Teaching of Psychology
2022, Vol. 0(0) 1–8

DOI: 10.1177/00986283211058069
journals.sagepub.com/home/top

The Use of Concrete Examples Enhances the Learning of Abstract Concepts; A Replication Study

Alexia Micallef[1], and Philip M. Newton[1]

Abstract

Background: Prior research suggests that the teaching of abstract concepts can be enhanced by the use of concrete examples, but there are few controlled studies.
Objective: To replicate key findings from experiment one from Rawson et al. (2015).
Method: Experiment participants studied definitions of abstract concepts from psychology, either with or without concrete examples. The replication differed from Rawson et al. by using a paid online participant pool, of non-psychology students, and a trimmed methodology focused on the key outcome.
Results: Concrete examples enhanced learning of abstract concepts. The critical finding was enhanced recognition of previously unseen examples matched to learned definitions, thus replicating the results of Rawson et al., with an effect size $d = 0.30$.
Conclusion: The use of concrete examples was found to enhance learning of abstract concepts when teaching concepts from psychology to non-psychology students using an online paid participant pool.
Teaching Implications: The teaching of abstract concepts in psychology could be helped by frequent use of concrete, real-world examples.

Keywords

learning science, teaching effectiveness, instructional methods, replication

Introduction

The evidence base for current teaching practice in Higher Education is mixed, and many questionable methods appear to persist despite years of research demonstrating their ineffectiveness (Newton et al., 2020; Newton & Salvi, 2020). Replication of results from key studies in education research is essential to confirm, or not, the generalisability of findings (Simons, 2014), perhaps even more important than the generation of new findings (Makel & Plucker, 2014). There has been, for many decades, calls for increased use of replication (e.g. (Good, 1992), but these calls remain unheeded, particularly in the fields of Psychology and Educational Psychology (Plucker & Makel, 2021), despite journals such as Teaching of Psychology explicitly inviting replications of research that replicates findings that have implications for the teaching of psychology.

The use of so-called 'concrete', 'illustrative' or 'real-world' examples has been repeatedly proposed as an evidence-based way of enhancing the learning of abstract concepts (e.g. Deans for Impact, 2015; Nebel, 2020; Weinstein et al., 2018). Abstract concepts are defined by not having a physical form and so can be difficult for learners to process and understand (Harpaintner et al., 2018). A concrete example is defined as a real-world, or illustrative, example of an abstract or declarative concept (Rawson et al., 2015). An example of an abstract concept from the current study is 'deindividuation', meaning the loss of a person's sense of individuality that results in a reduction of normal constraints against deviant behaviour. A concrete example of deindividuation might be when the anonymity of Internet chat rooms increases the willingness of an individual to be more candid than during face-to-face interactions. Since abstract concepts form a fundamental part of many disciplines of learning, including STEM subjects and Psychology (Hayes & Kraemer, 2017), there is a need for evidence-based approaches to facilitate their learning.

[1]Swansea University Medical School, Singleton Park Campus, Swansea, UK

Corresponding Author:
Philip M. Newton, Swansea University Medical School, Singleton Park Campus, Swansea, SA2 8PP UK.
Email: p.newton@swansea.ac.uk

A key part of the evidence base for the use of concrete examples is a study by Rawson et al. (2015) found that the use of concrete examples during learning significantly improved the conceptual learning of abstract concepts by psychology students. This study is widely cited in publications which give advice on evidence-based approaches to learning and teaching (e.g. Davidesco & Milne, 2019; Richmond et al., 2019; Wilson, 2019).

However, concrete examples that contain distracting features risk becoming 'seductive details' that may then *impair* learning (Harp & Mayer, 1998) and this may explain studies where concrete examples do not enhance learning (e.g. Kaminski et al., 2013). Prior research has identified the characteristics of the types of words used effectively as concrete examples (Caplan & Madan, 2016) and that participants respond to concrete words faster than to abstract words, suggesting that concrete words would be more useful than abstract words when providing contextual information (Xiao et al., 2012). Other studies have found that participants follow concrete examples of instructions more than the abstract instructions themselves (LeFevre & Dixon, 1986; Tse et al., 2016).

