Studies in Second Language Acquisition, **43** (2021), 1009–1039
doi:10.1017/S0272263120000431

Research Article

COMBINING EXPLICIT AND SENSITIVE INDICES FOR MEASURING L2 VOCABULARY LEARNING THROUGH CONTEXTUALIZED INPUT AND WORD-FOCUSED INSTRUCTION

Bert Vandenberghe *

KU Leuven

Maribel Montero Perez

Ghent University

Bert Reynvoet

KU Leuven

Piet Desmet

KU Leuven

Abstract

This study combines explicit (pen-and-paper) and sensitive (time-pressured) measures to gauge the impact of three instructional interventions (contextualized input with meaning-focused activities, contextualized input with word-focused activities, and decontextualized input with word-focused exercises) on the learning of 20 L2 French target verbs. Participants (N = 313, L1 = Dutch) completed a combination of explicit (form recognition, meaning recall, grammatical preference) and time-pressured sensitive tests (lexical decision, semantic relatedness, grammaticality judgment) as immediate and delayed posttests. Explicit posttests show the beneficial effects of word-focused instruction, and underline the efficiency of context for meaning-related knowledge. Sensitive posttests generally confirm the explicit results, but reveal differences between both word-focused conditions related to lexical processing and strength of knowledge. This study suggests that combining explicit and sensitive measures can provide a more complete picture about the effects of L2 vocabulary instruction and shows that contextualized and decontextualized word-focused instruction benefit vocabulary learning in a complementary way.

This work was funded by the Research Foundation – Flanders (FWO), grant G064116N.

* Correspondence concerning this article should be addressed to Bert Vandenberghe (itec, imec research group at KU Leuven), Etienne Sabbelaan 51, 8500 Kortrijk, Belgium. E-mail: b.vandenberghe@kuleuven.be.

INTRODUCTION

Mainstream language teaching methodology holds that second language (L2) instruction should result in learners capable of using language in real-life contexts (Loewen, 2015). Functional language use is key in dominant meaning-oriented pedagogical approaches like task-based language teaching (Erlam, 2016), and becoming an independent L2 user is promoted by the Common European Framework of Reference (CEFR) for languages (Council of Europe, 2001). Yet, previous L2 vocabulary research has shown that real-life language use requires solid lexical knowledge in all skill areas (Webb & Nation, 2017). Efficient L2 instruction and teaching should therefore incorporate a strong vocabulary component.

Solid vocabulary knowledge does not only include aspects related to vocabulary size (i.e., the number of words contained in the mental lexicon) and the strength of representation of different word knowledge components (pronunciation, orthography, meaning, etc.). It also refers to the ability to access the mental lexicon quickly and accurately (Pellicer-Sánchez, 2015). Yet, most previous research is exclusively based on explicit offline measures that provide ample opportunities for conscious retrieval of vocabulary knowledge, but lack the sensitivity to track speed of lexical access (Godfroid, 2020). Therefore, it has been argued that sensitive time-pressured measures may provide a deeper understanding of L2 vocabulary knowledge.

The aim of the present study is to investigate the value of combining explicit (pen-and-paper) and sensitive (reaction time) measures to assess the impact of L2 vocabulary instruction on the learning of 20 L2 French target verbs. Three prototypical instructional treatments in studies of vocabulary learning from written input will be compared: (a) contextualized input with meaning-focused activities only, (b) contextualized input with both meaning- and word-focused activities, and (c) decontextualized input with word-focused exercises only. To take into account the multidimensional nature of word knowledge, we will focus on form-, meaning-, and use-related word knowledge aspects.

MEASURING WORD KNOWLEDGE

WHAT IS INVOLVED IN WORD KNOWLEDGE?

Word knowledge is a multifaceted construct that can be broken down into partial knowledge aspects that relate to a word's form (e.g., spelling), meaning (e.g., word associates), and use (e.g., grammatical functions) (Nation, 2013). Hence, obtaining a clear picture of word knowledge requires the measurement of knowledge gains in a variety of aspects. Most studies, however, equated word knowledge with knowledge about the form-meaning link (Nation & Webb, 2011) and did not take into account the variety of aspects that word knowledge entails. An exception (for an overview, see González-Fernandéz & Schmitt, 2019) is a study by Webb (2007) in which the effects of word learning from either sentence contexts or paired associate learning were investigated. This study addressed 10 aspects of word knowledge (active and passive knowledge of orthography, syntax, association, grammar, and meaning) and showed that learning from sentence contexts did not lead to more use-related learning gains than paired-associate learning. González-Fernández and Schmitt (2019) examined the interrelatedness of recognition and recall of four word components (form-meaning

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

link, derivatives, multiple meanings, and collocations) and investigated whether the four components referred to one or multiple constructs. Results indicated that all components were strongly intercorrelated and that knowledge development followed an implicational scale showing, for instance, that recognition knowledge develops before recall knowledge for all components tested.

