Manuscript Title: Repetition and Incidental Learning of Multiword Units: A Conceptual Multisite Replication Study of Webb, Newton, and Chang (2013)

Author(s): Elke Peters[a], Eva Puimège[a], and Paweł Szudarski[b]

Author Affiliations:
[a]KU Leuven
[b]University of Nottingham

Author Twitter information [optional]:

Elke Peters: @Peters_Elke10

Eva Puimège: @EvaPuimege

Pawel Szudarski:@pawel_szudarski

Abstract

This multisite study replicates Webb, Newton, and Chang's (2013) study on the effect of repetition on incidental learning of multiword units (MWUs). Even though more researchers have started to investigate MWUs, most data have been collected from university students. Furthermore, the large effect of MWU repetition on learning reported by Webb et al. has not yet been corroborated. Data in our study were collected from two university samples (EFL students in Poland and Flanders) and one non-university sample (Flemish EFL learners in secondary schools). Unlike Webb et al., we adopted a counterbalanced within-participants design. Participants read and listened to a modified graded reader in which target MWUs occurred 1, 5, 10, or 15 times. In line with the initial study, we found a positive effect of repetition. However, the learning gains were smaller, and the number of repetitions needed was different. The findings were consistent across the university and non-university samples. The study concludes with a discussion of these findings in relation to both pedagogical implications and the benefits of multisite replication research.

Keywords: repetition, replication, multiword units, vocabulary, incidental learning, English

Author notes / acknowledgements

Elke Peters: conceptualization; methodology; investigation; writing – original draft preparation; writing – review and editing. **Eva Puimège**: methodology; investigation; formal analysis; writing – original draft preparation (Analyses; Results); writing – review and editing. **Paweł Szudarski**: methodology; investigation; writing – review and editing.

The following sentence must be included in this section: A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at https://oasis-database.org

Any acknowledgements or funding/grant support disclosure (optional): We would like to thank the reviewers and the guest editors for their helpful feedback on earlier versions of the manuscript. We are also grateful to the schools, teachers and participants who helped us with the data collection. In this regard, we would like to express our gratitude to Yves Stevens, Matylda Weidner, Emilia Wąsikiewicz-Firlej, Małgorzta Krzemińska-Adamek, Anna Kiszczak and Izabela Jarosz. A special thank-you is also due to Jessica Norledge for her assistance with making recordings.

The following sentence must be included in this section: Correspondence concerning this article should be addressed to [Elke Peters, KU Leuven, Faculty of Arts, Antwerp, Sint-Jacobsmarkt 49-51, 2000 Antwerpen, elke.peters@kuleuven.be].

Introduction

Language is to a large extent formulaic, that is, it consists of "strings of letters, words, or other elements […] that necessarily enjoy a degree of conventionality or familiarity among (typical) speakers of a language community or group, and that hold a strong relationship in communicative meaning" (Siyanova-Chanturia & Pellicer-Sánchez, 2019, p.5). While formulaic language consists of multiword (e.g., *spread the virus*) as well as single-word items (e.g., *wow, hello*) (Siyanova-Chanturia & Omidian, 2020, p.530), the focus in this manuscript is on multiword units (MWUs). There is a growing body of literature recognizing the importance of knowing MWUs for reading comprehension (Kremmel, Brunfaut, & Alderson, 2017), fluency (Tavakoli & Uchihara, 2019), and for achieving high levels of proficiency (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Crossley, Salsbury, & McNamara, 2015; Paquot, 2019). However, MWUs pose a challenge for language learners, because they make reading texts more difficult to comprehend, even if learners are familiar with all the single words making up the text (Martinez & Murphy, 2011). Further, they are a common source of errors in learners' spoken and written output (e.g., Laufer & Waldman, 2011; Nesselhauf, 2003).

In recent years, there has been increasing evidence that it is possible to pick up MWUs incidentally through reading (Pellicer-Sánchez, 2017), reading-while-listening (Webb & Chang, 2022), or watching (captioned) TV (e.g., Majuddin, Siyanova Chanturia, & Boers, 2021; Puimège & Peters, 2019b; Puimège & Peters, 2020; Puimège et al., 2023).

While repetition is positively related to incidental learning of single words (Uchihara, Webb, & Yanagisawa, 2019), the findings regarding MWUs have been mixed (Pellicer-Sánchez, 2017; Szudarski & Carter, 2016). A particularly interesting study on repetition is that by Webb et al. (2013) given its wide range of repetitions (1 to 15) and the large effects that it reported. Yet, no study has corroborated this large repetition effect. This means that the

generalizability of much published research into the effect of repetition on incidental learning of MWUs is problematic because of a lack of replication studies (see Toomer & Elgort, 2019, for an exception) and, also, an overreliance on university samples.

Replication research has been proposed as a way to improve the generalizability of findings into second language acquisition (SLA) (Marsden, Morgan-Short, Thompson, & Abugaber, 2018). With respect to the current study specifically, more research is needed with university as well as non-university samples to shed more light on how incidental learning of MWUs is affected by repetition effects as well as learner-related variables, like prior vocabulary knowledge and educational level. Yet, the number of replication studies in SLA is low "with fewer than one article in 400 being a replication study" (Marsden et al., 2018, p.394). This also holds true for the field of vocabulary studies, with only few attempts at replicating previous studies (Cobb, 2003; Crossley & Skalicky, 2017; Noreillie, Kestemont, Heylen, Desmet, & Peters, 2018; Toomer & Elgort, 2019). To address this issue, we aimed to replicate Webb et al. (2013) and investigate the effect of repetition on incidental learning of MWUs by two university samples (Polish-speaking and Dutch-speaking EFL university students) and one non-university sample (Dutch-speaking EFL learners [age 16-17] in secondary education in Flanders).

Background

Incidental learning of MWUs

There is a general consensus that in addition to deliberate learning of vocabulary, incidental learning should be "part of any L2 vocabulary learning program" (Webb et al., 2013, p. 92) for the simple reason that not all vocabulary can be taught explicitly. Research has shown that incidental vocabulary learning gains in one-off interventions, (e.g., reading a text, listening to a text, or watching one episode of a TV program) are typically smaller than in intentional

learning (e.g., Hulstijn, 2003; Webb, 2020). However, if learners engage with large amounts of foreign language (L2) input, e.g., in an extensive reading program (Webb & Chang, 2015) or outside of school (De Wilde, Brysbaert, & Eyckmans, 2020; Puimège & Peters, 2019a; Sundqvist, 2019), then larger vocabulary gains can accrue.

In usage-based accounts of language learning, input is regarded as a primary driving force of L2 acquisition (Ellis, 2006, 2012) and exposure frequency plays an important role in this process. In this light, it has been argued that incidental vocabulary acquisition is determined by characteristics of the input, such as the frequency of occurrence and salience of lexical items (Ellis, 2006). Specifically, it has been well established that repetition in written input enhances the learning of single words, even though the exact number of required repetitions seems to differ, and might depend on the word knowledge aspect tested, learners' proficiency and/or education level (see Uchihara, Webb, & Yanagisawa, 2019, for a meta-analysis). For instance, Uchihara et al.'s meta-analysis revealed that the effect of repetition was larger for university students than for learners in primary and secondary education. Similarly, Elgort and Warren (2014) found that the lower learners' proficiency level, the more repetitions were required for incidental learning to occur. In an eye-tracking study, Pellicer-Sánchez (2016) found that the effect of repetition on reading times for new words encountered during reading was slightly different for L1 and L2 readers. In L1 readers, a decrease in reading times was observed earlier (after the first encounter), which suggests that they familiarized with new words faster than L2 learners. Learning rates may thus vary depending on proficiency and prior vocabulary knowledge.

Research has also demonstrated that L2 learners are sensitive to the frequency patterns of MWUs in the input they are exposed to (Northbrook & Conklin, 2019). While formulaicity is pervasive in language as a whole, examples of specific MWUs are not (Cobb, 2019). Additionally, L2 learners may lack the necessary amount of exposure to the target language

to learn that words pattern together as part of a MWU (e.g., Durrant & Schmitt, 2010). Consequently, incidental learning of MWUs requires either large amounts of meaningful input or pedagogic interventions to increase the likelihood that students make learning gains (Boers, 2020).

In addition to highlighting the role of frequency, usage-based approaches to language learning hold that incidental vocabulary acquisition is also affected by the salience of lexical items (Ellis, 2006, 2012). This means that even with large amounts of input, non-salient MWUs may not easily become intake because they may not stand out for L2 learners (Boers, 2020). For instance, L2 learners might fail to notice MWUs which are made up of familiar, high-frequency constituent words, such as *pay attention, take a picture* (Boers, 2020). As a result, pedagogic interventions that draw learners' attention to different kinds of MWUs have been used to enhance the learning of such units (e.g., Puimège et al., 2023; Szudarski & Carter, 2016; Toomer & Elgort, 2019). Such a pedagogic intervention is input flooding, or seeding L2 texts with multiple examples of MWUs. The present paper examines the effectiveness of input flooding in the context of incidental learning of MWUs.

Repetition

Research into the learning of individual words suggests that if any incidental learning is to occur, the number of repetitions needs to be very high. Apart from Webb et al. (2013), four studies have investigated the effect of repetition in the input on learning MWUs (Durrant & Schmitt, 2010; Macis, 2018; Pellicer-Sánchez, 2017; Szudarski & Carter, 2016). However, no study has been able to corroborate Webb et al.'s (2013) large effect of repetition, as the learning gains from repetition were either smaller (Durrant & Schmitt, 2010; Macis, 2018) or non-existent (Pellicer-Sánchez, 2017; Szudarski & Carter, 2016). One explanation is probably that repetition was operationalized differently in the aforementioned studies (1

versus 2 in Durrant & Schmitt, 2010; 6 versus 12 in Szudarski & Carter, 2016; 4 versus 8 in Pellicer-Sánchez, 2017; 1, 5, 10 versus 15 in Webb et al.). It could be hypothesised that as many as 15 encounters are needed for repetition to have an effect on incidental learning of collocations.

Another explanation could be that the studies used different methodologies (a case study in Macis, 2018, short texts in Pellicer-Sánchez, 2017 and Szudarski & Carter, 2016, sentences in Durrant & Schmitt, 2010) and different types of MWUs (adjective-noun, verb-noun, adjective-pseudoword, low-frequency MWUs). Further, Webb et al. (2013) was the only study that used bimodal input (reading-while-listening), which may have made it easier to identify the MWUs because of the prosodic information in the spoken input (Lin, 2019; Malone, 2018; Webb & Chang, 2022). Further, the sample size also differs somewhat between these studies, ranging from 13 to 18 participants per experimental group. Macis's (2018) study was a case study with three participants.