Thus, the evidence base for the use of concrete examples in teaching would appear to be mixed. Given the potential significance of Rawson et al. for the teaching of psychology, but set against some mixed findings from other studies, we tested the replicability of the key finding from Rawson et al. Rawson gave their participants definitions of some abstract ideas from psychology, followed by multiple different concrete examples of those ideas. A control group received only the definitions, repeatedly. Both groups were then tested to determine whether they could match examples to definitions. The group which had received the concrete examples were better able to match definitions to examples including, critically, examples that they had not previously seen. Rawson et al. suggested their study was amongst the first study of its kind to use a 'no-example' control group, a design which considerably strengthened the conclusions but highlighted the paucity of well-controlled research into the application of this idea to learning and teaching, and further emphasised the need for replication of the findings from this key study.

In keeping with the principles of replication studies requested by Teaching of Psychology, we designed a constructive replication (Hüffmeier et al., 2016; Plucker & Makel, 2021) which tested replication of the key concept from Rawson et al. but designed in some differences to extend the study; we used an online labour market (www.prolific.co) to recruit participants rather than using an experimental computer laboratory setting (for more information on Prolific see (Palan & Schitter, 2018). We also deliberately used participants who were currently studying non-psychology subjects to determine whether the findings of Rawson et al. could be extended outside of psychology. We also designed a test of baseline knowledge in order to objectively determine whether participants learned more effectively using concrete examples. Rawson et al. also included a vocabulary test but did not report analyses based upon these data, so they were not included in our study.

We hypothesised that we would replicate the main finding of Rawson et al. (2015) that the use of concrete examples would improve the learning of abstract concepts, this time by students who were not studying Psychology as an honors degree.

Method

Participants

Participants ($n = 85$ per group, $N = 170$) were recruited using www.Prolific.co with the following demographic characteristics: current undergraduate students, studying any discipline other than Psychology. The mean age of the participants was 23.46 ± 0.36. The participant group was 42% female. Once the participants accepted, they were taken to the surveys, hosted on Qualtrics™ as described above. The criteria in Prolific were set to ensure that no participant was allowed to complete both surveys.

Materials and Procedure

The current study was a replication of the key findings from Experiment 1 of Rawson et al. (2015). Key similarities and differences are shown in Table 1 and explained further below. The steps are provided in chronological order. The detail of any differences is explained in the text below where relevant. Professor Rawson graciously provided the experimental materials from the 2015 study. Examples of abstract concepts, their definitions and a concrete example are in Table 2.

The Survey Structure

The design of the study instruments is illustrated in Table 3. We constructed two separate survey instruments using Qualtrics™. One instrument for the control 'Definitions only' group (DEF), the other for a group that received definitions followed by Concrete Examples (ConEx). Rawson et al. included a third group that received concrete examples followed by definitions but did not find a substantial difference between that group and the ConEx group and so we did not replicate that group.

Instructions

Participants were given the purpose of the study, which was to understand how best to help students learn, and a brief overview which included instructions to proceed at their own pace. Participants were asked not to use any external sources to aid their answers and told that they would be paid 4.00 GBP for any satisfactory contribution that they made. The study sections are shown below; each one followed immediately after the next, with no time limit on each section.

Table 1. Key Similarities and Differences Between the Current Study and Rawson et al. (2015).

	Rawson et al. (2015)	Current study
Study environment	'Computer'	Online
Participant pool	Students from two campuses	From www.prolific.co
Participant profile	Psychology students	Non-psychology students
Participant reward	Course credit or monetary compensation (value not stated)	Monetary compensation (4.00 GBP, 4.90 USD)
Number of groups	3	2
Number of abstract concepts in study trials	10	5
Number of concrete examples used per concept, in study trials	5	5
Baseline recall test	No	Yes
Mathematics filler test	Yes	Yes
Cue recall test	Yes – free text	Yes – MCQ (repeat of baseline recall test)
Example classification test	Yes – 100 questions (10 per concept)	Yes – 50 questions (10 per concept)
Number of studied concrete examples tested in example classification test	50	25
Number of novel concrete examples tested in example classification test	50	25
Post hoc 'Preliminary knowledge' questions	Yes	No
Demographics questionnaire	Yes	No
Vocabulary test	Yes	No

Table 2. Concrete Examples of the Concrete Examples used in this Study from Rawson et al. (2015).