We can conclude that studies that have adopted a multiple components approach have provided valuable insights. Hence, more multiple component studies are needed to better understand the effects of L2 vocabulary instruction. Moreover, it has been argued that a better understanding of the complex construct of L2 word knowledge would also entail pedagogical value. Multicomponent studies could indeed point to the aspects of L2 vocabulary learning that deserve teaching priority (Schmitt, 2019). Therefore, the present study focuses on the three components involved in Nation's L2 vocabulary knowledge framework, that is, form, meaning, and use.

EXPLICIT AND SENSITIVE MEASURES

Word knowledge develops incrementally throughout its different components as a result of recurrent exposures to lexical items (Schmitt, 2010). Concomitantly, cumulative speed gains in lexical processing are achieved (Frishkoff et al., 2011). Yet, measures that have the sensitivity to track speed of lexical processing have only scarcely been used in L2 vocabulary research (Pellicer-Sánchez, 2015). Indeed, the majority of L2 vocabulary studies, including the previously cited multicomponent studies, used pen-and-paper tests such as multiple-choice recognition tests or translation tests, that allowed for conscious thinking and attentional control (Godfroid, 2020).

The speed at which lexical items can be processed is considered to be an indicator of strength of lexical knowledge and an essential component of fluent language use (Godfroid, 2020; Pellicer-Sánchez, 2015). Therefore, sensitive measures are said to better represent the type of knowledge needed for fluent language use than traditional pen-and-paper measures (Elgort, 2018). Godfroid (2020) uses the term *sensitive measures* to refer to methods that have the sensitivity to track lexical processing speed by reaction time (RT) methodologies, whereby the element of time restriction is said to reduce the opportunities for conscious retrieval and attentional control (Elgort, 2018; Pellicer-Sánchez, 2015). The knowledge that sensitive measures may tap into can be of different nature. On the one hand, it can be readily available speeded-up knowledge (Pellicer-Sánchez, 2015) that is explicit in nature, that is, knowledge that learners are aware of and can retrieve consciously from memory (Loewen, 2015, p. 20). If this type of knowledge has undergone a qualitative change, in that it has become fast, effortless, reliable, and invariable, it is called automatized explicit knowledge (Godfroid, 2020). On the other hand, sensitive measures can also tap into tacit/implicit knowledge (Elgort, 2011; Sonbul & Schmitt, 2013), that is, knowledge that is not available to learners' conscious control or report (Elgort, 2018). In this article, we will use *explicit measures* to refer to traditional pen-and-paper measures that allow for conscious thinking and that tap into explicit knowledge. The term *sensitive measures* will be used for measures that intend to reduce the opportunities for attentional control by using an element of time restriction, and can tap into either speeded-up explicit, automatized explicit, or implicit/tacit knowledge.

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

Elgort (2011) investigated the effects of deliberate decontextualized learning of pseudowords through word lists and flashcards on the development of tacit lexical knowledge by using a priming paradigm. Priming techniques assume that words are recognized more quickly (i.e., a facilitation effect) when preceded by orthographically related (i.e., form priming, e.g., *junction-function*) or semantically related (i.e., semantic priming, e.g., *microwave-toaster*) words. The priming tasks showed that the treatment induced a facilitation effect for both formal and semantic priming. Hence, it was concluded that tacit lexical knowledge could emerge from decontextualized deliberate word learning. Accordingly, Elgort et al. (2018) studied the effect of a form- (i.e., word writing) and meaning-focused (i.e., deriving word meaning from context) treatment on the deliberate learning of low-frequency words and nonwords in two parallel experiments. In a speeded lexical decision task (LDT), participants were asked to decide as accurately and as quickly as possible whether a letter string was either a real L2 word or a nonword. Results showed that word-focused instruction not only was amenable to higher accuracy but also facilitated faster processing.

While the previously mentioned studies display the added value of sensitive measures, RT measurement failed to be informative in other studies. Pellicer-Sánchez and Schmitt (2012) assessed whether RT measures would be appropriate for scoring the Yes–No vocabulary test, assuming that certainty and accuracy would be reflected by fast RT, whereas slow RT would indicate hesitation and inaccuracy. Yet, they found no advantage for RT over traditional measures. Similarly, Fukkink et al. (2005) investigated whether time-pressured computerized L2 vocabulary learning (translation and gapped sentence exercise) would result in (a) faster lexical access during a LDT, (b) increased reading performance, and (c) text comprehension. It was found that the trained words were responded to faster during the experimental task, but this advantage was not transferable to reading speed and text comprehension. Finally, Sonbul and Schmitt (2013) used explicit traditional measures (i.e., multiple-choice form recognition and form recall) and a timed primed LDT to detect learning of collocations through two decontextualized conditions. Learning gains were revealed for the explicit measures but not for the sensitive measure.