Finally, it should be noted that participants in the four studies discussed here had diverse profiles, which may explain the different findings. The samples in both Durrant and Schmitt (2010) and Macis (2018) were university students in the UK, while Pellicer-Sánchez (2017) recruited adult EFL learners from a language school in the UK. The participants in Szudarski and Carter (2016) on the other hand were 18-year old EFL learners in a Polish secondary school. This highlights the need for research on repetition effects on incidental learning of MWUs with more diverse populations (Andringa & Godfroid, 2020). The same holds true for incidental vocabulary learning in general, as fewer studies have focused on non-university samples (for a few exceptions, see Laufer & Girsai, 2008; Pavia, Webb, & Faez, 2019; Serrano & Huang, 2018; Szudarski & Carter, 2016), such as learners in primary or secondary education. A notable exception is research focused on aural input (i.e., listening to texts or songs, audiovisual input), which has shown that learners in primary schools (e.g.,

d'Ydewalle & Van de Poel, 1999; Koolstra & Beentjes, 1999; Neuman & Koskinen, 1992; Pavia, Webb, & Faez, 2019) as well as secondary schools (e.g., Pavakanun & d'Ydewalle, 1992; Peters, 2019; Pujadas & Muñoz, 2019) can pick up new words by watching (subtitled or captioned) TV or listening to songs. Research also suggests that vocabulary gains are more pronounced with older learners than younger learners (De Vos, Schriefers, Nivard, & Lemhöfer, 2018; d'Ydewalle & Van de Poel, 1999). However, in their meta-analysis of vocabulary learning from spoken input, De Vos et al. (2018) rightly pointed out that the age effect could not be disentangled from proficiency, given that age coincided with education level (university, high school, elementary school, kindergarten). In other words, the older learners were more proficient and had longer experience with the L2 than younger learners.

The initial study

In light of the above, in order to determine whether the different research outcomes on repetition should be attributed to different approaches to testing repetition or the used methodologies, the effects of repetition should first be replicated with a university sample to examine whether we can confirm the effect of repetition on learning collocations. Because Webb et al. (2013) is an oft-cited study on the incidental learning of MWUs and one of the first examinations to have investigated the effect of repetition, our first aim is to replicate Webb et al. with two university samples, one in Poland and one in Flanders (Belgium). Second, it remains to be seen whether similar learning gains as in Webb et al. can be found with non-university-level learners (e.g., secondary school learners), as previous research has shown that the effect of repetition might depend on learners' profile (Elgort & Warren, 2014; Uchihara et al., 2019). Thus, this replication study with two university samples and one non-university sample allowed us to contribute to the reproducibility and generalizability of the research into the relationship between repetition and incidental learning of MWUs.

Webb et al.'s study adopted a pretest, treatment, posttest, between-participants design[1] to investigate the effect of repetition; 161 L1 Chinese university students of English in Taiwan were divided into 4 experimental groups (assigned to 4 different versions of a graded reader, each with a different number of occurrences of the target MWUs) and one control group, who only completed the tests. One week before the treatment, a pretest (one multiple-choice test of form recognition) was administered, after which the participants read and listened (= reading-while-listening or RWL) to one of four versions of a graded reader called *New Yorkers*. Graded readers are (simplified) books that are written for L2 learners. The participants encountered all 18 MWUs once, five, ten or 15 times, depending on the condition they were assigned to. As soon as the treatment ended, four posttests followed: a productive form test ("provide the collocate"), receptive form test (multiple-choice matching test), productive form-meaning test ("translate into English") and finally receptive form-meaning test ("translate into Chinese"). The MWUs were incongruent verb-noun MWUs, that is, items with a low degree of "word for word overlap between L1 and L2" (p. 103-104). Finally, it is also important to note that Webb et al. (2013) used written as well as spoken input in their reading-while-listening treatment, which may have positively affected the learning gains (see also Chen, 2021; Conklin, Alotaibi, Pellicer-Sánchez, & Vilkaitė-Lozdienė, 2020; Malone, 2018; Webb & Chang, 2022, for a discussion of the benefits of bimodal input).

Results revealed that MWUs can be learned incidentally from RWL and large gains were found in the receptive form test (Cohen's d = 4.15 for 15 repetitions; Cohen's d = 2.36 for 10 repetitions; Cohen's d = 1.08 for 5 repetitions; Cohen's d = 1.33 for 1 encounter). It should be noted however that some learning was also reported in the control group (Cohen's d = .40). Further, a MANOVA of the four posttests showed that repetition enhanced collocational learning (partial eta squared = .28), with 15 occurrences resulting in the highest scores and statistically being more effective than all the other conditions. Ten occurrences

resulted in more learning than five repetitions, but only in the productive tests, while five repetitions led to more learning in the receptive form test compared to one occurrence only, even though the effect size of 1 encounter was higher; this result may have been due to the greater precision (smaller standard deviation) in the observed learning gains following one rather than five repetitions. No differences were reported between one and 0 occurrences (control group). One-way analyses of variance of the four posttests indicated that the largest effect of repetition was found in the test where the participants had to supply the form of the target MWUs (partial eta squared = .56), while the smallest effect was found in the test where the participants had to recall the meaning of the English MWUs (partial eta squared = .33).

While Webb et al.'s findings suggest that repetition enhances incidental learning of MWUs, the study has three important limitations. Firstly, different tests were used during the pretest and posttest sessions, that is, the pretest did not include any measures of learners' productive knowledge. This was likely problematic methodologically because the target MWUs were made up of frequent words and hence some learners might have had knowledge at a receptive level of mastery that was undetected in the pretest. Also, learners might have been able to recall the meaning of collocations, but they may not have been able to link collocates in the multiple-choice test. Consequently, because the study did not establish whether this was the case, the learning reported might have resulted not only from the treatment but also from learners' previous knowledge.

Second, because of the input flooding, the reading materials were manipulated fairly substantially. As a result, the text length of the different versions (1, 5, 10, or 15 repetitions) was not the same. Further, the artificial nature of the texts (as a result of the text manipulation) might have had an impact on learners' motivation to read the texts. Horst, Cobb, and Meara (1998) pointed to the potential of text manipulation, for instance, by writing in additional repetitions, but they also argued that this should not be at the expense of the

integrity of the text. Consequently, it is important to explore a potential trade-off between repetition on the one hand and learners' motivation on the other, particularly if we want to make pedagogical recommendations regarding the benefits of input-flooding for incidental learning. Thirdly, the analyses did not take into account individual differences such as learners' prior vocabulary knowledge, which is known to affect the learning of multiword units (Puimège & Peters, 2020; Vilkaitė, 2017).

Rationale and research questions

In brief, there are a number of reasons why a replication of Webb et al. (2013) is warranted. First, the large, beneficial effects of repetition, which were reported in Webb et al. (2013), have not been corroborated. Second, like many studies in the field of SLA, research into incidental learning of MWUs suffers from a sampling bias (Andringa & Godfroid, 2020), which is why the effect of repetition should be investigated in a range of educational settings. Finally, it is worthwhile to address how learners' prior vocabulary knowledge affects the incidental learning of MWUs that occur repeatedly in the input, as previous research has shown that the effect of repetition may depend on learners' proficiency level.

With this in mind, the present study replicated Webb et al. (2013) with two university samples of EFL learners to assess the reproducibility of the original findings, and additionally with one non-university sample (EFL learners in grade 10 and 11) to investigate whether the original findings can be generalized to other educational settings, as previous research has shown that the effect of repetition might depend on the learner's profile (Elgort & Warren, 2014; Uchihara et al., 2019). Our aim was also to improve both the design and ecological validity of the study. Unlike Webb et al., we used a within-participants design, in which the repetitions were counterbalanced (see Godfroid, 2020, for a discussion of the advantages of repeated-measures, within-participants designs). This means that all participants received all

levels of the repetition variable, that is, 1, 5, 10, and 15 repetitions. The target collocations occurred once, 5, 10 or 15 times in the reading materials and participants were assigned to read and listen to one version of the text as shown in Table 1. For example, collocations encountered five times by one participant were encountered only once by another participant (see Table 1 and Appendix S1).

INSERT TABLE 1 APPROXIMATELY HERE

A within-participants design has the advantage of having more power without needing more participants (Godfroid, 2020; Nicklin & Vitta, 2021). Further, we focused on two knowledge aspects only (form recognition and form recall), but they were tested both prior to and after the treatment, increasing control over prior knowledge of the MWUs. Finally, we also took into account participants' prior vocabulary knowledge (i.e., learner-related variable) that might affect L2 learning.

Following Webb et al. (2013), we addressed the following research questions:

1. To what extent can MWUs be learned incidentally through reading-while-listening to a modified graded reader?

In line with Webb et al. (2013), we hypothesized that MWUs can be learned incidentally through reading-while-listening.

2. How many encounters are needed to incidentally learn the written form of MWUs through reading-while-listening?

We expected a positive relationship between repetition and learning (Uchihara et al., 2019). Given that recognition tests are easier than recall tests, we also hypothesized that fewer encounters would be needed for recognition than recall (e.g., Uchihara et al., 2019).

Additionally, we aimed to answer the following question which specifically focuses on sample characteristics:

3. Is the learning of MWUs affected by learners' profile in terms of education level, L1, and prior vocabulary knowledge?

Drawing on previous research (Puimège & Peters, 2020; Vilkaitė, 2017), we hypothesized that learners' prior vocabulary knowledge would be positively related to learning, and that it would moderate the effect of repetition on learning: learners with less prior vocabulary would need more repetitions than learners who know more words in general (e.g., Elgort & Warren, 2014). Further, our hypothesis was that there would be more learning in the university than in the non-university sample (e.g., de Vos et al., 2018; Uchihara et al., 2019) and that the results for Webb et al.'s (2013) university students in Taiwan would generalize to university students in Poland and Flanders.