Abstract idea: Fundamental attribution error: Definition: the tendency to believe that another person's behaviour is due to his/her disposition and to underestimate the impact of situations on his/her behaviour. Example: Stacy, a high school student, came home past her curfew. Her parents thought she was being careless, even though Stacy was actually late because she had to drive home slowly and cautiously due to thick fog on the roadway
Abstract idea: Deindividuation Definition: The loss of a person sense of individuality that results in a reduction of normal constraints against deviant behaviour. Example: Tribal warriors who depersonalise themselves with face paints or masks are more likely than those with exposed faces to kill, torture, or mutilate captured enemies

Baseline Recall Test. Both groups received five multiple-choice questions (MCQs) in the format 'Which of the following does the term 'X' describe?', with 'X' being one of the five abstract concepts. Ten definitions were given as potential answers, and participants were required to choose which of the 10 definitions they felt described the term 'X' best. Both groups received the same questions and potential answers for all five abstract concepts. This section was designed to determine baseline knowledge of the concepts to be compared to knowledge following the study trials. Rawson et al. (2015) did not include a baseline test but partially addressed this question by asking participants to self-report, *at the end of the experiment*, if they had previously studied (in class) any of the concepts. This subjective measure may be confounded by a priming effect wherein participants have been repeatedly asked about topics during the study itself. The self-report also relies on participants fully understanding the topics and so being able to report, accurately, whether they have previously studied them. Our approach is not without challenge since the baseline test may create a testing effect, but if the key finding can be replicated under both conditions (ours and Rawson's) then this will create confidence that it is not undermined by any confounds caused by the different approaches to testing baseline knowledge.

Study Trials. The DEF group were shown the definitions of the five abstract concepts six times over in random order, clicking through one at a time. The ConEx group were shown the same five definitions once, followed by the definitions again five times over, including a different concrete example alongside it each time, again making 30 trials also in random order. These were followed by a *Filler* section wherein participants were given 20 basic mathematics tasks (e.g. 'what is the square root of 64' and 'what is 9 x 12').

Table 3. Structure of the Two Survey Instruments used in this Study.

	Definition Only (DEF)	Concrete Example (ConEx)
Baseline recall test	MCQs of 5 definitions Question: 'Which of the following does the term X best describe?' where X is an abstract concept, with 5 definitions as answer options	
Study trials (Read only, 30 trials total)	Definitions of 5 different abstract concepts presented 6 times each, in random order	Definitions of 5 different abstract concepts presented once Definitions of same 5 concepts, each followed by one of five examples. Five times for each example, with a different definition each time. Random order
Filler test	20 Multiplication problems (~5 mins)	
Repeat of recall test	Repeat of baseline recall test	
Example classification test	Q: 'The following concrete example describes what abstract concept?' Answer options; 10 abstract concepts from Rawson et al., including the 5 from the ConEx study trials, random order. Examples: 25 from study Trials, 25 novel, 5 per concept, random order with none repeated [5 definitions x 10 examples + 2 attention check questions = 52 MCQs]	

Repeat of Baseline Recall. The baseline MCQ recall test was then repeated to test whether participants had learned the definitions alone, and if there was any difference in learning the definitions between the two groups. The format of this was different to Rawson who used free text to test recall of definitions.

Example Classification Test. A 50-question multiple-choice example classification test was given to both groups. This consisted of 10 questions for each of the five abstract concepts presented in the study trials. Questions were in the format of 'The following concrete example describes what abstract concept?', each followed by a list of all 10 abstract concepts from Rawson et al. as answer options. Twenty-five questions were concrete examples that had been presented in the study trials (five per concept). Twenty-five questions were novel concrete examples (five per concept). All participants received all 50 questions in random order.

The end of the survey asked participants if they understood the instructions provided. Participants were then thanked and debriefed regarding the study aims. Participants were paid 4.00 GBP (~4.90 USD at the time of the study [early 2020]).

We piloted the survey instruments twice, initially with six (three per instrument) local students. Following this, we added a progress bar. We then repiloted to 10 (five per instrument) paid participants on Prolific. Following this, we included the baseline recall test and two attention check questions in the Example Classification test (these were regular scenarios as in the other questions, but the text directed participants to pick a specific answer). The finalised surveys contained 114 items total.