In sum, while some previous studies indicate the potential of sensitive measures for L2 vocabulary research, other studies suggest that their role is not clear-cut. Therefore, more research that combines traditional and sensitive measures is warranted.

INSTRUCTIONAL TREATMENTS FOR L2 VOCABULARY LEARNING

While L2 vocabulary learning benefits from opportunities for incidental learning through repeated encounters in a variety of meaningful contexts, word-focused instruction (i.e., directing learners' conscious attention to new vocabulary in either communicative or noncommunicative situations; Laufer, 2017), plays an important role in L2 word learning (Webb & Nation, 2017). Yet, meaning-oriented language teaching methodologies, such as the strong versions of communicative teaching and task-based language learning, assume that L2 instruction is most efficacious through communication and functional language use (Ellis, 2009; Littlewood, 2014). However, more flexible versions of meaning-oriented approaches advocate opportunities for word-focused instruction, such as the implementation of an explicit focus on lexical items before, during and/or after

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

performing a meaningful tasks (e.g., Ellis, 2009; Van den Branden, 2016). Furthermore, Mason and Krashen (2010) argue in favor of meaning-oriented research that meets the conditions for successful vocabulary learning, such as interesting and motivating comprehensible input. To discuss the role of contextualized input and word-focused instruction, the relevant literature will be reviewed that corresponds to three prototypical instructional conditions: (a) contextualized input with meaning-oriented activities only, (b) contextualized input with both meaning-oriented and word-focused activities, and (c) decontextualized input with word-focused exercises.

CONTEXTUALIZED INPUT WITH MEANING-ORIENTED ACTIVITIES

Numerous L2 vocabulary studies have shown that exclusively meaning-focused conditions yield lower learning gains than word-focused conditions (Laufer, 2017). However, it has been argued that in those studies, meaning-oriented techniques that promote successful word learning have not been used in an optimal way. As such, in a critical review of File and Adams (2010), Mason and Krashen (2010) argue that the conditions in the reading-only treatment were not optimal for word learning, in that reading was not self-selected, readers had to follow along while the text was read out by the researcher, and the reading passage was very demanding, as only 78% of the words pertained to the 2,000 most frequent English words. Moreover, the authors state that the reading passages were not interesting to the participants. Indeed, topic interest and topic familiarity have shown to be beneficial to vocabulary learning (Lee & Pulido, 2017). Another factor that impacts vocabulary learning in meaning-oriented contexts is the number of encounters with new words. Pellicer-Sánchez and Schmitt (2010) used an authentic novel to investigate the relationship between frequency of occurrence and recognition of meaning and spelling, and recall of word class and meaning. It was found that in advanced L2 readers, substantial learning gains for all aspects occurred after 10 or more exposures to new vocabulary. Additionally, the presence of postreading activities may also impact vocabulary learning. While most studies supplemented reading with postreading comprehension activities, few studies used free meaning-focused postreading output activities, which could promote retention of text-relevant vocabulary (except for Rott, 2004). As such, the role of providing opportunities for using new words in guided and unguided meaning-oriented output activities requires further inquiry (Coxhead, 2011).

CONTEXTUALIZED INPUT WITH BOTH MEANING- AND WORD-FOCUSED ACTIVITIES

There is ample evidence that word-focused activities have beneficial effects on word learning (e.g., Laufer, 2017; Peters, 2012; Webb & Nation, 2017). The efficiency of word-focused activities can be explained by Hulstijn and Laufer's (2001) Involvement Load Hypothesis (ILH), stating that the more attentional involvement a vocabulary task requires, the more efficient it is for subsequent vocabulary learning. ILH is rooted in Craik and Lockhart's (1972) Depth of Processing Hypothesis (cited in, e.g., Schmitt, 2008) that rests on the idea that the more attention is allocated to processing new information, the better the opportunities for learning will be. According to ILH, activities with a high involvement load are predicted to be more effective for L2 vocabulary learning than activities with a low involvement load, whereby involvement varies as a

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

function of *need* (i.e., task completion requires using the word), *search* (i.e., task completion requires finding the word), and *evaluation* (i.e., deciding whether a word fits in the surrounding context). In two parallel experiments, Hulstijn and Laufer (2001) intended to find empirical evidence for ILH. Through a meaning recall test (i.e., provide an L1 translation or definition), it was found that results were better when targets were presented in a word list and had to be used in a writing task (49% and 67%) than reading with glosses (27% and 20%) or reading and completing a gapped text (29% and 40%).

Laufer and Girsai (2008) compared reading with comprehension questions, reading with vocabulary tasks and reading with translation tasks. The latter was found to induce high involvement and yielded the best results on a meaning recall and a form recall test (i.e., provide the L2 word form of an L1 meaning). Laufer and Rozovski-Roitblat (2015) compared the impact of number of encounters with new words in reading only, reading with a dictionary, and reading with word-focused activities. Four vocabulary tests were administered: cued form recall (first letter of L2 word was given), meaning recall (supplying L1 translation of an L2 form), form recognition (multiple choice), and finally meaning recognition (multiple-choice L1 translation of L2 word form). The authors found that reading with word-focused exercises yielded the highest scores on all tests and concluded that task type is more influential than number of exposures.

