Received: 26 April 2017 | Revised: 11 January 2018 | Accepted: 28 February 2018
DOI: 10.1002/acp.3410

SHORT PAPER - 5000 WORDS OR LESS

WILEY

The learning benefits of teaching: A retrieval practice hypothesis

Aloysius Wei Lun Koh | Sze Chi Lee | Stephen Wee Hun Lim

Department of Psychology, Faculty of Arts & Social Sciences, National University of Singapore, Singapore

Correspondence
Stephen Wee Hun Lim, PhD, Department of Psychology, Faculty of Arts & Social Sciences, National University of Singapore, Block AS4, Level 2, 9 Arts Link, Singapore 117570.
Email: psylimwh@nus.edu.sg

Funding information
National University of Singapore Faculty of Arts and Social Sciences Heads and Deanery Research Support Scheme, Grant/Award Number:
R-581-000-150-133

Summary

Teaching educational materials to others enhances the teacher's own learning of those to-be-taught materials, although the underlying mechanisms remain largely unknown. Here, we show that the learning-by-teaching benefit is possibly a retrieval benefit. Learners (a) solved arithmetic problems (i.e., they neither taught nor retrieved; *control* group), (b) taught without relying on teaching notes (i.e., they had to retrieve the materials while teaching; *teaching* group), (c) taught with teaching notes (i.e., they did *not* retrieve the materials while teaching; *teaching without retrieval practice* [*TnRP*] group), or (d) retrieved (i.e., they did not teach but only practised retrieving; *retrieval practice* group). In a final comprehension test 1 week later, learners in the teaching group, as did those in the retrieval practice group, outperformed learners in the TnRP and control groups. Retrieval practice possibly causes the learning benefits of teaching.

KEYWORDS

educational psychology, learning-by-teaching, retrieval practice, retrieval-based learning, the testing effect

1 | INTRODUCTION

The very act of teaching educational materials to others has been observed to enhance the teacher's own learning of those materials (e.g., Galbraith & Winterbottom, 2011; Hoogerheide, Deijkers, Loyens, Heijltjes, & van Gog, 2016; Roscoe & Chi, 2007, 2008). This is known as the *teaching effect* (i.e., learning by teaching). The learning-by-teaching literature is dominated by studies examining the learning benefits in the context of peer tutoring (see Fiorella & Mayer, 2013, for a discussion). A meta-analysis of educational outcomes from 65 independent evaluations of school tutoring programmes (Cohen, Kulik, & Kulik, 1982) revealed that peer tutors themselves gained a better understanding of the subject matter that they taught during the tutorials. The view is that peer tutors could learn from their teaching experiences as a result of reflective knowledge building via interactive processes of knowledge construction while teaching someone else (Roscoe & Chi, 2007, 2008). Furthermore, studies (Fiorella & Mayer, 2013, 2014) have shown that the act of teaching led to significantly better learning—retention of knowledge—as compared with preparing to teach but without actually teaching. For instance, Hoogerheide et al. (2016; Experiment 2) tasked participants to study a text on syllogistic reasoning problems. Following that, participants were randomly assigned to three groups, in which they either restudied the materials, explained the text in writing, or explained via a video. Participants who taught the materials to others via video reportedly learned better than did those who wrote about or simply restudied the materials.

Within the teaching process, a considerable amount of time and effort is devoted to retrieving the material acquired during teaching preparation. There are, therefore, grounds on which to speculate whether the learning benefits observed in the standard learning-by-teaching literature might not actually be attributed to retrieval practice. In other words, it may be the *retrieval* of educational materials involved during the teaching process that enhances learning. The benefits of retrieval practice (also known as the *testing effect*; Roediger & Karpicke, 2006b) on learning are well established and documented in the literature (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Butler & Roediger, 2007; Chan & McDermott, 2007; Kang, McDermott, & Roediger, 2007; McDaniel, Howard, & Einstein, 2009; see, also, Karpicke & Grimaldi, 2012, for a comprehensive review). In the standard retrieval-based learning paradigm, learners either studied educational materials repeatedly or studied and then retrieved the materials, before taking a final test through which their learning was assessed. For example, in the highly cited study by Roediger and Karpicke (2006a), students studied an expository text once and underwent three free recall tests about the material, studied the passage three times and took one test, or basically studied the passage four times. They then took a final retention test either 5 min or 1 week later. Students who studied the material repeatedly performed better when the retention test was administered immediately. The crucial finding, however, was that the knowledge of the studied materials was better retained over time (1 week later) by students who practised retrieving, as compared with those who merely restudied. This demonstrated the benefits of retrieval practice on longer term retention of educationally relevant knowledge (see, also, Carpenter, 2012; Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Karpicke, 2012; Rawson & Dunlosky, 2012; Roediger & Butler, 2011).

The benefits of retrieval practice have been observed in a variety of learning tasks, including free recall of word lists (e.g., Tulving, 1967), paired-associate learning (e.g., Carpenter, Pashler, & Vul, 2006), foreign language vocabulary learning (e.g., Carrier & Pashler, 1992), learning content from prose passages (e.g., Roediger & Karpicke, 2006a), video-based learning (e.g., Butler & Roediger, 2007; Johnson & Mayer, 2009; Yong & Lim, 2016), analogical problem solving (Wong, Ng, Tempel, & Lim, 2017), and, of particular relevance to the present study, science-based texts (e.g., Blunt & Karpicke, 2014; de Jonge, Tabbers, & Rikers, 2015; Dobson, 2013; Hinze & Wiley, 2011; Karpicke & Blunt, 2011). For instance, in Blunt and Karpicke (2014; Experiment 2), undergraduate students studied science texts about "enzymes" and "domains of life." After studying the materials, students performed tasks according to the learning conditions that they were randomly assigned to: (a) repeated study—concept mapping, (b) repeated study—paragraph writing, (c) retrieval practice—concept mapping, and (d) retrieval practice—paragraph writing. On a final test administered 1 week later, during which learners were assessed in terms of both their verbatim and inferential knowledge, those who actively retrieved the materials during learning, regardless of the activity format, performed better than did those who studied repeatedly. These data were corroborated by de Jonge et al. (2015; Experiment 2), in which undergraduate students studied an incoherent text on "black holes." They then either restudied the text or practised retrieving via a fill-in-the-blank test (see, also, Hinze & Wiley, 2011). After the learning phase, participants took a final test either 5 min or 1 week later. Engaging in retrieval practice promoted long-term knowledge retention.

