Summer 7-25-2019

Weighing up Exercises on Phrasal Verbs: Retrieval Versus Trial-And-Error Practices

BRIAN STRONG
strongbp@me.com

Frank Boers
fboers@uwo.ca

Citation of this paper:
Strong, B., & Boers, F. (2019). Weighing up exercises on phrasal verbs: Retrieval versus trial-and-error practices. *The Modern Language Journal, 103*(3), 562–579. https://doi.org/10.1111/modl.12579

Weighing up Exercises on Phrasal Verbs: Retrieval Versus Trial-And-Error Practices

ABSTRACT

EFL textbooks and internet resources exhibit various formats and implementations of exercises on phrasal verbs. The experimental study reported here examines whether some of these might be more effective than others. EFL learners at a university in Japan were randomly assigned to four treatment groups. Two groups were presented first with phrasal verbs and their meaning before they were prompted to retrieve the particles from memory. The difference between these two retrieval groups was that one group studied and then retrieved items one at a time, while the other group studied and retrieved them in sets. The two other groups received the exercises as trial-and-error events, where participants were prompted to guess the particles and were subsequently provided with the correct response. One group was given immediate feedback on each item, while the other group tackled sets of 14 items before receiving feedback. The effectiveness of these exercise implementations was compared through an immediate and a 1-week delayed post-test. The best test scores were obtained when the exercises had served the purpose of retrieval, although this advantage shrank in the delayed test (where scores were poor regardless of treatment condition). On average 70% of the post-test errors produced by the learners who had tackled the exercises by trial-and-error were duplicates of incorrect responses they had supplied at the exercise stage, which indicates that corrective feedback was often ineffective.

Keywords: phrasal verbs; retrieval; trial and error; feedback; errorless learning; interference

Phrasal verbs (e.g., *make up a story, turn down an offer, find out the truth*) are ubiquitous in English (Bolinger, 1971; Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011) and they are known to pose challenges that make many language learners shy away from using them, especially if single-word synonyms (e.g., *invent* for *make up*) are available (Dagut & Laufer, 1985; Liao & Fukuya, 2002; Siyanova & Schmitt, 2007). One of the reasons why mastering phrasal verbs can be so challenging is their semantic opacity. Many phrasal verbs are non-compositional in the sense that their meaning does not follow straightforwardly from adding up the meanings of the constituent words. For example, it cannot be obvious for a learner why the combination of the verb *carry* and the particle *on* should mean ‘continue’ or why the combination of *break* and *up* should mean ‘end a romantic relationship.’ In this regard, many phrasal verbs are akin to figurative idioms (e.g., Bolinger, 1971; Gairns & Redman, 2011; Kövecses & Szabó, 1996).

Another problem with phrasal verbs is that a single verb-particle combination can have multiple meanings (on average 5.6, according to Gardner & Davies, 2007), and so learners face the task of distinguishing between numerous form-meaning correspondences. In theory, a language course could give priority to phrasal verbs which, according to corpus research, are the ones most commonly used (Gardner & Davies, 2007; Liu, 2011; Liu & Myers, 2018), but even so this is likely to include a considerable number of form-meaning pairings. For example, Garnier and Schmitt (2015) developed a list of 150 verb-particle combinations and their most common meanings deemed worthy of prioritization in learning and teaching. While this selection can make the learning challenge appear less daunting, it nonetheless comprises close to 300 different meanings altogether. One might argue that at least learning the form (i.e., the composition) of phrasal verbs should be relatively easy, because many are made up of high-frequency verbs (e.g., *make, turn, give, come*) and a confined set of particles (e.g., *up, in, out, on*), which are also highly frequent words and thus likely to be familiar to the post-beginner learner. On the downside, these constituent words tend not to be used in their prototypical, concrete sense when they are part of a phrasal verb and are often semantically vague. This lack of semantic distinctiveness may render the constituents of phrasal verbs highly confusable. Especially the particles are likely to be prone to confusion, not only because their semantic contribution to the phrasal verb can seem quite arbitrary, but also because they are very short words that may lack formal distinctiveness (e.g., *in* and *on*).

Learners of English whose L1 does not have structural equivalents of phrasal verbs appear particularly slow at mastering them (e.g., Garnier & Schmitt, 2016) and appear particularly likely to avoid using them (Dagut & Laufer, 1985; Liao & Fukuya, 2002; but also see Cervantes & Gablasova, 2017). For learners whose L1 does have structural equivalents (e.g., speakers of other Germanic languages, such as Dutch), the task appears less daunting in relative terms (Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993), thanks to familiarity with the phenomenon at large and/or thanks to the availability of cognates, although these may be deceptive as well (e.g., the Dutch counterpart of *find out* means ‘invent’ instead of ‘discover’).

Given the broad recognition that learning phrasal verbs is challenging, it is not surprising that attention to this elusive part of the English lexicon is given in many EFL course books as well as internet resources for EFL learners and teachers. This is typically done by including sets of phrasal verbs with paraphrases of their meaning and by incorporating exercises with a focus on phrasal verbs. As we shall see, such exercises come in various formats and types of implementation. To date, there has been little research into their effectiveness, however. One exception is a study by Strong and Boers (2018), who first analyzed a collection of exercises in a corpus of 44 EFL textbooks and subsequently compared the learning gains obtained under two common implementations of such exercises.

The experiment we report in the present article is a conceptual replication intended to address questions that were left unanswered in that study.

BACKGROUND

Exercises on Phrasal Verbs in Textbooks and Internet Resources

Exercises on phrasal verbs and their implementation show variation along at least four dimensions: (a) the number of phrasal verbs tackled in the exercise, (b) the recurrence of the same phrasal verbs across exercises, (c) the nature of the operations to be performed by the learner, (d) the timing of feedback, and (e) the position of the exercise in a sequence of activities. In their analysis of 44 EFL textbooks, Strong and Boers (2018) found that, on average, a textbook exercise on phrasal verbs consists of 6.77 items, but there is considerable variation (*SD* = 2.42), with a range from two to 20 items in a single exercise. Given the proliferation of internet resources for EFL learning and teaching, it may be useful to explore if exercises on phrasal verbs available online resemble those in printed textbooks. We therefore simulated an EFL learner's search for online practice materials on phrasal verbs, by typing 'phrasal verbs' into an internet search engine (www.bing.com) and inspecting the first 10 EFL/ESL websites that came up and that offered freely accessible exercise material. We skipped websites such as online dictionaries which only provided lists of phrasal verbs and websites discussing phrasal verbs from a descriptive linguistics perspective. The hyperlinks to the 10 websites we examined are provided in Appendix A. Most of these websites also address teachers and offer them free downloadable exercise sheets for use in the classroom. In fact, most of them serve as platforms where practitioners share and disseminate lesson materials. In total, this simulation of a learner's browsing the internet for phrasal verb practice generated (in November 2018) a corpus of 204 freely accessible exercises on phrasal verbs. The average number per website was 19.78 (*SD* = 17.7; median = 15; min = 2; max = 26). Interestingly, the average number of phrasal verbs tackled per exercise in this corpus is considerably higher than what Strong and Boers (2018) found in their corpus of printed textbooks: 10.49 (*SD* = 4.83; median = 10; min = 4; max = 36), with 10 items per exercise being the most common (i.e., mode = 10). We can only speculate about the reasons for this, but the fact that, unlike printed publications, online resources are not constrained by page budgets may be one of them. In any case, what these data suggest is that learners who seek phrasal verbs practice online and teachers who look for downloadable phrasal verbs exercises are likely to be dealing with a substantial number of items in a single activity. It stands to reason that the number of items per exercise will influence the cognitive burden of the learning process, especially if many of the items are new to the learner.

As to the recurrence of the same phrasal verbs across different exercises, Strong and Boers (2018) noted that this is extremely rare in the textbooks they examined. Typically, a given phrasal verb is targeted in just one exercise per textbook. They acknowledge that only 'student books' were analyzed, while some of these are supplemented by workbooks, CDs and/or online resources which may include revision exercises. Still, as the student book is often the principal, if not the sole, resource used in language classrooms, it seems vital that the very little practice that *is* offered by that resource is optimally effective. The lack of systematic revision is also characteristic of the 10 websites we explored. Apart from one interactive quiz (gamestolearnenglish.com) and a couple of worksheets where the same set of phrasal verbs is the object of a sequence of different-format exercises, none of the websites examined appear to provide exercises with a view to re-engaging learners with previously tackled items. The collections of exercises on several websites are compilations of worksheets contributed by different individuals, and so any recurrence of the same phrasal verbs across different exercises seems accidental rather than planned, even where appearances do suggest an intentional sequencing. For example, when exercises labelled 'mixed' follow exercises focusing on specific sets of phrasal verbs (e.g., ones sharing the same verb; ego4u.com), this may give the impression these serve the purpose of revision work. However, on closer inspection, it turns out that the phrasal verbs targeted in the exercises labelled ‘mixed’ are different from those in the preceding exercises.

