. In the restudy-only group, the concepts were presented one at a time with the definition to restudy.[1] Two days later, students completed three final tests (example generation, definition cued recall, and multiple-choice application). Performance on the example generation test was of primary interest, as it provides strong conditions under which a benefit of example generation would be expected to emerge. All three experiments included this basic design, motivated by recent emphasis in the field on the importance of replication and recommendations that conclusions be based on cumulative outcomes involving multiple estimates of effect sizes (e.g., Braver et al. 2014; Lishner 2015; Maner 2014; Pashler and Harris 2012; Simons 2014).

In addition to the basic design (which was intended to provide estimates of the effectiveness of example generation relative to restudy only), each experiment also included other groups to investigate potential moderators of the effects of example generation on concept learning. To foreshadow, Experiment 1 examined whether the advantage of example generation over restudy is moderated by the timing of restudy. Prior research has shown that spaced versus massed restudy after other generative tasks can benefit learning (e.g., Butler et al. 2007; Butler and Roediger 2008; Pashler et al. 2007). All three experiments examined whether the effects of example generation are moderated by the quality of the examples generated. Only one prior study has examined example quality. Hamilton (1990) reported significant correlations of example quality with problem-solving performance and definition recall, which suggests that the benefit of example generation may depend on the quality of examples generated.

## Experiment 1

The primary purpose of Experiment 1 was to examine the extent to which example generation enhances declarative concept learning, relative to restudy only. Given the paucity of prior research on example generation in general and on potential moderators in particular, Experiment 1 also included additional groups to begin exploring conditions under which it may work best. Namely, we manipulated the timing of restudy (massed versus spaced), such that the opportunity to restudy the concept definition occurred either immediately after example generation for a particular concept or at a short delay (i.e., after the learner had generated an example for each concept). For purposes of comparison, the distribution of study time for the restudy-only group was also manipulated in a similar manner.

### Methods

**Participants and Design** Undergraduates who participated for course credit ($n = 133$) were randomly assigned to one of four groups defined by the factorial combination of kind of practice trial (example generation versus restudy only) and the timing of the restudy opportunity that followed each practice trial (spaced versus massed). Outcomes reported below are based on data from 110 participants who completed both sessions (23 did not return for session 2, 10 of these due to inclement weather). A power analysis conducted using G*Power 3.1.9.2

[1] We selected definition restudy as the activity for the comparison group (as opposed to no practice or some other encoding strategy) because it arguably provides the most appropriate business-as-usual comparison, given that restudy is the technique that students often report using most frequently during self-regulated learning (e.g., Hartwig and Dunlosky 2012; Karpicke et al. 2009; Susser and McCabe 2013). To foreshadow, experiments 2 and 3 also afforded comparison of example generation to another learning technique (retrieval practice).

(Faul et al. 2007) for a factorial ANOVA with power set at 0.80 and $\alpha = 0.05$ indicated that this sample size afforded sufficient sensitivity to detect medium-sized effects ($f = 0.27$).

**Materials and Procedure** We used materials developed by Rawson and Dunlosky (2011), which included a short instructional text adapted from General Psychology textbooks that defined and described eight key concepts about social attribution (e.g., "The just-world hypothesis is the strong desire or need people have to believe that the world is an orderly, predictable, and just place, where people get what they deserve"). We modified the instructional text for present purposes by removing any examples that were provided to illustrate the concepts (to eliminate the potential for students to simply reproduce provided examples instead of generating their own examples). An excerpt from the text is presented in the Appendix.

In both sessions, all instructions and tasks were administered via computer. Session 1 began with general task instructions. All participants were told that they would be asked to learn eight concepts and would be tested on their learning when they returned 2 days later (the nature of the final test was not specified). Participants were then presented with the instructional text for self-paced study. Next, the eight concepts were presented one at a time with the corresponding definition for a self-paced study trial.[2]

Participants were then given task instructions for the practice phase. Participants in the example-generation group were told that on each practice trial, they would have 45 s to come up with a real-world example for each concept. They were told to do their best to come up with an example that illustrated all the important ideas in the definition of the concept. They were also shown a sample concept definition from an unrelated topic (i.e., negative reinforcement) and a sample real-world example of the concept. Participants were also told that they would have opportunities to restudy each concept definition for 15 s each. Participants were told that they would repeat this cycle three times and that they should try to come up with a different example each time; the use of three repetitions was meant to give students sufficient practice to yield benefits without making the learning session too onerous. In the restudy-only group, participants were told that they would be presented with each concept definition to study six more times, and that some trials would last 45 s and others would last 15 s.

For all groups in the subsequent practice phase, each concept was presented for three practice trials and three restudy trials. A schematic of the sequence of practice trials and restudy trials is presented in Table 1. Each practice trial was 45 s and each restudy trial was 15 s; on each trial, a statement at the bottom of the screen informed participants about the allotted time for that trial. On practice trials in the example-generation group, participants were presented with a key concept term at the top of the screen along with a text field and the prompt to type in a real-world example that illustrated the meaning of the concept. On practice trials in the restudy-only group, a key concept term was presented at the top of the screen with the definition in a separate field below. On restudy trials in both groups, a key concept term was presented at the top of the screen with the definition in a separate field below.

[2] On average, participants spent 2.7 min (SE = 0.1) studying the text and 13.3 s (SE = 0.7) per concept definition; similar outcomes were observed in experiments 2 and 3 [text M = 2.8 min (SE = 0.1) for the text and M = 12.7 s (SE = 0.5) per concept definition in experiment 2; M = 2.7 min (SE = 0.1) for the text and M = 11.6 s (SE = 0.5) per concept definition in experiment 3]. Neither text study time nor concept study time differed significantly as a function of group in any experiment, and including these variables as covariates in analyses of final test performance did not qualitatively change any statistical conclusions.

**Table 1** Schematic of the sequence of trials for concepts 1–8 in each block of the practice phase of Experiment 1

| Massed restudy | Spaced restudy |
|---|---|
| Generate example or study concept 1 (45 s) | Generate example or study concept 1 (45 s) |
| Restudy concept 1 (15 s) | Generate example or study concept 2 (45 s) |
| Generate example or study concept 2 (45 s) | Generate example or study concept 3 (45 s) |
| Restudy concept 2 (15 s) | … |
| Generate example or study concept 3 (45 s) | Generate example or study concept 8 (45 s) |
| Restudy concept 3 (15 s) | Restudy concept 1 (15 s) |
| … | Restudy concept 2 (15 s) |
| Generate example or study concept 8 (45 s) | Restudy concept 3 (15 s) |
| Restudy concept 8 (15 s) | … |
| | Restudy concept 8 (15 s) |

In both groups, practice trials were presented in blocks, with each concept presented for one practice trial in each of three blocks (i.e., the sequence of trials shown in Table 1 was repeated three times). The same fixed item order was used in all blocks. For participants assigned to receive massed restudy, the restudy trial followed immediately after the practice trial for a given item within each block. For participants assigned to receive spaced restudy, all eight concepts were presented for a block of practice trials, followed by a separate block of one restudy trial for each concept, followed by the next block of practice trials, and so on.