DECONTEXTUALIZED INPUT WITH WORD-FOCUSED EXERCISES

To further explore the value of word-focused instruction, Laufer (2006) used the L2 grammar taxonomy of *Focus-on-Form* (i.e., meaning-oriented practice with opportunities for attending to linguistic forms), and *Focus-on-Forms* (i.e., decontextualized noncommunicative practice) to compare the impact of reading with dictionary use and word lists supplemented with word-focused exercises in both incidental and deliberate learning conditions. On the meaning recall posttests for incidental learning, it was found that word lists supplemented with word-focused exercises yielded better learning outcomes than reading with dictionary use. Notwithstanding the efficiency of decontextualized vocabulary learning (Schmitt, 2008), behaviorist paired-associate learning of word lists has pedagogically been disapproved of since the emergence of communicative approaches (Elgort, 2011). Similarly, translation practice has been discouraged in pedagogy but found support in L2 vocabulary studies (Hummel, 2010; Laufer & Girsai, 2008; Schmitt, 2008; Webb & Nation, 2017).

Although it is accepted that effective L2 vocabulary learning depends on attention and maximal engagement with lexical items (Schmitt, 2008), meaning-oriented paradigms are uncertain about the equilibrium between an explicit focus on linguistic features and authentic meaning-oriented instruction (Ellis, 2009; Littlewood, 2014). Hence, in the context of L2 vocabulary instruction, more research is warranted on the role of contextualized input and word-focused instruction.

RATIONALE AND RESEARCH QUESTIONS

The aim of the present study is to investigate the value of combining explicit (pen-and-paper) and sensitive (RT) measures to assess form-, meaning-, and use-related learning gains resulting from three instructional treatments: (a) contextualized input with

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

meaning-oriented but not word-focused activities [+CO−WF], (b) contextualized input with both meaning- and word-focused activities [+CO+WF], and (c) decontextualized input with word-focused exercises [−CO+WF].

The following research questions guided this study:

RQ1: What is the impact of [+CO−WF], [+CO+WF], and [−CO+WF] on L2 vocabulary learning as measured by explicit measures?

RQ 2: What is the impact of [+CO−WF], [+CO+WF], and [−CO+WF] on L2 vocabulary learning as measured by sensitive measures?

As previous findings point toward the efficiency of word-focused instruction (e.g., Laufer, 2017; Webb & Nation, 2017), we expect for RQ1 that both word-focused conditions will outperform the meaning-only condition. Moreover, given the efficiency of decontextualized instruction (e.g., Schmitt, 2008), we hypothesize that decontextualized word-focused instruction may fare better than contextualized word-focused instruction. For RQ2, previous findings (e.g., Elgort, 2011; Elgort et al., 2018) suggest that the decontextualized word-focused instruction will lead to more accurate and faster processing than the other conditions.

METHOD

PARTICIPANTS

Participants were 313 Flemish (age 15–16) intermediate learners (global CEFR B1 proficiency level) of L2 French. Participants were pretested on vocabulary size (VocSize), which was used as a proxy for L2 proficiency (Schmitt, 2010), and working memory span (WM), which was used as a measure of cognitive ability that plays a role in L1 and L2 word learning (Elgort et al., 2018). On the basis of the mean scores on these tests (see Table 3), four roughly homogenized groups consisting of four or five intact classes were composed, that is, the three experimental groups, [+CO−WF] ($n = 71$), [+CO +WF] ($n = 82$), [−CO+WF] ($n = 76$), and the control group ($n = 83$). The control group only took part in the tests to monitor the validity of the experimental tasks, learning between the treatment sessions and possible pre- to posttest effects (Nation & Webb, 2011). Participants were informed that the study dealt with optimizing learning materials. They were unaware that the focus was on vocabulary learning and that they would be tested afterward.

TARGET ITEMS

Targets were 20 real L2 French verbs and were selected as follows. Initially, 51 presumably unknown verbs were selected from French online news sites as potential target items. These verbs were tested in a meaning recognition test with last-year secondary school students ($N = 228$) who were comparable to the students who participated in the actual experiment in terms of age, years of instruction, and proficiency level. Next, another comparable group of secondary school students ($N = 239$) was asked to infer the meaning of the candidate targets in a variety of newspaper contexts. Finally, experienced L2 French teachers ($N = 22$) evaluated whether the meaning of 120 French verbs (including the potential targets) could be known by students at the end of secondary school. Results of

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

Target	***Translation***	**Target**	***Translation***
narguer	*to mock*	occulter	*to hide*
kiffer	*to like*	disjoncter	*to freak out*
fustiger	*to criticize*	huer	*to boo*
proliférer	*to spread*	relater	*to tell*
larguer	*to dump*	cartonner	*to be successful*
déclencher	*to trigger*	s'effondrer	*to break down*
attiser	*to instigate*	éclabousser	*to damage one's reputation*
subtiliser	*to take away*	divulguer	*to make generally known*
arnaquer	*to defraud*	écoper	*to get a punishment*
souiller	*to pollute*	estomaquer	*to shock*

FIGURE 1. Targets.