Exercises on phrasal verbs come in various formats, each requiring a specific operation to be performed by the learner. Some formats require the learner to match phrasal verbs with their meanings. This is done, for example, by asking learners to connect them to their correct paraphrase from a list of options or by asking learners to fill in the right phrasal verb from a set of options in a gapped sentence exemplifying its meaning. These are essentially meaning-recognition tests in that they furnish the form of the phrasal verbs as well as candidate paraphrases or sentential contexts. More challenging (but much less common) meaning-focused formats require learners to supply a paraphrase of a phrasal verb or to create a sentence with the phrasal verb that demonstrates its meaning. Such exercises qualify as meaning-recall tests. Other formats are more form-focused, in the sense that they include a focus on the composition of the phrasal verbs. Such formats involve determining which verbs combine with which particles to express a given meaning. They resemble form-recognition tests when the verbs and/or the particles are provided in a bank of words for the learner to choose from. The most challenging formats provide no sets of intact phrasal verbs, verbs or particles to select from, but present only paraphrases or gapped sentences. The latter can be considered form-recall tests. According to Strong and Boers’s (2018) counts, roughly 66% of the exercises in the textbooks they analyzed focus on the correspondence of intact phrasal verbs with their meaning (mostly using meaning-recognition formats). The remaining 44% require learners to produce phrasal verbs by making appropriate verb-particle combinations, and thus focus on their composition. Interestingly, the proportions are reversed in our corpus of exercises collected from websites. In this corpus, 47% of the exercises are about intact phrasal verbs and their meaning, while 53% of the exercises require the learner to combine verbs and particles.[1] The latter is almost invariably done by means of a gap-fill format, where the learner is asked to select or supply the missing verb (20%), the missing particle (61%), or pair up verbs and particles from lists of options (20%). Formats where learners are to decide what particle needs to be added to a given verb to express a certain meaning make up over 36% of the 204 exercises collected from the 10 websites. The second most common format (close to 29%) requires learners to complete gaps with intact phrasal verbs.

The fourth and fifth dimensions along which exercises on phrasal verbs vary regard not the exercise format as such, but how it is implemented. One difference in implementation is whether feedback follows immediately after each exercise response or whether it follows when a whole exercise is completed. Strong and Boers (2018) found that when feedback is given in textbooks (which is not always the case), this is usually in the form of an answer key in an appendix. Unless students check the answer key after each exercise response or receive feedback from their teacher after each response, they are likely to verify if their responses were correct only on completion of the whole exercise. The freely downloadable worksheets offered by the EFL websites we analyzed are similar: An answer key is available on a separate sheet, and so it seems to be expected that learners will compare their responses to this at the end of the activity. In one website (esl-lounge.com), learners need to click separate hyperlinks to access the answer keys. Some websites offer interactive exercises, however, and this creates opportunities for immediate feedback on each exercise response. This is done in one (perfect-english-grammar.com) by giving learners the choice to click on an answer key button on the side of every exercise item. In another (englisch-hilfen.de), the button to be clicked for the answer key to appear is placed below the complete exercise. Whether feedback is more effective when it is received immediately or with a certain delay is a controversial issue, which we will return to subsequently because it has a direct bearing on the experiment to be reported.

The primary factor of interest in this experiment, however, concerns yet another choice in the way phrasal verb exercises are implemented. One might expect an exercise to follow an activity where learners have engaged with the targeted items in a study episode or where they have at least encountered them (for instance, in a reading or listening text). In other words, one might expect the exercise to provide an opportunity for retrieval from memory of previously studied or encountered information. Surprisingly, Strong and Boers (2018) found that 62% of the exercises in the textbooks they analyzed are *not* preceded by examples or study material that could help students do the exercise. When a textbook unit does discuss a few examples as a way of raising students' awareness of phrasal verbs, the exercises that follow later do not necessarily target the items discussed and instead require the learners to make guesses about newly introduced phrasal verbs (e.g., Kay & Jones, 2009, pp. 42–45). Students are therefore often invited to guess the meaning of phrasal verbs which the textbook has not previously introduced (e.g., Clare & Wilson, 2002, p.40) or to provide the missing verbs or particles without prior examples of the phrasal verbs (e.g., Clare & Wilson, 2002, p. 40; Oxenden & Latham-Koenig, 2010, p. 60; Roberts, Clare, & Wilson, 2011, p. 141).[2]

If learners turn to EFL websites for help with phrasal verb learning, they may be even more likely to tackle exercises in a trial-and-error fashion. Of the 10 websites we examined, six only provide exercises, without study materials for learners to consult. One website (esl-lounge.com) does feature a list of recommended readings, including phrasal dictionaries, but the provided hyperlinks give access to the publishers' websites instead of freely accessible study material. Another website (englisch-hilfen.de) includes a link to a long list of phrasal verbs, but this is a list without clarification on the meaning of the phrasal verbs. One of the websites (ego4u.com) does provide links to lists of phrasal verbs and their meaning, but each of the lists focuses on a certain verb while most of the exercises on this website are organized by sets of particles instead. Among the 48 worksheets on the website agendaweb.com, we found only three documents where the exercise was preceded by explanations about the target items. Altogether, only one of the 10 websites we examined (perfect-english-grammar.com) demonstrably incorporates exercises as retrieval practice. On this website, learners are presented with sections of Garnier and Schmitt’s (2015) list of common phrasal verbs and their meaning before they are invited to embark on completion exercises.

What both Strong and Boers’s (2018) textbook analysis and this exploration of EFL websites suggest is that, unless learners have already acquired knowledge of the target phrasal verbs in some other way, they may well engage with phrasal verb exercises as trial-and-error events. In that case, the assumption seems to be that the correct knowledge will be established or consolidated at the feedback stage, after the trial-and-error exercise has piqued the learners’ curiosity about the correct responses. An experiment by Strong and Boers (2018) has cast doubt over the efficacy of this trial-and-error procedure when it is applied to phrasal verbs. In that experiment, Japanese EFL learners were given exercises on two sets of seven phrasal verbs implemented either as retrieval or as trial-and-error. The exercises presented the learners with a paraphrase of the target item and a short dialogue with a gap where they were required to supply the missing particle. In the retrieval condition, the exercise was preceded by study material on the seven target phrasal verbs. This comprised a paraphrase of their meaning and an example of their use in a short dialogue. In the trial-and-error condition, the same study material instead followed the exercise and thus served as feedback. In both an immediate and a 1-week delayed post-test, which used the same format as the exercises, the trial-and-error treatment generated significantly poorer results than the retrieval condition (41% vs. 56% correct responses in the delayed post-test). Item analyses revealed that 25% of the incorrect trial-and-error responses were duplicated in the delayed post-test, which suggests that feedback given on trial-and-error responses is often insufficient to supplant an initial response by the correct one in the learners' memory. The latter finding is reminiscent of some studies on the merits of course book exercises on verb-noun collocations (e.g., *make an effort*, *tell lies*), where corrective feedback was also often ineffective at preventing learners from making the same errors in the delayed post-test (Boers, Demecheleer, Coxhead, & Webb, 2014; Boers, Dang, & Strong, 2017). And yet, findings such as these appear at odds with ones obtained in certain experiments on word learning, which we discuss subsequently.

Word Learning Through Retrieval and Through Trial-and-Error

It is now well established that a retrieval effort after having studied information is more beneficial for the creation of durable memories than simply studying and then re-studying the same information (Karpicke, Lehman, & Aue 2014; Roediger & Butler, 2011; Smith, Roediger, & Karpicke, 2013). The benefits of this 'testing effect' (Roediger & Karpicke, 2006) have also been attested for L2 word learning. In Barcroft (2007), for example, L2 learners were asked to study word-picture pairs displayed on a screen. In one condition, the participants were subsequently presented the picture without its corresponding word and asked to recall the missing word. The word-picture pairs were then displayed again so the participants could verify the accuracy of their recall. In another condition the participants were simply asked to study the same word-picture pairs twice. Post-tests performance was found to be superior in the condition which included a retrieval component.

However, it is not only retrieval practice that has been shown to lead to better retention than merely studying and then re-studying material. There is mounting evidence in the realm of cognitive psychology to suggest that testing participants *before* presenting them with to-be-learned items influences how those items are learned (e.g., Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012; Kornell, Hays, & Bjork, 2009; Slamecka & Graf, 1978). The effect of pre-testing challenges the traditional view of learning held by many researchers who use pre-tests to assess participants' knowledge of to-be-learned items with the expectation that they do not affect learning (Kornell & Bjork, 2007). The benefits of pre-testing extend beyond the practice of informed guessing, where a test question provides sufficient contextual cues to help students infer the correct answer on their own. Test questions that are unlikely to be answered correctly on the basis of prior knowledge have also been shown to enhance retention of the correct responses subsequently presented to the learner (Richland, Kornell, & Kao, 2009). Implementing language exercises as trial-and-error events followed by feedback resembles this use of pre-testing in cognitive psychology.