Participants returned 2 days later to complete three final test measures, all of which were self-paced. On the example generation test, concept terms were presented one at a time and participants were prompted to type in a real-world example of each one (test instructions and task administration were similar to those used in the example generation group during session 1). On the cued recall test, concept terms were presented one at a time and participants were prompted to type in the definition of the concept. Participants were instructed that their responses did not need to be verbatim and that "you can use your own words, as long as you state the correct meaning of the concept. Do your best to recall as much of the meaning of each definition as possible." The order in which these two tests were completed was counterbalanced across participants in each group (across experiments, test order was not systematically related to patterns of performance on these two measures and thus, we do not discuss this nuisance variable further). For all participants, the last test included five alternative multiple-choice questions, with one question per concept. All questions were example-based comprehension questions, either presenting a scenario and asking students to identify which of five concepts it illustrated or presenting a concept term and asking students to identify which of five scenarios illustrated that concept.

**Scoring** In all three experiments, participants' example generation responses and cued recall responses were scored by trained raters. Each cued recall response was scored by one of the trained raters and was assigned a recall score based on the percentage of main ideas from the definition that the response contained. Responses were scored as correct if they included either verbatim restatements or paraphrases that preserved the meaning of the definition. Each example generation response was assigned a score of no credit, partial credit, or full credit (with corresponding values of 0, 50, and 100), based on the extent to which the example correctly illustrated the key components of the concept. Due to the much broader range of possible responses in the example generation task than in the cued recall test, all of the example generation responses were scored by two raters, and the two scores for each response were averaged. Interrater reliability across the sets of scores for each block of practice and for the final test in each experiment was acceptable (mean $\alpha = 0.84$, range 0.65–0.94 for practice blocks and $\alpha = 0.85$, 0.89, and 0.90 for final test).

## Results and Discussion

For the final test measures, split-half reliability was adequate for example generation ($\alpha = 0.69$) and for cued recall ($\alpha = 0.76$). Reliability was lower for the multiple-choice test ($\alpha = 0.45$). Outcomes for this measure must be interpreted with some caution, given that this secondary measure was administered after the example generation and cued recall tests for all participants. Results for this measure are reported in Table 2 for purposes of full reporting of outcomes (see Simmons et al. 2011), but we do not discuss them further. Cohen's *d* values were computed using pooled standard deviations (Cortina and Nouri 2000).

The outcomes of primary interest are reported in the left panel of Fig. 1. Performance on the final example generation test was greater when practice involved example generation versus restudy only, $F(1,106) = 4.94$, MSE = 277.65, $p = 0.028$, $\eta_p^2 = 0.04$. Generating examples during practice also improved performance on the final cued recall test, $F(1,106) = 14.07$,

**Table 2** Mean percent correct (and standard error) on the multiple-choice comprehension test in Experiments 1–3

| | M (SE) |
|---|---|
| Experiment 1 | |
| Restudy only (spaced) | 54 (4) |
| Example generation (spaced restudy) | 65 (3) |
| Restudy only (massed) | 49 (5) |
| Example generation (massed restudy) | 54 (4) |
| Main effect, kind of practice trial: $F(1,106) = 3.85$, $p = 0.052$, $\eta_p^2 = 0.035$ | |
| Main effect, timing of restudy: $F(1,106) = 3.85$, $p = 0.052$, $\eta_p^2 = 0.035$ | |
| Interaction: $F(1,106) < 1$, $\eta_p^2 = 0.005$ | |
| Experiment 2 | |
| Restudy only | 53 (3) |
| Example generation | 55 (4) |
| Open-book example generation | 61 (3) |
| Recall plus generation | 58 (4) |
| Recall only | 61 (4) |
| Main effect of group: $F(4,149) < 1$, $\eta_p^2 = 0.026$ | |
| Experiment 3 | |
| Restudy only | 60 (3) |
| Example generation | 61 (3) |
| Self-paced example generation | 59 (3) |
| Self-paced recall plus generation | 62 (3) |
| Self-paced recall only | 62 (3) |
| Main effect of group: $F(4,174) < 1$, $\eta_p^2 = 0.005$ | |

MSE = 328.62, $p < 0.001$, $\eta_p^2 = 0.12$; this outcome is perhaps surprising, given that the overall amount of time that the definitions were presented during the practice phase was greater in the restudy-only group than in the example generation group.

Concerning timing of restudy as a potential moderator, the benefit of example generation tended to be stronger when restudy of a definition was spaced versus massed after example generation (see estimates of effect size in Fig. 1), although the interaction was not significant for either final test ($F$s < 1.55). A follow-up test revealed significantly greater performance on the example generation final test for the example-generation group with spaced restudy versus the example-generation group with massed restudy, $t(53) = 2.07$, $p = 0.022$, $d = 0.56$.

Finally, we also explored the extent to which the benefit of example generation is related to the quality of examples generated during practice (mean performance on each example generation trial during practice is reported in Table 3). Collapsing across practice trials, the quality of examples generated during practice correlated strongly with performance on the example generation final test ($r = 0.75$, $p < 0.001$). Example quality also correlated with performance on the cued recall test ($r = 0.60$, $p < 0.001$). Bearing in mind the correlational nature of these outcomes, these findings are consistent with a *quality affects learning* hypothesis, which states that improving the quality of examples generated during practice will further enhance the benefit of example generation for declarative concept learning. The relatively modest level of example generation performance during practice certainly leaves room for improvement.

## Experiment 2

Experiment 2 was designed (a) to replicate the key outcomes of Experiment 1 demonstrating that example generation enhances declarative concept learning and (b) to extend them to further investigate example quality as a moderator of this effect. Concerning replication, Experiment 2 included two of the groups included in Experiment 1 (example generation with spaced restudy and the corresponding restudy-only comparison group).[3] Concerning extensions, Experiment 2 included three new groups, to answer the following questions: Why was the quality of examples generated during practice modest at best? And will support for generating higher-quality examples during practice in turn improve declarative concept learning? During example generation trials in Experiment 1, the definition of the concept was not presented until after example generation had been completed. Thus, learners would need to retrieve the meaning of a concept from memory when attempting to generate an example of that concept. If so, the quality of learners' examples may have been limited to the extent that they did not fully attempt to retrieve the meanings of the concepts or to the extent that they were unable to fully retrieve concept meanings. We briefly consider each of these possibilities in turn.

First, example quality may have been limited because learners did not fully attempt retrieval (hereafter referred to as the *retrieval likelihood hypothesis*). If so, example quality will be enhanced if learners are explicitly prompted to do so. To evaluate this hypothesis, Experiment 2

[3] Given that timing of restudy did not significantly moderate the benefit of example generation over restudy-only in experiment 1, we dropped the timing manipulation from experiment 2 to keep the design from becoming too unwieldy with the addition of other extension groups. Given the advantage of example generation involving spaced versus massed restudy, we used spaced restudy for all groups in experiments 2 and 3.