Analysis

We calculated the following scores for each participant; baseline recall test, repeat of baseline recall test, example classification test (25 studied), example classification test (25 novel), example classification test (all 50) and time taken to complete (seconds). Scores are reported as mean ± SEM. The majority of datasets were not normally distributed according to either a D'Agostino and Pearson omnibus test or a Kolmogorov–Smirnov test (Mishra et al., 2019) and so non-parametric tests were used throughout. A Wilcoxon matched pairs test was used when analysing paired data, while Mann–Whitney U tests were used for unpaired data. Two-tailed tests were used throughout. The cut off for significance was $p \leq 0.05$. Effect sizes were calculated using Cohen's d, as per Rawson et al. (2015).

Data Quality

There is concern that respondents in paid online surveys may answer randomly or use the minimal required behaviour (satisficing) (Anduiza & Galais, 2017), or that answers may come from 'bots' (Chmielewski & Kucker, 2020), although this is obviously more of a concern in surveys of subjective opinion. Rawson et al. (2015) removed participants who averaged under 4 seconds per item on the example classification task, on the basis that this suggested 'that participants were skipping quickly through many of the test items rather than attempting to answer each one' (Rawson et al., 2015, p. 491). Rawson et al. did not exclude participants on the basis of their performance. We therefore analysed the entire dataset initially to determine whether the main findings could be replicated, and then whether they were affected by the data quality analyses undertaken by Rawson.

Results

Baseline Recall Test and Repeat of Baseline Recall Test

Both groups showed a low baseline knowledge and showed an increase in performance following the study trials. These increases were statistically significant for both the participants in the DEF group (2.05 ± 0.15 vs. 3.56 ± 0.19, $W = 2193$, $d = 0.88$, $p < .0001$) and the ConEx group (2.49 ± 0.18 vs 3.55 ± 0.19, $W = 1524$, $d = 0.59$, $p < .001$). There was no difference

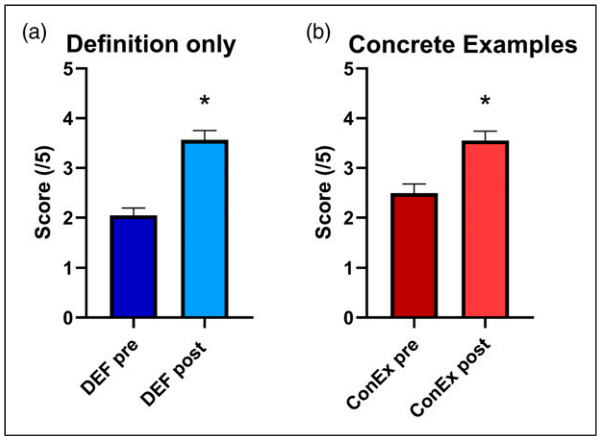

Figure 1. Both groups showed improved performance on recall tests following study trials, with no significant difference between groups either before or after the trials.

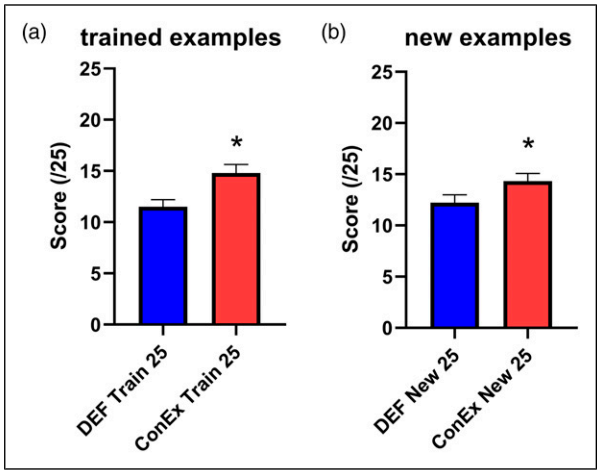

Figure 2. The group trained using Concrete Examples in the Study Trials were better able to identify concrete examples that they had seen previously (A) and concrete examples that were new to both groups (B).

between the DEF and the ConEx group either before (U = 3179, p = .1679, d = 0.29) or after (U = 3550, p = 0.837, d = 0.01) the study trials. As illustrated in Figure 1, both groups showed improved performance on recall tests following study trials, with no significant difference between groups either before or after the trials.