1.1 | The present study

The critical question we asked was whether the learning benefits observed in learning-by-teaching are possibly attributable to retrieval practice, given that learners, as they teach, had to primarily retrieve the materials. Students studied and prepared to teach a science text on the Doppler effect. After studying the text material once, they (a) solved arithmetic problems (*control* group), (b) taught the material without using teaching notes (i.e., actively retrieved the material during teaching; *teaching* group), (c) taught the material with teaching notes (i.e., there was no active retrieval during teaching; *teaching without retrieval practice* [*TnRP*] group), or (d) practised retrieving (without teaching) the material (*retrieval practice* group). Critically, a direct comparison of the teaching versus TnRP conditions would reveal whether the learning-by-teaching educational strategy works *only if it involved active retrieval* of the studied material. The total duration across all the conditions was matched. One week later, participants were assessed on their ability to explain key concepts regarding the educational material to which they were earlier exposed.

To the extent that the learning benefits arising from teaching are attributable to retrieval practice, we expected that final test performance in the teaching group would be as good as that in the retrieval practice group, relative to both the control and TnRP groups. Importantly, learners in the teaching group (as would those in the retrieval practice group) ought to outperform those in the TnRP group, even though both the teaching and TnRP groups had supposedly engaged in the act of teaching.

2 | METHOD

2.1 | Participants

One hundred and twenty-four undergraduate students (57 were female), aged between 19 and 25 (M = 21.7, SD = 1.63), from the National University of Singapore (NUS) were recruited for this study. Participants received either credits for an introductory psychology class or monetary compensation ($10 for an hour of participation). All students majoring in physics and/or have taken courses on the Doppler effect were excluded. This research was conducted with the appropriate ethics review board approval by the NUS, and participants have granted their written informed consent.

2.2 | Design

A between-subjects design was employed. The independent variable was learning strategy: (a) control (i.e., participants solved arithmetic problems, i.e., neither taught nor retrieved), (b) teaching (i.e., participants taught unaided from memory, i.e., with retrieval), (c) retrieval practice (i.e., participants did not teach but practised retrieving), and (d) TnRP (participants taught by reading a teaching script prepared by the experimenters beforehand). The dependent variable was the mean performance on the final comprehension test administered 1 week after. Each participant was randomly assigned to one of the four learning groups, with 31 participants in each group.

2.3 | Materials and tasks

2.3.1 | Demographics and prior knowledge questionnaire

Participants provided information relating to their age, gender, year of study, major of study, and faculty of study. They additionally indicated their prior knowledge of the Doppler effect on a 5-point Likert scale (1 = *very low*, 5 = *very high*) and checked off any of the following items that were applicable to them: (a) "I have taken a University course in Physics." (b) "I know what Hertz (Hz) means." (c) "I have used an oscilloscope." (d) "I know how radar works." (e) "I know the basic characteristics of sound waves." (f) "I know what relative motion is." (g) "I know what the red shift is." and (h) "I know what a sine curve is." The self-report on prior knowledge and the item checklist, combined, constituted the measure for participants' prior knowledge of the Doppler effect, with a total possible score of 13 (i.e., one point was allocated for each checked item on the list; the total score was derived by summing [a] the number of points arising from the checked items and [b] the knowledge rating of the Doppler effect).[1] Finally, participants rated how well they thought they would perform on a test after undergoing a lesson on the Doppler effect via a 5-point Likert scale: 1 = *very poorly*, 5 = *very well* (Fiorella & Mayer, 2013). This information was used to determine if the groups were equated in terms of their reported self-efficacies.

2.3.2 | Instructions for preparing to teach

All participants received the same set of prestudy instructions to induce teaching preparation. This way, participants across all groups prepared to teach (unaided, i.e., without reference to any notes) after they had studied the materials, when in fact only those in the teaching group would be asked to actually teach eventually.

2.3.3 | Doppler effect lesson

We adapted the lesson material from Fiorella and Mayer (2013), which conveyed foundational concepts of sound waves and discussed how the Doppler effect works. It comprised 585 words and five graphical figures.

2.3.4 | Arithmetic problems set

The set of arithmetic problems, adapted from Yong and Lim (2016), consisted of 55 multiplication questions.

2.3.5 | Free recall test

We designed the present free recall test in line with those used in standard retrieval-based learning studies (e.g., Dobson & Linderholm, 2015; Lim, Ng, & Wong, 2015; Roediger & Karpicke, 2006a; Yong & Lim, 2016), which consisted of instructions and sufficient blank space for participants to write on.

2.3.6 | Teaching script

We developed the teaching script from the Doppler effect lesson, which the participants in the TnRP group narrated verbatim during the teaching phase. The script was constructed on the basis of how a scripted lecture would be delivered in an actual educational setting. A brief pilot test was conducted for the purposes of adjusting the ease of readability of the script as well as its length (i.e., the time taken to read the script clearly and thoroughly was approximately 5 min). The finalized teaching script comprised 829 words.

2.3.7 | Guiding diagrams

We prepared guiding diagrams taken from the Doppler effect lesson and mounted them on the wall for the participants in the TnRP group to make reference to (e.g., point to using their fingers) as they taught (read verbatim) from their teaching script.