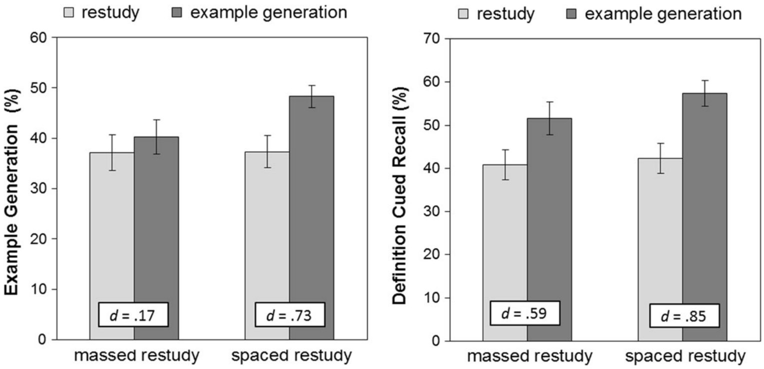


**Fig. 1** Performance on the final example generation test and the final cued recall test in Experiment 1. *Error bars* report standard error of the mean. Embedded values are Cohen's *d* for paired comparisons of interest

included a *recall-plus-generation* group in which learners were prompted to overtly retrieve a definition for a concept that then remained on the screen during example generation. This hypothesis predicts greater example quality in this group versus in the example-generation group.

Second, even if learners spontaneously attempt to retrieve definitions during example generation, they may be unable to fully retrieve some concept definitions, which in turn would limit the quality of their examples (hereafter referred to as the *retrieval failure hypothesis*). To estimate how well learners can retrieve concept definitions during practice, Experiment 2 included a *recall-only* group who retrieved definitions during practice. Inclusion of this group also afforded comparison of example generation to a more competitive learning technique—retrieval practice—that prior research has established as a potent strategy for enhancing learning of various kinds of material (for a recent review, see Rowland 2014).

Experiment 2 also included an *open-book generation* group in which the concept definition was provided on each example generation trial, to estimate the extent to which having the full definition available during example generation would enhance the quality of the examples generated. With the full definition available during example generation,

**Table 3** Mean percent correct (and standard error) on example generation trials during practice in Experiments 1–3

| | Trial 1 | Trial 2 | Trial 3 |
|---|---|---|---|
| Experiment 1 | | | |
| Example generation (massed restudy) | 30 (3) | 38 (3) | 40 (4) |
| Example generation (spaced restudy) | 36 (3) | 46 (3) | 49 (4) |
| Experiment 2 | | | |
| Example generation | 32 (4) | 36 (4) | 39 (3) |
| Open-book generation | 42 (4) | 47 (4) | 45 (3) |
| Recall plus generation | 16 (3) | 19 (3) | 23 (4) |
| Experiment 3 | | | |
| Example generation | 33 (4) | 40 (3) | 44 (3) |
| Self-paced example generation | 36 (4) | 43 (4) | 46 (4) |
| Self-paced recall plus generation | 37 (3) | 43 (4) | 43 (4) |

we predicted that learners would generate higher-quality examples than in the other two example generation groups. The quality of example generation under open-book conditions is also of interest for practical purposes; in naturalistic contexts, students would likely be able to refer back to their textbook or notes when engaging in example generation during independent study.

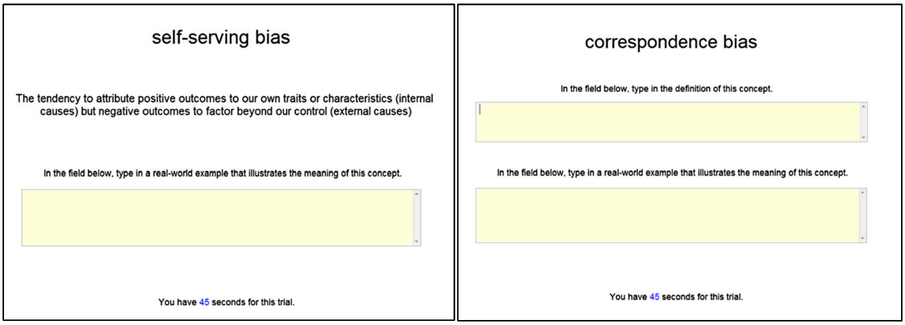
Finally, the extension groups also afford evaluation of the quality-affects-learning (QAL) hypothesis, which states that improving the quality of examples generated during practice will further enhance the benefit of example generation for declarative concept learning. To the extent that example quality is improved by prompting retrieval in the recall-plus-generation group or by providing the full definition in the open-book generation group, the QAL hypothesis predicts that final test performance will be greater in these groups than in the example generation group.

## Methods

**Participants and Design** Undergraduates who participated for course credit ($n = 163$) were randomly assigned to one of five groups defined by the kind of trial administered during practice (restudy only, example generation, open-book generation, recall plus generation, recall only). Outcomes reported below are based on data from 154 participants who completed both sessions. A power analysis for a one-way ANOVA with power set at 0.80 and $\alpha = 0.05$ indicated that this sample size afforded sufficient sensitivity to detect medium-sized effects ($f = 0.28$).

**Materials and Procedure** Materials were the same as in Experiment 1, except that we increased the number of questions on the final multiple-choice test from 8 to 16 (2 questions per concept), in an attempt to improve the internal reliability of the measure.

Procedures for the restudy-only and example-generation groups were the same as in Experiment 1 and involved spaced restudy as depicted in the second column of Table 1. The schedule of trials and procedure for the open-book generation group was the same as for the example generation group, except that the definition of the concept was presented on the screen during each example generation trial (see screenshot in left panel of Fig. 2). The schedule of trials and procedure for the recall-plus-generation group was the same as for the example generation group, except that each example generation trial also included a prompt to



self-serving bias

The tendency to attribute positive outcomes to our own traits or characteristics (internal causes) but negative outcomes to factor beyond our control (external causes)

In the field below, type in a real-world example that illustrates the meaning of this concept.

You have 45 seconds for this trial.

correspondence bias

In the field below, type in the definition of this concept.

In the field below, type in a real-world example that illustrates the meaning of this concept.

You have 45 seconds for this trial.

**Fig. 2** Screenshots illustrating the practice trials administered in the open-book generation group (*left panel*) and recall-plus-generation group (*right panel*) in Experiment 2

overtly recall the definition of each concept (see screenshot in right panel of Fig. 2). The prompts and editable fields for both tasks remained on the screen for the entire trial duration. Finally, the schedule of trials and procedure for the recall-only group was also the same, except that each 45-s practice trial only included a prompt to overtly recall the definition of the concept.

## Results and Discussion

For the final test measures, split-half reliability was acceptable for example generation ($\alpha = 0.75$), cued recall ($\alpha = 0.89$), and multiple-choice comprehension ($\alpha = 0.71$). Analyses below focus on planned comparisons appropriate for evaluating questions of primary interest, and we report one-tailed $p$ values when testing a priori directional predictions (Judd and McClelland 1989; Maner 2014).

**Does Example Generation Enhance Concept Learning?** For purposes of replication, the planned comparisons of primary interest involve the restudy group and the example generation group (i.e., the two leftmost bars in each panel of Fig. 3). As in Experiment 1, generating examples during practice (versus restudy only) improved performance on the final example generation test, $t(57) = 2.15$, $p = 0.018$, $d = 0.56$. Generating examples during practice also improved performance on the final cued recall test, $t(57) = 1.83$, $p = 0.036$, $d = 0.48$, despite the fact that the restudy-only group spent more time overall studying the definitions during the practice phase. These outcomes provide additional estimates of the effect of example generation on declarative concept learning.

Also, similar to Experiment 1, the quality of examples generated during practice was modest at best (see Table 3), which may have limited the benefit of example generation. Collapsing across practice trials, the quality of examples generated during practice correlated strongly with performance on the final example generation test ($r = 0.88$, $p < 0.001$; collapsing across all three example generation groups, $r = 0.67$, $p < 0.001$). Example quality also correlated with performance on the cued recall test ($r = 0.51$, $p = 0.004$; collapsing across all three example generation groups, $r = 0.20$, $p = 0.049$).

