Educ Psychol Rev
DOI 10.1007/s10648-014-9273-3

REPLICATION

The Power of Examples: Illustrative Examples Enhance Conceptual Learning of Declarative Concepts

Katherine A. Rawson · Ruthann C. Thomas · Larry L. Jacoby

© Springer Science+Business Media New York 2014

Abstract Declarative concepts (i.e., key terms with short definitions of the abstract concepts denoted by those terms) are a common kind of information that students are expected to learn in many domains. A common pedagogical approach for supporting learning of declarative concepts involves presenting students with concrete examples that illustrate how the abstract concepts can be instantiated in real-world situations. However, minimal prior research has examined whether illustrative examples actually enhance declarative concept learning, and the available outcomes provide weak evidence at best. In the three experiments reported here, students studied definitions of declarative concepts followed either by illustrative examples of those concepts or by additional study of the definitions. On a subsequent classification test in which learners were presented with examples and were asked to identify which concept the example illustrated, performance was greater for students who had studied illustrative examples during learning than for students who only studied definitions (*d*s from 0.74 to 1.67). However, the effects of illustrative examples on declarative concept learning depended in part on the conditions under which those examples were presented. Although performance was similar when examples were presented after versus before concept definitions (Experiments 1a–1b), classification accuracy depended on the extent to which examples of different concepts were interleaved and whether definitions were presented along with the examples (Experiment 2).

Keywords Declarative concepts · Examples · Concept learning · Classification

Declarative concepts are a common kind of information that students are expected to learn across many different grade levels and academic disciplines. We use the term *declarative concepts* here to refer to key terms with short (usually one to two sentences) definitions of the abstract concepts denoted by those terms. For example, a psychology student would learn

K. A. Rawson (✉)
Department of Psychology, Kent State University, P.O. Box 5190, Kent, OH 44242-0001, USA
e-mail: krawson1@kent.edu

R. C. Thomas
Hendrix College, Conway, AR, USA

L. L. Jacoby
Washington University in St. Louis, Saint Louis, MO, USA

Published online: 10 June 2014

about positive and negative reinforcement, a physics student would learn about kinetic and potential energy, and a biology student would learn about phenotypes and genotypes. Concepts of this sort are targeted in many concept inventories, which are taxonomies of foundational concepts within a particular topic or domain). Since introduction of the first concept inventory in physics (Force Concept Inventory; Hestenes, Wells, and Swackhamer 1992), a wide array of concept inventories has been developed for various scientific topics (e.g., geoscience, genetics, natural selection, electricity and magnetism, biology; for a review, see Libarkin 2008). Likewise, classroom instruction is often directed at teaching students about declarative concepts, and textbooks are often heavily populated by these concepts. For example, across chapters in three popular *Introductory Psychology* textbooks (Myers 2010; Schacter, Gilbert, and Wegner 2009; Zimbardo, Johnson, and McCann, 2012), the end-of-chapter concept lists included an average of 42 key concept terms (range 14–86 across 46 chapters). Declarative concepts represent an important part of the foundational knowledge that novice learners need to build upon, to acquire a cumulative knowledge base within a domain. Indeed, one of the best predictors of learning for new information within a domain is the amount of relevant prior knowledge an individual has within that domain (e.g., Hambrick 2003; Hambrick, Meinz, Pink, Pettibone, and Oswald 2010; Spilich, Vesonder, Chiesi, and Voss 1979).

Importantly, a defining characteristic of declarative concepts is that they have some level of abstraction that represents a type-token relationship with particular contexts, situations, or events in which that concept may be instantiated.[1] One reason that declarative concepts feature prominently in some courses is that they are applicable to enhancing understanding of and/or improving functioning in various real-world contexts. For example, the concept of positive reinforcement can be instantiated in the context of teaching children, training pets, improving adherence to exercise and diet regimens, and so on.

Concerning instruction of declarative concepts, a common feature of textbooks and of lectures is to introduce a declarative concept by presenting the definition and then to elaborate further by describing concrete examples of how that concept can be applied in one or more real-world situations (hereafter referred to as *illustrative examples*). An important learning goal is for students to be able to successfully apply declarative concepts to such contexts, which is a hallmark of conceptual learning. The implicit or explicit assumption of instructors and textbooks is that providing illustrative examples will support this kind of conceptual learning. However, as discussed further below, minimal evidence exists to support this assumption.

The importance of presenting examples for concept learning likely depends on the structure of the concept that is being conveyed. The distinction between well-defined and natural concepts (e.g., Murphy 2004) is particularly relevant. Category membership for well-defined concepts is all-or-none as a result of their reliance on a set of rules that specify features, corresponding to a definition, that are sufficient for including all instances of the concept and excluding all non-instances. The concept of "triangle" serves as a good example of a well-defined concept in that it can be defined via a set of necessary and sufficient features (i.e., a closed figure consisting of three straight lines). In contrast, for natural concepts, category membership is graded as a result of membership being determined on the basis of similarity to either a prototype (e.g., Rosch and Mervis 1975) or a collection of exemplars of the concept (e.g., Medin and Schaffer 1978; Nosofsky 1999). The concept "dog" is a natural category in that there are no features or combination of features that serve as a definition to allow inclusion of all instances and exclusion of all non-instances of the concept. Rather, classification is based

[1] This characteristic is also what distinguishes declarative concepts from declarative facts, which are non-abstract statements with truth values (e.g., George Washington was the first president of the USA; the capital of Ohio is Columbus).

on similarity with the result that category membership is graded (e.g., a Collie is a more typical instance of "dog" than is a Chihuahua). For natural categories, the examples or prototype derived from them serve as the basis for the concept, and thus examples are key for learning natural concepts.

As will be described, the declarative concepts used as materials in our experiments (topics of human judgment and decision making) correspond to natural categories rather than to well-defined categories. The parallel between natural categories and declarative concepts is useful for understanding why declarative concept learning may benefit from illustrative examples and thus why we expected to find evidence of the power of examples for conveying those concepts. Table 1 provides definitions and examples for some of the declarative concepts that were employed in the current research. To appreciate that the concepts are natural ones rather than well defined, consider the definition and first example given for the mere exposure effect. Suppose that example was changed to say that as a result of prior exposure you come to believe that most others will like the CD that was frequently played by your coworker. The changed example could be classified as an example of the mere exposure effect—you believe others will like the CD because you do as a result of being frequently exposed to it. Alternatively, the changed example could be classified as an example of the availability heuristic—you believe others will like the CD because an instance of somebody liking it readily comes to mind. Whereas classification of a well-defined object (e.g., a triangle) is unambiguous, this is not the case for classifying examples of declarative concepts (particularly from the domain of human judgment and decision making). More generally, a given declarative concept may have a one-to-many mapping with various surface features of contexts in which the concept applies. Likewise, several declarative concepts may have a many-to-one mapping with a particular surface feature, in that various concepts may have different applications within the same context. Examples afford structural alignment processes that support learning of the deep relational structures (Rittle-Johnson and Star 2011).

The primary goal of the current research was to determine whether examples facilitate learning of declarative concepts. Several theoretical frameworks support the expectation that

Table 1 Examples of concepts, definitions, and examples used in Experiments 1a, 1b, and 2

Availability heuristic: The tendency to estimate the likelihood that an event will occur by how easily instances of it come to mind.
• After the Columbine shootings and the extensive press coverage, people were more likely to overestimate the amount of teen violence and to fear school violence.
• Because it is easier to think of words that start with k than words with k as the third letter, people assume that more words start with k. In fact, however, many more words have k as the third letter.
Mere exposure effect: The phenomenon whereby the more people are exposed to a stimulus, the more positively they evaluate that stimulus.
• Imagine you are a computer programmer. Several cubicles away, one of your coworkers frequently plays a CD from the band, Elephants Abroad. You come to like this CD.
• Politicians spend a lot of time on fund-raising to buy advertising. They know that the repeated media exposure increases positive evaluation of their names for voters, which can be critical in their winning elections.
Door-in-the-face technique: A strategy to increase compliance based on the fact that refusal of a large request increases the likelihood of agreement with a subsequent smaller request.
• Carrie needs about $10 for a shopping trip, so she asks her mother for $50, but her mother refuses. Carrie then asks for $10 and her mother agrees.
• Your neighbor first asked you to take care of his dog and cat in your home. After you refused to do so, the neighbor might ask if you would at least water his plants, which you would agree to do.

they will do so. According to the *transfer-appropriate-processing* (TAP) framework, memory is enhanced to the extent that the cognitive processes engaged during encoding overlap with those engaged during retrieval (Morris, Bransford, and Franks 1977; Roediger et al. 1989). For example, in a study of analogical transfer (Needham and Begg 1991), learners encoded problem–solution scenarios during training, either with instructions to study each scenario for later recall (memory orientation) or with instructions to explain why the solution for each problem was correct (problem orientation). Subsequently, individuals were more likely to correctly solve analogous transfer problems if they had previously engaged in problem-oriented processing versus memory-oriented processing, even though memory for the training scenarios was better following memory-oriented processing versus problem-oriented processing. By extension to declarative concept learning, providing illustrative examples engages students in mapping the components of abstract concepts onto real-world referents during study. Thus, TAP would suggest that orienting students to the application of concepts during study will enhance the likelihood that students can successfully apply abstract concepts to real-world contexts subsequently.