The benefits of pre-testing for word learning were demonstrated in a series of experiments by Potts and Shanks (2014). Participants were asked to learn obsolete English words or words from a language they had no prior knowledge of (Euskara). In two conditions the participants were prompted to guess the meaning of the words before the meanings were given. These two pre-testing conditions differed in the likelihood of correct guesses: One used a multiple-choice format, while the other asked the participants to guess the meaning of the words freely. The participants in both trial-and-error conditions, but especially those who had been asked to freely guess the word meanings, performed better in a post-test than participants in a comparison condition who had been given the correct word meanings from the start and had been asked to study them. In a recent publication, however, Seabrooke, Hollins, Kent, Wills, and Mitchell (2019) have pointed out that the post-tests used in Potts and Shanks's (2014) did not necessarily measure whether the participants recalled the *association* of the new words and their meanings. These post-tests used a multiple-choice format where the foils (i.e., the incorrect options) were mostly new items, not included in the learning phase. It is therefore possible that the error-prone guessing benefited subsequent recognition of the target words without also benefiting recall of word-meaning pairings. A series of experiments by Seabrooke et al. (2019)—all approximate replications of Potts and

Shanks (2014)—indeed suggest that the benefits of error-prone meaning guessing before learning the correct word-meaning correspondences are less likely to emerge when multiple-choice post-test items include foils which the participants also encountered during the learning procedure (and which consequently stand an equally good chance of being recognized). Moreover, in a cued recall test (where the meaning of the target words was presented, and the participants were asked to produce the target words), it was the study-only (i.e., no meaning guessing) condition that fared the best. Seabrooke et al.'s (2019) findings thus cast doubt over the efficacy of an error-prone trial-and-error procedure for learning associations. This is highly relevant to the subject of the present article, because learning a phrasal verb requires establishing a dual association: (a) between the phrasal verb and its meaning and (b) between the verb and its particle.

Despite some inconsistencies, it appears from this literature review that not only retrieval procedures but also trial-and-error procedures may be preferable to the mere presentation of words and their meanings. However, very few studies to date have directly compared the effectiveness of retrieval and trial-and-error procedures. To some degree, two experiments by Warmington, Hitch and Gathercole (2013) and Warmington and Hitch (2014) could be considered ones that do include such a comparison. In these experiments, participants learned the (spoken) words for new objects, but in one condition they were asked to immediately repeat each word, while in another condition they were first prompted to guess the word (using the first letter as cue) before it was provided to them. Post-tests revealed better word recall in the first condition, indicating that an errorless encoding procedure is preferable to an error-prone trial-and-error procedure for word learning. That the errorless procedure in these experiments was found to be comparatively effective is somewhat surprising, because it required little effort on the part of the learners. It has been suggested, after all, that it is the cognitive effort invested in a retrieval act that helps to entrench knowledge (Bjork, 1994; Kornell, 2009; Pyc & Rawson, 2007). It is in that regard also that the experiments by Warmington and colleagues do not constitute an optimal comparison of retrieval and trial-and-error procedures. Moreover, the implementation of trial-and-error in these experiments differs from what is typically done in EFL textbooks (as will be elaborated subsequently), and so their results may only tentatively inform an evaluation of the latter.

A more direct comparison of retrieval and trial-and-error procedures is the experimental study by Strong and Boers (2018) already mentioned previously, and which specifically concerned exercises on phrasal verbs. In the retrieval condition sets of seven phrasal verbs were first studied and then recalled in the exercise, whereas in the trial-and-error condition the same sets of seven phrasal verbs were first tackled in the exercise before they were studied. The post-test results showed a clear advantage of the former sequence (i.e., using the exercise for retrieval practice). It is worth noting that the relatively small number of items to be studied in one go helped the participants in the retrieval condition to supply a high proportion of correct exercise answers (85%). In the trial-and-error condition, by contrast, the exercise responses were highly error prone (with just 5% correct answers). While there are some parallels with the experiments by Warmington and colleagues, where one condition was error-free and the other error-prone, there are also notable differences. Because the items were tackled in sets of seven in Strong and Boers (2018), retrieval is likely to have been more effortful, and feedback on the trial-and-error exercises was received with a certain delay.

Whether feedback is given immediately or with a delay may matter according to the memory research in cognitive psychology (e.g., Butler, Karpicke, & Roediger, 2007; Metcalfe, Kornell, & Finn, 2009), although it is not yet clear whether immediate or delayed feedback is preferable for intentional L2 word learning (Nakata, 2015). On the one hand, it has been argued (e.g., Mory, 2004, for review) that giving immediate feedback on errors risks creating an association in memory between the initial erroneous response and its correct alternative, so that the former may interfere with retrieval of the correct alternative in future. Given a sufficient interval between an erroneous response and the presentation of the correct alternative, the initial response may be forgotten and will thus not compete in memory with the correct alternative—so the argument goes. On the other hand, in the case of exercise responses that a learner finds plausible, forgetting these in the interval before receiving delayed feedback is not so likely. If so, it is perhaps advisable to immediately point out the errors to the learner lest they linger in memory unchecked and might become harder to eradicate as time passes.

Whether immediate or delayed feedback is the more appropriate choice is likely to depend on the nature of the items to be learned as well. For example, even though the meaning of a given phrasal verb does not usually follow straightforwardly from the literal meanings of its constituents, the fact that those constituents as such will tend to be familiar may enable the learner to try and make a reasoned guess at the item’s meaning. If so, it is conceivable that learners will feel more committed to the plausibility of their interpretation than if it were a totally wild guess. This is a different situation from being presented (in a laboratory experiment) with an unknown L2 word form (without any contextual cues) and being asked to freely guess its meaning (e.g., Potts & Shanks, 2014), because in the latter case the respondent knows quite well that the guess will almost certainly be wrong and thus not worth holding in memory in the first place. The fact that the number of available particles (*in, up, out*, etc.) is quite small and that some occur in a wide range of phrasal verbs may enable learners to choose a particle with a degree of confidence that their response is at least plausible—even if only statistically so. Sometimes, a certain particle is also semantically more compatible with a given meaning, owing to general ‘conceptual’ metaphors (e.g., Kövecses, 2010). For example, a learner may reason that the particle *up* fits the meaning expressed by *brighten up* better than, say, *down,* because *up* is often associated with positive emotions and *down* with negative ones. In other words, although the precise combination of verbs and particles to express a certain meaning is unpredictable, it is not always fully arbitrary (e.g., Lindstromberg, 2010). Delaying feedback on the assumption that learners will swiftly forget a wrong hunch where the hunch seemed plausible on semantic grounds might not be advisable either. In any case, Strong and Boers's (2018) observation that over 25% of the delayed post-test errors of their trial-and-error group were duplicates of the particle errors they had made in the exercises indicates that the delayed feedback procedure did not work wonders in their experiment. It is therefore worth investigating whether *immediate* feedback on such errors is more effective.

Summing up, while the retrieval condition in Strong and Boers (2018) was found to lead to the better post-test results, it is not clear whether the benefits of retrieval could have been different had the procedure been implemented in an errorless fashion by asking the participants to retrieve each item one at a time instead of per set. Neither is it clear whether the results for the trial-and-error condition could have been different had immediate feedback been given on each individual exercise response rather than after completion of a set of items. In what follows, we report a conceptual replication of Strong and Boers's (2018) experiment intended to shed light on these issues. Ultimately, finding answers to these questions may inform materials writers', teachers', and learners' decisions about the design and use of exercises on phrasal verbs and possibly, by extension, other multi-word items.

THE PRESENT STUDY: GENERAL DESIGN AND RESEARCH QUESTIONS

Like Strong and Boers (2018), the present study investigated the influence of learning method as a factor on the learning of phrasal verbs. In the retrieval method, participants were asked to memorize phrasal verbs before recalling the particles in an exercise (as will be elaborated subsequently), whereas participants in the trial-and-error method were asked to try the exercise and then memorize the phrasal verbs when these were explained as feedback on the exercise. We also considered the contribution of an additional factor: the number of items tackled in an exercise in correspondence with the number of items studied before or after the exercise. More specifically, we were interested in what influence would be exerted on the learning and retention of phrasal verbs when participants in the retrieval procedure studied one phrasal verb at a time followed by an exercise that prompted its immediate recall compared to studying several phrasal verbs in a row and then being prompted to recall them together in the exercise. Likewise, we were curious to know what would happen when participants in the trial-and-error procedure completed an exercise item on one phrasal verb and then immediately received feedback on their response compared to getting feedback after completing an exercise on several phrasal verbs.

Input materials and exercises were identical across the four treatment groups. The difference between participants in the retrieval condition was learning phrasal verbs one at a time or 14 in a row, and the difference between those in the trial-and-error condition was also learning the same phrasal verbs one at a time or 14 in a row. Post-tests were administered at the end of the learning session and 1 week later.