Method

Participants

Our primary goal was an approximate replication of Webb et al. (2013). To this end, first- and second-year university students majoring in English were recruited from the Polish site and first- and second-year university students not majoring in English from the Flemish site. These groups were similar to Webb et al.'s study in terms of educational level and level of English proficiency. Our second aim was to extend the findings to a non-university sample,

that is, EFL learners in Flanders who were in grade 10 or 11 (age 16-17). This group differs from the university students in Poland and Flanders with respect to educational level. We hypothesized that they would also differ from the university students in proficiency, but this difference was not found in the results of the Vocabulary Levels Test (see Table 3). It should be noted that the inclusion criterion for all participants was their score on the 2000-word level of the Vocabulary Levels Test, VLT (Schmitt, Schmitt, & Clapham, 2001) to ascertain that participants were able to comprehend the vocabulary in the graded reader. The cut-off score was 26/30.

Researchers at each site collected data independently, following the same procedures. Informed consent was obtained from all participants. To determine the sample size, we conducted a power analysis. First, we built a lme4 model in the simr package (Green & MacLeod, 2016) in R (version 3.4.3), using artificial data. The model included fixed effects for repetition, prior vocabulary knowledge (VLT score), L1, education level, and pretest score, an interaction term between repetition and VLT score, and random effects for participants and items. We then ran simulations of the model to obtain power estimates for different sample sizes. The results indicated that we needed a minimum of 140 participants per test to obtain a power of 80% to be able to detect a significant effect of Repetition on Gain score. It should be noted that we ran one statistical model per test, i.e., one model for the form recall test and one for the form recognition test (see also Data collection instruments below). This meant that we needed to recruit about 47 EFL university students in Poland, 47 EFL university students in Flanders, and 47 EFL students in secondary schools in Flanders per test format. Given that there might be some data loss due to absence of participants or participants not obtaining a score of 26/30 on the 2K-level of the VLT, we planned to invite approximately 160 participants per test, or 320 participants in total, to ensure that we would have data of 140 participants taking part in both the pretest and posttest session. A detailed

description of the power analysis procedure, as well as the R code, are available in the OSF (https://osf.io/uh7sw/?view_only=ff8cada925c144089cadbd77b091f69b).

Following data collection, 297 participants in total completed at least one experimental session. Of these 297 participants, 114 learners did not finish the experiment. Data from another six participants were removed because they failed to follow the instructions, and, as a result, completed a posttest that did not match the pretest format. Finally, data from 15 participants whose score was below 26/30 on the 2,000 level (*n* = 2 in the Polish sample, *n* = 5 in the Flemish university sample, and *n* = 8 in the Flemish secondary school sample) were excluded from the analysis to ensure that all participants could understand the vocabulary in the reading materials (see Schmitt et al., 2001, for the criterion of mastery of a level).

This resulted in a final dataset of 162 participants in total, across sites and education levels, 80 of whom completed the recognition pre- and posttest, and 82 of whom completed the recall pre- and posttest. The final sample size per site and education level is presented in Table 2 below.

INSERT TABLE 2 APPROXIMATELY HERE

Site 1: Poland

Eighty-nine participants were recruited from English and linguistics majors at several universities in Poland (age 19 onwards). All of these participants were L2 learners of English as a foreign language who had learned English through formal instruction since primary or secondary school. Similarly to the learners in Webb et al. (2013), Polish participants were university students and therefore they represented the same education level. In terms of proficiency in English, previous research (Szudarski, 2019) showed that L1-Polish first- and second-year university students majoring in English had an average score of 28.4/30 on the

2,000 word level of the VLT, which we assumed would be sufficient for adequate comprehension of the reading texts.

Site 2: Flanders

Our aim was to recruit participants from the first and second year at university who were not majoring in English (e.g., business students, law students, communication students). Our second sample in Flanders was EFL learners from grade 10 and 11 of the *algemeen secundair onderwijs* (academic track in secondary education) in Flanders (the Dutch-speaking region in the northern part of Belgium).

The sample of Flemish university students consisted of 35 participants. Most of the university students (age 18-34, M = 20.87, SD = 3.18) had had formal English instruction since age 13 or 14 (n = 23), with 9 participants reporting having had English instruction since age 10-12, two participants reporting English instruction starting at 15 or 18, and one participant having had English instruction since age 7. The participants' education level was comparable to that of the Polish site and the initial study. The number of hours of current English instruction varied between learners, with most reporting zero (n = 24) or 2 (n = 10) hours of English per week, one participant reporting 8 hours of English per week (Japanese studies), and one participant indicating having 20 hours of English per week (physiotherapy). Using the VLT results from a previous data collection (n = 217; 2000-word level = 27.93, with 199 participants obtaining a score of 26/30 or more), we predicted that these learners would be able to read the text.

The 38 non-university participants were EFL learners (aged 16-17) in grade 10 or 11. English is a compulsory subject in secondary education in Flanders. Participants normally have two to three 50-minute classes per week. Despite fewer years of instruction compared to Webb et al.'s (2013) study, participants in the present study were expected to have a minimum score of 26 out of 30 on the 2,000 word level of the VLT as a result of large

amounts of exposure to extramural English (Peters, 2018; Peters et al., 2019). Previous research has indeed shown that Flemish secondary school students tend to know the 2,000 most frequent words in English receptively (Peters et al., 2019)[2].

INSERT TABLE 3 APPROXIMATELY HERE

Design

We replicated Webb et al.'s (2013) study by adopting a pretest-posttest design and by using the same texts and some of the test instruments. Unlike the initial study, however, we used a counterbalanced, within-participants design. One of the advantages of within-participants designs is that they have more power because individual differences can be better controlled for i.e., accounted for statistically through the estimation of random effects parameters. In this design, all participants were exposed to all experimental conditions, that is, the four repetition conditions of 1, 5, 10 or 15 repetitions (see also Table 1 and Appendix S1). Further, the design allowed us to better control for text length. Consequently, there were no differences between participants in terms of treatment, unlike in Webb et al. (2013), where the text with 15 repetitions condition took much longer to read and listen to.

All participants read a modified graded reader, *New Yorkers,* while simultaneously listening to a recording of this text. The four repetition levels were counterbalanced across the participants and text versions. The four repetition levels in the input were as follows: (1) 1 occurrence of the MWU, (2) five repetitions, (3) ten repetitions, and (4) 15 repetitions. Table 1 summarizes the design (see also Appendix S1).

To control for the effect of repeated testing, in addition to 17 target MWUs, we also included nine distractor items on the tests that did not occur in the graded readers (see Appendix S2). This was different from the initial study, which had a control group who only

took the tests and was not exposed to the reading materials and the target items. The distractor items were treated as the 0-occurrences baseline to which the four repetition levels were compared.

Reading materials

We used the same reading materials from Webb et al. (2013) (see Appendix S1). Webb et al.'s participants read and listened (RWL) to a 700-headword stage-2 graded reader called *New Yorkers*. The graded reader consists of high-frequency words, that is, words from Nation's (2004) first and second 1,000 word lists. These words were considered appropriate for the participants, as they constituted a minimal vocabulary load. There were four stories within the reader (one original story was omitted), with the target MWUs embedded throughout the texts in an effort to manipulate the number of occurrences. Items occurred once, five, ten or fifteen times throughout the texts.

In this study, four different versions were created to counterbalance the frequency of occurrence of the items within groups (see Table 1 and Appendix S1). Importantly, given that the target MWUs were verb-noun phrases, different grammatical forms (both present and past) were used (e.g. *broke the silence, break her silence*). The distance between node words and collocates also varied, reflecting language users' authentic encounters with naturally occurring discourse. This means for instance that some MWUs were adjacent, while others were not (e.g., *she didn't have to* <u>*cut corners*</u> versus *without* <u>*raising*</u> *unpleasant* <u>*questions*</u>). It should be noted that in previous research adjacency has not been found to affect learning (Vilkaitė, 2017) and therefore this factor was not considered. A native speaker of British English, who read the texts at a natural pace and placed no special emphasis on the target items, recorded the aural versions of the text.

Target items

We used 17 of the 18 target items as in Webb et al. (2013) in the pre- and posttests (see Table 4). The MWUs consisted of high-frequency single word components. Unlike in the initial study, in which the items had a low degree of L1-L2 congruency, some items were congruent for either the Flemish or Polish sample[3]. Congruent MWUs can be translated word-for-word and therefore it can be predicted that participants might have been able to provide the correct form of the target MWUs without having learned them from reading the grader reader. Because the present study was a replication study, we aimed to be as close to the original design as possible, but to address this issue, congruency was taken into account in a secondary analysis (see endnote 1) as a covariate. Further, we argue that given that L2 learners encounter both congruent and incongruent MWUs in real life, the inclusion of both types of MWUs was ecologically valid.

The target items and their Polish and Dutch translations, their frequencies, t-scores and MI scores, which are measures of association strength, are shown in Table 4.

INSERT TABLE 4 APPROXIMATELY HERE

Data collection instruments

The data collection instruments in the initial study consisted of the VLT (Schmitt et al., 2001), one pretest and four posttests. Webb et al. used four written posttests to measure the learning of MWUs: a productive form test, followed by a receptive form test (labelled form recognition test in the current study), then a productive form-meaning test (labelled form recall translation test in the current study), and finally a receptive form-meaning test (i.e., meaning recall test). However, Webb et al. only pretested one aspect of lexical mastery,

namely form recognition, which makes it difficult to draw any conclusions regarding the learning gains for other aspects of lexical knowledge. Batteries of tests are often used to give a more complete picture of the learning process, as more word knowledge aspects are targeted (Webb, 2007), but administering multiple tests can also result in a test effect. Webb et al. also considered this effect and treated it as their rationale for using only one measure on the pretest and argued that four pretests "would have alerted participants to the purpose of the study and may have also contributed to learning" (p.112).

Because we wanted to test more than one knowledge aspect, but also avoid a test effect, we split the group of participants into halves (see also Peters & Webb, 2018). Each group took either the form recall or form recognition test, as pretest and posttest, so each group was tested on one knowledge aspect only (see below). This means that unlike Webb et al., we focused on two and not four knowledge aspects; that is, form recall and form recognition[4]. To avoid any ambiguity, we will use the terms *form recognition* and *form recall* instead of receptive form and productive form-meaning tests respectively (see also Laufer & Goldstein, 2004).

Form recognition test: Participants' knowledge of the target MWUs prior to the treatment was measured in a written form recognition test, which focused on learners' ability to recognize the correct form of the MWUs by matching the node word with its collocate (see Appendix S3). Each test item consisted of the node word (verb) and five options: four nouns and the *I don't know* option. The latter was added to minimize guessing.