The expectation that illustrative examples will facilitate conceptual learning also aligns with central assumptions of contemporary theories of text comprehension concerning the levels of representation afforded by a nominal unit of text (Kintsch 1998; van den Broek 2010; Zwaan and Radvansky 1998). In brief, the *surface level* includes a verbatim representation of the linguistic information (i.e., the particular words and syntax) contained in the text. The *textbase* includes an amodal semantic representation of the ideas explicitly stated in the text. The *situation model* goes beyond the textbase by integrating ideas from the text with general world knowledge to form a multimodal representation of the situation being described in the text. Comprehension theorists further assume that performance on memory-based tasks largely depends on the textbase, whereas performance on tasks requiring inference and application depend more heavily on the situation model. By extension, providing the definition of a declarative concept affords a textbase representation of the ideas contained in that definition, whereas providing an illustrative example would afford the encoding of a situation model by describing a real-world situation that instantiates those ideas and that affords integration with general world knowledge.

To what extent does the available evidence confirm the expectations that follow from these theories (and the intuitive expectations of instructors and textbook publishers)? Surprisingly minimal prior research has directly examined the effects of illustrative examples on learning of declarative concepts, and the outcomes are not compelling. Hamilton (1990) presented undergraduates with a short instructional text that introduced four declarative concepts (positive and negative reinforcement, positive and negative punishment). The text included the definition and four examples for each concept, and all students responded to several adjunct questions over the text. Half of the students were then presented with three additional examples of each concept, whereas the other half were told to come up with three of their own examples. All students then completed a recall test for the definitions, a classification test in which they were presented with novel examples of the concepts and asked to identify which concept each example illustrated, and then a problem-solving test describing two classroom scenarios for which students were to explain how they would reduce a disruptive behavior. Providing examples (versus having students generate their own) produced a small but significant effect on problem-solving performance, but groups did not differ on either of the other measures. One possible explanation for the weak effects is that the functional difference between the two groups in exposure to illustrative examples was minimal, given that all participants were exposed to illustrative examples during initial study of the text, and 84 % of the examples

generated in the comparison group were acceptable. In other words, the functional dosage for the two groups was 7.0 versus 6.5 examples for each concept.

In a quasi-experimental study by Griffin (1993), students in four sections of an educational psychology course received the same lectures in class on interpreting different types of criterion-referenced and norm-referenced test scores. Each section was then assigned to complete a worksheet that either provided examples of each concept or prompted students to come up with examples of each concept. Students worked together in small groups in class to complete the worksheet and then individually completed a final test that involved classifying novel examples. Final test performance in the four course sections did not significantly differ.

Taken together, the outcomes of these studies are not particularly encouraging with respect to the potential benefits of providing illustrative examples to foster conceptual learning of declarative concepts. However, the comparison group in these studies also received some exposure to illustrative examples of the concepts in the initial instructional materials and/or in the learning task assigned to that group. Arguably, a more appropriate comparison would involve a group who was neither exposed to nor generated illustrative examples during the learning phase.

No prior research on declarative concept learning has compared illustrative examples to a condition that does not include examples. Indirect evidence comes from related literatures on learning other kinds of concepts. In work on mathematical concept learning, fourth graders were more accurate at identifying acceptable instances of the well-defined concept of "equilateral triangle" on a final test when they were presented with a definition plus examples and non-examples during learning versus the definition only (Klausmeier and Feldman 1975). Reed and Bolstad (1991) asked learners to solve "work" problems (i.e., computing the productivity of two workers with different rates and/or time spent on task together). Learners were provided with general procedural instructions about how to solve work problems, either with or without a worked example including a detailed solution for solving a simple problem. Performance on a subsequent transfer test was greater when examples were versus were not provided during learning. In contrast, other related literatures have shown negative effects of providing concrete instances on analogical transfer (e.g., Kaminski, Sloutsky, and Heckler 2013) and negative effects of providing elaborative details on learning main points from text material (e.g., Reder and Anderson 1982).

In sum, although illustrative examples are commonly used in practice and several theoretical accounts support the prediction that they will enhance conceptual learning of declarative concepts, it remains an open question. The few studies that have included illustrative examples of declarative concepts found minimal effects, but these studies did not include a no-example comparison condition. Outcomes from research on other kinds of concept learning are suggestive but somewhat mixed and do not directly establish the effects of illustrative examples on declarative concept learning.

Accordingly, the primary goal of the current research was to investigate the extent to which illustrative examples enhance conceptual learning of declarative concepts beyond presentation of definitions alone. Based on the theoretical accounts described above, we predicted a facilitative effect of examples over definitions alone. As earlier stated, the declarative concepts employed in the current experiments were taken from the domain of human judgment and decision making and correspond to natural concepts. Many concepts in this domain have widespread application to many real-world contexts, and thus it is particularly well-suited for present purposes. Additionally, these concepts are often accompanied by illustrative examples in instructional materials. For example, in the Social Psychology chapter (which included the concepts examined here) of a top-selling *Introductory Psychology* textbook (Myers 2010), 40

out of 43 concepts included in the chapter were accompanied by one or more illustrative examples.

All three experiments included the same two groups, a *definition-then-examples* group and a *definition-only* group. In the definition-then-examples group, learners studied the definitions of declarative concepts in the first block of trials and then were presented with illustrative examples of each concept in the next five blocks of trials. In the definition-only group, no illustrative examples were provided, and learners simply studied the definition in each block of trials. In all three experiments, the primary outcome was performance on a final test in which learners were presented with examples and were asked to identify which concept the example illustrated. Example classification is a widely used measure of conceptual learning within the literature on concept learning (e.g., Allen and Brooks 1991; Brooks, Norman, and Allen 1991; Di Vesta and Peverly 1984; Murphy 2004). Similar to much of the prior research on other kinds of concept learning, half of the examples presented for classification here were novel (i.e., had not been presented during the learning phase). Classification of novel examples provides a particularly stringent test of concept learning, in that it requires understanding of the concept and cannot be based on overlap of surface features alone (because the new example contexts have minimal to no overlap with the definition or with other examples used during learning). The key prediction is that classification accuracy will be greater in the definition-then-examples group than in the definition-only group.

The secondary purpose of each experiment was to explore potential moderators of the effect of illustrative examples on conceptual learning. Additional aspects of the design relevant for this purpose will be described in each experiment below.

Experiments 1a and 1b

As outlined above, learners in the definition-then-examples group studied the definition of each concept prior to studying illustrative examples of each concept. This order of presentation mimics the conventional approach in classroom instruction and textbook materials, in which a concept definition is explicitly stated and then illustrated by examples. For example, in the Social Psychology chapter of Myers (2010), illustrative examples were presented after definitions for 34 of 40 (85 %) concepts. In contrast, an alternative approach suggested by recent work on guided discovery learning would involve presenting examples prior to the definition. A particular form of guided discovery involves engaging learners in exploratory activity prior to direct instruction. Exploratory activity prior to instruction versus after instruction has been shown to enhance children's conceptual learning in mathematics (e.g., DeCaro and Rittle-Johnson 2012) and physics (e.g., Schwartz et al. 2011). By comparison here, presenting examples prior to the definition versus after the definition may further enhance conceptual learning of declarative concepts.