The general research question we seek an answer to is this:

Do the two factors for learning phrasal verbs lead to different success rates in an immediate and a 1-week delayed post-test?

More specific research questions are:

RQ1: Is the retrieval condition better at enhancing the learning of phrasal verbs than the trial-and-error condition?

RQ2: Is the comparatively easy one-at-a-time study plus retrieval procedure as effective as the more effortful 14-in-a-row study plus retrieval procedure? Are successfully retrieved items under the latter, more effortful procedure, better retained in the longer run?

RQ3: Is immediate feedback on each trial-and-error response as effective as delayed feedback (i.e., feedback on a set of 14 responses)? Is duplication of exercise errors in the post-tests more frequent in one than the other trial-and-error condition?

Participants

The participants were 145 second-year university students learning English as a foreign language at a university in Japan. They were from five parallel classes, but within each class they were randomly assigned to the four treatment groups. At the start of the semester, the students sat the TOEIC Bridge Test. Their average score was 157 out of 180 (*SD* = 11.5). They were familiar with the 2,000 most frequent words in English, as gauged by the Vocabulary Levels Test (VLT; Schmitt, Schmitt, & Clapham, 2001). Their average score on that test (at the 2,000-word level) was 27.5 out of 30 (*SD* = 1.2). The reason for administering the VLT was to determine whether participants were likely to be familiar with the words that make up the target phrasal verbs and the paraphrases of their meaning (as will be elaborated subsequently). All the participants had studied English only in Japan prior to the experiment. It is worth noting that Japanese is a language without phrasal verbs, and that these learners could therefore be expected to find learning this facet of English quite challenging.

Materials

Twenty-eight phrasal verbs were selected as target items and divided into two sets of 14 (see Appendix B). Since some of the target phrasal verbs have the same initial letter (e.g., *catch on, carry on, call off, chip in, crack on*), attempts were made to balance the sets with these items by assigning half to one set and half to the other. The number of target items assigned to each set in the present study is greater than that used in Strong and Boers (2018). The motive for this was to enhance the difference in experience between the learning procedures (i.e., between the one-at-a-time and the 14-in-a-row conditions). We acknowledge that targeting as many as 14 items in a single exercise is not very common, but it nonetheless falls within the range of up to 20 items encountered by Strong and Boers (2018) in their examination of textbooks and clearly within the range of up to 36 items in our exploration of EFL websites.

Prior to the experiment, 29 students were randomly selected from the student population to sit a cued production test (the same test that would be used to assess learning in the experiment) on the 28 phrasal verbs to gauge the extent to which the students who would be participating in the experiment might already be familiar with them. None of the students' responses turned out correct, and this was taken as an indication that the students in the actual experiment would also be very unlikely to have productive knowledge of the target items prior to the experiment. The use of this procedure was preferred over pre-testing the actual participants, because taking a pre-test would have amounted to a trial-and-error event and would thus have compromised the distinction between the retrieval and the trial-and-error conditions.

According to the BNC-COCA CORE-4-word list, 20 verbs of the selected phrasal verbs belong to the first 1,000-level frequency band, seven to the second frequency band, and just one verb (*nod*) belongs to the third frequency band. All the particles belong to the first 1,000 frequency band. In the study materials, the phrasal verbs were accompanied by paraphrases of their meaning (e.g., *hang out – spend time with friends*; see Appendix B). Of the words used in these paraphrases, 63 belong to the first 1,000 frequency band and just eight belong to the second 1,000 frequency band. Given the participants' scores on the VLT, it is reasonable to assume that they were familiar with the constituents of the target phrasal verbs as well as the vocabulary used in their paraphrases.

The same paraphrases served as cues in the exercises and the post-tests. In the exercises, the students were only required to supply the missing particle (e.g., *hang* ____ – *spend time with friends*). This imitates a format which was found to be relatively common in EFL textbooks and websites. Another motive for including the verb in the exercise prompts was that this would prevent learners from proffering synonymous phrasal verbs instead of the ones targeted in the experiment. For instance, when prompted with the paraphrase *to use all of something*, instead of answering with *run out* (the phrasal verb meant to be elicited), a participant in the trial-and-error condition might suggest *use up*, and such alternative responses would not have enabled us to determine if the participant also knew the actual target item.

In the post-tests, however, the students were required to produce the complete phrasal verbs (e.g., _____ _____ – *spend time with friends*). The rationale for this was that the ability to use a phrasal verb entails knowledge of the intact unit, not just one constituent, and so an evaluation of exercise procedures needs to examine whether that goal is reached. It is also a format that requires learners to connect the phrasal verb to its meaning—otherwise the paraphrase could hardly cue it— rather than merely remembering a verb-particle sequence, without understanding it.

Design

The experiment used a two (learning method: retrieval vs. trial-and-error) × two (number of phrasal verbs: one-at-a-time vs. 14-in-a-row) between-participants factorial design. Twenty-nine students were randomly assigned to one of the four groups: one-at-a-time retrieval group, 14-in-a-row retrieval group, one-at-a-time trial-and-error group, and 14-in-a-row trial-and-error group.

Procedure

The learning activities and the immediate post-tests concerning the two sets of 14 phrasal verbs took place in a 90-minute class. The delayed post-test followed 1 week later. All the activities and the tests were presented on computers. At the beginning of class, participants were told that they were going to learn phrasal verbs and then take a memory test. They were sent a link that, when clicked on, randomly assigned them to one of the four treatment groups.

For the one-at-a-time retrieval group, participants were asked to learn the association between a phrasal verb with a paraphrase of its meaning (e.g., *hang out – to spend time with friends*). The form-meaning correspondence was presented for 15 seconds. After time elapsed, they were asked to complete a gap-fill exercise in which the particle was replaced by an underlined space (e.g., *hang __ – to spend time with friends*). Participants had 15 seconds to type the missing particle. The 14-in-a-row retrieval group followed a similar procedure, except that these participants were asked to learn the correspondence between 14 phrasal verbs and paraphrases of their meaning within 210 seconds before the same 14 items were presented in the gap-fill exercise with the particles deleted. Participants had 210 seconds to type the 14 particles in the text boxes where they were missing.

Participants in the one-at-a-time trial-and-error group were asked to complete the same gap-fill exercise given to those in the one-at-a-time retrieval condition. That is, they were shown an incomplete phrasal verb with an underlined space for the missing particle along with a paraphrase of the meaning of the phrasal verb for 15 seconds (e.g., *hang __ – to spend time with friends*). If a participant correctly produced the missing particle, a green tick was displayed as feedback. If the response was wrong, a red cross was placed next to it and the correct particle was provided. Participants had 15 seconds to study the feedback. The 14-in-a-row trial-and-error group followed a similar procedure, except that these participants were shown 14 verbs with empty spaces for the missing particles along with paraphrases of the meanings of the phrasal verbs. They had a total of 210 seconds to type the missing particles in the text boxes and the same amount of time was given to study the feedback, which was identical in nature to the one-at-a-time procedure but given only when the whole exercise on the 14 items was completed.

After completing the procedure for 14 target items, the students were given a 10-minute distractor task in which they answered trivia questions in their first language and solved simple math problems. After the distractor task, participants completed the (near-) immediate post-test on the 14 phrasal verbs, with the items presented in a randomized order. Next, they moved on to the second set of 14 phrasal verbs and followed the same procedure as before. The participants were not notified that they were going to sit the same test in the following week. The delayed test was identical to the immediate test, except that all 28 phrasal verbs were tackled in a single series, with the item order randomized again.

The amount of time necessary to complete the activities had been established through piloting with same-profile students. The total amount of time invested was kept the same across the four learning conditions to ensure a fair comparison of the efficacy of the procedures. Qualtrics, a web-based tool for data collection, was used to display the study materials, exercises and tests, and to record the participants' responses.

Data Analysis

Exercise and post-test responses were scored dichotomously (correct/incorrect). No credit was awarded for a response with an incorrect particle (e.g., *in* vs. *on*) because it would have been impossible to distinguish a spelling mistake from a wrong choice of word. Minor spelling mistakes were accepted for verbs (e.g., *cary* instead of *carry*; *spent* instead of *spend*) as long as they did not hinder recognition of the target verb. The exercise responses were examined as well. This was to ascertain that the success rates at the exercise stage were different between the retrieval and the trial-and-error groups, as expected, and to tally the proportion of incorrect post-test responses that were duplicates of exercise errors (see RQ3).

Scores on the post-tests were analyzed separately using mixed-effects logistic regression models with the glmer function in the lme4 package (Bates, Maechler, Bolker, Walker, 2013) in the R environment (R Development Core Team, 2017).[3] The learning method predictor comprised two levels: retrieval and trial-and-error. The number of phrasal verbs predictor had two levels: one-at-a-time and 14-in-a-row. The fixed factors of participants' VLT scores and class membership were initially included but then removed (using a backward stepwise method) because they did not improve the models. The random effects in the models were the participants and the items. Regarding the particles, we subsequently also compared, per condition, the proportion of test errors that were duplicates of exercise errors with the proportion of other test errors. For this, we used one-sample *z*-tests.