Throw	a) light	b) name	c) risk	d) clock	e) I don't know
Remember	a) room	b) money	c) time	d) decision	e) I don't know

We altered Webb et al.'s instructions from "Circle the words which go together in a sentence" into "Circle the words which often go together", because the former might result in

acceptable free combinations, such as *remember the decision* or *break the desk* (see Appendix S3). The same test was administered as the posttest, but the items were presented in a different order and, similarly to the pretest, learners were encouraged not to guess blindly. The reliability was α = .76 (pretest) and α = .79 (posttest).

Form recall: In this test, participants had to translate the MWUs from their L1 into English. Participants were prompted to use the MWUs that occurred in the reading texts. We kept the same instruction as in the initial study: *Write the English translations in the blanks. All of the answers are at least two words: a verb and a noun. [and for the posttest only] All of the English translations were present in the stories you have read*. As the example below shows, participants were given a cue in their L1 (*aan de eisen voldoen* in Dutch and *zaspokajać popyt* in Polish) and were expected to produce the MWU *meet demand*. We did not provide the first letter of the constituent words in the target MWUs because this was not done in the initial study either.

aan de eisen voldoen _____

Zaspokajać popyt _____

We used the same form recall test as the posttest, but the items appeared in a different order. The reliability of the form recall test was α = .60 (pretest) and α = .62 (posttest).

In line with Webb et al., we used the VLT (Schmitt et al., 2001) to assess learners' prior vocabulary knowledge, as research has shown that prior vocabulary knowledge might affect the amount of learning (Puimège & Peters, 2019b, 2020; Vilkaitė, 2017). The VLT, which is a frequency-based vocabulary test (Nation, 1983; Schmitt et al., 2001), gives an estimate of learners' knowledge of single words at four frequency levels (2,000 most frequent or 2K, 3,000 most frequent or 3K, 5,000 most frequent or 5K, and 10,000 most frequent word

families in English or 10K) and for the Academic Word List (Coxhead, 2000). The test has a matching format, in which 30 items per test frequency level have to be matched to their definition, totalling 150 items. The VLT has been shown to be a valid and reliable vocabulary levels test for the targeted participants in the present study (Puimège & Peters, 2020; Schmitt et al., 2001). The VLT's reliability was $\alpha = .97$.

Comprehension

In addition to the vocabulary tests, participants were asked to complete a short comprehension task and answer three easy questions per text to verify their global understanding and to give the participants a clear reading goal. The questions did not involve any knowledge of the target MWUs. This is different from the initial study, in which comprehension was not tested (see Appendix S4 for the comprehension questions). Our plan was to exclude data of participants with a score lower than 8/12 (2 out of 3 questions correct per text), as they may have not properly understood the text or may not have read the texts seriously. However, this was not done because of the small sample size. Out of 162 participants, 137 (= 85%) responded correctly to at least 10 out of 12 comprehension questions. The descriptive results of the comprehension task are reported in Appendix S6.

Questionnaire

Given that the reading materials were manipulated to contain 1, 5, 10 or 15 repetitions of 18 target MWUs, we administered a questionnaire in order to determine whether the learners had noticed the target items and also to tap into their perceptions and the ecological validity of the reading materials. In line with Godfroid et al. (2018), we ran an exploratory analysis (ANOVA or Kruskal-Wallis, depending on the distribution of the data) to verify whether participants who noticed the repetition of items performed better on the posttests than

participants who did not. We also asked questions about participants' contact with English outside of school (see Appendix S5 for the questionnaire).

Procedure

We adopted the same procedure as the initial study. The data collection procedure consisted of two sessions. In session 1, all participants took the pretest (i.e., either the form recognition or the form recall test), the VLT, and completed the consent form. One week later, participants read and listened to the experimental texts. They were told that each reading text would be followed by a comprehension task, but they were not informed about the upcoming vocabulary tests. Immediately after reading a text, participants did the comprehension task before moving on to the next text. After all texts had been read, the participants took the unannounced vocabulary posttest (corresponding to the pretest format, i.e., either the form recall or the form recognition test). The second session ended with a questionnaire tapping into learners' perception of the learning treatment. At the end of the experiment, participants were debriefed about the aim of the study.

The data were collected completely online through the experiment builder *Gorilla*. This means that the participants took the MWU tests, read the texts, and did the comprehension task online. The Polish participants completed the tasks and tests during regular contact hours, while most Flemish participant did the experiment at home on a voluntary basis. Given that an English lesson in secondary schools in Flanders is 50 minutes, the non-university sample would not have been able to take the posttest immediately after the treatment in the same English lesson if the data had been collected during their regular class time. By having participants do the experiment at home, we could ensure that the procedure was comparable for both Flemish sites.

Scoring and analyses

Pre- and posttest items were scored dichotomously. We tolerated spelling mistakes in the form recall tests. The initial study used *t*-tests and MANOVAs to analyse the data. In order to answer our research questions, we ran two multilevel logistic regressions, one for each test (form recall and form recognition), with item-level gain score as the binary outcome, and repetition, learners' prior vocabulary knowledge (VLT score), education level, L1, and pretest score as predictors. To tap into vocabulary growth, we used a binary gain score as the dependent variable (see also Vanhove, 2021). In cases where the pretest and posttest score for an item were identical (0-0 or 1-1), the gain score was 0. If the pretest score was 0 and the posttest score was 1, the gain score was 1. In cases where the pretest score was 1 and posttest score was 0, the gain score was 0. However, a potential issue was the small learning rate overall, which led to a low degree of variation in the dependent variable Gain score. As a result, we ran mixed effects models with posttest score as the dependent variable, in order to find out to what extent posttest scores varied across conditions, while controlling for total pretest score. We used the glmer function (lme4 package, version 1.1-18; Bates, Maechler, Bolker, & Walker, 2015) in R (version 3.4.3; R Core Team, 2012) (see also Siyanova-Chanturia & Omidian's (2020) plea for mixed effects modelling when researching MWUs). Models included random intercepts for participants and items, with repetition condition (four dummy variables, reference level = zero repetitions/control condition), the participants' VLT score, pretest score, education level (university, non-university, reference level = non-university), and L1 (reference level = Flemish) as predictors. We also added an interaction term between Repetition and VLT score, to account for the possibility that the strength of the relationship between repetition and learning would depend on learners' prior vocabulary knowledge (e.g., Elgort & Warren, 2014). The VLT scores and pretest scores were centered around the grand mean.

We first constructed the baseline models containing only random intercepts for items and participants, before adding the fixed effects. Finally, random slopes were added at item and participant level for the variable Repetition.

Baseline model:

Gain score ~ 1 + (1|Item) + (1|Participant)

Model including random intercepts and fixed effects:

Gain score ~ Repetition*VLT score + Education level + L1 + Pretest score + (1|Item) + (1|Participant)

Model including random intercepts, fixed effects and random slopes:

Gain score ~ Repetition*VLT score + Education level + L1 + Pretest score + (1|Item) + (1|Participant) + (1|Item : Repetition) + (1|Participant : Repetition)

The final models were evaluated by comparing them to the null models by means of a likelihood ratio test. The models were also subjected to model criticism (Baayen, 2008) and potentially harmful outliers were removed before refitting the models. We reported the *B* estimates, standard errors, *z*-values, *p*-values (significance level set at .05), and odds ratios of the fixed effects, as well as the ICC and AIC values.

Because of the reduced data set (see Participants), the planned mixed effects models did not converge. The following model converged for both test formats:

Posttest score ~ Repetition + VLT score + L1 + Education level + Pretest score + (1|Participant) + (1|Item)

The changes to the design, data collection instruments, analyses, and procedure are summarized in Table 5.

INSERT TABLE 5 APPROXIMATELY HERE

Results

The results of the pre- and posttests are presented in Tables 6 and 7. Additional descriptive results, including confidence intervals, can be found in Appendix S6. Learning gains in both the form recognition and form recall test were low. In the form recognition test, only 174 out of 1360 items (80 participants*17 items) had a Gain score of 1. In the form recall test, only 243 out of 1394 items were learned (82 participants*17 items). At recognition level, this was likely due to a ceiling effect in the pretest: 1030 out of 1360 items were known before the treatment took place. At recall level, the opposite was true: only 366 out of 1394 items were known, and very few of the unknown collocations were learned (see also Table 8).

The results of both models (see Tables 9 and 10 below) indicate that, like in the initial study, MWU were learned incidentally and that repetition had a significant effect on learners' posttest scores, with significantly higher odds of a correct response for 5, 10, and 15 exposures, compared to the 1-exposure condition.

INSERT TABLE 6 APPROXIMATELY HERE

INSERT TABLE 7 APPROXIMATELY HERE

INSERT TABLE 8 APPROXIMATELY HERE

For the form recognition test, pairwise comparisons indicated that there were significant differences in predicted posttest score between 1 exposure and 5 exposures (B =, -0.573, p = .018), between 1 exposure and 10 exposures (B = -1.022, p < .001), and between 1 exposure and 15 exposures (B = -0.798, p = .002). The effect was strongest in the 10 exposures condition: items that appeared 10 times in the input were estimated to be 2.78 times more likely to be recognized in the posttest compared to items that appeared only once. The effect

sizes (odds ratios, see Table 9) were slightly lower in the 5 and 15 exposures conditions. Please note that unlike Cohen's d, odds ratios are not interpreted in terms of small, medium, or large. However, there were no significant differences between any of the other repetition levels. Likewise, at the level of form recall, there were significant differences between 1 exposure and 5 exposures (B = -0.443, p = .016), between 1 exposure and 10 exposures (B = -0.626, p < .001), and between 1 and 15 exposures (B = -0.858, p < .001). None of the other pairwise comparisons were statistically significant. The strongest effect was found in the 15 exposures condition, where estimated odds of a correct posttest score were 2.36 times higher compared to the 1 exposures condition (see odds ratios in Table 10).

Pretest score and VLT score were also significant predictors of form recall and form recognition posttest scores. Learners' L1 (Polish or Dutch) only predicted posttest scores in the form recall test, with slightly higher odds of a correct response in the Polish sample. Education level did not significantly predict posttest scores in either test format.

INSERT TABLE 9 APPROXIMATELY HERE

INSERT TABLE 10 APPROXIMATELY HERE

Figures 1 and 2 show the predicted probabilities of a correct response in the two posttests, for each level of Repetition. Figure 1 suggests that the odds of knowing an item at the level of form recognition were at ceiling, in particular for the 5, 10, and 15 exposure levels. Figure 2 indicates that the predicted odds of knowing an item at the level of form recall were much lower (between 20 and 40%), and suggests that predicted probability of knowing an item in the posttest increased gradually with an increase in the number of repetitions.