In contrast, Di Vesta and Peverly (1984) hypothesized that learning would be enhanced by presenting concept definitions prior to examples, based on the idea that the definition would orient learners to process the key attributes and relations in the examples and that examples prior to the definition "would lead to poorest performance because initial encoding might incorporate erroneous inferences" (p. 110). In their study, undergraduates learned artificial concepts (e.g., "*belk*, to disguise something by directing attention away from it, originally used by magicians") through initial presentation of the definition and several practice cued-recall trials with restudy. In addition, examples of each concept were presented either prior to the learning phase (i.e., examples-then-definition group) or after the learning phase (i.e., definition-then-examples group). Two days later, learners were presented with novel examples

and non-examples. For each item, learners were asked a yes/no question about whether it was an example of one of the previously learned concepts, and if yes, to name which concept it illustrated. On the yes/no decision component of the test, the two groups did not differ in false alarms to non-examples but did differ in hits for examples (with a higher hit rate for the definition-then-examples group). However, the two groups did not differ in classification accuracy for those examples. In the current research, inclusion of the examples-then-definition group affords a conceptual replication of DiVesta and Peverly's study to further explore the extent to which the effect of illustrative examples depends on the sequence of presentation.

Finally, as a first step toward exploring the generalizability of any effects of illustrative examples on declarative concept learning, we sampled students from two different universities. Experiment 1a involved undergraduates from a large public university, and Experiment 1b involved undergraduates at an elite private university. The student populations at these two universities differ considerably in background characteristics relevant to educational outcomes (including high school GPA, entrance examination scores, first-generation student status, SES, etc.), affording an opportunity to examine the effects of illustrative examples with heterogeneous samples. In these and subsequent experiments, our particular interest was in the effect of illustrative examples on initial learning of unknown concepts (rather than on relearning of already known concepts). Thus, we focused analyses on students who reported minimal prior familiarity with the declarative concepts included in the experimental materials. This criterion is also consistent with the emphasis in prior studies described above. In Hamilton's (1990) study, students with high pre-experimental familiarity with concepts were excluded from analyses, and Di Vesta and Peverly (1984) used artificial concepts to avoid prior knowledge of concepts. To foreshadow, outcomes for students with higher levels of pre-experimental familiarity are reported in the Appendix and will be briefly discussed in the "General Discussion" section.

In sum, the primary goal of Experiments 1a and 1b was to test the prediction that conceptual learning of declarative concepts will be enhanced by illustrative examples (relative to definitions only), and the secondary goal was to explore whether the effect differs depending on the order of presentation (definition-then examples versus examples-then-definition).

Methods

Participants and Design Participants in Experiment 1a were undergraduates enrolled at Kent State University (n=131; 73 % female) who participated for course credit. Participants in Experiment 1b were undergraduates enrolled at Washington University (n=176; 57 % female) who participated for course credit or monetary compensation. In each experiment, we oversampled to ensure a sufficient number of participants with minimal prior familiarity with the target concepts included in the experimental materials. Characteristics of students in the overall sample for each experiment are summarized in Table 2. In each experiment, students were randomly assigned to one of three groups (definition only, definition then examples, or examples then definition).

Materials and Procedure Materials included ten concepts from the topic of human judgment and decision making. The item set for each concept included a one-sentence definition and ten examples. The examples for each concept were excerpted from various undergraduate psychology textbooks and thus represented authentic educational materials. Each example illustrated the target concept in the context of a concrete, real-world situation. Table 1 includes a

Table 2 Means for descriptive measures of student samples in each experiment

	Experiment 1a	Experiment 1b	Experiment 2
Age (years)	20.5 (0.4)	19.8 (0.2)	19.6 (0.2)
Education (years completed)	13.0 (0.1)	14.3 (0.1)	12.4 (0.1)
Vocabulary (% correct)	68.8 (1.0)	85.9 (0.7)	67.6 (0.8)

Note. Vocabulary = Shipley Vocabulary Test (Zachary 1986). Standard errors of the mean are reported in parentheses

sample of concepts, definitions, and examples (the full set of materials is available from the first author). The ten concepts were randomly divided into two sets of five, for counterbalancing purposes described further below. The ten examples for each concept were also randomly divided into two sets of five for counterbalancing described below.

All tasks and instructions were administered by a computer program, and participants worked at individual computer carrels. All participants were told that they would be asked to learn ten concepts from psychology and would later be tested on their memory and comprehension of the concepts. Participants in the definition-only group were then told that they would study each concept definition six times. Participants in the definition-then-examples group were told that they would study each definition once and would then study five examples of each concept. Participants in the examples-then-definition group were told that they would study five examples of each concept and would then study each definition once.

Participants in the definition-only group then completed six blocks of study trials for half of the concepts. On each trial, the concept name was presented at the top of the screen with the definition in a field below. Participants were informed that they had up to 60 s to study the definition. If they finished studying sooner, they could click a button on the screen to advance to the next trial. Within each block, concepts were presented in random order, with the constraint that the trials for a given concept in consecutive blocks were separated by at least one other concept. After completing all six blocks for the first set of concepts, the second set of concepts was presented for six blocks of study trials in the same manner. Assignment of concept set to presentation order was counterbalanced across participants (counterbalancing in the final sample was approximate due to exclusion of participants described below).

The procedure for participants in the definition-then-examples group was the same as in the repeated definition group, except that each trial in blocks 2–6 involved presentation of a different example of each concept instead of the definition. The concept name was presented along with each example. The set of five examples to be presented was approximately counterbalanced across participants; examples within the presented set were randomly assigned blocks 2–6. The procedure for participants in the examples-then-definition group was the same, except that the examples and concept names were presented in blocks 1–5 and definitions were presented in block 6.[2]

After all presentation trials had been completed, participants solved multiplication problems for 5 min as a filler task. The final test then began with a cued recall test of definitions for five

[2] Examples for each concept were presented in separate blocks to align the schedule of presentation trials in the example groups with the schedule of presentation trials in the definition-only group. In the definition-only group, presentation of the definition for each concept was spaced across blocks, because a wealth of prior research has shown that massed restudy often has minimal to no benefit for learning (for reviews, see Cepeda, Pashler, Vul, Wixted, and Rohrer 2006; Dunlosky et al. 2013). Thus, massed restudy would have provided a very weak comparison group.

concepts. On each trial, a concept name was presented at the top of the screen, and participants typed in as much of the definition as they could remember. In both example groups, the instructions emphasized that participants should type in definitions and not examples.

After the cued recall test, participants completed an example classification test. On each trial, an example of one of the ten concepts was presented at the top of the screen. The bottom of the screen presented all ten concept names, with a button next to each concept (order of concept names was randomized anew for each participant). Participants were prompted to click a button to indicate which concept the example illustrated, at which point they were moved on to the next trial. The classification test included 100 trials, with ten examples for each concept. These ten trials consisted of the five examples that had been presented during the study phase (hereafter referred to as *studied* examples) and the five examples that had not been presented (hereafter referred to as *novel* examples). The order of trials was randomized anew for each participant.

After the classification test, participants completed a cued recall test for the remaining five concepts that were not tested prior to classification. Assignment of concept set to be tested prior versus after classification was approximately counterbalanced across participants. The purpose of splitting the cued recall test was to examine whether test order influenced performance on either measure. Neither cued recall nor classification performance significantly differed for concepts sets tested on cued recall before versus after classification, and thus we do not discuss this variable further. All test trials (cued recall and classification) were self-paced.

After the final test, participants were shown a list of the ten concept names and asked to indicate for each one whether they had learned that concept in a psychology class prior to the experiment (this variable was used for participant selection, described below). Participants then completed a demographics questionnaire and a general vocabulary test including 40 multiple-choice questions in which they were presented with a word and asked to select the best synonym from among four choices.

Results

Data for four participants in Experiment 1a and four participants in Experiment 1b were excluded from analyses based on evidence of non-compliance with task instructions. During the classification task, these participants had response times faster than 1 s for more than 30 % of the items (compared with 0.6 % of items in the remainder of the sample) and/or had mean response times less than 4 s per item (compared with M=12 s per item in the remainder of the sample). Rapid response times during the classification task suggest that participants were skipping quickly through many of the test items rather than attempting to answer each one (particularly given that the length of the test prompts was 49 words on average).