TABLE 1. Overview of the three treatments

	Contextualized not word-focused [+CO−WF]	Contextualized word-focused [+CO+WF]	Decontextualized word-focused [−CO+WF]
Day 1	Input: news site Meaning-oriented comprehension activities	Input: news site Word-focused comprehension activities	Input: word list Word-focused exercises
Day 2	Input: new site Free production	Input: news site Guided production	Input: word list Translation

the recognition and inferencing test, as well as the ratings from the teachers, were combined (Supplement 1) to obtain the final set of 20 target verbs (Figure 1).

TREATMENTS

The study consisted of three conditions: two contextualized conditions (with and without word-focused activities) and one decontextualized condition (Table 1). Each treatment consisted of two sessions that took place during regular classroom hours on two consecutive days

CONTEXTUALIZED CONDITIONS

As can be seen from Table 1, a lifelike online news site that included 10 articles based on authentic news items was created for both contextualized conditions ([+CO−WF] and [+CO+WF]). The articles were conceived in such a way that they would facilitate word learning. First, the items were selected to be interesting for the target audience (Lee & Pulido, 2017). Some examples of topics were *plastic pollution in the oceans* and *parents who drug their unmanageable children*. Second, each article (average length: 201.2 words, *SD* = 16.85) contained two targets that each occurred three times in the text. We assumed that reading the texts on two consecutive days, followed by repeated encounters and opportunities for retrieval during the activities, would result in at least 10 exposures per target (Pellicer-Sánchez & Schmitt, 2010). To avoid unnatural redundancy due to the repeated use of the targets, all texts were checked by a native speaker.

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

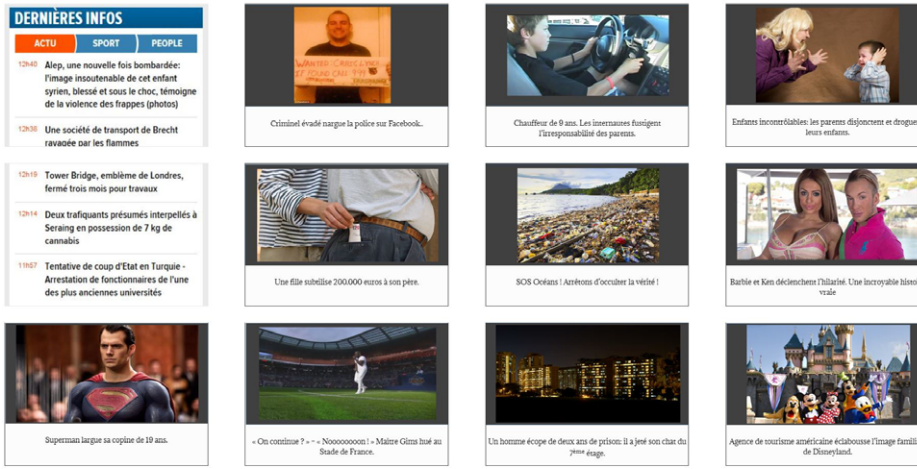

FIGURE 2. Homepage of the news website.

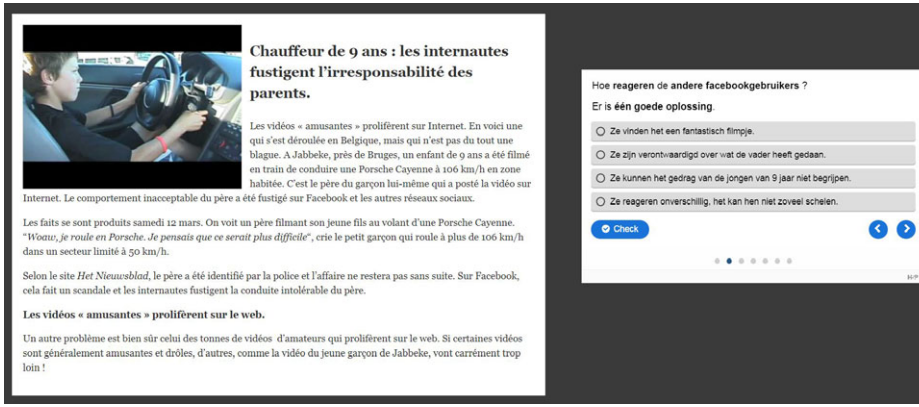

FIGURE 3. Example website interface of a meaning-oriented activity.