RESULTS AND DISCUSSION

Performance on the Exercises

During the exercise phase, students were asked to provide the missing particle of each target phrasal verb. While students in the retrieval condition had studied the target items before the exercise, those in the trial-and-error condition had not. Therefore, it is not surprising that exercise scores were higher in the retrieval condition (88%) than the trial-and-error condition (18%). In the retrieval condition, performance was better for the one-at-a-time group (92%) than for the 14-in-a-row group (84%). This finding was expected considering that the one-at-a-time retrieval was near-immediate, while the 14-in-a-row group faced the challenge of avoiding cross-interference from 13 other items from the set they had just studied. In the trial-and-error condition, the exercise scores were seldom correct: 16% for the one-at-a-time group and 19% for the 14-in-a-row group.

Correct Responses on the Immediate Post-test

Table 1 displays the average proportion of target phrasal verbs correctly supplied in the post-tests for all treatment groups. These scores show the effects of the two predictors for each treatment group in the immediate and delayed post-test.

<INSERT TABLE 1 ABOUT HERE>

TABLE 1

Average Proportion of Correct Responses in the Exercise, Immediate and Delayed Post-Test

Treatment		Exercise		Immediate Post-test		Delayed Post-test	
		M	*SD*	*M*	*SD*	*M*	*SD*
Retrieval	One phrasal verb	92%	14%	76%	25%	18%	15%
	14 phrasal verbs	84%	11%	78%	22%	13%	11%
Trial-and-error	One phrasal verb	16%	17%	42%	20%	13%	14%
	14 phrasal verbs	19%	12%	51%	28%	10%	12%

Note. *n* = 29 per treatment group.

We found that in the immediate post-test, participants in the retrieval condition recalled close to 80% of all the target PVs, while those in the trial-and-error condition recalled less than half (47%). Comparing the two treatment groups within the retrieval condition, the difference in scores between the one-at-a-time group (76%) and the 14-in-a-row group (78%) was minimal. By contrast, within the trial-and-error condition, the 14-in-a-row group (51%) outperformed the one-at-a-time group (42%). Overall, the average difference between the retrieval and the trial-and-error conditions (33 percentage points) was greater than the average difference between the one-at-a-time and the 14-in-a-row conditions (5.5 percentage points).

When the data were entered into our model (see Table 2), the results indicated that there was a significant main effect of learning method (i.e., retrieval vs. trial and error), with a large effect size (z = 3.10, p < .001, d = 1.71 [1.10, 2.32]). No significant effect emerged for number of phrasal verbs (i.e., one at a time vs. 14 in a row) or for a learning method × number of phrasal verbs interaction (z = 0.55, p = .310, d = 0.30 [−0.28, 0.89] and z = −0.68, p =.386, d = −0.38 [−1.23, 0.48], respectively).

<INSERT TABLE 2 ABOUT HERE>

TABLE 2

Performance on the Immediate Post-Test: Fixed and Random Effects Summary

	Fixed effects					
Parameter	Estimate	*SE*	*z*	*p*	Odds ratio	95% CI
Intercept	−0.53	0.43	−1.25	.211	0.69	[0.32, 1.54]
Number of phrasal verbs	0.55	0.54	1.02	.310	1.26	[0.58, 2.73]
Learning method	3.10	0.56	5.51	< .001	15.80	[7.21, 34.74]
Number of phrasal verbs × Learning method	−0.68	0.79	−0.87	.392	0.51	[0.11, 2.36]

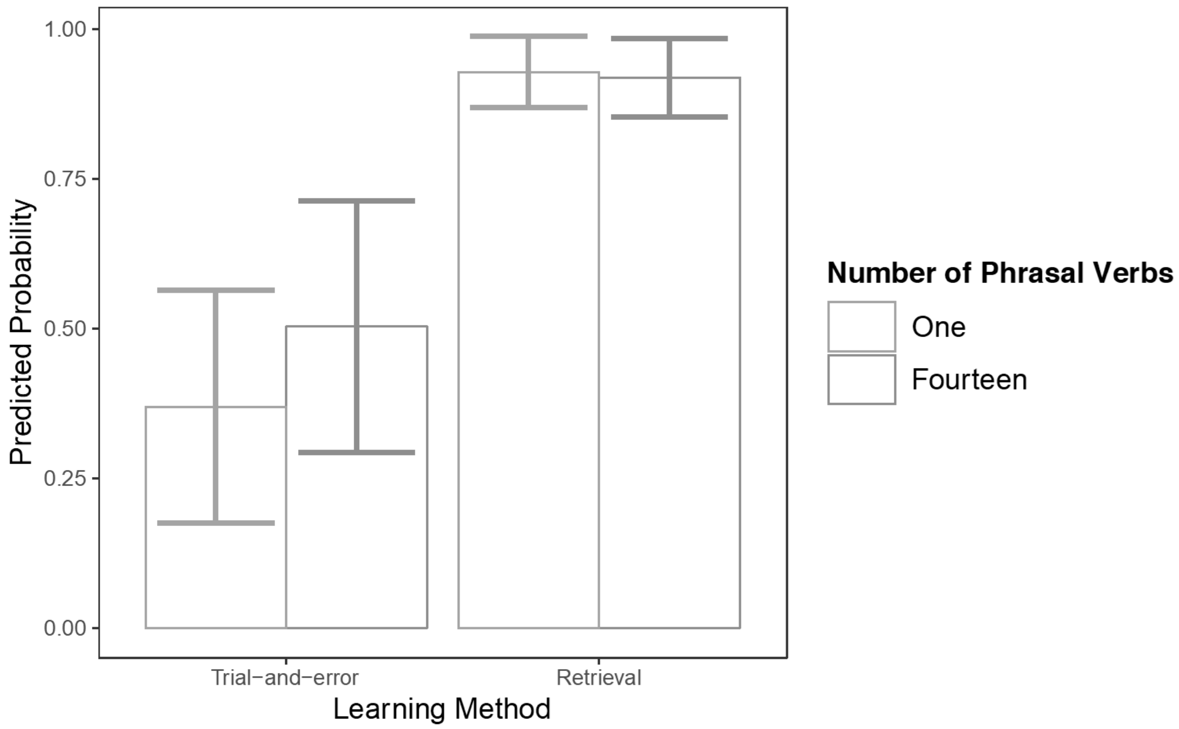
Note. Conditional r^2 value was .68. Random effect *SD*s were 1.99 for Subjects and 1.02 for Items

Figure 1 depicts the effect size calculated from the model in terms of the predicted probabilities (along with 95% pointwise confidence intervals) of obtaining a correct response for all conditions in the immediate post-test. The predicted probability of recalling a phrasal verb in the retrieval condition (92%) was approximately double that in the trial-and-error condition (45%). It is worth noting the 95% pointwise confidence intervals, which indicate where the true effect is estimated to lie. In the retrieval condition, the spread of the lower and upper values is relatively narrow, suggesting that not only is the effect size quite precise, but the treatment is very likely to facilitate learning, at least in the short term. By contrast, in the trial-and-error condition, the lower limit is far below chance (26%), which raises concern over the pedagogical value of this treatment for learning phrasal verbs.

<INSERT FIGURE 1 ABOUT HERE>

FIGURE 1

The Predicted Probability of Providing a Correct Response on the Immediate Post-Test as a Function of Learning Method and Number of Phrasal Verbs



Correct Responses on the Delayed Post-test

On the 1-week delayed post-test (see Table 1), performance was poor in general and especially poor in the 14-in-a-row trial-and-error group (with only 10% correct responses). On average, test scores dropped by as many as 43 percentage points between the immediate and delayed post-test. However, the average number of target items recalled was still higher in the retrieval condition (15%) than the trial-and-error condition (12%). In addition, participants in the one-at-a-time-condition (on average 15%) fared better than participants in the 14-in-a-row condition (on average 12%).

The mixed effects logistic regression model for the delayed post-test performance (see Table 3) did not find a significant main effect of learning method (i.e., retrieval vs. trial and error) nor of number of phrasal verbs (one at a time vs. 14 in a row; *z* = 0.65, *p* = .130, *d* = 0.36 [−0.85, 0.48] and *z* = −0.26., *p* = .557, *d* = −0.14 [−0.62, 0.33], respectively). No statistically significant learning method × number of phrasal verbs interaction was found either (*z* = −0.34, *p* = .584, *d* = −0.19 [−0.85, 0.48]).