2.3.8 | Postexperimental questionnaire

This questionnaire consisted of phenomenological items measuring participants' subjective experiences during, and opinions towards, the learning phase. Participants were requested to report how much they agreed with each of seven statements on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*): (a) "I felt the subject matter was difficult." (b) "I enjoyed learning about the Doppler effect." (c) "I would like to learn this way in the future." (d) "I feel like I have a good understanding of how the Doppler effect works." (e) "After this lesson, I would be interested in learning more about the Doppler effect." (f) "I found the lesson about the Doppler effect to be useful to me." and, finally, (g) "I felt stressed while I was learning about the Doppler effect." In addition, participants rated the amount of mental effort they invested while learning about the Doppler effect on a 7-point Likert scale (1 = *very low effort*, 7 = *very high effort*). At the end, participants recorded any additional comments they had about the study.

2.3.9 | Instructions for final comprehension test

Participants were informed of the test 1 week later, its expected duration, and format and nature of the test questions via the instructions formulated by the experimenters.

2.3.10 | Final comprehension test

The final comprehension test comprised six free-response questions designed to assess participants' abilities to explain key concepts relating to the lesson on the Doppler effect (Fiorella & Mayer, 2013): (1) "Explain how the Doppler effect works." (2) "How could you increase the intensity of the Doppler effect?" (3) "Would the Doppler effect occur if the source was stationary and the observer was moving? Why or why not?" (4) "What would happen to the Doppler effect if the observer was moving at the same speed and in the same direction as the source? Explain your answer." (5) "How would an observer experience sound if the speed of the source was traveling faster than the speed of sound?" and, finally, (6) "What would happen to the Doppler effect if the source and observer were both moving towards each other on a parallel path, at a constant speed? Explain your answer." The six free-response questions focused either on recall of facts (e.g., (1) "Explain how the Doppler effect works.") or on application in novel situations (e.g., (6) "What would happen to the Doppler effect if the source and observer were both moving towards each other on a parallel path, at a constant speed? Explain your answer."). A standardized marking rubric was applied in scoring all test responses from participants (see Appendix A). On the basis of the marking rubric, each point was awarded for a response on the basis of both its (a) *content* (i.e., memory representation of concepts) and (b) *quality* (i.e., demonstration of understanding). Specifically, in such a question that focused on recall of facts as Q1, participants were expected to (a) remember the concepts accurately (i.e., "The Doppler effect involves sound and movement.") and (b) understand those concepts adequately (i.e., "The Doppler effect is due to the change in how sounds are perceived due to movement.") when providing their responses. Similarly, in such a

[1]This procedure for scoring prior knowledge was identical to that used by Fiorella and Mayer (2013).

question that focused on application in novel situations as Q6, participants had to (a) remember the concepts accurately (i.e., "The Doppler effect would be more intense than if one was stationary.") and (b) understand those concepts adequately (i.e., "Sound waves would change in frequency at a higher rate hence the Doppler effect would be more intense than if one was stationary.") when providing their responses. One point was awarded for each response only when *both* the criteria for (a) content and (b) quality were satisfied. The resultant scores from the final comprehension test thus served as a measure of both (a) *correct content* and (b) *adequate quality* in participants' responses. The maximum possible score on the comprehension test was 25 points.

2.4 | Apparatus

Two smart phones with video-recording capabilities—Apple iPhone 6 and Xiaomi Redmi 1S—were used to record the participants in the teaching group (as they taught unaided) and in the TnRP group (as they taught with the script).

2.5 | Procedure

Participants underwent the first part of the experiment individually in cubicles. The experimenter verbally introduced the experiment briefly and handed out the consent forms to participants. Participants also filled out the demographics and prior knowledge questionnaire.

Next, the experimenter handed out an instructions sheet with the goal of inducing teaching preparation in all participants. Specifically, participants were informed that they had a total of 10 min to study and prepare to teach a lesson on the Doppler effect, in a way as if they were teaching the material to a realistic audience and during which their teaching would be video-recorded. Participants were also told that they could take notes during the study phase on a blank sheet of paper provided, although they knew that both the lesson material and any notes recorded had to be surrendered after the 10-min study phase, as they were expected to teach the material unaided. The experimenter verbally reiterated the instructions where required and clarified any doubts that participants might have.

After the study phase, the experimenter collected the lesson material and any notes recorded by the participants. For the subsequent 5 min, the participants across the different conditions completed their respective tasks: solving arithmetic problems (control group), delivering a video lecture unaided by any teaching script (teaching group), taking a free recall test (retrieval practice group), and delivering a video lecture with reference to a teaching script (TnRP group). All participants across groups were told to fully utilize the time allotted to them.

Participants in the control group undertook a filler task that involved solving arithmetic problems for 5 min. The participants were told to prioritize accuracy over speed per se; thus, they did not have to feel pressured to finish all of the questions.

Each participant in the teaching group taught individually for 5 min and was video-recorded. Participants were requested to stand while teaching and were offered the option of using a blank flipchart mounted on the wall if they felt that would enhance the quality of their teaching.

The retrieval practice group took a free recall test for 5 min, writing down as much information as they could remember from what they had earlier studied (see, e.g., Blunt & Karpicke, 2014; Lim et al., 2015; Roediger & Karpicke, 2006a; Yong & Lim, 2016). Participants were allowed to write in point form and/or any intelligible short forms (e.g., f for frequency and λ for wavelength), write the points in any order, and draw diagrams to support their responses.

The participants in the TnRP group taught individually for 5 min and were video-recorded. Similar to participants in the teaching group, they were required to stand while teaching. The critical difference was that participants in the TnRP group taught by reading from the teaching script verbatim and, on the basis of designated prompts printed on the script, making references to the guiding diagrams mounted on the wall (rather than retrieving and drawing the diagrams by themselves). There was thus no need (grounds) for participants to retrieve any information from their memory while teaching.