Example Classification Test. The ConEx group performed significantly better than the DEF group on the Example Classification test for both the 25 studied examples (DEF 11.51 ± 0.69, ConEx 14.81 ± 0.82, U = 2625, p = 0.0019, d = 0.46) and the 25 novel examples (DEF 12.21 ± 0.77, ConEx 14.32 ± 0.76, U = 2982, p = 0.0489, d = 0.30). As illustrated in Figure 2, the group trained using Concrete Examples in the Study Trials were better able to identify concrete examples that they had seen previously (A) and concrete examples that were new to both groups (B).

Time Taken

There was a significant difference in the time taken to complete the study. The participants in the ConEx group took longer (2304 seconds ± 98) than the participants in the DEF group (2100 ± 102), d = 0.22. This difference was significant according to a Mann–Whitney U test (p = 0.0458, U = 2972) and again replicates results from Rawson et al. (2015). Spearman rank correlation test revealed a significant positive correlation between time, and score on the 50-item example classification test (r = 0.289, p < .001).

Data Quality

We were able to successfully replicate the main findings of Rawson without the need to eliminate participants and so data are reported from all participants. Rawson et al. removed 2.6% (8/307) participants who they considered completed the task too quickly. Removing the 2.6% (four) participants who completed our task quickest strengthened our main finding of interest (performance on the untrained 25 items in the example classification test), the p-value was reduced (from 0.0489 to 0.0284) and the effect size increased (from 0.3 to 0.34), in line with work suggesting that elimination of responses from outlier participants can increase effect sizes and reduce error (Berinsky et al., 2016). We also analysed performance on the twenty mathematics filler questions, for all participants. The mean score was high (18.25 ± 0.18), suggesting good engagement, with no difference between groups according to a Mann–Whitney U test (p = 0.1544, U = 3170).

Discussion

Consistencies and Inconsistences with Rawson et al. (2015)

The aim of this study was to test a key finding from Rawson et al. (2015) that learning abstract concepts using concrete examples is more effective than using definitions alone. We replicated the key finding, with an effect size of 0.30 for the novel examples that had not been seen by either group. This is smaller than the effect sizes obtained by Rawson et al. who conducted the study on two different populations, with effect sizes of 0.40 and 1.22, respectively. Both studies used the same outcome to measure performance and generated effect size (a MCQ-based classification test). Our study used participants who were students from a discipline other than psychology, which may partially explain the smaller effect size.

As described in the introduction, much of the current literature on concrete examples focuses on explanations based in cognitive psychology to explain the concrete examples effect (Rawson et al., 2015; Weinstein et al., 2018). In both our study and Rawson et al., participants in the concrete examples took longer to complete the study. Rawson et al. conducted ANOCOVA analysis which suggested that performance in the examples group was unrelated to time-on-task. We are unable to replicate that specific analysis as our set up did not allow for

us to separate out time spent on the study trials. However, we did find that the concrete examples group took significantly longer to complete the study overall. We also found a significant correlation between time taken to complete the study and performance on the 50-item example classification test. This correlation was modest, but suggests that the effect of concrete examples may, in our study at least, be partially explained by time-on-task. The additional reading material required to engage with the concrete examples may force learners to spend more time studying the concepts. From our pragmatic perspective (seeking to enhance and improve teaching), this competing explanation is not a concern as the net result is an improvement in learning, but it does suggest potentially important avenues for further research; is the effect simply explained by time-on-task, and/or do students find the examples more engaging than the definitions, or easier to visualise, or to relate to their own experience? Understanding these details will help us with future applications of the principle.

Our study also demonstrates again that key findings from experimental educational psychology can be replicated using an online labour market. There are a number of advantages and disadvantages to such participant pools; for example, they can deliver a broader demographic range than traditional undergraduate student pools, data can be generated very quickly and the participants are completely anonymous. This facilitates the undertaking of replication of education research studies such as the one undertaken here. However, the literature on online participant pools contains concerns about their use which should be considered when discussing the current findings. For example, there are concerns about the use of bots to earn money (Chmielewski & Kucker, 2020). In education research studies such as theses, it is also possible that participants might use external sources to answer questions rather than recalling from their own knowledge, although this should be less of a concern in the current study since participants are paid a flat fee for engaging with the task rather than performance-based remuneration dependent upon the number of correct answers.