Concerning the number of concepts that participants reported learning in prior coursework, the distribution was highly skewed in Experiment 1a and bimodal in Experiment 1b. Given our interest in the effect of illustrative examples on initial learning of unknown concepts, outcomes reported below include data from participants who reported having learned three or fewer of the target concepts in prior coursework. In Experiment 1a, 66 % of participants reported having learned three or fewer concepts (for number of concepts reported within this subset, M=0.9, median and mode=0; for the remaining participants, M=6.3, median and mode=6). In Experiment 1b, 45 % of participants reported having learned three or fewer concepts (for number of concepts reported within this subset, M=0.9, median and mode=0; for the

remaining participants, M=7.7, median=8, mode=10). Outcomes for the remaining participants in each experiment are reported in Table A1 of Appendix.

For each outcome measure in each experiment, we tested a set of planned contrasts appropriate to our primary and secondary research questions. To evaluate our primary question concerning the effect of illustrative examples, the first contrast compared performance in the definition-only group to performance in the two example groups (i.e., definition-then-examples and examples-then-definition groups). To evaluate our secondary question concerning the effect of providing examples before or after direct instruction of definitions, the second contrast compared performance between the two example groups.

For tests of the a priori directional prediction that examples will enhance classification performance, we report one-tailed p values (Judd and McClelland 1989); two-tailed p values are reported for analyses of the remaining secondary outcomes. Effect sizes reported are Cohen's d (Cortina and Nouri 2000). Split-half reliability estimates (corrected using Spearman-Brown prophecy formula; Carmines and Zeller 1979) for the three performance measures ranged from 0.85 to 0.89 in Experiment 1a and from 0.74 to 0.85 in Experiment 1b.

Classification Performance For each participant, we computed the percentage of classification items that were answered correctly for studied examples and for novel examples. Mean performance for both measures is reported in Fig. 1, for Experiment 1a (left panel) and Experiment 1b (right panel). On both measures in both experiments, performance in all groups was significantly greater than chance (10 %), all ts>14.20.

Concerning our primary question of interest, illustrative examples markedly improved conceptual learning of declarative concepts. For classification of studied examples, the two example groups outperformed the definition-only group in Experiment 1a [$t(85)$=3.12, p=0.001, d=0.74] and in Experiment 1b [$t(77)$=6.06, p<0.001, d=1.43]. Classification performance for the studied examples may in part reflect associative memory for the concept names that were presented with these examples during learning. Consistent with this possibility, classification performance was greater for studied versus novel examples [collapsing across the two example groups, $t(61)$=6.97, p<0.001, d=0.31 in Experiment 1a and $t(50)$=3.75, p<0.001, d=0.37 in Experiment 1b]. However, the pattern of performance for the novel examples provides strong evidence that illustrative examples enhanced conceptual learning:

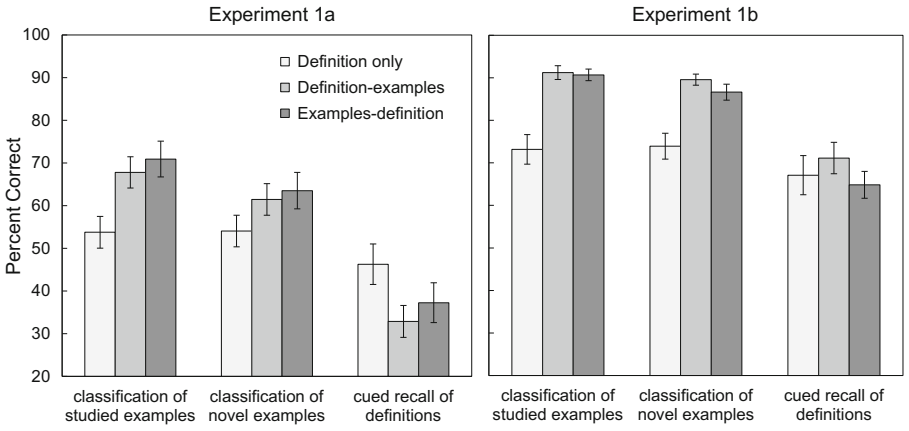

Fig. 1 Mean percent correct on three performance measures in Experiments 1a–1b as a function of group. *Error bars* are standard errors of the mean

Concerning classification performance for novel examples, performance was greater in the two example groups than in the definition-only group in Experiment 1a [$t(85)=1.67$, $p=0.049$, $d=0.40$] and in Experiment 1b [$t(77)=5.17$, $p<0.001$, $d=1.22$].

Concerning our secondary question of interest, performance in the two example groups did not significantly differ in either Experiment 1a or in Experiment 1b for studied examples [$t(60)=0.56$, $d=0.14$ and $t(49)=0.25$, $d=0.07$, respectively] or for novel examples [$t(60)=0.37$, $d=0.09$ and $t(49)=1.29$, $d=0.36$, respectively]. Thus, the effects of illustrative examples were not moderated by the order in which examples and definitions were presented.

Secondary Outcome Measures For cued recall, definitions were broken down into three to four idea units, and responses were scored based on the percentage of idea units recalled that were either verbatim restatements or paraphrases that preserved the meaning of the idea unit. Partial credit was given for responses that included some but not all of the correct meaning of the definition. Two raters were trained to complete the scoring, and 10 % of the protocols served as a training set. Each rater scored the training set, and reliability was quite high ($r=0.89$). Each remaining protocol was then scored by one of the raters.

For cued recall of definitions (see Fig. 1), performance was greater in the definition-only group than in the two example groups in Experiment 1a [$t(84)=2.09$, $p=0.020$, $d=0.50$] but not in Experiment 1b [$t(77)=0.19$, $d=0.04$]. Performance in the two example groups did not significantly differ in either Experiment 1a or in Experiment 1b [$t(59)=0.74$, $d=0.19$ and $t(48)=1.30$, $d=0.37$, respectively].

Informal comparisons across Experiments 1a and 1b reveal that performance was generally higher in Experiment 1b, which is not surprising given that participants in that experiment were from a university that employs admission standards that are more stringent than those employed by the university attended by participants in Experiment 1a. More interesting, the gain in classification performance from presenting examples was offset by a cost in cued recall of definitions in Experiment 1a but not in Experiment 1b. In Experiment 2, we further examine whether presenting examples rather than repeating definitions produces a cost for recall of definitions.

Finally, mean study time during the learning phase is reported in Table 3. To revisit, we equated the number of trials in each group but not nominal trial duration. In brief, equating nominal time on task does not guarantee equivalent functional time on task (e.g., participants likely do not attend to a stimulus during the entire enforced period of time); permitting self-paced study is preferable because it affords examination of functional time on task. In Experiment 1a, a 6 (block) × 3 (group) ANOVA yielded a significant main effect of block [$F(5, 420)=96.63$, mean squared error (MSE)$=29.55$, $p<0.001$, $\eta_p^2=0.54$], no effect of group ($F=1.29$), and a significant interaction [$F(10,420)=7.34$, MSE$=29.55$, $p<0.001$, $\eta_p^2=0.15$]. In Experiment 1b, the main effects of block and group were both significant [$F(5, 380)=77.93$, MSE$=21.84$, $p<0.001$, $\eta_p^2=0.51$ and $F(2, 76)=3.52$, MSE$=408.86$, $p=0.035$, $\eta_p^2=0.09$], as was the interaction [$F(10, 380)=6.51$, MSE$=21.84$, $p<0.001$, $\eta_p^2=0.15$].

As apparent from inspection of the pattern of means in Table 3, the interaction in each experiment largely reflects longer study times in the example groups than in the definitions-only group in later blocks of practice. However, these differences in study time are unlikely to explain the effect of examples on conceptual learning. In an analysis of covariance (ANCOVA) controlling for study time in each block of trials, classification performance was still significantly greater for the two example groups than for the definition-only group for studied examples [$F(1, 79)=10.52$, MSE$=412.88$, $p=0.002$, $\eta_p^2=0.12$ in Experiment 1a and $F(1, 70)=14.41$, MSE$=163.38$, $p<0.001$, $\eta_p^2=0.17$ in Experiment 1b] and for novel examples [$F(1, 79)=4.41$, MSE$=429.40$, $p=0.039$, $\eta_p^2=0.05$ in Experiment 1a and $F(1, 70)=9.10$, MSE$=138.14$, $p=0.004$, $\eta_p^2=0.12$ in Experiment 1b].