After participants completed their respective tasks, the experimenter administered the postexperimental questionnaire. Upon completing the questionnaire, participants were thanked and reminded about the second part of the study. Participants were not informed of the final comprehension test in advance.

One week later, the participants returned for the second part of this experiment, sat away from one another at individual (partitioned) desks, and read the instructions individually for the final comprehension test on the Doppler effect. The experimenter reiterated the instructions and answered any administrative queries. Each test question on the final comprehension test was timed and completed independently from other questions. Three min were allowed in answering the first question, and 2 min for each of the remaining five questions.[2] Participants were guided via short interim instructions on the test paper to keep working on each question until the allocated time expired. After the test was completed, the experimenter debriefed and thanked the participants.

3 | RESULTS

3.1 | Participant characteristics[3]

Separate one-way analyses of variance (ANOVAs) were conducted, with learning strategy as the factor, on age, prior knowledge, and self-efficacies reported by the participants, in order to assess statistical equivalence of these parameters across the four groups. Also, chi-square analyses were conducted on gender and year of study, with learning strategy as the factor, to determine whether the distributions of gender and year of study were comparable across groups. All tests revealed no significant differences across groups on the respective variables, all $ps > .05$.

[2]This approach was identical to that adopted by Fiorella and Mayer (2013).

[3]The analyses were carried out the same way as did Johnson and Mayer (2009).

TABLE 1 Mean ratings of the postexperimental questionnaire items

Groups	Ratings							
	Difficulty	Enjoyment	Learn	Understand	Interest	Usefulness	Stress	Effort
Control	3.03 (1.43)	5.00 (1.21)	4.71 (1.19)	5.23 (1.06)	4.26 (1.32)	4.87 (1.23)	3.45 (1.69)	4.39 (1.45)
Teaching	3.06 (1.15)	5.13 (1.28)	5.00 (1.55)	4.97 (1.49)	4.71 (1.66)	4.84 (1.44)	3.45 (1.65)	4.23 (1.43)
Retrieval Practice	2.35 (1.60)	5.16 (1.53)	4.71 (1.49)	5.58 (1.52)	4.19 (1.25)	4.55 (1.65)	3.58 (2.05)	4.45 (1.43)
TnRP	2.58 (1.29)	4.77 (1.36)	4.23 (1.75)	5.48 (1.36)	4.10 (1.33)	4.55 (1.46)	3.74 (1.84)	4.81 (1.51)

Note. Standard deviations appear in parentheses. Participants were rated on a 7-point Likert scale. TnRP = teaching without retrieval practice.

3.2 | Reading time

Although a pilot test was conducted to ensure that the narration of the teaching script (by the participants in the TnRP group) could be completed within approximately 5 min, individual differences in reading speeds did arise inevitably, resulting in varying reading durations. Consequently, participants' reading times were submitted to a single-sample t test, in order to determine if the mean reading time significantly differed from the stipulated 5 min. The analysis showed that the mean reading time of the participants in the TnRP group (M = 4.98, SD = 0.446) was comparable with the intended duration of 5 min during the experimental phase, $t(30) = -0.235$, $p = .816$.

3.3 | Phenomenological item analyses

The ratings of the postexperimental phenomenological items were submitted to separate one-way ANOVAs, with learning strategy as the factor. There were no significant differences across groups as a function of all questionnaire items (*enjoyment*, *learn*, *understand*, *interest*, *usefulness*, *stress*, and *effort*), all $ps > .05$. The mean ratings of the postexperimental questionnaire items appear in Table 1.

3.4 | Scoring

Two independent coders scored 15% of the comprehension test scripts (19 out of 124 scripts) on the basis of the standard marking rubric, and the Pearson product–moment correlation (r) between the two sets of scores was .99, $p < .001$. The small proportion of disagreements was resolved via discussions between the two coders to arrive at 100% consensus. As the interrater agreement was high, the remaining test scripts were scored by one of the coders subsequently.[4]

3.5 | Organization of studied information

Preliminary analyses were conducted to elucidate differences, if any, in the way participants organized the studied information during their teaching or their retrieval across the different conditions. To facilitate these analyses, the sequence of the content in the Doppler effect lesson was first decomposed into its constituent components; the lesson was structured as follows: (a) a brief introduction on Doppler effect; (b) definitions and characteristics of sound waves, wave frequency, and wavelength; and (c) the application of the Doppler effect in real-life situations (see Appendix B). The way in which individual participants across the teaching, TnRP, and retrieval practice groups actually organized the sequence of information during their teaching, or their retrieval, was then compared against that of the prescribed model (three-part) sequence; these comparisons would have illuminated any deviation in the way participants organized the studied information during their teaching or their retrieval. A binary coding scheme was used; that is, participants either deviated or did not deviate from the prescribed model structure. It was observed that all participants in the TnRP and retrieval practice conditions modelled exactly after our prescribed model sequence during teaching or retrieval; all except two participants in the teaching condition modelled after the model sequence. In particular, these two participants did not introduce the Doppler effect; one of them started the teaching session with defining wave frequency, whereas the other began with a general statement "we hear a lot of sounds every day" and then proceeded with defining sound waves, although notably, the remainder sequence of each of their teaching sessions adhered with the prescribed model sequence (i.e., although they missed introducing the Doppler effect, they faithfully proceeded with definitions and characteristics of sound waves, wave frequency, and wavelength, and then the application of the Doppler effect in real-life situations). Notwithstanding, subsequent main analyses were conducted (a) *with* versus (b) *without* these two participants ("deviant learners") in the teaching condition, and effects persisted across both sets of analyses (see Section 3.6).