The elimination of responses from online participants who fail data quality checks can increase effect sizes and reduce error (Berinsky et al., 2016), but can be a subjective process, and risks inducing biases into the dataset (Berinsky et al., 2014; Vanette, 2017), particularly in a study such as ours where there are a number of objective measures which may indicate that participants are not fully engaged (performance on both recall tests, classification tests, the maths filler tests, time to complete the surveys, answers to attention check questions and the question which asked participants whether they had understood the instructions). We were able to reproduce the findings of Rawson without discarding participants, using a sample size that was larger than that used by Rawson but still modest. Discarding participants using criteria defined by Rawson et al. increased the effect size for the key finding, as expected. In addition, an analysis of the filler test data suggests that our participants were engaged with the task

Educational Implications

Our findings, combined with those of Rawson and others, suggest that the teaching, and learning, of abstract ideas in psychology can be enhanced when those ideas are paired with concrete examples. An obvious application of this is for instructors to use concrete examples in their teaching. A less obvious, yet potentially even more powerful, application of this principle is to ask students to generate their own concrete examples, or to explain why a particular example is/is not a good illustration of a specific abstract idea. These approaches pair the concrete examples effect with the principles of retrieval practice and elaboration, both of which are also effective, evidence-based teaching techniques (Weinstein et al., 2018). Asking students to generate their own examples is an effective teaching and learning strategy (Rawson & Dunlosky, 2016), although students may struggle to judge the quality of their self-generated examples, even with feedback (Zamary et al., 2016) A final suggestion is for instructors to reflect on their expert understanding of these abstract ideas versus that of their students, for whom these ideas may be new and so truly abstract. The introduction of esoteric, abstract ideas and the associated jargon is potentially a source of high cognitive load that can delay learning for novices but may be missed by an instructor for whom these ideas and terms are part of their everyday language (Sweller, 2010).

Conclusion

Limitations and Future Research

The findings of our study warrant further investigation into the effect of concrete examples, for example, in other disciplines where there is an abundance of abstract concepts (STEM, Law), and even using additional abstract concepts in psychology. Perhaps most importantly, there is a need to test the effect on concrete examples in a more naturalistic teaching setting; the effect size found here is modest and may be partially due to the unnatural paradigm, particularly for the definitions only group.

It would also be valuable to study the implementation of the concrete examples effect alongside other evidence-based teaching strategies. A number of evidence-based strategies have been developed, including some, such as retrieval practice, whose effectiveness has been demonstrated in naturalistic teaching settings (Rowland, 2014). Most interventions tested in education research have a modest effect size, in keeping with that achieved here (Schneider & Preckel, 2017). The impact of combining multiple interventions into an evidence-based programme of study has not been fully studied.

We were able to replicate the main finding of Rawson et al. (2015) that the use of concrete examples was more effective tool for the study of abstract concepts, when compared to studying definitions alone. Our replication extended the findings of Rawson, by using non-psychology students, participating via an online labour market.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, *29*(3), 497–519. https://doi.org/10.1093/ijpor/edw007

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739–753. https://doi.org/10.1111/ajps.12081

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2016). Can we turn shirkers into workers? *Journal of Experimental Social Psychology*, *66*, 20–28. https://doi.org/10.1016/j.jesp.2015.09.010

Caplan, J. B., & Madan, C. R. (2016). Word imageability enhances association-memory by increasing hippocampal engagement. *Journal of Cognitive Neuroscience*, *28*(10), 1522–1538. https://doi.org/10.1162/jocn_a_00992

Chmielewski, M., & Kucker, S. C. (2020). An MTurk Crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Davidesco, I., & Milne, C. (2019). Implementing cognitive science and discipline-based education research in the undergraduate science classroom. *CBE—Life Sciences Education*, *18*(3), es4. https://doi.org/10.1187/cbe.18-12-0240

Deans for Impact. (2015). *The science of learning*. Deans for Impact. https://deansforimpact.org/resources/the-science-of-learning/ (accessed July 17 2017).