Table 3 Mean study time in each block of trials as a function of group in each experiment

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Experiment 1a						
Definition only	34.4 (3.0)	20.0 (2.4)	16.2 (2.3)	14.8 (2.3)	12.5 (2.1)	11.0 (2.1)
Definition then examples	26.7 (1.9)	24.2 (1.4)	22.0 (1.8)	19.2 (1.4)	18.7 (1.3)	15.7 (1.1)
Examples then definition	30.1 (2.3)	24.8 (2.0)	21.9 (1.9)	20.8 (2.0)	18.8 (1.5)	14.9 (1.7)
Experiment 1b						
Definition only	26.2 (2.7)	14.3 (2.4)	11.8 (1.9)	9.4 (1.6)	8.1 (1.4)	6.6 (1.2)
Definition then examples	23.7 (2.5)	22.4 (2.8)	19.1 (2.4)	16.5 (1.8)	16.4 (1.6)	15.2 (1.2)
Examples then definition	23.7 (1.9)	18.0 (1.4)	17.1 (1.4)	16.0 (1.5)	15.3 (1.1)	12.6 (1.4)
Experiment 2						
Definition only	31.6 (2.9)	21.2 (3.3)	17.7 (2.8)	13.8 (2.2)	10.8 (2.1)	10.2 (1.7)
DE interleaved	30.2 (2.1)	24.5 (1.7)	22.7 (1.5)	21.1 (1.4)	20.4 (1.3)	19.4 (1.4)
DE blocked	26.9 (2.5)	22.6 (2.4)	20.4 (2.4)	20.2 (2.0)	18.4 (1.8)	17.6 (2.1)
DE interleaved with definition	30.1 (2.5)	30.5 (2.2)	25.1 (1.9)	23.8 (2.1)	19.9 (1.7)	19.9 (1.9)
DE blocked with definition	25.1 (1.9)	22.7 (1.9)	20.6 (1.5)	18.5 (1.6)	17.6 (1.6)	16.6 (1.5)

Note: Study time is reported in seconds. Standard errors of the mean are reported in parentheses

Experiment 2

Given that Experiment 1 provides the first demonstration that illustrative examples enhance conceptual learning for declarative concepts, an important goal of Experiment 2 was to replicate this key outcome (regarding the importance of directly replicating novel findings, see Pashler and Harris 2012). Accordingly, Experiment 2 included the same definition-only and definition-then-examples groups as in Experiment 1. For economy of design, we dropped the examples-then-definition group because the two example groups did not significantly differ and because this schedule is less representative of how examples are typically incorporated into classroom instruction and textbook materials.

Experiment 2 was also designed to provide important extensions beyond Experiment 1, to explore other potential moderators of the effects of examples. The two moderators examined in Experiment 2 were motivated by the conditions in which examples are commonly presented in practice. In the definition-then-examples group in Experiment 1, students studied each definition for a set of concepts prior to studying examples for those concepts. As noted above, presentation of definitions prior to examples is a common practice. However, the particular presentation schedule used here departs from typical practice in two respects. First, a set of concepts is often taught in a sequence such that one concept definition is introduced, followed by examples of that concept, before proceeding on to introduction of the next concept. That is, presentation of the definition and examples for a given concept are commonly *blocked* (as opposed to *interleaved* with definitions and examples of other concepts, as in Experiment 1). For example, in the Social Psychology chapter of Myers (2010), illustrative examples were included in the same paragraph as the definition for 30 of 40 concepts (examples were included in an immediately adjacent paragraph for the remaining ten concepts). Second, the definition of a concept is commonly available for students to refer back to as needed while studying examples of that concept (as opposed to withholding the definition during presentation of the examples as in Experiment 1). To explore the extent to which either or both of these conditions moderate the effects of examples, we manipulated the presentation schedule

(blocked versus interleaved) and the availability of definitions (examples with or without concurrent definitions) in Experiment 2.

What might be predicted with respect to the effects of these two factors? Concerning presentation schedule, prior research strongly supports the prediction that interleaving will be more effective than blocking for enhancing concept learning. The advantage of interleaved versus blocked schedules has been shown in several other conceptual or complex learning tasks, including inductive learning of artists' painting styles, categorization of bird species, and learning how to solve different kinds of math problems (for recent reviews, see Dunlosky, Rawson, Marsh, Nathan, and Willingham 2013; Rohrer 2012). One theoretical account of interleaving that emerges from this work is that interleaving is particularly effective for enhancing discrimination learning (Taylor and Rohrer 2010). For present purposes, interleaving illustrative examples may enhance declarative concept learning by helping students discriminate between various related concepts and understand differences in how they map onto real-world contexts. Outcomes reported by Di Vesta and Peverly (1984) provide initial evidence for an advantage of interleaving versus blocking examples. As described earlier, Di Vesta and Peverly presented students with artificial concepts along with examples. Among other factors, their study included a manipulation of the sequence of examples (interleaved versus blocked). On the final test 2 days later, interleaving improved hit rates on the yes/no decision component of the test and enhanced subsequent classification accuracy.

Concerning definition availability, predictions for the effects of this factor are less straightforward, as the presence of the definition with the example may have both negative and positive effects on learning. In this vein, presence of the definition may lead students to focus more on studying the definition than the example and/or may produce relatively shallow processing of the example. Absence of the definition may lead students to attempt retrieval of the definition on each trial, and a wealth of research has established that retrieval practice enhances learning (for reviews, see Dunlosky et al. 2013; Rawson and Dunlosky 2012). Alternatively, attempting to retrieve the relevant definition from memory and then hold it in mind while studying the example may exceed the processing capacity of many learners. Presence of the example may thus free processing resources for mapping the ideas in the definition to the provided example, which may enhance conceptual learning. Given that the availability of definitions has several plausible and potentially offsetting effects on learning, our examination of this factor is largely exploratory.

Methods

Participants and Design Because of the difficulty of obtaining a sufficient number of Washington University students who had minimal prior learning for the target concepts, we limited data collection for Experiment 2 to undergraduates at Kent State University, who participated for course credit (n=197; 60 % female). Student characteristics are summarized in Table 2. Students were randomly assigned to one of five groups. Four versions of the definition-then-example (hereafter referred to as *DE*) group were defined by a 2 (blocked versus interleaved) × 2 (with or without concurrent definition) factorial design. We also included a definition-only group, for purposes of replicating the primary outcomes of Experiment 1a.

Materials and Procedure Materials were the same as in Experiment 1a. Procedures for the definition-only group and the DE-interleaved group were the same as in Experiment 1a. The procedure in the other three example groups was the same as in the DE-interleaved group with the following exceptions. In the DE-interleaved-with-definition group, the definition was

presented on the screen along with the concept name and example on each trial of blocks 2–6. In the DE-blocked group, the presentation schedule began with a study trial for the definition of one concept, followed by five study trials that each included a different example of that concept. The definition of the next concept was then presented, followed by five study trials for the examples of that concept, and so on for the remaining eight concepts. The procedure for the DE-blocked-with-definition group was the same except that the definition was presented on the screen along with the concept name and example on each trial.

The only other procedural change involved the inclusion of an additional secondary measure. After all study trials had been completed and prior to the filler task, participants were told that some of the questions on the upcoming test would present new real-world examples of the concepts they had just learned and that they would be asked to identify (by selecting from the list of concept names) which concept each example illustrates. For each concept, participants were asked "How confident are you that you will be able to accurately identify real-world examples that illustrate the following concept?" Participants indicated their judgment by moving a pointer on a sliding scale with the end points labeled "0 % confident" and "100 % confident." We refer to these subsequently as *concept learning judgments* (cf. category learning judgments; Jacoby, Wahlheim, and Coane 2010).

Results

Data for 13 participants were excluded from analyses based on evidence for non-compliance with task instructions. During the classification task, these participants had response times faster than 1 s for more than 30 % of the items (compared with 1.4 % in the remainder of the sample) and/or had mean response times less than 4 s per item (compared with M=14 s per item in the remainder of the sample). Concerning the number of concepts that participants reported learning in prior coursework, the distribution was again highly skewed. Outcomes reported below include data from participants who reported having learned three or fewer of the target concepts in prior coursework (69 % of participants; for number of concepts reported within this subset, M=0.7, median and mode=0; for the remaining 31 % of participants, M=6.4, median=6, mode=5). For archival purposes, outcomes for the remaining participants are reported in Table A1 in the Appendix.