Third, targets were essential for comprehension and relevant to the message (Peters et al., 2009; Peters, 2012). For instance, understanding the target *narguer* (to mock) was essential for understanding an article about an escaped criminal using social media to mock the police. Fourth, Lextutor (https://www.lextutor.ca) was used to check the lexical coverage of all texts. Results revealed that 95.82% (SD = 0.70) of the words were high frequency vocabulary, that is, pertaining to the 3,000 most frequent words (Schmitt & Schmitt, 2014). Finally, pilot testing of the texts confirmed that students (N = 71) of the same age enrolled in the same L2 program had no difficulties in understanding the articles.

The homepage (Figure 2) of the news site used in the contextualized conditions showed 10 news items. By clicking on each item, participants were directed to a split-screen page with the article on the left-hand side, and activities on the right-hand side (Figure 3). The types of activities were different for both contextualized conditions (for a detailed

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

FIGURE 4. Example website interface of a writing activity.

overview of the activities in the contextualized conditions, see Supplement 2). Participants in both contextualized treatments could continuously consult the articles while performing the activities.

Participants in [+CO−WF] read the articles and performed two types of activities. On day one, they answered multiple-choice comprehension questions about the content of each article (e.g., yes/no statements). On day two, participants were asked to reread each news item and to post a short free comment of about five lines per article (Figure 4). Using the targets in the output activity was not required nor suggested.

Participants in [+CO+WF] were presented with the same news site. On day one, they answered multiple-choice comprehension questions that were both meaning- and word-focused. For instance, in the article entitled "Scandale: une agence de tourisme américaine *éclabousse* la reputation de Disneyland" (Scandal: a travel agency *damages* the reputation of Disneyland), it was asked which statement would best fit the title: On *sauve/dégrade/améliore* la réputation de Disneyland (The reputation of Disneyland is *saved/harmed/improved*). On day two, participants were asked to reread each news item and to post a short comment of about five lines per article in which they were required to use the targets.

DECONTEXTUALIZED CONDITION

Participants in the decontextualized group were provided with four vocabulary exercise blocks including four word lists related to (multi)media. Each word list contained five targets and eight to nine filler verbs, a translation and a sample sentence. By clicking on each block, they were directed to a split-screen page with a word list on the left-hand side and interactive exercises on the right-hand side (Figure 5). On day one, participants made the following exercises for each of the four training blocks: a drag-and-drop exercise about the form-meaning link, a fill-in-the-gap exercise on synonyms, a drag-and-drop exercise on verb structures, a picture-matching exercise, and an odd-man-out exercise. On day two, they were presented with the same word lists and were asked to translate 10 short sentences per block (L1 to L2), containing both targets and filler verbs. As in the contextualized conditions, participants in the decontextualized condition could continuously consult the word lists while performing the word-focused exercises.

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

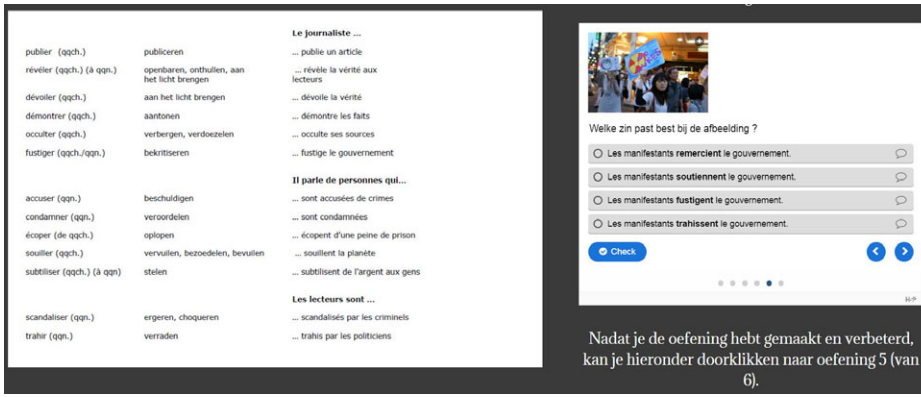

FIGURE 5. Example website interface decontextualized condition.

All instructions in the activities (both contextualized and decontextualized) were given in learners' L1 (Dutch) to ensure learners' comprehension. Participants performed the tasks individually at their own pace, but were aware of the overall time frame of 80 minutes (Elgort, 2011). Piloting the materials had shown that sessions took 60 to 70 minutes. Participants were given a checklist to ensure that they completed all tasks. To have a rough indication of the time needed for task completion, participants were asked to write down their respective start and end time of the treatment. Accessing the website outside of the learning sessions was impossible. All sessions were led by the first author.