<INSERT TABLE 3 ABOUT HERE>

TABLE 3

Performance on the Delayed Post-Test: Fixed and Random Effects Summary

	Fixed effects					
Parameter	Estimate	*SE*	*z*	*p*	Odds ratio	95% CI
Intercept	−3.24	0.44	−7.39	< .001	0.04	[0.02, 0.09]
Number of phrasal verbs	−0.26	0.44	−0.59	.557	0.65	[0.33, 1.83]
Learning method	0.65	0.43	1.51	.130	1.92	[0.82, 4.47]
Number of phrasal verbs × Learning method	−0.34	0.61	−0.55	.584	0.71	[0.21, 2.37]

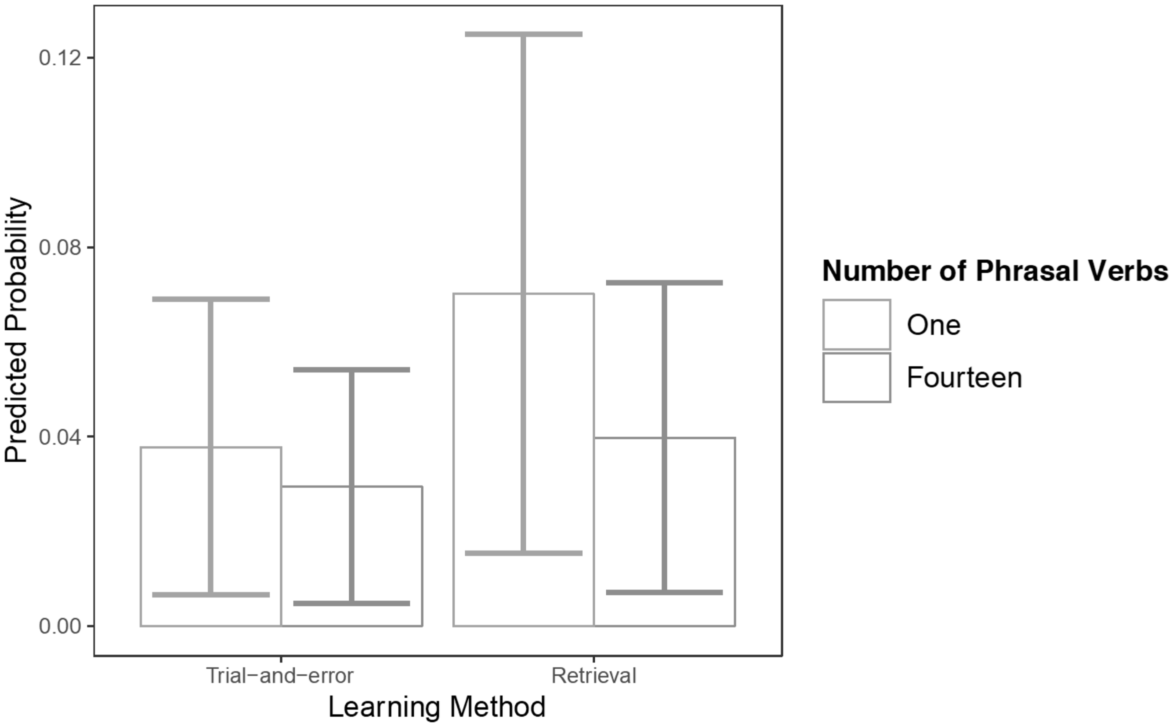
Note. Conditional r^2 value was .59. Random effect *SD*s were 1.47 for Subjects and 1.58 for Items

As shown in Figure 2, the predicted probability of a correct response in the retrieval condition (5%) was almost the same as that in the trial-and-error condition (3%). For the one-at-a-time condition (5%) and the 14-in-a-row condition (4%) the predicted probabilities were also very similar. Looking at the confidence intervals, the spread between the upper and lower values is quite large in each condition, with the lower limits showing a floor effect. Thus, the findings of the delayed post-test raise serious concerns over the benefits of *all* the conditions tried here to foster longer-term retention of phrasal verbs.

<INSERT FIGURE 2 ABOUT HERE>

FIGURE 2

The Predicted Probability of Providing a Correct Response on the Delayed Post-Test as a Function of Learning Method and Number of Phrasal Verbs



The results on the delayed post-test are inconsistent with those of Strong and Boers (2018), who found that the retrieval condition led to significantly higher scores than the trial-and-error condition not only on the immediate post-test but also on the 1-week delayed post-test, where the success rates were 56% and 41% for the retrieval condition and the trial-and-error condition, respectively. One thing to bear in mind here is that the post-tests in this study required recall of the full phrasal verbs, while at the exercise stage the students had been asked to supply only the particles. The post-tests thus posed a dual challenge: recalling not only the particle but also the verb. This is different from Strong and Boers (2018), because in that study the post-test format required recall only of the particle. That the post-test in the present experiment was more challenging may help to explain the greater attrition rate between the immediate and the delayed post-test in comparison to what was observed in Strong and Boers (2018). The fact that in both the retrieval condition and the trial-and-error condition a considerable number of participants scored close to zero in the delayed post-test inevitably also reduced the chances of finding a statistically significant difference between the two conditions. In Strong and Boers (2018), such a floor effect in the delayed test was not observed (probably thanks to the less challenging test format) and the between-condition difference in the delayed post-test performance remained significant. The different outcome in comparison with Strong and Boers (2018) may also be due to the different numbers of items tackled per exercise. In Strong and Boers (2018), all the participants were presented with sets of seven phrasal verbs, which may have posed a sufficient challenge while nonetheless avoiding too heavy a learning burden. In the present study, the phrasal verbs were tackled either one at a time, which may have invited insufficient effort to reap the rewards of retrieval (Bjork, 1994), or 14 in a row, which was perhaps overly challenging and increased the risk of inter-item interference in the longer run.

Arguably, the exercise format prompted more thinking about the particles than about the verbs (which were part of the prompts), and so it is worth investigating whether it was perhaps especially the students' failure to recall the verbs that helps to explain the overall poor delayed post-test results. In the immediate post-test, the students failed to correctly supply on average 38% verbs as compared to 42% particles. This indicates that it was not failure to recall the verbs specifically that underlies failed post-test responses. In the delayed post-test, these proportions were 87% and 82%, respectively. At first glance, this suggests a rate of forgetting that is especially steep for the verbs. On the other hand, since particles constitute a much smaller class than verbs, the probability of supplying a correct particle through guessing will be higher. If so, it is possible that the verb and particle constituents of the phrasal verbs were forgotten at similar rates between the immediate and the delayed post-test, but the slightly higher number of correct particles in the delayed post-test simply reflects a greater chance of lucky guessing. As to the comparison of learning conditions, tallies of post-test responses where only the verb was correct parallel those of fully correct phrasal verb responses: students in the retrieval conditions recalled on average 77% and 16% of the target verbs in the immediate and delayed post-test, respectively, and this compares to only 47% and 12% for the trial-and-error conditions.

Duplication of Particle Errors

Let us now turn to the more specific question regarding the effectiveness of corrective feedback on trial-and-error responses: how effective was this feedback at overriding initial incorrect responses? The proportion of particle errors students made in the tests that were identical to the particle errors they made in the exercise is reported in Table 4. (Recall that the exercise format required the students to supply only the particles, and so a similar analysis of verb errors is not possible.)

TABLE 4

Proportion of Duplicated Particle Errors (Out of Total Number of Test Errors)

Group	Immediate	Delayed
One-at-a-time retrieval	0%	5%
14-in-a-row retrieval	35%	17%
One-at-a-time trial and error	67%	67%
14-in-a-row trial and error	73%	73%

On the immediate post-test, a greater number of duplicate errors were made by participants in the trial-and-error condition than those in the retrieval condition. This is in part because the one-at-a-time retrieval group did not produce any duplicate errors, which is not surprising since that group hardly made any exercise errors in the first place and so hardly any error duplication was possible. It is therefore more meaningful to examine the 14-in-a-row retrieval group in comparison to the two trial-and-error groups. The 14-in-a-row retrieval group produced fewer duplicate errors than new errors ($z = -4.78$, $p < .001$, 95% CI [29%, 41%]). In contrast, participants in the one-at-a-time trial-and-error group and the 14-in-a-row trial-and-error group produced a greater number of duplicate errors than new errors ($z = 7.94$, $p < .001$, 95% CI [67%, 71%] and $z = 10.22$, $p < .001$; 95% CI [67%, 76%], respectively).

On the delayed post-test, the one-at-a-time retrieval group and the 14-in-a-row retrieval group still produced fewer duplicate errors than other errors ($z = -25.12$, $p < .001$, 95% CI [1%, 4%]; $z = -17.79$, $p < .001$, 95% CI [15%, 20%] respectively), whereas the one-at-a-time trial-and-error group and the 14-in-a-row trial-and-error group continued to produce a greater number of duplicate errors than other errors ($z = 7.42$, $p < .001$; 95% CI [60%, 67%]; $z = 9.12$, $p < .001$, 95% CI [63%, 70%], respectively).

What these data indicate is that the feedback received on the incorrect trial-and-error responses very often failed to prevent the same responses from re-emerging in the post-tests. The data furnished no statistically compelling evidence in support of either immediate feedback (in the one-at-a-time trial and error) or delayed feedback (in the14-in-a-row trial and error; see also Table 1), although it is noteworthy that the rate of duplicate errors was the highest in the latter condition.