For each outcome measure, we conducted two sets of analyses appropriate to the purposes of Experiment 2. To examine the direct replication of outcomes from Experiment 1a, the first analysis compared performance in the definition-only group and the DE-interleaved group. To evaluate the extent to which the effect of examples depends on the schedule of presentation and/or the availability of the definition during study, the second analysis compared the four example groups. Split-half reliability estimates for the three performance measures ranged from 0.80 to 0.89.

Classification Performance Classification performance for both studied and novel examples in all groups was significantly greater than chance (10 %), all ts>7.84. Concerning replication of key outcomes from Experiment 1a, performance for the DE-interleaved and definition-only groups is reported in Fig. 2. Classification performance for studied examples was greater for the DE-interleaved group than for the definition-only group ($t(55)$=3.59, p<0.001, d=0.95). Classification performance for novel examples was also greater for the DE-interleaved group than for the definition-only group ($t(55)$=3.40, p<0.001, d=0.90). Thus, we replicated both of the key outcomes from Experiment 1a.

Concerning extension beyond our initial findings, to what extent does the effect of examples depend on the presentation schedule and/or the availability of definitions during

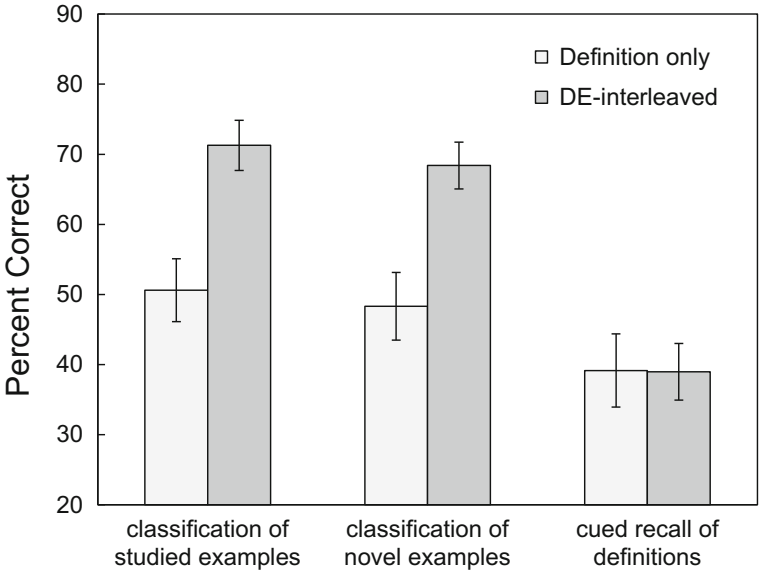

Fig. 2 Mean percent correct on three performance measures in Experiment 2 for the two groups that provide a direct replication of Experiment 1a. *Error bars* are standard errors of the mean

study? Classification performance for studied examples for the four DE groups is reported in the left panel of Fig. 3 (note that the leftmost bar in each panel reports the same outcomes for the DE-interleaved group as in Fig. 2 and is repeated here to facilitate comparison with the other DE groups). A 2 (presentation schedule: blocked versus interleaved) × 2 (definition: with or without concurrent definition) factorial ANOVA yielded only a significant interaction ($F(1, 93)=6.04$, MSE=544.18, $p=0.016$, $\eta_p^2=0.06$ (Fs<1 for main effects)). As shown in the right panel of Fig. 3, a similar pattern emerged for classification of novel examples ($F(1, 93)=7.58$, MSE=481.69, $p=0.007$, $\eta_p^2=0.08$ (Fs<1.46 for main effects)).

Concerning our predictions for the effect of presentation schedule, as expected, interleaving produced a sizeable advantage over blocking in the absence of definitions [for studied

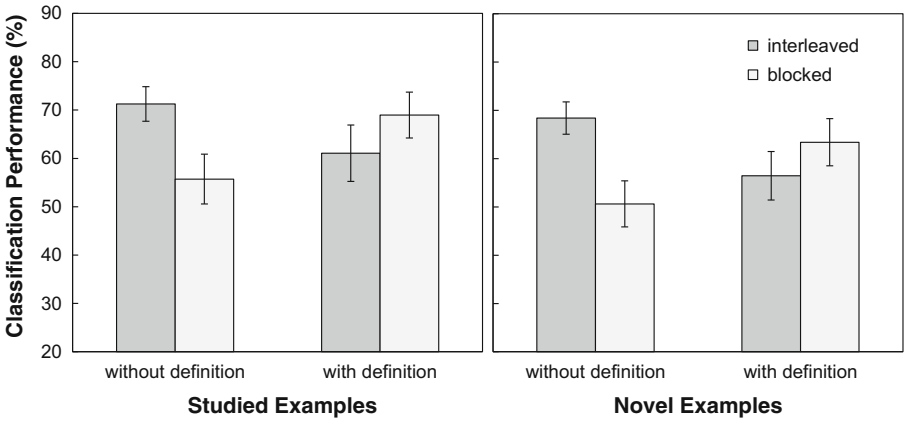

Fig. 3 Mean percent correct for classification of studied and novel examples for the four DE groups in Experiment 2. *Error bars* are standard errors of the mean

examples, $t(46)=2.40$, $p=0.011$, $d=0.70$; for novel examples, $t(46)=3.07$, $p=0.004$, $d=0.90$]. However, no interleaving effect emerged when definitions were present, with non-significant trends in the opposite direction [for studied examples, $t(47)=1.13$, $d=0.32$; for novel examples, $t(47)=1.02$, $d=0.29$].

Presenting definitions with examples tended to enhance classification performance when trials were blocked [$t(43)=1.78$, $d=0.53$ for studied examples; $t(43)=1.81$, $d=0.54$ for novel examples], but trends were in the opposite direction when trials were interleaved [$t(50)=1.67$, $d=0.46$ for studied examples; $t(50)=2.10$, $d=0.58$ for novel examples]. However, given that we made no a priori predictions for reasons noted earlier, use of a conservative correction for multiple comparisons is warranted, and none of these comparisons reached statistical significance with $\alpha=0.01$.

Secondary Outcome Measures For cued recall of definitions, the DE-interleaved and definition-only groups did not significantly differ (see Fig. 2; $t(55)=0.03$). Thus, in contrast to Experiment 1a but consistent with the pattern in Experiment 1b, the gain in classification performance from presenting examples was not offset by a cost in cued recall of definitions.

In a 2×2 ANOVA of cued recall in the four DE groups, only the interaction was significant ($F(1, 93)=4.05$, MSE=513.11, $p=0.047$, $\eta_p^2=0.04$ (Fs<1 for main effects)). Paralleling the qualitative pattern for the classification measures, cued recall was numerically (but not significantly) greater following interleaving versus blocking when definitions were not available during study [39.0 % (standard error (SE)=4.0) versus 30.2 % (SE=3.3), $t(46)=1.59$, $d=0.46$], with a non-significant trend in the opposite direction when definitions were available [33.6 % (SE=5.4) versus 43.5 % (SE=5.1), $t(47)=1.34$, $d=0.38$].

Mean study time during the learning phase is reported in Table 3. Comparing the two replication groups, a 6 (block) × 2 (group: DE-interleaved versus definition only) ANOVA yielded significant main effects of block [$F(5, 275)=37.49$, MSE=54.43, $p<0.001$, $\eta_p^2=.41$] and group [$F(1, 55)=5.05$, MSE=513.28, $p=0.029$, $\eta_p^2=.08$] and a significant interaction [$F(5, 275)=4.50$, MSE=54.43, $p=0.001$, $\eta_p^2=.08$]. The interaction largely reflects longer study times in the DE-interleaved group than in the definition-only group in later blocks of practice. However, these differences in study time are unlikely to explain the effect of examples on conceptual learning. In ANCOVAs controlling for study time in each block of trials, classification performance was still significantly greater for the DE-interleaved group than for the definition-only group for studied examples [$F(1, 49)=7.48$, MSE=460.32, $p=0.009$, $\eta_p^2=.13$] and for novel examples [$F(1, 49)=6.11$, MSE=478.56, $p=0.017$, $\eta_p^2=.11$].