INSERT FIGURE 1 APPROXIMATELY HERE

INSERT FIGURE 2 APPROXIMATELY HERE

Questionnaire

The questionnaire results (see Appendix S5) indicated for both sites that) more than 50% of participants strongly agreed with the statement “I noticed that some word combinations occurred several times”. Further, 125 out of 162 learners named at least one MWU that re-appeared in the text. Four out of 35 participants who commented on the reading texts mentioned that the repetition of MWUs was annoying, or as one Polish participant put it: “I'd suggest spreading the same collocations more in the text, it can get annoying when you hear "cut corners" every 3/5 sentences”. However, 13 other participants commented that they found the stories enjoyable or fun to read, which suggests that the repetition was not distracting for all learners.

To verify whether participants’ noticing of the repetition of items affected their posttest scores, we performed a Kruskal-Wallis test, which indicated that there were no significant differences in total posttest scores between learners who rated the statement “I noticed that some word combinations occurred several times” differently ($H(4) = 8.51$, $p = .07$, $\eta^2 = 0.05$). Figure 3 below presents average posttest scores for each of the response categories (1 = strongly disagree, 5 = strongly agree). Finally, the majority of participants (54%) (strongly) disagreed with the statement “I expected a vocabulary posttest”. Approximately 34% of participants (strongly) agreed with this statement."

INSERT FIGURE 3 APPROXIMATELY HERE

Table 11 summarizes the similarities and differences in findings between the initial and the replication study.

INSERT TABLE 11 APPROXIMATELY HERE

Discussion

The initial study by Webb et al. (2013) investigated the effects of repetition on learning MWUs during reading-while-listening in a between-participants design with Taiwanese university students. They compared the learning gains in five groups: no repetition of the target MWUs (control group), 1, 5, 10 , and 15 occurrences of the MWUs. Their findings showed that repetition had a large effect on the learning of MWUs, and that the learning gains tended to increase when the number of repetitions increased with a large effect size for the 15 repetitions.

The present study aimed (1) to replicate Webb et al.'s (2013) study with two new samples of university students, in Poland and in Flanders, and (2) to extend the findings to secondary school learners in Flanders. We adopted a counterbalanced within-participants design in which all participants were exposed to all conditions (1, 5, 10, and 15 occurrences of the target MWUs). Our findings showed that repetition was beneficial for the learning of MWUs. However, the learning gains were small and we did not find any differences between the 5, 10 or 15 repetitions. These findings were consistent across the two research sites (Poland and Flanders) and the two samples (university and secondary education). In addition, we found that learners' prior vocabulary knowledge, as measured by the VLT, and their pretest score predicted their posttest score.

Incidental learning of multiword units

The present study confirms Webb et al.'s findings that MWUs can be learned incidentally when EFL students read (and listen to) texts. Further, the effect was found in both the form

recognition and form recall test and this finding was replicated across two sites (Poland and Flanders). Nevertheless, the gains in the present study were small, which means that the large effect of repetition of the initial study was not corroborated.

There may be several explanations for the lower learning gains. First, we observed a ceiling effect in the form recognition pretest, so there was little room for learning in that test. Second, there were differences in design between the initial and replication study (between vs. within-participants design), which was a trade-off we made to increase methodological rigor (see Table 5). Uchihara et al. (2019) also found larger effects in between-participants than in within-participants designs. Further, as suggested by one reviewer, the participants in the initial study may have been more alerted about the vocabulary learning aim of the study because of the between-participants design, especially those in the 10 or 15-repetition group. They encountered each target item 10 or 15 times. The within-participants design in the present study may have reduced this. Unlike the initial study, the present study included a comprehension task as a way to include more true conditions of incidental vocabulary learning, with learners' focus being focused on reading for content. Yet, it should be added that this interpretation remains speculative.

The second aim of our study was to extend Webb et al.'s findings to another learner profile by examining L2 MWUs learning by EFL learners in secondary education. Our results were consistent with those of Webb et al. (2013). However, the learning gains were small and in line with those found in the university sample of our study.

The effect of repetition on learning MWUs

Webb et al. (2013) found a positive effect of repetition. They showed that the 15-repetition condition had a large and significant effect on learning MWUs, while there was no difference between 5 repetitions and 1 occurrence. Further, there was a significant difference between

10 and 5 occurrences (but only in the form recall test) and between 10 and 1 occurrence. The present study's findings partially support the initial study. Like Webb et al., we found a beneficial effect of repetition. This finding adds to the growing body of work on the effects of frequency on L2 learning (Uchihara et al., 2019). However, we could not replicate the number of encounters needed to incidentally learn MWUs, as we only found a significant difference between one and more than one (5, 10, 15) occurrence. This held true for both the Polish and Flemish university students and for both test formats, confirming the benefits of multisite research. Further, we hypothesized that fewer repetitions would be needed for form recognition than for form recall, but this hypothesis was not borne out. This shows that learning L2 MWUs is not simply a matter of increasing the number of encounters with target items, because there are many other factors likely to influence the learning process (see also Szudarski, 2017).

The second aim of our study was to extend Webb et al.'s findings to a new population. The results of the secondary education EFL learners showed that there was a repetition effect, which accords with the initial study. However, we only found a significant difference between one and more than one (5, 10, 15) occurrence. This result contrasts Webb et al.'s findings, but holds true across the Polish and Flemish university samples. Overall, the findings of all the samples in the present study seem to be more consistent with previous studies that did not find significant differences between repetitions (e.g., Pellicer-Sánchez, 2017 comparing 4 versus 8 encounters, or Szudarski and Carter, 2016 comparing 6 versus 12 encounters). Interestingly, as regards the effects of repetition, more than half of the participants in this replication indicated they had noticed that some MWUs occurred several times. However, the results of the Kruskal Wallis tests suggest that their noticing did not seem to affect their learning.

The relationship between learner variables and learning

Our third research question focused on the effect of education level, L1, and prior vocabulary knowledge. The analyses showed that learning gains were predicted by learners' prior vocabulary knowledge and their pretest score, across both the two university samples and the secondary education sample. The present study thus adds to the growing body of research that the amount of incidental vocabulary learning is affected by the number of words already known by learners (Puimège & Peters, 2020; Vilkaitė, 2017). Further, the participants with Polish as their L1 had higher odds of a correct response in the form recall test compared to the Flemish learners, which may be explained by the fact that they were English majors. Another explanation may be that they took part in the online experiment while being in class, while the Flemish participants participated online from their homes. This may have affected the level of engagement of the Flemish participants. Patterson and Nicklin (2023), who compared different data collection procedures, also found that participants in the online condition were less engaged than participants in the in-person condition. No other differences in terms of educational level (university or secondary school) or L1 were found. The fact that we did not find a difference between the university and non-university sample may be explained by their level of prior vocabulary knowledge, which was very similar regardless of educational level (university vs. secondary school). It seems that learners' prior vocabulary knowledge is a more important predictor of learning than their educational level (see also De Vos et al., 2018).

Pedagogical implications

The present study confirms previous research findings that MWUs can be learned incidentally when EFL learners engage with input and that repetition is beneficial for learning. However, the gains in the form recall test were very small. The results of the form

recall test may have more pedagogical value than those of the form recognition test, with accurate production of MWUs likely being more important to language learners than the ability to recognize such phrases. Further, it remains unclear how many repetitions are needed for successful learning of MWUs. The present study seems to suggest that five repetitions may suffice for some learning to occur, but gains are likely to be small. In light of our findings, it may well be that more than 15 repetitions are needed when MWUs are to be learned productively from mere exposure only and high gains are expected. Webb et al. proposed, that "if an approach were taken to include useful collocations in graded reading schemes, there may be little need to teach collocations explicitly for learners actively taking part in extensive reading programs" (p.111). The small learning gains in the present study do not seem to lend evidence to their pedagogical recommendation. Practically, this means that either more repetitions are needed or that incidental learning activities should be supplemented with activities explicitly targeting the learning of MWUs (for an example, Szudarski 2012), especially in foreign language learning settings, in which exposure to the foreign language may be limited. Finally, our questionnaire findings showed that great care should be taken when manipulating or flooding texts, as some learners' reactions toward the flooding intervention were negative. Horst et al. (1998) already warned of a potential trade-off effect of text manipulation in terms of learners' motivation. Future research could address the suitability of the learning materials more consistently, for example in relation to topics covered, the length of experimental texts, and the number of occurrences of target items, to increase the ecological validity of research findings.

Limitations and conclusion

Even though we aimed to improve the design of the initial study, this replication also has a number of limitations. The first limitation is the sample size of the Flemish university

students and Flemish secondary school learners; because of participant attrition, our sample size was smaller than originally projected. Patterson and Nicklin (2023) argue that in-person data collection may be characterized by logistical constraints, which is why we moved to online data collections. However, fully online studies, in which participants can do the experiment at home at a time that suits them, may be less appropriate when the study consists of more than one session given the large data loss after the first session. A second limitation of the present replication is that it is not straightforward to directly compare the findings of the initial and replication study given the differences in the designs (see Table 5). However, the methodological changes were introduced to improve the original methodology, to control for more variables (e.g., time-on-task) and to benefit from a multi-site approach.

In spite of these limitations, Webb et al.'s effect of repetition could be replicated in the present study, even though the effect was smaller and the number of repetitions needed was different. Given that many other factors may play a role (e.g., type of MWU, learner characteristics, text context), we argue that the aim of finding the exact number of repetitions necessary for learning MWUs may not be a fruitful research direction. Further, the findings of our study are robust across two settings, two L1s, and two educational levels, showing the benefits of multisite research for increasing the generalizability of L2 findings. Finally, we found that participants' prior vocabulary knowledge was an important predictor of the number of MWUs learned.

Endnotes

1.A delayed posttest was included in the design of Webb et al. (2013), but the results were not reported.

2. Peters et al. (2019) did not use the VLT, but unpublished VLT results corroborate these findings (*n*=37 in grade 11; 26.97/30 on the 2,000 level).

3. We ran an additional exploratory analysis (see S7) taking into account congruency and transparency (see S6 for a discussion of how congruency and transparency were determined). Both congruency and transparency have been found to affect the learning of MWUs (Puimège & Peters, 2020). However, because our main focus was not on item-related variables, we included this secondary analysis in the supplementary online materials (see S7) and not in the main text (see also Vanhove, 2021).