MEASUREMENT INSTRUMENTS

Explicit Measures

To address form-, meaning-, and use-related word knowledge components, vocabulary gains were measured by means of three pen-and-paper tasks (Figure 6). The *form recognition test* assessed receptive knowledge about the word form. Participants needed to indicate the real French target verb amongst three legally spelled pseudoverb variants. They could indicate their certainty on a five-point scale, ranging from "very sure" to "very unsure." To discourage guessing, an "I don't know" option was provided (e.g., Pellicer-Sánchez & Schmitt, 2010). To assess knowledge about the meaning of the target verbs, a *meaning recall tests* was administered in which participants could demonstrate receptive knowledge of the meaning of the words by giving a translation, using the word in a sentence, explaining the word, and so forth (e.g., Webb, 2007). With respect to use-related word knowledge, we assessed receptive knowledge about the grammatical patterns in which the target verbs occur. In the *grammatical preference test*, the targeted type of grammaticality was the verb's complement structure. Participants were provided with four sentences in which targets were either correctly or incorrectly used (e.g., Chen & Truscott, 2010). In order to discourage guessing, an "I don't know" option was provided (Nation & Webb, 2011), as well as a "none of the above is correct" option (Ionin & Zyzik, 2014). There was one correct option per target.

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

Form recognition – Indicate which of the 4 words is the correctly spelled French word. Indicate also how sure you are about your answer. If you have no idea, indicate honestly "I don't know".

					Very unsure	Unsure	Sure	Very sure	I do not know
	larguer	guarler	élarguir	guarlir					

Meaning recall – Do you know what these words mean ? You can show in any way you want that you understand the meaning of the given words, e.g., by giving a translation, a description, a synonym, "I think it has to do with…", "I think it means something like …", using in a sentence, etc.

Huer :

Grammatical preference – Which sentence is correctly constructed ? There is only one good answer. If you don't know the answer, then tick off "I don't know".

[ ] Ils narguent avec les forces de l'ordre.
[ ] Ils narguent les forces de l'ordre.
[ ] Ils se narguent des forces de l'ordre.
[ ] Ils narguent contre les forces de l'ordre.
[ ] None of the sentences is correct.
[ ] I don't know

FIGURE 6. Overview of explicit measures.

Sensitive Measures

To address speed of access to form-, meaning-, and use-related word knowledge components, we administered three time-pressured tasks (E-prime 2.0, Psychology Software Tools, http://pstnet.com).

Speed of access to the word form was tested trough a LDT in which participants had to decide whether a letter string was an existing L2 French word or not (e.g., Elgort et al., 2018). After an exercise block (12 trials followed by feedback), each trial started with a 1,500 ms fixation cross, followed by stimulus presentation until response. Each response was followed by a 1,000 ms blank screen. Stimuli were presented in random order and consisted of 10 targets (e.g., *huer*, "to boo"), 10 matched verbs (e.g., *tuer*, "to kill") pertaining to the most frequent 1,000 French words (i.e., the 1K frequency band, based on Lonsdale & Le Bras, 2009), 20 matched pseudoverbs (e.g., **fuer*), and 20 fillers from various parts of speech. Speed of access to meaning-related word knowledge was assessed through a *semantic relatedness task* (SEMREL) in which participants had to decide whether prime-target pairs (e.g., *kiffer–aimer*, "to like–to love") were semantically related or not (e.g., Frishkoff et al., 2011). After an exercise block (10 trials followed by feedback), each trial started with a 1,500 ms fixation cross, followed by 1,000 ms prime presentation and 1,000 ms target presentation until response. Each response was followed by a 1,000 ms blank screen. The test consisted of 10 related and unrelated pairs containing

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

the target verbs, 10 related and unrelated pairs containing well-known words, and 10 filler pairs. Stimuli were presented pseudorandomly to ensure an interval between related and unrelated pairs containing the target verb. Finally, with respect to use, speed of access to the grammatical patterns in which the target verbs occur was tested through a *grammatical judgement task* (GJT) in which participants had to decide whether sentences were well-formed or not (e.g., Godfroid et al., 2015). The targeted type of grammaticality was the verb complement structure (e.g., *Je kiffe la musique électronique* vs. *Je kiffe *à la musique électronique*, "I like electronic music" vs. "I like *to electronic music"). Sentences were matched for length and subject-verb-complement structure. Stimulus presentation was cut off at 3,250 ms, that is, mean RT in the pilot study plus 1 SD (Bley-Vroman & Masterson, 1989). After an exercise block (16 trials followed by feedback), each experimental trial started with a 1,500 ms fixation cross, followed by sentence presentation until response or 3,250 ms, and a 1,000 ms blank screen. Stimuli consisted of 10 correct and 10 incorrect sentences including the targets, an identically conceived set of sentences containing K1 verbs, and 20 filler sentences. Stimuli were presented pseudorandomly to ensure a presentation interval between correct and incorrect sentences containing the targets.