Table 4 Mean judgment magnitudes and judgment accuracy for concept learning judgments in Experiment 2

	CLJ magnitude	CLJ accuracy	
		Studied examples	Novel examples
Definition only	66.9 (3.4)	0.26 (.07)	0.26 (.07)
DE interleaved	76.5 (2.3)	0.39 (.06)	0.38 (.07)
DE blocked	68.9 (3.6)	0.28 (.10)	0.37 (.09)
DE interleaved with definition	70.9 (3.9)	0.17 (.10)	0.19 (.08)
DE blocked with definition	69.8 (3.9)	0.30 (.09)	0.29 (.09)

Note: Standard errors of the mean are reported in parentheses. CLJ magnitude is the mean judgment on a scale from 0 to 100. CLJ accuracy is the mean intraindividual correlation between judgments and classification performance for the ten concepts; accuracy was computed separately for classification of studied examples and for classification of novel examples

CLJ concept learning judgments

Comparing the four example groups, a 6 (block) × 2 (presentation schedule: blocked versus interleaved) × 2 (definition: with or without concurrent definition) ANOVA yielded only significant main effects of block [$F(5, 465)=54.02$, MSE=24.39, $p<0.001$, $\eta_p^2=.37$] and presentation schedule [$F(1, 93)=4.22$, MSE=382.97, $p=0.043$, $\eta_p^2=.04$], all other Fs<2.09. Students spent more time with interleaving than with blocking, but the increase was relatively modest (collapsing across blocks, 23.9 versus 20.5 s per trial, respectively).

Finally, outcomes for the concept learning judgments are reported in Table 4. For each participant, we computed the mean judgment across the ten concepts. The first column of Table 4 reports the mean across individual values in each group. The effect of group in a one-way ANOVA was not significant ($F<1.21$). For each participant, we also computed judgment accuracy as the intraindividual gamma correlation between judgments and classification performance across the ten concepts. We computed two correlations for each participant, one between the participant's judgments and classification for studied examples and another between the participant's judgments and classification for novel examples. Means across intraindividual correlations are reported in Table 4.

Within the literature on metacognition, an issue of perennial interest concerns the extent to which students can accurately monitor how well they have learned information (Dunlosky and Metcalfe 2009), and interest in the extent to which students can accurately monitor their conceptual learning has recently emerged in this literature (Jacoby, Wahlheim, and Coane 2010). Judgment accuracy did not significantly differ as a function of group for either studied examples or novel examples (Fs<1). Collapsing across groups, judgment accuracy was significantly above zero but still relatively modest for both studied and novel examples [studied: $M=0.29$, SE=0.04, $t(117)=7.70$, $p<0.001$; novel: $M=0.30$, SE=0.04, $t(116)=8.31$, $p<0.001$]. This level of judgment accuracy for concept learning is consistent with the modest level of accuracy reported by Jacoby et al. (2010) for individuals learning bird categories ($M=0.28$). Although examination of concept learning judgments (CLJ) accuracy in the current study was largely exploratory, these outcomes provide further evidence that students are somewhat limited in the extent to which they can accurately assess their own conceptual learning.

General Discussion

Although presenting illustrative examples is a common pedagogical device used for instruction of declarative concepts, minimal research has examined the effects of illustrative examples on conceptual learning. The current work reports the first examination of declarative concept learning with versus without illustrative examples. Across all three experiments, a consistent pattern emerged: Providing illustrative examples enhanced conceptual learning relative to only providing concept definitions, as evidenced by more accurate classification of both studied and novel examples (with ds ranging from 0.74 to 1.67).

However, the effect of illustrative examples on declarative concept learning depends in part on the conditions under which those examples are presented. Although similar levels of performance obtained when examples were presented before versus after students studied the concept definitions (Experiments 1a–1b), classification accuracy did depend on the extent to which examples of different concepts were interleaved and whether definitions were presented along with the examples (Experiment 2). When definitions were not presented, an advantage of interleaving over blocking emerged, as has been shown in prior research on other kinds of concept learning. However, no interleaving effect emerged when definitions were present. An important implication of the latter outcome is that the common practice of

presenting illustrative examples in blocked fashion would substantially attenuate the benefits of examples unless the definitions are also available.

The latter outcome also departs from the typical pattern reported in prior interleaving research. However, no prior interleaving research has involved explicit presentation of definitional information during interleaved practice. The extent to which interleaving effects on other kinds of concept learning are diminished by presentation of definitional information is an interesting question for future interleaving research. The current finding also has implications for theoretical accounts of interleaving effects. As mentioned earlier, the prevailing theoretical account of interleaving effects states that interleaving is particularly effective for enhancing discrimination learning (Taylor and Rohrer 2010). However, this account does not afford a straightforward explanation for why the effects of interleaving would be moderated by the presence of definitional information. One possible explanation concerns the different kinds of processing involved in concept learning. Concept learning involves both inter-concept discrimination (learning to distinguish tokens of different types) and intra-concept similarity (identifying similarities between tokens of the same type). One possibility is that learners tend to focus on inter-concept discrimination in the absence of definitional information (as in all prior interleaving research), which would be facilitated by interleaved presentation. In contrast, presentation of definitional information may engender focus on intra-concept processing, which would be facilitated by blocked presentation (cf. the *multifactor account* of generation and testing effects, which assumes that various conditions can influence the extent to which limited processing resources are directed at item-specific encoding versus interitem relational encoding; Peterson and Mulligan 2013). Although speculative, this account points to an interesting direction for further theoretical research that would be informative to both the interleaving literature and the example-based learning literature.

Given the paucity of research examining the effects of illustrative examples on declarative concept learning, further research is clearly needed to establish the generality of the effects demonstrated here, as well as to explore other potential moderators of these effects. For example, although the current research focused on conceptual learning for students with minimal prior familiarity with the to-be-learned concepts, outcomes for students with higher familiarity (reported in the Appendix) suggest that knowledge level may be another potential moderator of the effects of illustrative examples. Whereas lower-familiarity students had consistently better classification accuracy in the definition-then-examples group than in the definitions-only group, minimal differences emerged for higher-familiarity students. Performance in Experiment 1b was near ceiling, and the sample sizes in Experiments 1a and 2 were relatively small, so these outcomes should be interpreted with caution. Additionally, lower-familiarity and higher-familiarity students likely differed on dimensions other than familiarity with the concepts.[3] Nonetheless, the diminished effects for higher-familiarity students are consistent with patterns observed in the related literature on worked examples. In brief, research on various kinds of problem-solving (e.g., algebra, geometry, physics) has shown

[3] Lower-familiarity and higher-familiarity students in each experiment did not significantly differ in age, education, or vocabulary (all *p*s 0.18–0.98, except for a significant 4 % difference in vocabulary and a 0.7-year difference in education favoring the higher-familiarity subset in Experiment 1a), although they may have differed on other factors not measured here. Differences in concept familiarity were related to the time of semester in which participants completed the experiment in Experiment 1a and Experiment 2, with 77 and 62 % of lower-familiarity participants completing the experiment in the first half of the semester versus only 36 and 33 % of higher-familiarity participants in the first half of the semester. Both of these experiments involved samples drawn from the Kent State participant pool, the majority of which consists of students enrolled in General Psychology (in which the relevant content domain tends not to be covered until later in the semester). In Experiment 1b, 72 and 70 % of lower-familiarity and higher-familiarity participants completed the experiment in the first half of the semester, but the Washington University participant pool includes a much large proportion of advanced undergraduates who likely had completed coursework in which they may have previously encountered the experimental concepts.

that final test performance is greater when novice learners are presented with worked examples during problem-solving practice versus problem-solving practice without worked examples. However, several studies found that the benefit of worked examples is diminished or even reversed for high-knowledge learners (for a review of *expertise reversal effects*, see Kalyuga, Rikers, and Paas 2012). Thus, systematic investigation of the effects of illustrative examples as a function of learner characteristics may be a fruitful direction for future research.

Other important directions for future research concern exploring the power of examples with other materials, measures, and under other learning conditions. Note that the declarative concepts used here were representative of natural categories rather than well defined. The possibility remains that illustrative examples may have weaker effects on conceptual learning for concepts that more closely resemble well-defined categories. Another outstanding issue concerns the extent to which the power of examples may be moderated by characteristics of the illustrative examples used, including example quantity, quality, prototypicality, and variability (e.g., with examples drawn from more versus fewer real-world contexts).