3.6 | Main analyses

A one-way ANOVA revealed a significant effect of learning strategy on final comprehension test performance, both with the inclusion of the two deviant learners in the teaching group, $F(3, 120) = 5.82$, $MSE = 9.56$, $p = .001$, $\eta^2 = .127$, and without them, $F(3, 118) = 5.60$, $MSE = 9.71$, $p = .001$, $\eta^2 = .125$. Specifically, participants in both the teaching group (with the two deviant learners, M = 13.3, SD = 2.69, vs. without them, M = 13.3, SD = 2.78) and the retrieval practice group (M = 13.6, SD = 2.80) outperformed those in the control group (M = 10.9, SD = 3.07), $t(60) = -3.30$, $p = .002$, $d = 0.839$; $t(58) = -3.16$, $p = .003$, $d = 0.817$; and $t(60) = -3.59$, $p = .001$, $d = 0.912$, respectively, reflecting the effectiveness of both teaching and retrieval practice as educational strategies. Importantly, the performance of participants across the teaching (with the two deviant learners, M = 13.3, SD = 2.69, vs. without them, M = 13.3, SD = 2.78) and *retrieval practice* (M = 13.6, SD = 2.80) groups was comparable, $t(60) = 0.370$, $p = 0.713$, $d = 0.095$

[4]This procedure was identical to that used by both Roediger and Karpicke (2006a) and Yong and Lim (2016).

and $t(58) = 0.397$, $p = 0.693$, $d = 0.104$, respectively, suggesting that both teaching and retrieval practice promoted learning to the same extent. Notably, the final performance of participants across the TnRP ($M = 11.5$, $SD = 3.70$) and control ($M = 10.9$, $SD = 3.07$) groups was comparable, $t(60) = -0.616$, $p = 0.540$, $d = 0.156$. This suggests that teaching from a script without retrieving the to-be-taught material did not enhance learning.

Most crucially, participants in the teaching group (with the two deviant learners, $M = 13.3$, $SD = 2.69$, or without them, $M = 13.3$, $SD = 2.78$), as did those in the retrieval practice group ($M = 13.6$, $SD = 2.80$), outperformed participants in the TnRP group ($M = 11.5$, $SD = 3.70$), $t(54.8) = 2.30$, $p = .026$, $d = 0.583$; $t(55.6) = 2.21$, $p = .031$, $d = 0.568$; and $t(55.9) = 2.57$, $p = .013$, $d = 0.653$, respectively. Altogether, this set of results revealed that the benefits observed in the learning-by-teaching strategy are attributable to retrieval practice; that is, the robust learning-by-teaching strategy works but *only* when the teaching involves *retrieving* the to-be-taught materials.

To ascertain whether prior knowledge influenced the present effects, a between-subjects analysis of covariance was conducted. The independent variable was learning strategy (control, retrieval practice, teaching, and TnRP), and the dependent variable was final test scores, with prior knowledge entered as a continuous covariate. The main effect of prior knowledge did not reach significance, regardless whether the two deviant participants were included, $F(1, 116) = 1.26$, $MSE = 9.78$, $p = .26$, $\eta^2 = .011$, or excluded, $F(1, 114) = 1.21$, $MSE = 9.94$, $p = .27$, $\eta^2 = .011$, suggesting that prior knowledge did not predict test performance. The Learning Strategy × Prior Knowledge interaction did not reach significance also, regardless whether the two deviant participants were included or excluded, both Fs < 1, suggesting that prior knowledge did not influence the effect of learning strategy on test performance. These observations likely stemmed from the lack of variability in participants' prior knowledge across all conditions, given that prior knowledge has primarily been controlled for at the start of the experiment (i.e., all students majoring in physics and/or have taken courses on the Doppler effect were excluded from the experiment). To confirm this interpretation, two separate one-way ANOVAs were conducted (in which the two deviant learners were either included or excluded, respectively), with learning strategy (control, retrieval practice, teaching, and TnRP) as the independent variable and prior knowledge as the dependent variable. Indeed, no significant effects emerged, both Fs < 1, implicating a lack of variability in prior knowledge in participants across conditions.

Additionally, to ascertain that the present effects persisted across all six test questions attempted by the participants who adopted different learning strategies, two separate one-factor multivariate ANOVAs (MANOVAs) were performed (in which the two deviant learners were either included or excluded, respectively), with learning strategy (teaching, TnRP, and retrieval practice) as the independent variable and participants' performance on the six test items as dependent variables. The MANOVAs indicated that the effects of learning strategy on test performance persisted across all six test items, regardless whether the two deviant learners were included, $F(12, 170) = 1.38$, $p = .18$, Wilks $\Lambda = .83$, $\eta^2 = .089$, or excluded, $F(12, 166) = 1.34$, $p = .20$, Wilks $\Lambda = .83$, $\eta^2 = .088$, suggesting that the present effects persisted across all test questions.

4 | DISCUSSION

We pursued the idea that the learning benefits observed in the standard learning-by-teaching literature could be attributed to retrieval practice, on the basis that successful teaching normally predicates on the teacher's ability to, first, remember and retrieve the to-be-taught material. We expected to observe good final test performances in both the teaching group and the retrieval practice group, in comparison with the control and TnRP groups. The crucial hypothesis was that learners in the teaching group, as well as those in the retrieval practice group, would outperform learners in the TnRP group. Our data supported the present predictions. The findings are compatible with those of Fiorella and Mayer (2013) and further illuminate the critical role of retrieval underpinning the learning-by-teaching educational strategy. Furthermore, it is noteworthy that learners in the teaching and retrieval practice groups did not differ from those in the TnRP and control groups in terms of any of their participant (e.g., prior knowledge and self-efficacies) or phenomenological (e.g., perceived task difficulty and effort) characteristics.