CONCLUSIONS, IMPLICATIONS AND LIMITATIONS

The principal aim of this study was to compare the effectiveness of two common uses of exercises on phrasal verbs: a retrieval procedure that involved retrieving previously studied material and a trial-and-error procedure comprising a ‘pre-test’ followed by the study material as feedback. A secondary aim was to compare two variants of these procedures, where the phrasal verbs were tackled (studied and then recalled or pre-tested and then studied) either one-at-a-time or in sets. The finding of the near-immediate post-test showed that the retrieval condition was superior to the trial-and-error condition, and this treatment effect was substantial (*d* = 1.71), supporting the idea that (at least in the case of phrasal verbs and with the exercise format used here) successful retrieval of a studied item fosters better learning than does studying feedback following a failed attempt to guess the target item. On the delayed post-test, however, the advantage of retrieval over trial and error was much less pronounced (*d* = 0.36), and a sharp loss in knowledge was observed for both the retrieval and trial-and-error condition. This finding shows that initial knowledge of phrasal verbs can quickly deteriorate without follow up activities to consolidate it.

In both the retrieval and trial-and-error procedures, phrasal verbs were introduced one at a time or several in one go. In the retrieval condition, recalling an item immediately after studying it can be expected to be virtually error-free. The upside of error-free retrieval is the absence of any potential interference effects from committing errors. The downside, however, is that retrieval tends to be effortless, which is thought to be less effective for longer term retention than effortful retrieval (Bjork, 1994). By contrast, attempting to recall several items from short-term memory is likely to involve considerable cognitive effort on the part of the participant. The upside then is that the retrieval effort increases retention of the items successfully recalled. The downside, however, is that it also increases the chances of committing errors, and these errors will likely persist without remedial feedback. It is important to bear in mind, though, that the experiment involved only a single retrieval round—we cannot tell from our data if the retrieval practice with sets of phrasal verbs would lead to superior retention in the long term if it were repeated. However, as mentioned, Strong and Boers (2018) found little evidence of repeated practice with the same phrasal verbs in their textbook analysis (see also Boers, Dang & Strong, 2017, for similar findings regarding exercises on other multiword expressions) and neither did our exploration of EFL websites.

In the trial-and-error procedure, the comparatively easy one-at-a-time condition did not enhance the learning of phrasal verbs more than the more effortful 14-in-a-row trial-and-error condition did. This finding suggests that attempting to correct an error with immediate feedback does not necessarily reduce the potential detrimental effect of that error on subsequent attempts to learn the correct answer. That so many of the failed post-test responses in the trial-and-error condition were duplicates of incorrect responses given by the students at the exercise stage can be interpreted in different ways. It is possible that the feedback was not elaborate enough for the correct responses to displace the incorrect initial responses—the feedback merely presented the students with the same input but included the intact phrasal verb. The correct phrasal verb could perhaps be made more memorable if the feedback included an explanation of how the particle contributes to its overall meaning (e.g., Boers, 2013, for a review of this approach). However, the feedback provided on exercises in textbooks is typically a simple answer key (often to be found in an appendix to the book) and therefore not dissimilar from how we implemented feedback in the present experiment. A complementary explanation for the reiteration of incorrect responses may be that these students were familiar with the basic senses of the particles (i.e., when these items are used as spatial prepositions) as well as the basic senses of the (high-frequency) verbs used in the experiment. This may sometimes (for items such as *pop in* and *brighten up*) have prompted 'reasoned' guesses of what particle was compatible with the meaning prompt. Not surprisingly, given the complexity of form-meaning correspondences that phrasal verbs are notorious for, many guesses were unsuccessful, however. Participants may then understandably have found it hard to appreciate how the particle offered in the feedback made 'better sense' than the one they had supplied themselves in the exercise. An additional explanation for the re-emergence of exercise errors in the delayed post-test is that students simply forgot not only the feedback but also their initial exercise responses. If so, they may have resorted to the same 'reasoned' guessing as they did when they first tried the exercise and may thus have ended up with the same erroneous choices.

The retrieval procedure clearly reduced the likelihood of error at the exercise stage and this helps to explain why fewer of the mistakes on the post-tests were duplicates of exercise errors. This explanation also holds for the proportional difference in duplicate errors between the one-at-a-time retrieval condition and the 14-in-a-row retrieval condition. Learners in the 14-in-a-row retrieval group produced a more substantial proportion of duplicate errors than those in the one-at-a-time retrieval group. This outcome likely resulted from retroactive interference (e.g., Kulhavy, 1977), which made it harder to distinguish between a previously studied target phrasal verb and the generated error. The harmful effect of this interference was less likely to impact learners in the one-at-a-time retrieval group because they produced so few errors in the exercise phase. For this reason, the one-at-a-time retrieval condition appears more desirable than the 14-in-a-row retrieval condition, at least in the short term.

The findings of this study have implications for the design and implementation of exercises on phrasal verbs. For one, it appears more judicious to implement such exercises as retrieval practice than trial-and-error events. This means that it is advisable for textbooks and other pedagogic materials, such as online EFL resources, to present learners with relevant input on the target items first instead of pre-testing learners in the hope of reaping benefits from piquing their curiosity. For another, it is probably even more judicious to implement retrieval practice on phrasal verbs in a way that minimizes the risk of error, and this can be done by keeping the number of items to be learned in one go small, because studying and retrieving a small number of phrasal verbs will be less error-prone than targeting a large number in one go. In that regard, the average number of about seven items per exercise found by Strong and Boers (2018) as part of their analysis of EFL textbooks is clearly more recommendable than the 14 items per set used in the present study. As explained, the latter number (which nonetheless falls within the range found in EFL textbooks and EFL websites) was meant to ensure that the four experimental treatments were experienced as different by the participant-learners, but it would be interesting in an approximate replication to compare the one-at-a-time retrieval procedure to, for example, a seven-in-a-row procedure. More generally, it is worth designing retrieval-exercise conditions where retrieval is almost guaranteed to be successful but nonetheless sufficiently challenging for the learner (in keeping with the notion of desirable difficulty).

The decision to use as many as 14 items per exercise set is but one of several limitations of this study which call for caution about the generalizability of the findings. Another limitation is the fact that no feedback was given on the exercise responses supplied under the retrieval condition. As explained, this design choice was motivated by the wish to keep total time invested identical across the experimental conditions.

Follow-up studies would be welcome where additional activities with the same phrasal verbs are included. This could include not only feedback but also additional retrieval practice in which learners attempt to recall the same items again. In fact, learners should ideally test themselves repeatedly until no more mistakes are made and the desired knowledge is firmly consolidated in long-term memory (e.g., Nakata, 2017). However, the purpose of the present investigation was to evaluate what is done in many EFL textbooks and online resources and these do not tend to incorporate much repeated retrieval practice in their design. Of course, that need not prevent learners from seeking repeated retrieval practice themselves by re-doing the same exercises. In that sense, the findings from this line of research could not only inform the design of materials but also how teachers and learners can make better use of materials that are already at their disposal.

It would be interesting in this regard to explore by means of classroom observations, interviews or surveys how teachers and learners habitually go about using textbook exercises, worksheets, and online quizzes. After all, the way activities are presented and sequenced in available resources need not determine how they are implemented by individuals. Teachers who, according to such inquiries, tend to follow the textbook by the letter may then be encouraged to modify the sequencing so that exercises serve a retrieval instead of a pre-testing purpose. Where a textbook fails to provide any input on the target items either before or after the exercise, teachers can supply this input themselves by pre-teaching these items or by guiding their students to collect relevant information themselves (e.g., from dictionaries) prior to tackling the exercise. A textbook is no straightjacket.

It also needs to be acknowledged that the present investigation considered but one exercise format. Replication studies might reveal different outcomes if other formats were used. The same holds true for the post-test format: a meaning recognition test instead of a cued recall test might generate different results. All the same, that the delayed post-test yielded poor scores overall (with just over 18% successes after the learning condition that worked ‘the best’) calls the usefulness of the tried learning procedures into question. Perhaps a case should be made for the further exploration of alternative approaches to the teaching and learning of phrasal verbs, including approaches alluded to earlier which stimulate deeper cognitive involvement with the form-meaning links (e.g., Boers, 2000; Kövecses & Szabó, 1996; Lindstromberg, 2010; Strong, 2013; White, 2012; Yasuda, 2010), but which have remained conspicuously absent from mainstream textbooks to date.

Finally, the extent to which the findings can be generalized to learners of English at large remains a question as well. For students whose L1 has structural equivalents (e.g., Dutch and German), exercises on phrasal verbs might be less error-prone, for instance. Moreover, also when students share the same L1, individual differences possibly play a part so that some learners may be more susceptible than others to side effects of trial-and-error procedures. For example, in a study where L2 English learners were asked to infer the meanings of new words from short contexts before these meanings were presented to them, Elgort (2017) found a negative effect of incorrect inferences on subsequent meaning recall *especially* for participants whose proficiency was comparatively low. Although such an interaction with proficiency level was not attested in our experiment (possibly because the participant group was quite homogenous in that respect), the roles of proficiency and of learner traits more generally merit further investigation as part of this line of inquiry.