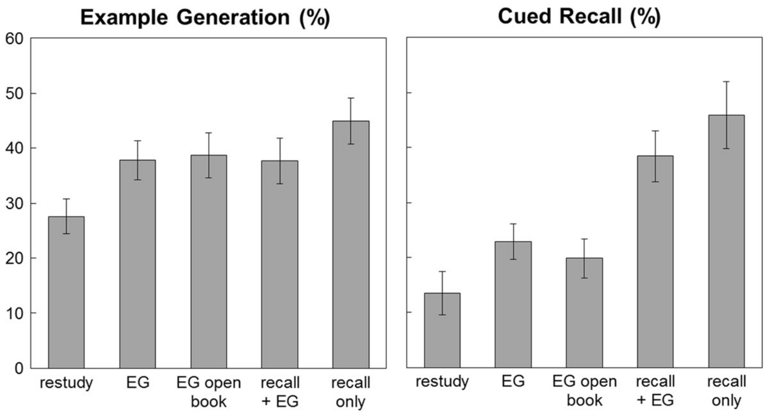


**Fig. 3** Performance on the final example generation test and the final cued recall test in Experiment 2. *EG* example generation during practice trials. *Error bars* report standard error of the mean

**Why was Example Quality During Practice Limited?** According to the retrieval-likelihood hypothesis, example quality is limited because learners do not fully attempt retrieval of concept definitions during example generation. This hypothesis predicts that example quality will be enhanced if learners are explicitly prompted to retrieve each definition. Contrary to this expectation, example quality was lower in the recall-plus-generation group than in the example generation group [see Table 3; collapsed across trials, $t(60) = 3.63$, $d = 0.92$]. The mundane explanation for this unexpected outcome is that the 45 s allotted on each trial was insufficient for participants in the recall-plus-generation group to complete both tasks; that is, after attempting to retrieve the definition, participants likely had minimal time left for generating an example. To foreshadow, Experiment 3 included groups to remedy this methodological limitation to afford better evaluation of the retrieval-likelihood hypothesis.

Even if learners spontaneously attempt to retrieve definitions during example generation, the retrieval-failure hypothesis states that the quality of their examples may nonetheless be limited to the extent that they are unable to fully retrieve the meaning of the concept. Consistent with this hypothesis, the level of successful recall during practice for the recall-only group was relatively modest (see Table 4). Also consistent with this hypothesis, providing the full definition during example generation enhanced the quality of the examples that learners generated (see Table 3). Collapsing across trials, example quality was greater in the open-book generation group than in the example-generation group [45 versus 36 %, $t(61) = 1.93$, $p = 0.03$, $d = 0.49$].

**Does Improving Example Quality Enhance Concept Learning?** The quality-affects-learning (QAL) hypothesis states that improving the quality of examples generated during practice will enhance the benefit of example generation for declarative concept learning. Given that providing the full definition in the open-book generation group improved example quality, the QAL hypothesis predicts that final test performance will be greater in this group than in the example-generation group. Disconfirming this prediction, enhancing the quality of examples during practice did not yield a concomitant improvement in final test performance (Fig. 3): Performance on the final example generation test was similar for the open-book generation group and the example-generation group [$t(61) = 0.16$, $d = 0.04$]; performance on the final cued recall test was numerically lower in the open-book generation group than in the example-generation group [$t(61) = 0.63$, $d = -0.16$]. As further evidence, the unexpectedly low example quality in the recall-plus-generation group serendipitously affords another test of the QAL hypothesis. Despite the lower example quality in this group, performance on the final example generation test did not differ for the recall-plus-generation group versus the example-generation group, $t(60) = 0.02$, $d = 0.01$. The contrast between the recall-plus-generation group

**Table 4** Mean percent correct (and standard error) on recall trials during practice in Experiments 2 and 3

| | Trial 1 | Trial 2 | Trial 3 |
|---|---|---|---|
| Experiment 2 | | | |
| Recall plus generation | 18 (3) | 33 (4) | 41 (5) |
| Recall only | 27 (4) | 40 (4) | 49 (5) |
| Experiment 3 | | | |
| Self-paced recall plus generation | 33 (4) | 47 (4) | 53 (5) |
| Self-paced recall only | 32 (4) | 48 (4) | 61 (4) |

and the open-book generation group is even more striking: Despite the substantial difference between these two groups in example quality during practice ($d = 1.42$), the difference in performance on the final example generation test was minimal ($d = 0.04$).

## Experiment 3

Experiment 3 was designed (a) to replicate key outcomes from Experiments 1 and 2 so as to provide additional estimates of effect sizes and (b) to further test the hypotheses concerning example quality and its effects on declarative concept learning. Concerning replication, Experiment 3 included the example-generation and restudy-only groups included in Experiments 1 and 2. As in the previous two experiments, the practice trials for these two groups were experimenter-paced (45 s per trial). Experiment 3 also included three groups similar to those in Experiment 2 (example generation, recall plus generation, and recall only), except that the practice trials were self-paced. To revisit, the 45 s allotted on each practice trial in Experiment 2 was likely insufficient for participants in the recall-plus-generation group to complete both tasks, which limited the extent to which outcomes in this group were informative for evaluating the retrieval-likelihood hypothesis. The self-paced recall-plus-generation group in Experiment 3 thus affords a stronger test of this hypothesis. The self-paced example-generation group and the self-paced recall-only group provided estimates of the time required to complete each component task.

These groups also provided other outcomes relevant to evaluating the focal hypotheses. Namely, the recall-only group afforded another test of the retrieval-failure hypothesis by examining how well learners can retrieve the meanings of the concepts during practice with unlimited time to do so. To evaluate the QAL hypothesis, the self-paced example generation group provided an appropriate comparison for the self-paced recall-plus-generation group. Namely, to the extent that example quality is improved by prompting retrieval in the self-paced recall-plus-generation group, the QAL hypothesis predicts that final test performance will be greater in the self-paced recall-plus-generation group than in the self-paced example-generation group.

### Methods

**Participants and Design** Undergraduates who participated for course credit ($n = 191$) were randomly assigned to one of five groups defined by the kind of trial administered during practice, which included two experimenter-paced groups (restudy only, example generation) and three self-paced groups (example generation, recall plus generation, recall only). Outcomes reported below are based on 179 participants with data for both sessions (eight did not return for session 2 and data were lost for four due to experimenter error). A power analysis for a one-way ANOVA with power set at 0.80 and $\alpha = 0.05$ indicated that this sample size was sufficient to detect medium-sized effects ($f = 0.26$).

**Materials and Procedure** Materials were the same as in Experiment 2. Procedures for the experimenter-paced restudy-only and example-generation groups were the same as in Experiment 2. Procedures for the three self-paced groups were the same as in the corresponding groups in Experiment 2, except no time limit was imposed for the practice trials (i.e., example generation or cued recall trials). Additionally, in the recall-plus-generation group, participants were first prompted to type in their recall of the definition. After submitting their recall response, the example generation prompt and response field were then shown; the participant's recall response remained on the screen. Restudy trials for all groups were still experimenter-paced at 15 s each.