4. The following two test formats from the initial study were not used in our replication study: the productive form test and the meaning recall test (labelled *receptive form-meaning test* in the initial study). The productive form test could be considered a memory test because participants had to provide collocates (nouns) for the 17 nodes (verbs) without any cue. Given that language learners struggle more with productive knowledge of MWUs than receptive knowledge (e.g., Laufer & Waldman, 2011), we decided not to focus on meaning knowledge. A second reason not to use the meaning recall test is to avoid a potential test learning effect within the pretest session as well as from the pretest to the treatment (see Puimège & Peters, 2020). Thirdly, it needs to be remembered that the control group in the initial study, who were not exposed to the target items in the text, obtained high scores on this test, which might indicate that participants could guess the meaning from the constituent words.

References

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in Applied Linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033

Baayen, R. (2008). *Analyzing linguistic data. A practical introduction to statistics using R.* Cambridge, UK: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Boers, F. (2020). Factors affecting the learning of multiword items. In S. Webb (Ed.), The Routledge handbook of vocabulary studies (pp.143-157). New York: Routledge.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, *3*, 245–261.

Chen, Y. (2021). Comparing incidental vocabulary learning from reading-only and reading-while-listening. *System, 97*, 102442. https://doi.org/10.1016/j.system.2020.102442

Cobb, T. (2003). Analyzing Late Interlanguage with Learner Corpora: Québec Replications of Three European Studies. *Canadian Modern Language Review/ La Revue Canadienne Des Langues Vivantes*, *59*(3), 393–424. https://doi.org/10.3138/cmlr.59.3.393

Cobb, T. (2019). From corpus to CALL: The use of technology in teaching and learning formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language. A second language acquisition perspective* (pp.192-210). Routledge.

Columbus, G. (2013). In support of multiword unit classifications: Corpus and human rating data validate phraseological classifications of three different multiword unit types. *Yearbook of Phraseology*, *4*(1), 23-43. https://doi.org/10.1515/phras-2013-0003

Conklin, K., Alotaibi, S., Pellicer-Sánchez, A., & Vilkaitė-Lozdienė, L. (2020). What eye-tracking tells us about reading-only and reading-while-listening in a first and second language. *Second Language Research,* 36(3) 257-276. https://doi.org/10.1177/0267658320921496

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. Retrieved from http://onlinelibrary.wiley.com/doi/10.2307/3587951/abstract

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*(5), 570–590. https://doi.org/10.1093/applin/amt056

Crossley, S. A., & Skalicky, S. (2017). Examining lexical development in second language learners: An approximate replication of Salsbury, Crossley & McNamara (2011). *Language Teaching*, 1–21. https://doi.org/10.1017/S0261444817000362

de Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input. *Language Learning*, *68*(4), 906–941. https://doi.org/10.1111/lang.12296

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2020). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition*, *23*,171-185.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*(2), 163–188. https://doi.org/10.1177/0267658309349431

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194. https://doi.org/10.1093/applin/aml015

Ellis, N. C. (2012). Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics*, *32*, 17–44. https://doi.org/10.1017/S0267190512000025

Godfroid, A. (2020). *Eye tracking in second language acquisition and bilingualism. A research synthesis and methodological guide*. New York: Routledge.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*(2), 207–223.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. Doughty & M. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 349–381). Malden, MA: Blackwell.

Kremmel, B., Brunfaut, T., & Alderson, J. C. (2017). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics*, *38*(6), 848–870. https://doi.org/10.1093/applin/amv070

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, *29*(4), 694–716. https://doi.org/10.1093/applin/amn018

Laufer, B., & Waldman, T. (2011). Verb-Noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lin, P. (2019). Formulaic language and speech prosody. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language. A second language acquisition perspective* (pp.78-94). Routledge.

Macis, M. (2018). Incidental Learning of Duplex Collocations from Reading: Three Case Studies. *Reading in a Foreign Language*, *30*(1), 48–75.

Malone, J. (2018). Incidental vocabulary learning in SLA. Effects of frequency, aural enhancement, and working memory. *Studies in Second Language Acquisition*, *40*(3), 651–675. https://doi.org/10.1017/S0272263117000341

Majuddin, E., Siyanova-Chanturia, A., & Boers, F. (2021). Incidental acquisition of multiword expressions through audiovisual materials. *Studies in Second Language Acquisition*, 1–24. https://doi.org/10.1017/S0272263121000036

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in

Second Language Research: Narrative and Systematic Reviews and Recommendations for the Field. *Language Learning*, *68*(2), 321–391. https://doi.org/10.1111/lang.12286

Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, *45*(2), 267–290. https://doi.org/10.5054/tq.2011.247708

Nation, P. (1983). Testing and teaching vocabulary. *Guidelines, 5*, 12–25.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3–13). Amsterdam: John Benjamins.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, *24*(2), 223–242. https://doi.org/10.1093/applin/24.2.223

Nicklin, C., & Vitta, J. P. (2021). Effect-driven sample sizes in second language instructed vocabulary acquisition research. *The Modern Language Journal*. https://doi.org/10.1111/modl.12692

Noreillie, A., Kestemont, B., Heylen, K., Desmet, P., & Peters, E. (2018). Vocabulary knowledge and listening comprehension at an intermediate level in English and French as foreign languages An approximate replication study of Stæhr (2009). *ITL - International Journal of Applied Linguistics*, *169*(1), 212–231. https://doi.org/https://doi.org/10.1075/itl.00013.nor

Northbrook, J., & Conklin, K. (2019). Is what you put in what you get out ? — Textbook-derived lexical bundle processing in beginner English learners. *Applied Linguistics*, *40*(3), 816–833. https://doi.org/10.1093/applin/amy027

Otwinowska, A., Foryś-Nogala, M., Kobosko, W., & Szewczyk, J. (2020). Learning orthographic cognates and non-cognates in the classroom: Does awareness of cross-

linguistic similarity matter? *Language Learning*, 1–47. https://doi.org/10.1111/lang.12390

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Patterson, A. S., & Nicklin, C. (2023). L2 self-paced reading data collection across three contexts: In-person, online, and crowdsourcing. *Research Methods in Applied Linguistics*, *2*(1), 100045. https://doi.org/10.1016/j.rmal.2023.100045

Pavakanun, U., & d'Ydewalle, G. (1992). Watching foreign television programs and language learning. In F. L. Engel, D. G. Bouwhuis, T. Bosser, & G. d'Ydewalle (Eds.), Cognitive modelling and interactive environments in language learning (pp. 193–198). Berlin: Springer.

Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learing through listening to songs. *Studies in Second Language Acquisition*, *41*(4), 745–768. https://doi.org/10.1017/S0272263119000020

Pellicer-Sánchez, A. (2017). Learning L2 collocations incidentally from reading. *Language Teaching Research*, *21*(3), 381–402. https://doi.org/10.1177/1362168815618428

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: do Things Fall Apart ? *Reading in a Foreign Language*, *22*(1), 31–55.

Peters, E. (2018). The effect of out-of-class exposure to English language media on learners ' vocabulary knowledge. *ITL - International Journal of Applied Linguistics*, *169*(1), 142–168. https://doi.org/https://doi.org/10.1075/itl.00010.pet

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audio-visual input. *TESOL Quarterly*, *53*(4), 1008–1032. https://doi.org/10.1002/tesq.531

Peters, E., Noreillie, A.-S., Heylen, K., Bulté, B., & Desmet, P. (2019). The impact of instruction and out-of-school exposure to foreign language input on learners' vocabulary knowledge in two languages. *Language Learning*, 1–36. https://doi.org/10.1111/lang.12351

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, *40*(3), 551–577. https://doi.org/10.1017/S0272263117000407

Puimège, E., & Peters, E. (2019a). Learners ' English vocabulary knowledge prior to formal instruction: The role of learner-related and word-related variables. *Language Learning*, 1–35. https://doi.org/10.1111/lang.12364

Puimège, E., & Peters, E. (2019b). Learning L2 vocabulary from audiovisual input: an exploratory study into incidental learning of single words and formulaic sequences. *The Language Learning Journal*, *0*(4), 1–15. https://doi.org/10.1080/09571736.2019.1638630

Puimège, E., & Peters, E. (2020). Learning formulaic sequences through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, *42*(3), 525–549. https://doi.org/10.1017/S027226311900055X

Puimège, E., Montero Perez, M., & Peters, E. (2023). Promoting L2 acquisition of multiword units through textually enhanced audiovisual input: An eye-tracking study. *Second Language Research*, *39*(2), 471–492. https://doi.org/10.1177/02676583211049741

Pujadas, G., & Muñoz, C. (2019). Extensive viewing of captioned and subtitled TV series: a study of L2 vocabulary learning by adolescents. *The Language Learning Journal*, *47*(4), 479–496. https://doi.org/10.1080/09571736.2019.1616806

Pujadas, G., & Muñoz, C. (2020). Examining adolescent EFL learners' tv viewing comprehension through captions and subtitles. *Studies in Second Language Acquisition*, *42*, 551–575. https://doi.org/10.1017/S0272263120000042

Serrano, R., & Huang, H. Y. (2018). Learning vocabulary through assisted repeated reading: How much time should there be between repetitions of the same text? *TESOL Quarterly*, *52*(4), 971–994. https://doi.org/10.1002/tesq.445

Siyanova-Chanturia, A., & Omidian, T. (2020). Key issues in researching multiword items. In S. Webb (Ed.), The Routledge handbook of vocabulary studies (pp.529-544). New York: Routledge.

Siyanova-Chanturia, A., & Pellicer-Sánchez, A., (2019). Formulaic language. Setting the scene. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language. A second language acquisition perspective* (pp.1-15). Routledge.

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: acquisition of collocations under different input conditions. *Language Learning*, *63*(1), 121–159. https://doi.org/10.1111/j.1467-9922.2012.00730.x

Sundqvist, P. (2019). Commercial-off-the-shelf games in the digital wild and L2 learner vocabulary. *Language Learning & Technology*, *23*(1), 87–113.

Szudarski, P. (2012) Effects of meaning- and form-focused instruction on the acquisition of verb-noun collocations in L2 English. *Journal of Second Language Teaching and Research*, *1* (2), 3–37.