In adherence to the standard learning by teaching literature, the present study operationalized teaching as having participants stand while conveying the material, with the option of using flipcharts if these were deemed helpful in enhancing the teaching quality (see Fiorella & Mayer, 2013, 2014). Roscoe and Chi (2007, 2008) argue, however, that effective teaching (and learning) predicates on the nature of teacher–student interactions, including the quality of the explanations, answers, and feedback provided by the teacher (see, also, Cohen, 1986; Gartner, Kohler, & Riessman, 1971; King, Staffieri, & Adelgais, 1998). Specifically, the teacher's gains depend in part on whether he or she engages in *reflective knowledge building* (i.e., the extent to which teachers reflect on their own understanding of the material and integrate it with their own prior knowledge while teaching their students). In contrast, learning is, it has been argued, less likely to emerge when teachers engage in *knowledge telling*, where they simply summarize the material without integrating it with their prior knowledge (Mayer, 2005, 2009; Wittrock, 1989). Although the question of how we might best operationalize "teaching (and learning)" is beyond the intent and scope of our study, future studies should certainly assess the importance of retrieval practice across a variety of teaching scenarios and activities.

On the one hand, that we found no observable differences across the teaching versus retrieval practice groups suggests the possibility that both teaching and retrieval practice have been effective owing to, for instance, reflective knowledge building. On the other hand, it is prudent to recognize that having similar performances in two conditions does not necessarily denote similar or the same mechanisms at work. Future studies should test all of these (competing) interpretations directly.

Relatedly, teaching arguably involves a number of cognitive processes beyond retrieval. These include planning what material to (vs. not to) present, thinking about how to frame and express the material, and attempting to provide organizational structure (see, e.g., Nestojko, Bui, Kornell, & Bjork, 2014). In theory, any of these processes, which do not necessarily require retrieval, may boost comprehension performance. Additionally, it is conceivable that other potentially beneficial elaborative processes, such as self-explanation, were impeded in the TnRP condition, as the teaching script was provided. Furthermore, it is possible that there are synergistic benefits of encoding to teach versus actually teaching. Thus, the retrieval condition that best evaluates whether the learning benefit is fundamentally a retrieval benefit might be a retrieval condition in which learners are given “standard” learning instructions and then retrieval practice. All of these interesting prospects potentially contribute toward the broader goal of understanding the extent to which retrieval is indeed a critical component in a learning-by-teaching benefit.

That teaching can enhance learning particularly when the teaching involves active retrieval of the to-be taught materials not only can have important implications for the teaching-by-learning literature but also can inspire new ways of designing real-world educational activities involving teaching. For example, in order to insure that students and tutors learn and retain the educational material that they have prepared and presented in class, they ought to internalize the to-be-presented material prior to communicating it to an audience, rather than rely on study notes during the presentation process.

There is a solid body of work showing that teaching educational materials promotes the learning of those materials, as discussed earlier in Section 1. This study attempted to extend that literature by exploring a possible underlying mechanism—retrieval practice (see, e.g., Dunlosky et al., 2013, for a comprehensive review)—of the learning-by-teaching strategy in promoting long-term learning, and has the potential to fundamentally transform the way in which both researchers and educators have previously understood and viewed the teaching-by-learning strategy.

ACKNOWLEDGEMENT

This work was supported by a National University of Singapore Faculty of Arts and Social Sciences Heads and Deanery Research Support Scheme Grant (R-581-000-150-133) awarded to S.W.H.L.

CONFLICT OF INTEREST

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

AUTHOR CONTRIBUTION

The initial research concept came from S. W. H. L. and was developed by all authors. A. W. L. K. and S. C. L. performed the testing and data collection. A. W. L. K. and S. C. L. performed the data analysis and interpretation under the supervision of S. W. H. L.. A. W. L. K. and S. C. L. drafted the manuscript, and S. W. H. L. provided critical revisions. All authors approved the final version of the manuscript for submission.

ORCID

Stephen Wee Hun Lim http://orcid.org/0000-0003-3636-7587

REFERENCES

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. https://doi.org/10.1002/acp.1391

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849–858. https://doi.org/10.1037/a0035934

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279–283. https://doi.org/10.1177/0963721412452728

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*(5), 826–830. https://doi.org/10.3758/bf03194004

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, *20*(6), 633–642. https://doi.org/10.3758/bf03202713

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 431–437. https://doi.org/10.1037/0278-7393.33.2.431

Cohen, J. (1986). Theoretical considerations of peer tutoring. *Psychology in the Schools*, *23*, 175–186.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, *19*(2), 237–248. https://doi.org/10.3102/00028312019002237

de Jonge, M., Tabbers, H. K., & Rikers, R. M. J. P. (2015). The effect of testing on the retention of coherent and incoherent text material. *Educational Psychology Review*, *27*(2), 305–315. https://doi.org/10.1007/s10648-015-9300-z

Dobson, J. L. (2013). Retrieval practice is an efficient method of enhancing the retention of anatomy and physiology information. *Advances in Physiology Education*, *37*(2), 184–191. https://doi.org/10.1152/advan.00174.2012

Dobson, J. L., & Linderholm, T. (2015). Self-testing promotes superior retention of anatomy and physiology information. *Advances in Health Sciences Education*, *20*(1), 149–161. https://doi.org/10.1007/s10459-014-9514-8

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*(1), 4–58. https://doi.org/10.1177/1529100612453266

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology*, *38*(4), 281–288. https://doi.org/10.1016/j.cedpsych.2013.06.001

Fiorella, L., & Mayer, R. E. (2014). Role of expectations and explanations in learning by teaching. *Contemporary Educational Psychology*, *39*(2), 75–85. https://doi.org/10.1016/j.cedpsych.2014.01.001

Galbraith, J., & Winterbottom, M. (2011). Peer-tutoring: What's in it for the tutor? *Educational Studies*, *37*(3), 321–332. https://doi.org/10.1080/03055698.2010.506330

Gartner, A., Kohler, M. C., & Riessman, F. (1971). *Children teach children: Learning by teaching*. New York: Harper & Row.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, *19*(3), 290–304. https://doi.org/10.1080/09658211.2011.560121

Hoogerheide, V., Deijkers, L., Loyens, S. M., Heijltjes, A., & van Gog, T. (2016). Gaining from explaining: Learning improves from explaining to fictitious others on video, not from writing to them. *Contemporary Educational Psychology*, *44-45*, 95–106. https://doi.org/10.1016/j.cedpsych.2016.02.005

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*(3), 621–629. https://doi.org/10.1037/a0015183

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*(3), 157–163. https://doi.org/10.1177/0963721412443552

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-based learning: A perspective for enhancing meaningful learning. *Educational Psychology Review*, *24*(3), 401–418. https://doi.org/10.1007/s10648-012-9202-2

King, A., Staffieri, A., & Adelgais, A. (1998). Mutual peer tutoring: Effects of structuring tutorial interaction to scaffold peer learning. *Journal of Educational Psychology*, *90*, 134–152.