## Results and Discussion

For the final test measures, split-half reliability was acceptable for example generation ($\alpha = 0.75$) and cued recall ($\alpha = 0.87$) but somewhat lower for multiple-choice comprehension ($\alpha = 0.67$).

**Does Example Generation Enhance Concept Learning?** For purposes of replication, the planned comparisons of primary interest involve the experimenter-paced restudy-only and example-generation groups (i.e., the two leftmost bars in each panel of Fig. 4). Experiment 3 showed minimal benefit of generating examples during practice (versus restudy only) for performance on the final example generation test, $t(71) = 0.12$, $d = 0.03$. Consistent with outcomes of Experiments 1 and 2, generating examples during practice showed a trend for improving performance on the final cued recall test, $t(71) = 1.14$, $p = 0.129$, $d = 0.27$, despite the fact that the restudy-only group spent more time overall studying the definitions.

As in the previous experiments, the quality of examples generated during practice was modest (see Table 3). Collapsing across practice trials, the quality of examples generated during practice was strongly correlated with performance on the example generation final test ($r = 0.80$, $p < 0.001$; collapsing across all three example generation groups, $r = 0.78$, $p < 0.001$). Example quality was also correlated with performance on the cued recall test ($r = 0.56$, $p < 0.001$; collapsing across all three example generation groups, $r = 0.61$, $p < 0.001$).

**Why was Example Quality During Practice Limited?** According to the retrieval-likelihood hypothesis, example quality is limited because learners do not fully attempt retrieval of concept meanings during example generation; this hypothesis predicts that example quality

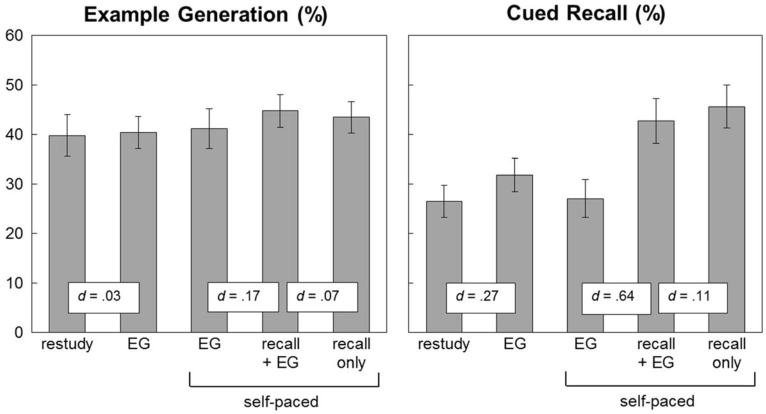


**Fig. 4** Performance on the final example generation test and the final cued recall test in Experiment 3. *EG* example generation during practice trials. *Error bars* report standard error of the mean

will be enhanced if learners are explicitly prompted to do so. Contrary to this expectation, example quality did not differ for the recall-plus-generation group versus the self-paced example generation group [see Table 3; collapsed across trials, $t(67) = 0.11$]. This outcome was obtained despite learners being given as much time as needed to complete both retrieval and example generation on each practice trial (see Table 5).[4]

Even if learners spontaneously attempt to retrieve definitions during example generation, the retrieval-failure hypothesis states that the quality of their examples may nonetheless be limited to the extent that they are unable to fully retrieve the meaning of the concept. Consistent with this hypothesis, the level of successful recall during practice for the recall-only group was relatively modest (see Table 4), even with unlimited time to complete retrieval.

**Does Improving Example Quality Enhance Concept Learning?** To revisit, the quality-affects-learning (QAL) hypothesis states that improving the quality of examples generated during practice will further enhance the benefit of example generation for declarative concept learning. Given that prompting retrieval of definitions in the recall-plus-generation group did not affect example quality, the QAL hypothesis predicts that final test performance will be similar in the self-paced example-generation group versus the recall-plus-generation group. Performance in these two groups did not significantly differ on the final example generation test [$t(67) = 0.69$, $d = 0.17$], although the recall-plus-generation group showed a significant advantage on the final cued recall test [$t(67) = 2.66$, $p = 0.005$, $d = 0.64$].

## General Discussion

Motivated by the paucity of research on example generation, the current work systematically examined the extent to which example generation enhances students' learning of declarative concepts and investigated potential moderators of its effects.

### Does Example Generation Enhance Learning?

All three experiments compared the effects of example generation to restudy only, and we observed some variability in the estimates of effect size. Such variability would be expected (e.g., Simmons et al. 2011; Stanley and Spence 2014)[5] and is one reason for the recent emphasis on the importance of replication for basing conclusions on multiple estimates of effect sizes. To this end, we adopted the *continuously cumulating meta-analysis* (CCMA) approach recommended by Braver et al. (2014), who note that the approach can be used "to combine internal replications of multi-study articles. This would be far more informative than simply reporting whether each single study succeeded or failed" (p. 340). CCMAs comparing performance for example generation versus restudy only (specifically, the experimenter-paced groups with spaced restudy that were common across all three experiments) are reported in

[4] Cognitive overload from completing both tasks was unlikely, given that learners were prompted to complete the two tasks sequentially, spent a similar amount of time for each component task as learners in the single-task groups (example generation only or recall only; see Table 5), and produced responses of similar quality as learners in the single-task groups (see Tables 3 and 4).

[5] Simmons et al. (2011) particularly recommend that the field "should be more tolerant of imperfections in results … Underpowered studies with perfect results are the ones that should invite extra scrutiny."

**Table 5** Mean seconds per task (and standard error) across practice trials in Experiment 3

| Group | Task | |
|---|---|---|
| | Example generation | Definition recall |
| Self-paced example generation | 40 (3) | n/a |
| Self-paced recall plus generation | 39 (3) | 37 (3) |
| Self-paced recall only | n/a | 31 (2) |

Table 6 for the example generation final test and for the cued recall final test. Overall, these outcomes support the conclusion that example generation yields moderate improvements in learning of declarative concepts relative to restudy (i.e., the normative business-as-usual study technique used by students). Specifically, example generation enhanced learners' ability to apply declarative concepts (example generation final test: pooled $d = 0.39$, 95 % confidence interval (CI) = 0.10–0.68) as well as memory for the meanings of the concepts (cued recall final test: pooled $d = 0.49$, 95 % CI = 0.19–0.78).