Good, R. (1992). Editorial. The importance of replication studies. *Journal of Research in Science Teaching*, *29*(3), 209–209. https://doi.org/10.1002/tea.3660290302

Harpaintner, M., Trumpp, N. M., & Kiefer, M. (2018). The semantic content of abstract concepts: A property listing study of 296 abstract words. *Frontiers in Psychology*, *9*, 1748. https://doi.org/10.3389/fpsyg.2018.01748

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology*, *90*(3), 414–434. https://doi.org/10.1037/0022-0663.90.3.414

Hayes, J. C., & Kraemer, D. J. M. (2017). Grounded understanding of abstract concepts: The case of STEM learning. *Cognitive Research: Principles and Implications*, *2*(1), 7. https://doi.org/10.1186/s41235-016-0046-z

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, *66*, 81–92. https://doi.org/10.1016/j.jesp.2015.09.009

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*(1), 14–29. https://doi.org/10.1037/a0031931

LeFevre, J.-A., & Dixon, P. (1986). Do written instructions need examples? *Cognition and Instruction*, *3*(1), 1–30. https://doi.org/10.1207/s1532690xci0301_1

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: replication in the education sciences. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189X14545513

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*, *22*(1), 67–72. https://doi.org/10.4103/aca.ACA_157_18

Nebel, C. (2020). Considerations for applying six strategies for effective learning to instruction. *Medical Science Educator*, *30*, 9–10. https://doi.org/10.1007/s40670-020-01088-8

Newton, P. M., Da Silva, A., & Berry, S. (2020). The case for pragmatic evidence-based higher education: A useful way forward? *Frontiers in Education*, *5*, 271. https://doi.org/10.3389/feduc.2020.583157

Newton, P. M., & Salvi, A. (2020). How common is belief in the learning styles neuromyth, and does it matter? A pragmatic systematic review. *Frontiers in Education*, *5*, 270. https://doi.org/10.3389/feduc.2020.602451

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Plucker, J. A., & Makel, M. C. (2021). Replication is important for educational psychology: Recent developments and key issues. *Educational Psychologist*, *56*(2), 90–100. https://doi.org/10.1080/00461520.2021.1895796

Rawson, K. A., & Dunlosky, J. (2016). How effective is example generation for learning declarative concepts? *Educational Psychology Review*, *28*(3), 649–672. https://doi.org/10.1007/s10648-016-9377-z

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*(3), 483–504. https://doi.org/10.1007/s10648-014-9273-3

Richmond, A., Cranfield, T., & Cooper, N. (2019). Study tips for medical students. *BMJ*, *365*, k663. https://doi.org/10.1136/sbmj.k663

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565–600. https://doi.org/10.1037/bul0000098

Simons, D. J. (2014). The value of direct replication: *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5

Tse, C. Y. A., Wong, A., Whitehill, T., Ma, E., & Masters, R. (2016). Examining the cognitive demands of analogy instructions compared to explicit instructions. *International Journal of Speech-Language Pathology*, *18*(5), 465–472. https://doi.org/10.3109/17549507.2015.1112834

Vanette, D. (2017). *Using attention Checks in your surveys may harm data quality.* Qualtrics. https://www.qualtrics.com/blog/using-attention-checks-in-your-surveys-may-harm-data-quality/ (accessed 25 December 2018).

Weinstein, Y., Madan, C. R., & Sumeracki, M. A. (2018). Teaching the science of learning. *Cognitive Research: Principles and Implications*, *3*(1), 2. https://doi.org/10.1186/s41235-017-0087-y

Wilson, G. (2019). Ten quick tips for creating an effective lesson. *PLOS Computational Biology*, *15*(4), e1006915. https://doi.org/10.1371/journal.pcbi.1006915

Xiao, X., Zhao, D., Zhang, Q., & Guo, C. (2012). Retrieval of concrete words involves more contextual information than abstract words: Multiple components for the concreteness effect. *Brain and Language*, *120*(3), 251–258. https://doi.org/10.1016/j.bandl.2011.09.006

Zamary, A., Rawson, K. A., & Dunlosky, J. (2016). How accurately can students evaluate the quality of self-generated examples of declarative concepts? Not well, and feedback does not help. *Learning and Instruction*, *46*, 12–20. https://doi.org/10.1016/j.learninstruc.2016.08.002