Lim, S. W. H., Ng, G. J. P., & Wong, G. Q. H. (2015). Learning psychological research and statistical concepts using retrieval-based practice. *Frontiers in Psychology: Educational Psychology*, *6*, 1484. https://doi.org/10.3389/fpsyg.2015.01484

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). New York, NY: Cambridge University Press.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York, NY: Cambridge University Press.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, *20*(4), 516–522. https://doi.org/10.1111/j.1467-9280.2009.02325.x

Nestojko, J. F., Bui, D. C., Kornell, N., & Bjork, E. L. (2014). Expecting to teach enhances learning and organization of knowledge in free recall of text passages. *Memory & Cognition*, *42*, 1038–1048. https://doi.org/10.3758/s13421-014-0416-z

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*(3), 419–435. https://doi.org/10.1007/s10648-012-9203-1

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

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, *77*(4), 534–574. https://doi.org/10.3102/0034654307309920

Roscoe, R. D., & Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science*, *36*(4), 321–350. https://doi.org/10.1007/s11251-007-9034-5

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*(2), 175–184. https://doi.org/10.1016/s0022-5371(67)80092-6

Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist*, *24*, 345–376.

Wong, S. S. H., Ng, G. J. P., Tempel, T., & Lim, S. W. H. (2017). Retrieval practice enhances analogical problem solving. *The Journal of Experimental Education*, *42*, 1038–1048. https://doi.org/10.1080/00220973.2017.1409185

Yong, P. Z., & Lim, S. W. H. (2016). Observing the testing effect using Coursera video-recorded lectures: A preliminary study. *Frontiers in Psychology*, *6*, 2064. https://doi.org/10.3389/fpsyg.2015.02064

How to cite this article: Koh AWL, Lee SC, Lim SWH. The learning benefits of teaching: A retrieval practice hypothesis. *Appl Cognit Psychol*. 2018;32:401–410. https://doi.org/10.1002/acp.3410

APPENDIX A.

MARKING RUBRIC FOR FINAL COMPREHENSION TEST

1. **Explain how the Doppler effect works.**
   - A. Doppler effect: change in how sounds are perceived due to movement.
   - B. Two characteristics of sound waves: frequency and wavelength.
   - C. Definition of wave frequency: number of waves in a given time period.
   - D. Definition of wavelength: distance between adjacent waves.
   - E. Relationship between frequency and wavelength (e.g., higher frequency and shorter wavelength).
   - F. Relationship between frequency/wavelength and pitch (e.g., shorter wavelength and higher perceived pitch).
   - G. When source is stationary, waves at same frequency/wavelength in all directions.
   - H. When source approaches *(3 possible points)*
     1. Shorter wavelengths
     2. Higher frequency
     3. Higher perceived pitch
   - I. When sources passes by *(3 possible points)*
     1. Longer wavelengths
     2. Lower frequency
     3. Lower perceived pitch
2. **How could you increase the intensity of the Doppler effect?**
   - A. Increase speed of source towards observer (or increase in frequency/pitch).
   - B. Observer move towards the source.
3. **Would the Doppler effect occur if the source was stationary and the observer was moving? Why or why not?**
   - A. Yes (with no explanations, no points awarded).
   - B. Because there is still movement, which changes the way sound is perceived.
   - C. Because the perceived frequency/wavelength/pitch of the sound waves still changes.

4. **What would happen to the Doppler effect if the observer was moving the same speed and in the same direction as the source? Explain your answer.**
   A. No Doppler effect; same as if they were stationary *(with no explanations, no points awarded).*
   B. No change in perceived frequency/wavelength of the sound waves.
5. **How would an observer experience sound if the speed of the source were travelling faster than the speed of sound?**
   A. Sound would be delayed; experience sound after source has passed.
   B. Sonic boom; loud "cracking" sound; like an explosion; very high pitched [0.5].
   C. Sound waves heavily compressed.
6. **What would happen to the Doppler effect if the source and observer were both moving towards each other at on a parallel path at a constant speed? Explain your answer.**
   A. Doppler effect would be more intense than if one was stationary *(with no explanations, no points awarded).*
   B. Sound waves would change in frequency/wavelength at a higher rate.

APPENDIX B.

DOPPLER EFFECT LESSON

Background

Almost everyone has experienced the Doppler effect, though perhaps without knowing what caused it. For example, imagine you are standing on a street corner as a fire truck approaches with its siren blaring. The perceived pitch of the siren will sound higher as it comes closer to you. Then, as it passes by, the pitch will sound lower. This is one of many examples of the Doppler effect: The change in how sounds are perceived due to movement.

Sound Waves

Why does this change occur? Movement changes the way different characteristics of sound waves are perceived, and therefore, how the sound is perceived. Sound waves have two primary characteristics: frequency and wavelength. As we will see, movement causes changes in how we perceive the frequency and length of sound waves, which ultimately impact how we perceive the sound. First, let us briefly go over each of these characteristics.