With that said, restudy is not a particularly effective learning technique as compared to more active processing techniques (for a review, see Dunlosky et al. 2013), and thus it provides a relatively conservative baseline. In contrast, comparing example generation to the recall-only group sets a higher bar, given the sizeable literature establishing retrieval practice as a particularly potent learning technique (for recent reviews, see Dunlosky et al. 2013; Rowland 2014). Example generation did not fare as well when compared to retrieval practice. Performance on the example generation final test tended to be lower after example generation than after retrieval practice (for Experiments 2–3, pooled $d = -0.20$, 95 % CI = −0.54–0.13), despite the overlap in the processes engaged during practice and final test for the example generation group. Less surprising, performance on the cued recall final test was considerably lower after example generation than after retrieval practice (for Experiments 2–3, pooled $d = -0.47$, 95 % CI = −0.81 to −0.12). Example generation joins other encoding techniques

**Table 6** Continuously cumulating meta-analysis (CCMA) outcomes for example generation versus restudy only in Experiments 1–3

| | Mean Diff | $S_{\text{pooled}}$ | $t$ | $p$ (2-tail) | Cohen's $d$ | $Z$ |
|---|---|---|---|---|---|---|
| Example generation final test: | | | | | | |
| Experiment 1 | 11 | 15 | 2.68 | 0.010 | 0.73 | 2.58 |
| Experiment 2 | 10 | 18 | 2.15 | 0.036 | 0.56 | 2.07 |
| Experiment 3 | 1 | 23 | 0.12 | 0.897 | 0.03 | 0.13 |
| CCMA results | | | | 0.009 | 0.39 | 2.76 |
| Cued recall final test: | | | | | | |
| Experiment 1 | 15 | 18 | 3.08 | 0.003 | 0.85 | 2.94 |
| Experiment 2 | 9 | 19 | 1.83 | 0.072 | 0.48 | 1.80 |
| Experiment 3 | 5 | 20 | 1.14 | 0.258 | 0.27 | 1.14 |
| CCMA results | | | | 0.001 | 0.49 | 3.39 |

Outcomes are reported for the experimenter-paced groups with spaced restudy in each experiment. Mean Diff = mean difference between groups in the percentage correct on the final test. Effect size homogeneity tests were nonsignificant for example generation [$Q(2) = 4.15$, $p = 0.13$] and for cued recall [$Q(2) = 2.33$, $p = 0.31$]

that have also been outperformed by retrieval practice (including concept mapping, elaboration, imagery, and highlighting; Coane 2013; Goossens et al. 2014; Karpicke and Blunt 2011; Neuschatz et al. 2005; Wooldridge et al. 2014).

In sum, example generation outperformed restudy but underperformed retrieval practice in comparisons of relative effectiveness. Why was the level of learning achieved with example generation intermediate between these two techniques? One possible explanation involves the extent to which the three techniques evoke retrieval practice. Example generation may evoke enough covert retrieval of concept meanings to surpass restudy, which does not afford covert retrieval (cf. Blunt and Karpicke 2014, showed that closed-book concept mapping is effective because it involves some covert retrieval practice). In contrast, example generation likely evokes less retrieval than do overt retrieval practice (as suggested by differences in performance on the final cued recall test for example-generation versus recall-plus-generation groups in Experiments 2 and 3). If so, one might ask whether example generation per se confers any benefits to learning beyond retrieval itself. Outcomes for the open-book example generation group in Experiment 2 provide evidence that example generation per se does benefit learning. The overt presentation of the concept definition precluded retrieval practice during example generation, but this group outperformed the restudy group nonetheless.

## What Moderates the Effect of Example Generation on Declarative Concept Learning?

In addition to providing estimates of the relative effectiveness of example generation for concept learning, these experiments also investigated potential moderators of the effects. Experiment 1 demonstrated trends for a stronger effect of example generation when restudy of a concept's definition was spaced rather than massed. This pattern is consistent with previous research showing benefits of spaced versus massed restudy after other generative tasks (e.g., Butler et al. 2007; Butler and Roediger 2008; Pashler et al. 2007).

The other moderator of interest was the quality of examples generated during practice, which was modest. All three experiments provided evidence relevant to evaluating the quality-affects-learning (QAL) hypothesis, which states that the effect of example generation on declarative concept learning depends on the quality of examples generated during practice. Consistent with this hypothesis, all three experiments revealed strong correlations between example quality and final test performance. These correlational outcomes motivated experimental tests of a relatively straightforward prediction—namely, that improving the quality of example generation during practice will enhance learning outcomes. Inconsistent with the QAL hypothesis, experimental manipulations that improved example quality did not improve learning. In particular, the level of example quality differed dramatically across the three groups involving example generation in Experiment 2, whereas final test performance differed minimally across these groups. These initial outcomes suggest that the benefit of example generation for concept learning may arise from the processes involved in attempting to generate an example rather than from the successful generation of an example per se. This possibility is consistent with outcomes from research on other generative tasks (e.g., elaborative interrogation) in which learning outcomes are often (but not always) related more strongly to whether a learner attempted to generate a response than to the quality of the generated response (e.g., Woloshyn and Stockley 1995; Wood et al. 1994). With that said, more research is needed to further investigate the potential contribution of example quality to the effects of example generation on concept learning.

## Theoretical Implications

Given the paucity of research on example generation, it is perhaps not surprising that no theoretical accounts have been proposed as direct explanations of the effects of example generation. However, various broader theoretical frameworks may provide insight into the possible bases of the effects of example generation. These accounts are not mutually exclusive and together suggest that example generation may have multiple routes to enhancing conceptual knowledge. We briefly consider each of these perspectives.

According to schema-based theories (e.g., Chi et al. 1981, 2012; Gick and Holyoak 1983; Kalyuga et al. 2001; Paas and Van Merriënboer 1994), schemata are conceptualized as "cognitive structures that enable problem solvers to recognize problems as belonging to particular categories requiring particular operations to reach a solution" (Paas and Van Merriënboer 1994, p. 123). Schemata are assumed to support successful comprehension and application in many domains, including reading, mathematics, and problem solving. Importantly, schema theory can also be applied to students' representations of abstract concepts, with the idea that example generation may facilitate learning by supporting schema acquisition or schema acquisition. Example generation may support schema acquisition to the extent that it encourages students to go beyond mere encoding of a definition for a concept to develop a knowledge structure that represents how the components of the concept are related to another and the kinds of problems or situations to which the concept applies. Example generation may support schema application to the extent that it provides students with practice at applying an abstract concept to concrete scenarios, which may further strengthen their understanding of how the concept is applied in appropriate situational categories. The moderate advantage of example generation over restudy only on the example generation final test is consistent with these possibilities.

The central tenet of the transfer-appropriate-processing (TAP) framework is that test performance increases as a function of the match between encoding processes and the processes required by the test (for evidence and reviews, see Blaxton 1989; Roediger 1990; for applications to text materials, see Einstein et al. 1990). As applied here, the TAP framework suggests that the benefit of example generation over restudy on the example generation final test reflects greater overlap in the processes engaged during learning and test. The TAP framework also accounts for the advantage of the recall-plus-generation and recall-only groups over the remaining groups on final cued recall tests in Experiments 2–3. Although these particular outcomes are consistent with TAP, the overall pattern of outcomes is more difficult for TAP to explain. For example, the estimated benefit of example generation over restudy was numerically larger for the cued recall final test than for the example generation test (Table 6), despite the lower degree of overlap between the processes engaged during learning and test. Potentially more troublesome for TAP, performance on the example generation final test tended to be lower for learners who generated examples during practice than for learners who recalled definitions during practice.

Finally, according to general theories of self-regulated learning (e.g., Dunlosky and Ariel 2011; Winne and Hadwin 1998), students use monitoring to control their learning, and substantial evidence now indicates that students invest more effort in studying materials that they believe they have not yet learned or do not understand (e.g., Metcafle and Finn 2008; Metcalfe and Kornell 2005; Nelson et al. 1994). In the present context, attempting to generate an example may have helped students monitor how well they understood a given concept, which in turn may have supported more effective control during subsequent restudy. Findings in related literatures are also consistent with this possibility. For example, van Loon et al.

(2013) had children study idioms only or study idioms and then generate sentences in which the idioms were used. Children who generated sentences regulated study more effectively (for related findings from other generative tasks, e.g., de Bruin et al. 2011; Thiede et al. 2005).