ACKNOWLEDGMENTS

We are grateful for the financial support from the Faculty of Humanities and Social Sciences (grant # 207988) of Victoria University of Wellington which enabled the first author to conduct the experiment in Japan. We would like to thank the students for their participation, Ariel Sorensen for assisting with the data collection, and Lisa Woods for her advice on statistical analyses. We extend our gratitude also to two anonymous reviewers and to Tatsuya Nakata for their insightful and detailed feedback on earlier versions of this article.

NOTES

1. We also found a small number of exercises with a focus on word order (e.g., *give up hope* vs. **give hope up*; *get your message across* vs. **get across your message*). However, as they do not concern the learning of form-meaning correspondences (i.e., the knowledge aspect of interest here), they were excluded from our counts.

2. It is noteworthy that the textbooks referred to for these examples of trial-and-error implementations of the exercises do *not* provide answer keys, thus leaving it up to the teacher to check over students' responses or relying on the students' diligence to look up the phrasal verbs in a dictionary.

3. Students were forewarned a memory test would follow on completing the exercises, but they were not informed they would be tested again 1 week later. Since the immediate and delayed post-tests were therefore likely to be experienced differently, separate analyses seemed justified.

The R code used to analyze the data is the following: model <-glmer (data = dataframe, outcome ~ fixed effect + (1|random effect) + (1|random effect), family = binomial, control = glmerControl (optimizer = "bobyga")).

REFERENCES

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57,* 35–56.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2013). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Boers, F. (2000). Metaphor awareness and vocabulary retention. *Applied Linguistics*, 21, 553–571.

Boers, F. (2013). Cognitive Linguistic approaches to second language vocabulary: Assessment and integration. *Language Teaching, 46*, 208–224.

Boers, F., Dang, T. C. T, & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises: A partial replication of Boers, Demecheleer, Coxhead and Webb (2014). *Language Teaching Research, 21*, 362–380.

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research, 18,* 54–74.

Bolinger, D. L. M. (1971). *The phrasal verb in English*. Cambridge, MA: Harvard University Press.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: Massachusetts Institute of Technology Press.

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

Cervantes, I. M., & Gablasova, D. (2017). Phrasal verbs in L2 spoken English: The effect of L2 proficiency and L1 background. In V. Brezina & L. Flowerdew (Eds.), *Learner corpus research: New perspectives and applications* (pp. 28–46). London: Bloomsbury.

Clare, A., & Wilson, J. (2002). *Language to go: Upper intermediate student's book*. Harlow, UK: Pearson Education Limited.

Dagut, M., & Laufer, B. (1985). Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition*, *7,* 73–80.

Elgort, I. (2017). Incorrect inferences and contextual word learning in English as a second language. *Journal of the European Second Language Association, 1,* 1–11.

Gairns, R., & Redman, S. (2011). *Idioms and phrasal verbs: Intermediate*. Oxford: Oxford University Press.

Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly, 41*, 339–360.

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research, 19*, 645–666.

Garnier, M., & Schmitt, N. (2016). Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System, 59*, 29–44.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition*, *40*, 505–513.

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition, 40*, 514–527.

Hulstijn, J., & Marchena, E. (1989). Avoidance: Grammatical or semantic causes? *Studies in Second Language Acquisition, 11*, 241–255.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation*, *61*, 237–284.

Kay, S. & Jones, V. (2009). *New inside out: Upper intermediate student's book.* Oxford, UK: MacMillan Education Limited.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology, 23*, 1297–1317.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*, 219–224.

Kornell, N., Hays, M. J. & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 35*, 989–998.

Kövecses, Z. (2010). *Metaphor: A practical introduction* (2nd ed.). Oxford: Oxford University Press.

Kövecses, Z., & P. Szabó, P. (1996). Idioms: A view from cognitive semantics. *Applied Linguistics, 17*, 326–355.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211–232.

Laufer, B., & Eliasson, S. (1993). What causes avoidance in L2 learning: L1-L2 differences, L1-L2 similarity, or L2 complexity? *Studies in Second Language Acquisition, 15*, 35–48.

Liao, Y., & Fukuya, Y. (2002). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning, 54*, 193–226.

Lindstromberg, S. (2010). *English prepositions explained* (2nd ed.). Philadelphia/Amsterdam: John Benjamins.

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly, 45*, 661–688.

Liu, D., & Myers, D. (2018). The most-common phrasal verbs with their key meanings for spoken and academic written English: A corpus analysis. *Language Teaching Research* (Online First).

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition, 37*, 1077–1087.

Mory, E. H. (2004). Feedback research revisited. In D. H. Jonassen (Ed.), *Handbook of research on educational communications and technology*. 2nd edition (pp. 745–783). Mahwah, NJ: Lawrence Erlbaum.

Nakata, T. (2015). Effects of feedback timing on second language vocabulary learning: Does delaying feedback increase learning? *Language Teaching Research, 19*, 416–434.

Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition, 39*, 653–679.

Oxenden, C., & Latham-Koenig, C. (2010). *New English file: Advanced student's book*. New York: Oxford University Press.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143*, 644–667.

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition, 35*, 1917–1927.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15*, 243–257.

Roberts, R., Clare, A., & Wilson, J. (2011). *New total English: Intermediate student's book*. Harlow, UK: Pearson Education.

Roediger III, H. L., & Butler, A. C. **(**2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20–27.

Roediger III, H. L., & Karpicke, J. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18,* 55–88.

Seabrooke, T., Hollins, T. J., Kent, C, Wills, A. J., & Mitchell, C. J. (2019). Learning from failure: Errorful generation improves memory for items, not associations. *Journal of Memory and Language, 104*, 70–82.

Siyanova, A., & Schmitt, N. (2007). Native and non-native use of multi-word vs. one-word verbs. *International Review of Applied Linguistics in Language Teaching, 45,* 119–139.

Slamecka, N., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604.

Smith, M. A., Roediger, H. L., & Kapicke, J. D. (2013) Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1712–1725.

Strong, B. (2013). A cognitive semantic approach to L2 learning of phrasal verbs. *The Language Teacher, 37*, 27–31.

Strong, B., & Boers, F. (2018). The error in trial and error: Exercises on phrasal verbs. *TESOL Quarterly* (online early view).

Warmington, M., & Hitch, G. J. (2014). Enhancing the learning of new words using an errorless learning procedure: Evidence from typical adults. *Memory, 22,* 582–594.

Warmington, M., Hitch, G. J., & Gathercole, S. E. (2013). Improving word learning in children using an errorless technique. *Journal of Experimental Child Psychology, 114,* 456–465.

White, B. J. (2012). A conceptual approach to the instruction of phrasal verbs. *The Modern Language Journal*, *96*, 419–438.

Yasuda, S. (2010). Learning phrasal verbs through conceptual metaphor. *TESOL Quarterly*, *44*, 250–273.

APPENDIX A

EFL Websites Explored for Phrasal Verbs Practice (Accessed November 2018)

https://www.english-grammar.at/index.htm

https://www.agendaweb.org/verbs/phrasal-verbs-worksheets-lessons

http://eslpdf.com/index.html

https://www.ego4u.com/en/cram-up/grammar/phrasal-verbs

http://www.esl-lounge.com/student/phrasal-verbs-exercises.php

https://www.englishgrammar.org/category/verbs/

https://www.perfect-english-grammar.com/phrasal-verbs.html

https://intercambioidiomasonline.com/2017/09/16/how-to-learn-and-use-phrasal-verbs-free-ebook-in-this-post/

https://www.englisch-hilfen.de/en/exercises_list/phrasal.htlm

https://www.gamestolearnenglish.com/phrasal-verbs/

APPENDIX B

Phrasal Verbs Targeted in the Learning Experiment

Phrasal Verbs Set A

Back down – to decide not to do something; Catch on – to become very popular quickly; Carry on - to continue to do something; Dive in – to start to eat food; Figure out – to understand something; Get out – a secret becomes known; Hang out – to spend time with friends; Hold up – to cause a delay; Make up – to create a story; Nod off – to fall asleep for a short time; Own up – to admit that you have done something wrong; Pop in – to visit for a short visit; Rip off – to charge someone too much money; Screw up – to make a serious mistake.

Phrasal Verbs Set B

Boil down – to give the most important information; Brighten up – to become happier; Brush up – to improve your skill; Call off – to cancel a something; Chip in – to give money; Crack on – to continue doing something as quickly as possible; Give in – to accept that you cannot win; Head off – to go somewhere; Open up – to talk about your personal feelings; Pass away – to die; Stick out – to be easy to notice because of being different; Run out – to use all of something; Turn off – to lose interest; Wrap up – to finish something.