Wave Frequency

Wave frequency refers to the number of waves passing through a given point during 1 s. It corresponds to how we perceive the pitch of a sound: If waves occur at a high frequency, they will produce a high pitch; if waves occur at a low frequency, they will produce a low pitch. For example, the cry of a baby has a relatively high pitch, whereas the sound of thunder has a relatively low pitch. The reason these two sounds are perceived differently is because they have different wave frequencies. Figures 1 and 2 below illustrate the difference between low and high frequency sound waves.

Wavelength

Closely related to wave frequency is wavelength. Wavelength refers to the distance between adjacent waves (see figures above). As you might expect, longer waves require more time to travel a given distance than do shorter waves. Consequently, longer sound waves have a lower frequency and a lower pitch. On the other hand, short sound waves have a higher frequency and higher pitch.

How the Doppler Effect Works

The Doppler effect is about movement influencing how the frequency and length of sound waves are perceived. To illustrate this, imagine a bug jiggling on the surface of a pond. If the bug is stationary, the waves on the surface of the water around it will be at the same frequency and length in all directions, as in Figure 3 below. Now suppose that the bug begins moving to the right. The waves it produces become shorter and more frequent to the right of the bug and longer and less frequent to the left of the bug, as shown in Figure 4 below.

Now, let us relate the bug example to how the Doppler effect occurs in sound waves. Imagine again that a fire truck is approaching with its siren blaring, as illustrated in Figure 5 below. As the fire truck approaches, the sound waves between the fire truck and the girl become shorter and more frequent, resulting in a higher perceived pitch. As the fire truck drives by, the sound waves between the fire truck and the girl are longer and less frequent. As a result, the girl perceives the pitch as getting lower. This is because the movement of the fire truck causes changes in how the sound is perceived. This influence of movement on perceived sound is the core principle of the Doppler effect.

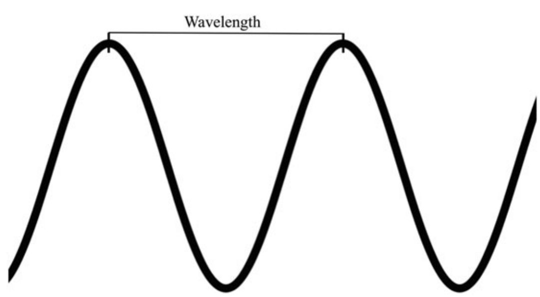

FIGURE 1 Low-frequency sound waves

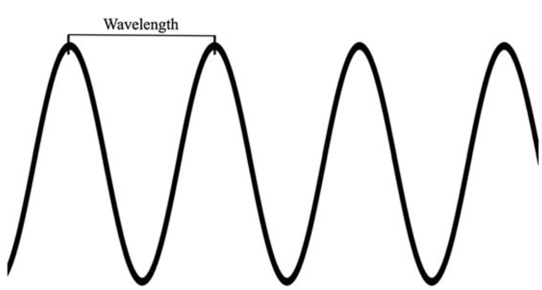

FIGURE 2 High-frequency sound waves

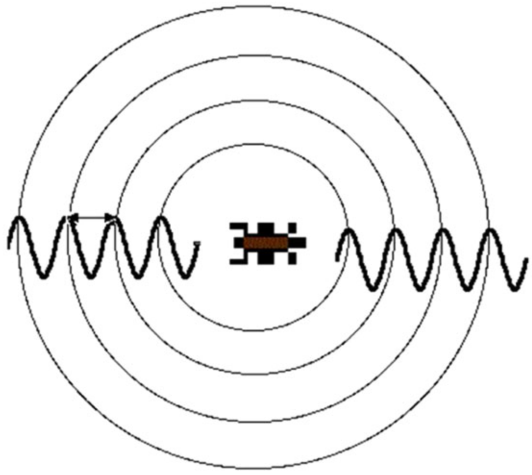

FIGURE 3 Stationary bug [Colour figure can be viewed at wileyonlinelibrary.com]

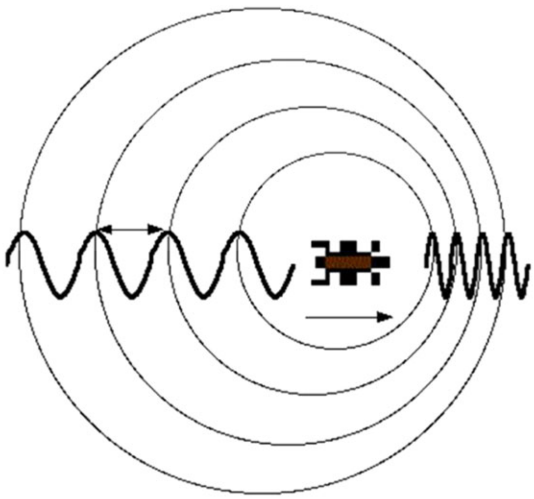

FIGURE 4 Bug moving to the right [Colour figure can be viewed at wileyonlinelibrary.com]

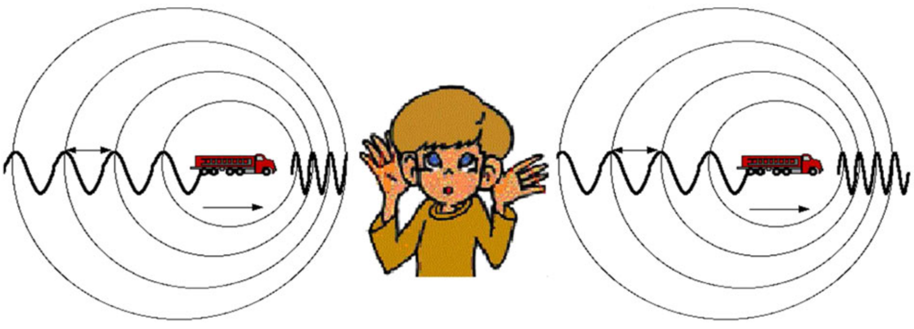

FIGURE 5 The Doppler effect of sound waves [Colour figure can be viewed at wileyonlinelibrary.com]