Szudarski, P. (2017). Learning and teaching L2 collocations: Insights from research. *TESL Canada Journal*, *34*(3), 205–216. https://doi.org/10.18806/tesl.v34i3.1280

Szudarski, P. (2019). Using a mixed-methods approach to examine the lexical and collocational development of EFL learners. Paper presented at Vocab at Leuven. KU Leuven Belgium, 1-3.07.2019.

Szudarski, P. (2020). Effects of data-driven learning on enhancing the phraseological knowledge of secondary-school learners of L2 English. In P. Crosthwaite (Ed.) *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge., pp 133-149

Szudarski, P., & Carter, R. (2016). The role of input flood and input enhancement in EFL learners' acquisition of collocations. *International Journal of Applied Linguistics (United Kingdom)*, *26*(2), 245–265. https://doi.org/10.1111/ijal.12092

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*(2), 506–547. https://doi.org/10.1111/lang.12384

Toomer, M., & Elgort, I. (2019). The development of implicit and explicit knowledge of collocations: A conceptual replication and extension of Sonbul and Schmitt (2013). *Language Learning*, *69*(2), 405–439. https://doi.org/10.1111/lang.12335

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559–599. https://doi.org/10.1111/lang.12343

Vanhove, J. (2021).Towards simpler and more transparent quantitative research reports. *ITL-International Journal of Applied Linguistics*, *172* (1), 3-25. https://doi.org/10.1075/itl.20010.van

Venables, W. N., & Ripley, B. D. (2003). *Modern applied statistics with S-Plus* (4th ed.). New York, NY: Springer.

Vilkaitė, L. (2017). Incidental acquisition of collocations in L2. Effects of adjacency and prior vocabulary knowledge. *ITL - International Journal of Applied Linguistics*, *168*(2), 248–277. https://doi.org/10.1075/itl.17005.vil

Webb, S. (2020). Incidental vocabulary learning. In S. Webb (Ed.), The Routledge handbook of vocabulary studies (pp.225-239). New York: Routledge.

Webb, S., & Chang, A. C. S. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, *19*(6), 667–686. https://doi.org/10.1177/1362168814559800

Webb, S., & Chang, A.C.S. (2022). How does mode of input affect the incidental learning of collocations? *Studies in Second Language Acquisition*, 44 (1), 35-56. doi:10.1017/S0272263120000297

Webb, S., Newton, J., & Chang, A. C. S. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91–120. https://doi.org/10.1111/j.1467-9922.2012.00729.x

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. **Target items and reading materials**.

Appendix S2. **Distracter items**.

Appendix S3. **Tests**.

Appendix S4. **Comprehension task**.

Appendix S5. **Questionnaire**.

Appendix S6. **Descriptive results.**

Appendix S7. **Secondary analysis with congruency and transparency**.

Tables

Table 1

	1 repetition	5 repetitions	10 repetitions	15 repetitions
Version 1	MWU 1-5	MWU 6-10	MWU 11-14	MWU 15-18
Version 2	MWU 15-18	MWU 1-5	MWU 6-10	MWU 11-14
Version 3	MWU 11-14	MWU 15-18	MWU 1-5	MWU 6-10
Version 4	MWU 6-10	MWU 11-14	MWU 15-18	MWU 1-5

Table 2. Sample size per site and education level for each test format.

	Form recognition		**Form recall**	
	Secondary education	*University*	*Secondary education*	*University*
Polish site	NA	43	NA	46
Flemish site	23	14	15	21

Table 3. Average scores and standard deviations (in parentheses) for each level of the VLT per sample.

	2K (Max=30)	**3K (Max=30)**	**5K (Max=30)**	**10K (Max=30)**	**Academic (Max=30)**	**Total (Max=150)**
L1 = Polish (n = 89)	28.82 (1.25)	27.65 (2.69)	25.04 (4.03)	15.25 (7.17)	25.22 (6.29)	122 (17.28)
L1 = Dutch (high school, n = 38)	28.92 (1.24)	27.42 (2.65)	23.05 (4.67)	13.87 (6.60)	25.76 (2.98)	119.0 (15.61)
L1 = Dutch (university, n = 35)	29.06 (1.08)	27.26 (3.27)	23.51 (5.20)	13.97 (7.40)	26.06 (4.78)	119.9 (19.84)

Table 4

Target MWUs

MWUs	**Dutch translation**	**Polish translation**	**COCA frequency**	**t-score**	**MI score**
blow nose	neus snuiten	wydmuchać nos	662	10.08	5.18
break silence	stilte verbreken	przerwać ciszę	1276	13.36	4.40
buy time	tijd winnen	zyskać na czasie/grać na zwłokę	1864	5.44	0.32
cut corner	de kantjes ervan aflopen	iść na skróty	637	15.55	3.24
face fact	feiten onder ogen zien	stawić czoło faktom	622	12.7	0.57
grant wish	wens vervullen	spełnić życzenie	274	5.55	3.14
lose touch	contact verliezen	tracić kontakt	877	13.48	2.53
make mind	beslissing nemen	zdecydować się	3794	26.74	0.91
meet demand	aan eis voldoen	zaspokoić popyt	2050	28.39	3.67
pull string	invloed gebruiken	pociągać za sznurki	657	11.86	5.32
raise question	vragen stellen	zadać pytanie/ poruszyć kwestię	9272	26.04	4.45
reach decision	beslissing bereiken	podjąć decyzję	713	12.42	2.05

read thought	gedachten lezen	odczytać myśli	285	8.38	1.27
remember time	terugdenken aan	pamiętać moment/czas kiedy	3778	22.24	1.04
run risk	risico lopen	ponosić ryzyko	1696	26.62	2.26
spread word	verder vertellen	rozprzestrzenić informacje	1403	24.25	4.18
throw light	licht werpen	rzucić światło	361	15.54	1.05

Table 5

List of changes made to the initial study

	Identical	**Change**
Design	No	Within-participants design instead of between-participants design; participants split into two groups, each taking one test format only, that is, form recall or form recognition. The study, thus, consisted of two data sets, one focusing on form recall and one focusing on form recognition.
Reading materials	No	The same graded readers were used, but because of the within-participants design, changes were made to the texts.
Target items	No	17 instead of 18 items
Test of prior vocabulary knowledge (VLT)	Yes	/
Pretest	No	Each knowledge aspect was pretested. Participants took either a form recognition pretest and posttest, or a form recall pretest and posttest.
Posttests	No	Two instead of four posttests, only form recall and form recognition.

Procedure	No	Two sessions instead of three sessions because no delayed posttests were included. Two tasks added to the second session: a comprehension task and a questionnaire.
Analyses	No	A multilevel logistic regression analysis to control for learner-related variables. In a secondary analysis, also item-related variables were taken into account (see Appendix S7).

Table 6. Average scores and standard deviations (in parentheses) for the form recognition test.						
	Polish university sample		*Flemish high school sample*		*Flemish university sample*	
	Pretest	**Posttest**	**Pretest**	**Posttest**	**Pretest**	**Posttest**
0 exposures (max. = 10)	7.21 (1.23)	7.09 (1.16)	6.57 (1.38)	6.44 (0.97)	6.93 (1.23)	6.36 (1.45)
1 exposure (max. = 5)	3.55 (1.14)	3.58 (1.10)	3.45 (1.26)	3.28 (1.38)	3.19 (0.87)	3.10 (0.91)
5 exposures (max.= 5)	3.26 (1.08)	3.81 (0.87)	2.75 (1.23)	3.29 (1.06)	3.30 (1.10)	3.59 (1.31)
10 exposures (max.= 5)	3.28 (1.26)	3.97 (0.90)	3.18 (1.25)	3.51 (1.00)	3.79 (0.89)	4.21 (0.75)
15 exposures (max.= 5)	3.53 (1.32)	4.02 (0.89)	2.84 (0.95)	3.41 (1.21)	2.97 (1.19)	3.39 (0.75)
All target items (max.= 17)	13.33 (2.41)	15.14 (1.81)	11.32 (2.91)	12.71 (2.65)	12.44 (2.68)	13.38 (2.87)
Control items (max.= 9)	7.21 (1.25)	7.09 (1.17)	6.14 (1.65)	6.12 (1.29)	6.50 (1.67)	6.19 (1.52)
Total (max. = 26)	20.53 (3.15)	22.23 (2.71)	17.46 (4.05)	18.82 (3.71)	18.94 (4.22)	19.56 (4.10)

Table 7. Average scores and standard deviations (in parentheses) for the form recall test.						
	Polish university sample		*Flemish high school sample*		*Flemish university sample*	
	Pretest	**Posttest**	**Pretest**	**Posttest**	**Pretest**	**Posttest**
0 exposures (max. = 10)	3.54 (1.69)	3.30 (1.43)	3.2 (1.38)	3.27 (1.78)	3.14 (1.76)	3.57 (1.71)
1 exposure (max. = 5)	1.26 (0.96)	1.56 (1.00)	0.77 (1.03)	1.00 (1.19)	0.98 (1.05)	1.19 (1.29)
5 exposures (max. = 5)	1.57 (0.97)	1.89 (1.05)	1.02 (0.96)	1.71 (1.28)	1.36 (1.14)	1.55 (1.13)
10 exposures (max. = 5)	1.32 (1.04)	2.14 (1.25)	0.66 (0.94)	1.27 (1.18)	0.76 (0.97)	1.22 (0.97)
15 exposures (max. = 5)	1.06 (0.98)	2.07 (1.22)	1.09 (0.99)	1.99 (1.33)	1.27 (1.10)	2.10 (1.25)
All target items (max. = 17)	5.02 (1.80)	7.35 (2.29)	2.89 (1.53)	5.44 (2.43)	3.79 (1.84)	5.33 (2.16)
Control items (max. = 9)	3.46 (1.73)	3.25 (1.45)	3.11 (1.41)	3.22 (1.70)	2.88 (1.85)	3.33 (1.79)
Total (max. = 26)	8.48 (3.09)	10.6 (3.17)	6.00 (2.59)	8.66 (3.58)	6.67 (2.85)	8.67 (3.42)

Table 8. Scores for target and control items in form recall test						
	Pretest			Posttest		
	Target (max = 17)	Control (max = 9)	Total (max = 26)	Target (max = 17)	Control (max = 9)	Total (max = 26)
Polish university	5.02 (1.80)	3.46 (1.73)	8.48 (3.09)	7.35 (2.29)	3.25 (1.45)	10.6 (3.17)
Flemish university	3.79 (1.84)	2.88 (1.85)	6.67 (2.85)	5.33 (2.16)	3.33 (1.79)	8.67 (3.42)
Flemish high school	2.89 (1.53)	3.11 (1.41)	6.00 (2.59)	5.44 (2.43)	3.22 (1.70)	8.66 (3.58)

Table 9. Best-fitting models for posttest scores: recognition

Predictors	*Odds Ratios*	*std. Error*	*CI*	*p*
(Intercept)	0.00	0.00	0.00 – 0.02	**<0.001**
Repetition = 5	1.77	0.43	1.10 – 2.85	**0.018**
Repetition = 10	2.78	0.72	1.67 – 4.63	**<0.001**
Repetition = 15	2.22	0.56	1.35 – 3.65	**0.002**
VLT score	1.03	0.01	1.01 – 1.04	**<0.001**
Education level	1.13	0.30	0.67 – 1.90	0.659
L1	1.46	0.38	0.88 – 2.44	0.144
Pretest score	1.24	0.05	1.15 – 1.33	**<0.001**
Random Effects				
σ^2	3.29			
$\tau_{00\ Participant}$	0.03			
$\tau_{00\ Item}$	2.55			
ICC	0.44			
$N_{Participant}$	80			
N_{Item}	17			
Observations	1360			
Marginal R^2 / Conditional R^2	0.195 / 0.549			

Note. The reference level of Repetition is 1 (1 exposure), the reference level of L1 is Dutch, and the reference level of Education level is Secondary school.