An important assumption of self-regulated learning theories is that the effectiveness of control is constrained by monitoring accuracy—if monitoring informs control, then inaccurate monitoring will lead to ineffective control, which in turn will compromise learning. In the current research, learners may not have been aware of the relatively low quality of their generated examples and thus may not have adjusted their control strategies during restudy. Consistent with this possibility, students are substantially overconfident in the quality of the examples they generate during practice (Zamary et al. 2015). These outcomes suggest the interesting possibility that the benefits of example generation may not be constrained by example quality per se but rather by learners' lack of awareness of the quality of their self-generated examples.

The assumption of self-regulated learning theories that monitoring accuracy influences performance also provides an interesting explanation for why performance in the example generation groups was in between performance for the restudy-only and the recall-only groups. Restudy is not a generative task and thus provides learners with relatively impoverished cues for monitoring their current state of learning. As a result, encoding during later restudy trials may have been less effectively controlled by learners in the restudy-only group than by learners in the example generation group. Example generation and recall are both generative tasks and thus provide potentially more diagnostic cues for learners to infer their current state of learning. However, monitoring accuracy is likely lower when learners attempt to evaluate the quality of generated examples versus the quality of retrieved definitions. Consistent with this possibility, the degree of overconfidence for example quality judgments reported by Zamary et al. (2015) was considerably greater than overconfidence for judgments of definition recall observed in prior research (e.g., Dunlosky et al. 2005, 2011; Rawson and Dunlosky 2007), even though these studies involved similar samples, judgment scales, and forms of feedback. Cross-experiment comparisons notwithstanding, these outcomes suggest that monitoring is less accurate with example generation versus retrieval practice, which may lead to less effective control and thus lower learning. Consistent with this possibility, improvements across practice trials in Experiments 2 and 3 were less pronounced for example generation (Table 3) than for cued recall (Table 4).

## Limitations and Future Directions

Even with the addition of the three experiments reported here, the body of research on example generation is meager and many open questions remain. Some key directions for further research are also suggested by limitations of the current work. For example, given that all three experiments involved the same set of concepts, establishing generalizability beyond these materials would be useful. Additionally, examining the effects of example generation on a broader range of outcome measures would also be informative. To revisit, the current research focused on final tests involving example generation, primarily to provide a strong condition under which a benefit of example generation would be expected to emerge. Additionally, and more important for education, example generation reflects one of several learning goals for students who are learning declarative concepts—namely, that they are able to apply those concepts for practical purposes in novel contexts (e.g., a student learning about operant conditioning concepts such as positive and negative reinforcement would ideally be able to apply those concepts in novel contexts in daily life such as training pets, disciplining children, etc.). Examining the effects of example generation on measures of transfer would also be additionally informative. Although all three experiments involved a multiple-choice comprehension test that nominally involved transfer beyond the practice tasks, outcomes were not readily interpretable due to lower levels of reliability and to concerns about test order. Accordingly, future research involving a broader range of tests of transfer would be valuable.

Concerning comparisons of the effectiveness of example generation to restudy and to retrieval practice, more research investigating the underlying bases for the relative effectiveness of these techniques would be useful. For example, students who engaged in generative tasks (example generation and retrieval practice) may have been more motivated to engage in deeper processing of the conceptual material than students who only restudied concept definitions repeatedly. The current research did not include any outcomes that bear directly on this possibility, but relatively straightforward tests of this hypothesis could be pursued in future research. Research comparing the relative effectiveness of example generation to other learning techniques would also be valuable. For example, another appropriate comparison technique involves a condition in which learners study provided examples versus generate their own examples. This comparison would be pedagogically relevant, given that another common form of example-based learning involves providing students with illustrative examples of declarative concepts (Rawson et al. 2015).

Whereas the current research focused on example quality as a potential moderator of the effects of example generation, other moderators are worthy of investigation. For example, whereas we operationalized example quality as the extent to which a given example correctly illustrated the meaning of a concept, the effectiveness of example generation may depend on the degree of example variability (i.e., the number of different domains in which an individual is able to generate an example). Given that task instructions in the current study prompted students to generate a different example on each trial, the range of example variability was restricted in the current data set, but example variability could easily be examined in future research as a correlational predictor in which task instructions are less restrictive. Alternatively, example variability could be manipulated by instructing students to generate examples in the same versus different contexts. Hamilton (1990) found that learning was enhanced to a greater extent when students generated examples in the same domain versus in three different domains. However, this experiment involved immediate tests and is the only one to date to examine example variability as a moderator.

Another key moderator of interest involves students' level of initial learning. Many learners in the current research likely had an incomplete understanding of the concepts after studying the instructional materials and thus may not have been well equipped to successfully generate examples. Unfortunately, this situation may be representative of the initial level of learning achieved by at least some students for at least some course content after their initial exposure to concepts either in a lecture or from reading an assigned passage in a textbook. A plausible hypothesis to evaluate in future research is that example generation will be more effective for learners with a higher versus lower level of initial learning (or relevant prior knowledge more generally). This hypothesis is consistent with the *expertise reversal effect* demonstrated in research on worked examples; in brief, lower-knowledge learners benefit more from studying worked-out examples of problems (e.g., in algebra or geometry) versus attempting problem solving on their own, whereas higher-knowledge learners tend to benefit more from problem solving versus worked examples (for a review, see Kalyuga et al. 2012).

The possibility that example generation effects may be moderated by initial learning level also suggests other interesting directions for future research. In particular, the effectiveness of example generation may depend on the use of various scaffolds to support students' initial or ongoing level of concept learning. For example, although we removed provided examples of the concepts from the short instructional text used here to avoid interpretive difficulties (as discussed earlier), providing students with illustrative examples during initial instruction may enhance their understanding of the concepts, which in turn might support more successful implementation of the example generation technique. Consistent with this possibility, research on worked examples has consistently demonstrated that providing learners with worked examples prior to problem solving is more effective than problem solving alone (for a review, see Renkl 2014). Another scaffold that may moderate the effectiveness of example generation concerns feedback. Providing students with feedback on the quality of their generated examples may enhance ongoing learning of the concepts and may improve their successful implementation of the example generation technique.

## Practical Implications and Conclusions

Example generation has practical appeal as a technique that students can use while studying because it requires minimal support from instructors. Identifying learning activities that students can successfully use outside of class are particularly important for practical purposes, given that techniques that require intensive amounts of training or that place undue burden on instructors (e.g., to construct practice activities or materials) are less likely to be adopted. Furthermore, example generation is a technique that students report using regularly when studying (Gurung 2005; Gurung et al. 2010; Weinstein et al. 2013). The question is, should students be using this technique? The current outcomes suggest the tentative answer is "yes." Example generation is at least more effective than restudy, which is normatively the most common technique that students report using during self-regulated learning. Nevertheless, more work is needed before prescriptive conclusions can be confidently made, and to that end, the current work provides an empirical and theoretical foundation to support further investigations of this potentially promising but underexplored learning technique.

**Acknowledgments** The research reported here was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior Collaborative Award.