Concerning measures, as noted earlier, we used the classification task to measure conceptual learning because it is commonly used in the broader literature on concept learning. For practical purposes, this measure was also advantageous in that it taps a key learning goal for students—namely, that students can identify when concepts are applicable in real-world contexts. For theoretical purposes, this measure also afforded a straightforward test of the predictions of the transfer-appropriate processing framework, according to which performance is enhanced to the extent that the cognitive processes engaged during encoding overlap with those engaged during test. Although the nominal tasks during learning and the classification test were not identical (i.e., students in the example groups were explicitly told which concept the example illustrated during learning versus had to infer which concept the examples illustrated during test, particularly for novel examples), the learning task afforded practice with seeing how the concepts could be mapped onto aspects of real-world contexts. However, the functional overlap between the learning task and the final test may have benefited the example groups in ways other than enhancing the overlap of conceptual processes (e.g., although the particular items included in the novel classification task had not been practiced earlier, learners in the example groups may still have benefited from overall familiarity with a task that involved presentation of examples). With that said, other literatures have shown that task overlap per se does not always confer benefits for subsequent performance. For example, Karpicke and Blunt (2011) demonstrated that retrieval practice was more effective than concept mapping practice for enhancing performance on a subsequent concept mapping test, and Paas and Van Merriënboer (1994) demonstrated that studying worked examples of geometry problems was more effective than solving geometry problems for enhancing performance on a subsequent problem-solving test. Likewise, in the current experiments, one might reasonably have expected the definitions-only condition to have conferred greater benefits for definition cued recall, relative to the example groups. Nonetheless, an important direction for future research involves examining the effects of illustrative examples using other measures of conceptual learning (e.g., example generation, problem-solving).

Concerning learning conditions, it is potentially important that interleaving does not appear to be effective when the definition is provided. This intriguing outcome warrants independent replication and further investigation. Other important conditions worthy of future research include the timing of example presentation (e.g., distributed across sessions rather than within a single session) and the retention interval between studying examples and subsequent tests of conceptual learning. Further exploration of retention interval would be particularly informative, given that the experiments reported here involved a relatively short delay. In contrast, long-term retention of student learning is of practical interest, and thus exploring the longer-term effects of illustrative

examples on declarative concept learning is an important future direction. By comparison, retention interval has been shown to moderate the benefits of other learning techniques. For example, practice testing and spacing both have larger effects after longer versus shorter retention intervals (for reviews, see Dunlosky et al. 2013). These findings suggest the intriguing possibility that the benefits of illustrative examples may be even larger with longer retention intervals.

Conclusions

The current work provides the first definitive demonstration that presenting students with illustrative examples can significantly enhance conceptual learning for declarative concepts. Establishing the effects of illustrative examples is important given that this pedagogical device is frequently used in practice, despite the fact that no prior research had demonstrated its efficacy—indeed, the only outcomes available from prior research suggested that illustrative examples had minimal to no effects on declarative concept learning. The current work also provides a point of departure for several interesting directions for future research to further establish and explain the effects of illustrative examples. As noted above, many important factors remain to be explored (including characteristics of learners, concepts, examples, and learning conditions). Thus, the current work reveals only the tip of the iceberg with respect to the power of examples to enhance conceptual learning.

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

Table 5 Performance on primary dependent variables for participants who indicated pre-experimental familiarity with four or more concepts

	Classification of examples		Cued recall
	Studied	Novel	
Experiment 1a			
Definitions only (n=17)	64.9 (6.1)	65.9 (5.8)	58.2 (7.2)
Definitions then examples (n=10)	69.6 (7.1)	64.4 (8.1)	42.7 (8.3)
Examples then definitions (n=17)	70.7 (4.6)	65.6 (4.7)	42.0 (4.9)
Experiment 1b			
Definitions only (n=24)	90.0 (1.7)	89.3 (2.0)	81.0 (3.5)
Definitions then examples (n=29)	93.8 (1.5)	92.8 (1.6)	75.8 (3.2)
Examples then definitions (n=44)	89.9 (1.9)	86.6 (2.2)	65.8 (3.0)
Experiment 2			
Definitions only (n=7)	50.3 (7.5)	47.2 (9.6)	39.0 (11.0)
DE interleaved (n=12)	44.6 (7.1)	38.8 (7.1)	19.1 (4.6)
DE blocked (n=15)	57.3 (4.2)	52.2 (3.2)	35.2 (5.0)
DE interleaved with definitions (n=10)	57.8 (7.5)	53.6 (6.5)	39.0 (7.2)
DE blocked with definitions (n=14)	62.2 (6.2)	58.3 (7.0)	39.5 (7.0)

Note: Standard errors of the mean are reported in parentheses

References

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General, 120*, 3–19.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnosis task. *Journal of Experimental Psychology: General, 120*, 278–287.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks: Sage.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380.

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

DeCaro, M. S., & Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. *Journal of Experimental Child Psychology, 113*, 552–568.

Di Vesta, F. J., & Peverly, S. T. (1984). The effects of encoding variability, processing activity, and rule-examples sequence on the transfer of conceptual rules. *Journal of Educational Psychology, 76*, 108–119.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Beverly Hills: Sage.

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.

Griffin, M. M. (1993). Do student-generated rational sets of examples facilitate concept acquisition? *Journal of Experimental Education, 61*, 104–115.

Hambrick, D. Z. (2003). Why are some people more knowledgeable than others? A longitudinal study of knowledge acquisition. *Memory & Cognition, 31*, 902–917.

Hambrick, D. Z., Meinz, E. J., Pink, J. E., Pettibone, J. C., & Oswald, F. L. (2010). Learning outside the laboratory: Ability and non-ability influences on acquiring political knowledge. *Learning and Individual Differences, 20*, 40–45.

Hamilton, R. (1990). The effect of elaboration on the acquisition of conceptual problem-solving skills from prose. *Journal of Experimental Education, 58*, 5–17.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*, 141–158.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1441–1451.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach*. New York: Harcourt Brace Jovanovich.

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.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2013). The cost of concreteness: The effect of nonessential information on analogical transfer. *Journal of Experimental Psychology: Applied, 19*, 14–29.

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

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Klausmeier, H. J., & Feldman, K. V. (1975). Effects of a definition and a varying number of examples and nonexamples on concept attainment. *Journal of Educational Psychology, 67*, 174–178.

Libarkin, J. (2008). *Concept inventories in higher education science.* National Research Council, Promising Practices in Undergraduate STEM Education Workshop 2.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207–238. 31 pgs.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519–533.

Murphy, G. (2004). *The big book of concepts*. Cambridge: MIT Press.

Myers, D. G. (2010). *Psychology: Ninth Edition*. New York: Worth Publishers.

Needham, D. R., & Begg, I. M. (1991). Problem-oriented training promotes spontaneous analogical transfer: Memory-oriented training promotes memory for training. *Memory & Cognition, 19*, 543–557.

Nosofsky, R. M. (1999). Examplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 700–708.

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.

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1287–1293.

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review, 24*, 419–435.

Reder, L. M., & Anderson, J. R. (1982). Effects of spacing and embellishment on memory for the main points of a text. *Memory & Cognition, 10*, 97–102.

Reed, S. K., & Bolstad, C. A. (1991). Use of examples and procedures in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 753–766.

Rittle-Johnson, B., & Star, J. R. (2011). The power of comparison in learning and instruction: Learning outcomes supported by different types of comparisons. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (Vol. 55, pp. 199–222). San Diego: Elsevier.

Roediger, H. L., Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 3–41). Hillsdale: Erlbaum.

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review, 24*, 355–367.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Schacter, D. L., Gilbert, D. T. & Wegner, D. M. (2009). *Psychology*. New York: Worth Publishers.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology, 103*, 759–775.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18*, 565–583.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology, 24*, 837–848.

van den Broek, P. (2010). Using texts in science education: Cognitive processes and knowledge representation. *Science, 328*, 453–456.

Zachary, R. (1986). *Shipley Institute of living scale revised manual*. Los Angeles: Western Psychological Services.

Zimbardo, P. G., Johnson, R. L. & McCann, V. (2012). *Psychology: Core Concepts (Seventh Edition)*. Upper Saddle River, NJ: Pearson.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185.