Table 10. Best-fitting models for posttest scores: recall

Predictors	*Odds Ratios*	*std. Error*	*CI*	*p*
(Intercept)	0.04	0.03	0.01 – 0.14	**<0.001**
Repetition = 5	1.56	0.29	1.09 – 2.23	**0.016**
Repetition = 10	1.87	0.35	1.30 – 2.68	**0.001**
Repetition = 15	2.36	0.43	1.65 – 3.37	**<0.001**
VLT score	1.01	0.00	1.00 – 1.02	**0.022**
Education level	1.03	0.22	0.67 – 1.58	0.896
L1	1.52	0.26	1.09 – 2.11	**0.013**
Pretest score	1.08	0.03	1.03 – 1.14	**0.003**
Random Effects				
σ^2	3.29			
$\tau_{00\ Participant}$	0.05			
$\tau_{00\ Item}$	1.19			
ICC	0.27			
$N_{Participant}$	82			
N_{Item}	17			
Observations	1394			
Marginal R^2 / Conditional R^2	0.060 / 0.318			

Note. The reference level of Repetition is 1 (1 exposure), the reference level of L1 is Dutch, and the reference level of Education level is Secondary school.

Table 11

Comparison of results between the initial and replication study

	Webb et al.		Replication	
	Form recognition	Form recall	Form recognition	Form recall
Effect of input	☑	☑	☑	☑
Effect of repetition	☑	☑	☑	☑
1 vs. 5	☒	☒	☑	☑
1 vs. 10	☑	☑	☑	☑
1 vs. 15	☑	☑	☑	☑
5 vs. 10	☒	☑	☒	☒
5 vs. 15	☑	☑	☒	☒
10 vs. 15	☑	☑	☒	☒

Figures

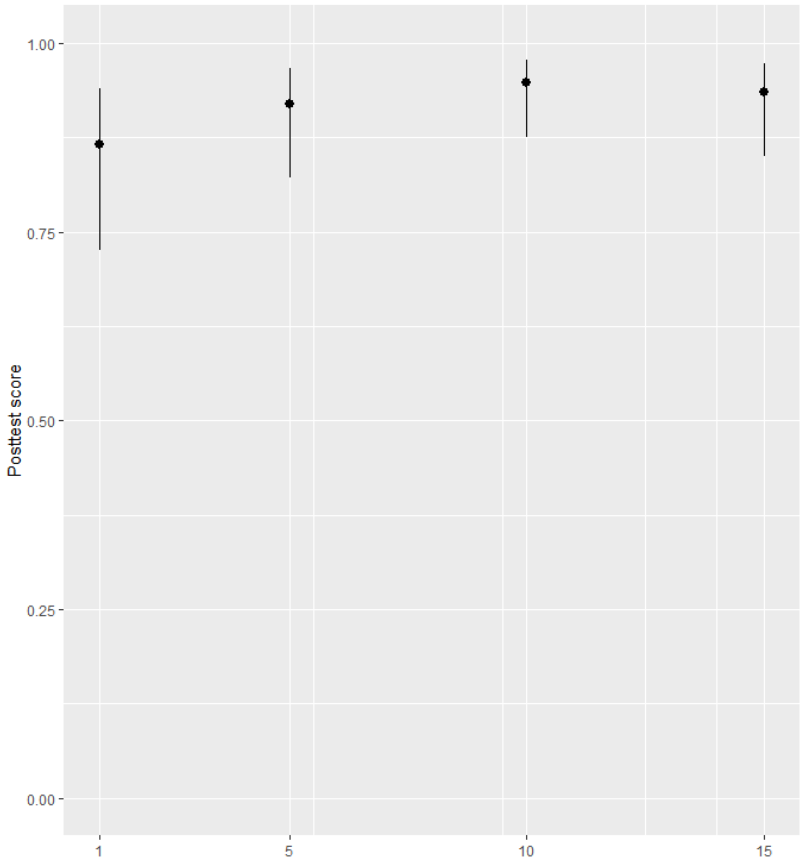

Figure 1. Predicted probabilities of knowing an item in the recognition posttest, for each level of the variable Repetition (1, 5, 10, and 15 exposures).
Note: bars indicate 95% confidence intervals

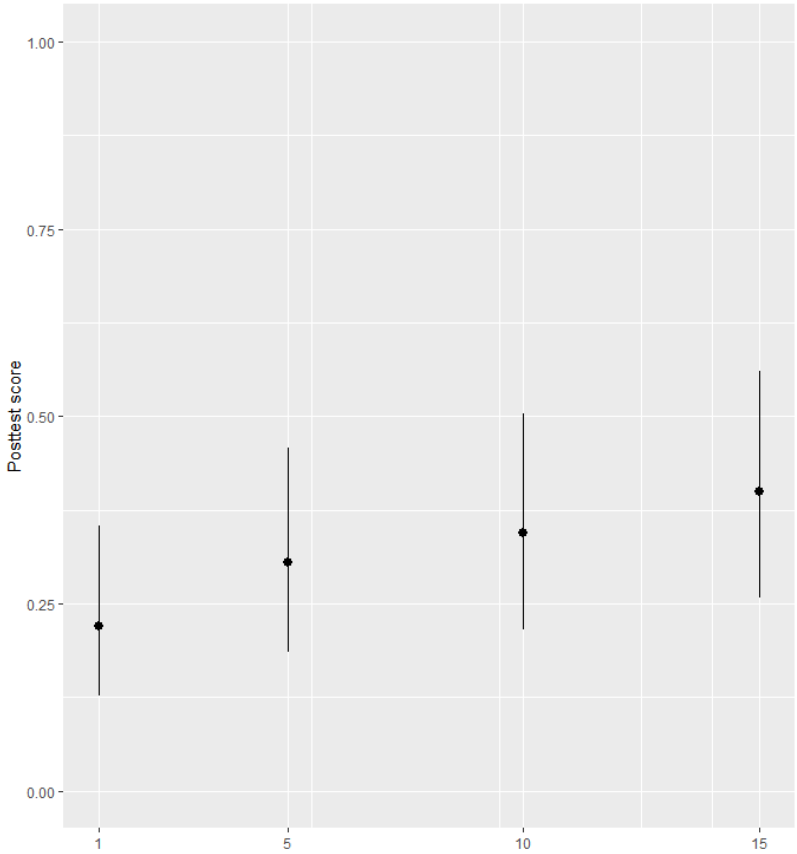

Figure 2. Predicted probabilities of knowing an item in the recall posttest, for each level of the variable Repetition (1, 5, 10, and 15 exposures).
Note: bars indicate 95% confidence intervals

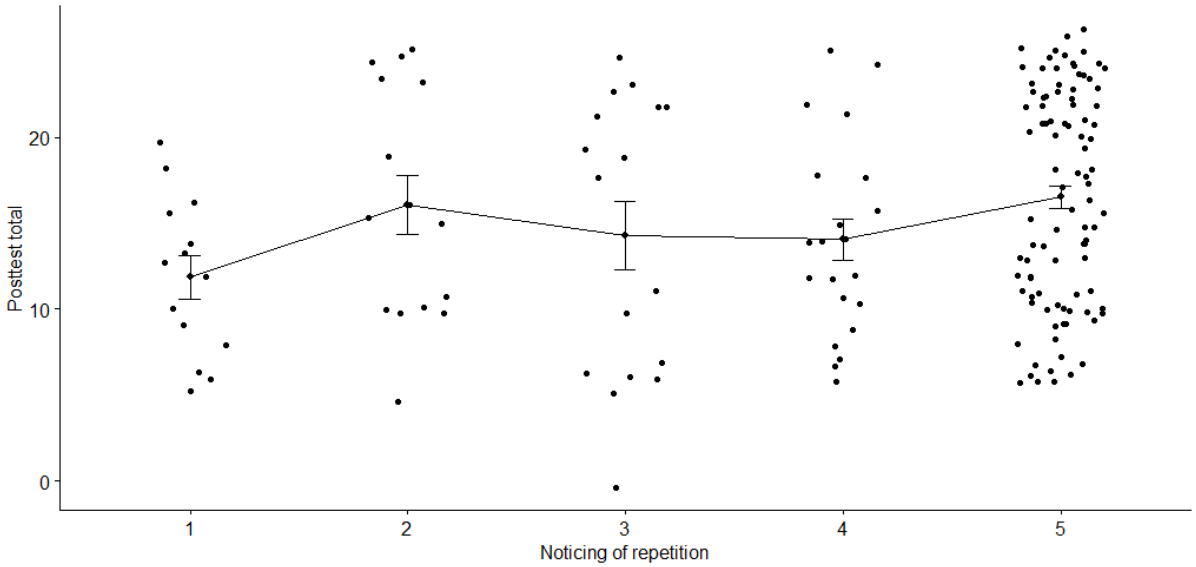

Figure 3. Average total posttest score per response category in the Noticing of Repetition question (1 = strongly disagree, 5 = strongly agree).