## Appendix: Excerpt from Text and Concept Definitions Used in Experiments 1–3

Although attribution often involves the logical kind of reasoning just described, this is not always the case. In fact, it is subject to several kinds of biases. One of the most important is known as the *correspondence bias*, which is the tendency to attribute other people's behavior to internal causes to a greater extent than is actually justified while underestimating the effect of the situation. The correspondence bias can lead us to false conclusions about others. Another bias in our attributions concerns our own behavior. The *self-serving bias* is the tendency to attribute positive outcomes to our own traits or characteristics (internal causes) but negative outcomes to factor beyond our control (external causes). Finally, the *just-world hypothesis* refers to the strong desire or need people have to believe that the world is an orderly, predictable, and just place, where people get what they deserve. This influences our attributions because when we encounter evidence suggesting that the world is not just, we sometimes persuade ourselves that no injustice has occurred.

Correspondence bias

The tendency to attribute other people's behavior to internal causes to a greater extent than is actually justified while underestimating the effect of the situation

Self-serving bias

The tendency to attribute positive outcomes to our own traits or characteristics (internal causes) but negative outcomes to factor beyond our control (external causes)

Just-world hypothesis

The strong desire or need people have to believe that the world is an orderly, predictable, and just place, where people get what they deserve

## References

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612–637.

Blaxton, T. A. (1989). Investigating dissociations among memory measures: support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 657–668.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology, 106*, 849–858.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*, 333–342.

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects multiple-choice testing. *Memory & Cognition, 36*, 604–616.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*, 273–281.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*, 153–170.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Chi, M. T. H., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science, 36*, 1–61.

Coane, J. H. (2013). Retrieval practice and elaborative encoding benefit memory in younger and older adults. *Journal of Applied Research in Memory and Cognition, 2*, 95–100.

Cortina, J. M., & Nouri, H. (2000). *Effect size for ANOVA designs*. Thousand Oaks: Sage.

de Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*, 294–310.

Dornisch, M., Sperling, R. A., & Zeruth, J. A. (2011). The effects of levels of elaboration on learners' strategic processing of text. *Instructional Science, 39*, 1–26.

Dunlosky, J & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In B. Ross (Ed), *Psychology of Learning and Motivation*, *54*, 103–140.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*, 551–565.

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *Quarterly Journal of Experimental Psychology, 64*, 467–484.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4–58.

Einstein, G. O., McDaniel, M. A., Owen, P. D., & Cote, N. C. (1990). Encoding and recall of texts: the importance of material appropriate processing. *Journal of Memory and Language, 5*, 566–581.

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

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition, 3*, 177–182.

Gorrell, J., Tricou, C., & Graham, A. (1991). Children's short- and long-term retention of science concepts via self-generated examples. *Journal of Research in Childhood Education, 5*, 100–108.

Gurung, R. A. R. (2005). How do students really study (and does it matter)? *Teaching of Psychology, 32*, 239–241.

Gurung, R. A. R., Weidert, J., & Jeske, A. (2010). Focusing on how students study. *Journal of the Scholarship of Teaching and Learning, 10*, 28–35.

Hamilton, R. J. (1989). The effects of learner-generated elaborations on concept learning from prose. *The Journal of Experimental Education, 57*, 205–217.

Hamilton, R. J. (1990). The effect of elaboration on the acquisition of conceptual problem-solving skills from prose. *The Journal of Experimental Education, 59*, 5–17.

Hamilton, R. J. (1997). Effects of three types of elaboration on learning concepts from text. *Contemporary Educational Psychology, 22*, 299–318.

Hamilton, R. J. (1999). The role of elaboration within a text processing and text adjunct context. *British Journal of Educational Psychology, 69*, 363–376.

Hamilton, R. J. (2004). Material appropriate processing and elaboration: the impact of balanced and complementary types of processing on learning concepts from text. *British Journal of Educational Psychology, 74*, 221–237.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*, 126–134.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: a model-comparison approach*. San Diego: Harcourt Brace Jovanovich.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579–588.

Kalyuga, S., Rikers, R., & Paas, F. (2012). Educational implications of expertise reversal effects in learning and performance of complex cognitive and sensorimotor skills. *Educational Psychology Review, 24*, 313–337.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborate studying with concept mapping. *Science, 331*, 772–775.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: do students practice retrieval when they study on their own? *Memory, 17*, 471–479.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: a distribution-based bifurcation model. *Journal of Memory and Language, 65*, 85–97.

Lishner, D. A. (2015). A concise set of core recommendations to promote the dependability of psychological research. *Review of General Psychology, 19*, 52–68.

Maner, J. K. (2014). Let's put our money where our mouth is: if authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science, 9*, 343–351.

Metcafle, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choices. *Psychonomic Bulletin & Review, 15*, 174–179.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*, 463–477.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5*, 207–213.

Neuschatz, J. S., Preston, E. L., Toglia, M. P., & Neuschatz, J. S. (2005). Comparison of the efficacy of two name-learning techniques: expanding rehearsal and name-face imagery. *American Journal of Psychology, 118*, 79–101.

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*, 122–133.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: choices and consequences. *Psychonomic Bulletin & Review, 14*, 187–193.

Rawson, K. A. (2012). Why do rereading lag effects depend on test delay? *Journal of Memory and Language, 66*, 870–884.

Rawson, K. A., & Dunlosky, J. (2007). Improving self-evaluation of learning for key concepts in expository texts. *European Journal of Cognitive Psychology, 19*, 559–579.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: how much is enough? *Journal of Experimental Psychology: General, 140*, 283–302.

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend upon time of test. *Journal of Educational Psychology, 97*, 70–80.

Rawson, K. A., Thomas, R. C., & Jacoby, L. L. (2015). The power of examples: illustrative examples enhance conceptual learning of declarative concepts. *Educational Psychology Review, 27*, 483–504.

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*, 1–37.

Roediger, H. L. (1990). Implicit memory: retention without remembering. *American Psychologist, 45*, 1043–1056.

Roediger, H. L. I. I. I., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, e22*, 1359–1366.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*, 76–80.

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: are yours realistic? *Perspectives on Psychological Science, 9*, 305–318.

Susser, J. A., & McCabe, J. (2013). From the lab to the dorm room: metacognitive awareness and use of spaced study. *Instructional Science, 41*, 345–363.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1267–1280.

van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merriënboer, J. J. G. (2013). The effect of delayed-JOLs and sentence generation on children's monitoring accuracy and regulation of idiom study. *Metacognition and Learning, 8*, 173–191.

Weinstein, Y., Lawrence, J. S., Tran, N., & Frye, A. A. (2013). *How and how much do students study? Tracking study habits with the diary method.* Poster presented at the annual meeting of the Psychonomic Society, Toronto, Canada.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Mahwah: Lawrence Erlbaum Associates.

Woloshyn, V. E., & Stockley, D. B. (1995). Helping students acquire belief-inconsistent and belief-consistent science facts: comparisons between individual and dyad study using elaborative interrogation, self-selected study and repetitious-reading. *Applied Cognitive Psychology, 9*, 75–89.

Wood, E., Willoughby, T., Kaspar, V., & Idle, T. (1994). Enhancing adolescents' recall of factual content: the impact of provided versus self-generated elaborations. *Alberta Journal of Educational Research, 40*, 57–65.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: a cautionary note. *Journal of Applied Research in Memory and Cognition, 3*, 214–221.

Zamary, A., Rawson, K. A., & Dunlosky, J. (2015). How accurately can students evaluate the quality of self-generated examples of declarative concepts? Not well, and feedback does not help. *Submitted manuscript.*