PROCEDURE

Pretests

To test prior knowledge of the targets, participants completed a pretest containing 30 French verbs (20 targets and 10 fillers) in which they were asked whether they were familiar with each verb. If so, they were asked to demonstrate knowledge about the meaning of each verb by providing a translation, using the verb in a sentence, using *I think it has to do with*, and so forth. Following this test, participants' L2 proficiency was established through an adapted version (Noreillie, 2019) of the Vocablab test (Peters et al., 2019) consisting of 120 meaning recognition multiple-choice items (four distractors in Dutch and an *I don't know-option*), covering the 1K–4K frequency bands (30 items per frequency band). Third, to determine participants' WM, a computerized version of the OSPAN-test was administered (De Neys et al., 2002). In this test, participants were presented with a series of math problems (e.g., IS $(7 \times 2) - 4 = 10$?). In between each math problem, high-frequency Dutch words were presented shortly (800 ms) and had to be remembered. At the end of each sequence, participants were asked to write down the words in the order of appearance. Finally, a background questionnaire was administered to detect whether any of the participants had French as L1. Learners' language background was also double-checked with their teachers.

Learning Sessions

Each condition consisted of two treatment sessions that took place during regular classroom hours on two consecutive days. Participants performed the tasks individually at their own pace. Participants were given a checklist to ensure that they completed all tasks. They also wrote down their respective start and end time of the treatment to have a

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

TABLE 2. Distribution of targets over explicit and sensitive posttests

	Explicit			Sensitive		
	Form	Meaning	Use	Form	Meaning	Use
1st half of the group		Set A			Set B	
2nd half of the group		Set B			Set A	

rough indication of the time needed for task completion. Accessing the website outside of the learning sessions was impossible. All sessions were led by the first author.

Posttests

Six surprise vocabulary posttests (cf. supra) were administered at the end of the second learning session as well as 2 weeks later. Test sessions took 45 minutes. Tests were carefully sequenced to avoid testing effects, based on Nation and Webb (2011). For the explicit measures, the form recognition test was administered first because providing the L2 word form would not inform learners about the meaning, nor about grammatical use. Accordingly, the grammatical preference test was administered as last test so that the sentence contexts could not prime participants with meaning-related knowledge. For the sensitive measures, the same logic was adopted. LDT was administered first because word form recognition is not informative to meaning (SEMREL) and grammatical use (GJT). To avoid that sentence contexts would be informative for meaning-related knowledge, GJT was administered as final test. The pilot study had revealed that it was impossible to take six different tests within a 45 minutes time frame, which was considered the maximum duration to prevent test fatigue. It was therefore decided to split-up target words in two sets (A and B), which were counterbalanced within groups, that is, one half of each experimental group was tested on set A for the explicit tests and on set B for the sensitive measures, and vice versa for the other half of the group (Table 2).

SCORING AND ANALYSES

Scoring

For the explicit tests, responses on the *form recognition*, *meaning recall*, and *grammatical preference tests* were scored binomially (0 or 1). Participants received 1 point when they indicated the correctly spelled verb for *form recognition*, provided any correct translation or answer that reflected the meaning of the word for *meaning* (tests were scored by two raters, interrater reliability = 99.24 %), and indicated the correctly built sentence for *grammatical use*. For the sensitive tests, accuracy (0 or 1) and RT on correct responses were recorded for LDT, SEMREL, and GJT.

Analyses

Linear mixed-effects models (*lme*) were used to analyze the data. We used R (R Core Team, 2018) and *lme4* (Bates et al., 2015) to compare [+CO−WF], [+CO+WF], and

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

[−CO+WF]. Mixed-effects models have the advantage of including fixed factors and random effects related to participants and items (Linck & Cunnings, 2015). We used the *glmer* function (generalized linear mixed-effects models) for dichotomous accuracy data and the *lmer* function for continuous RT data. A mixed-effects model was computed for each dependent measure according to the following procedure: We started with a null model including the dependent measure, Treatment as fixed factor, and Participants and Items as crossed random variables. The most adequate statistical model was fitted to the data by adding covariates that accounted for participant and item characteristics (i.e., vocabulary size, working memory, gender, and word length expressed as the total number of letters). Given the theoretical importance of vocabulary size for word learning (Schmitt, 2010), we checked interactions between VocSize and Treatment. The estimation method was Restricted Maximum Likelihood for models created with *lmer* and Maximal Likelihood for models created with *glmer*. The significance of fixed effects for models created with *lmer* was with *lmerTest* (Kuznetsova et al., 2017), *p*-values were based on *t*-tests with Satterthwaite's degrees of freedom. For models created with *glmer*, *p*-values were based on Wald's *z*. The final statistical model for each dependent variable consisted of factors and interactions that significantly contributed to the model and improved the model fit according to the AIC-index (Akaike Information Criterion). To compare [+CO−WF], [+CO+WF], and [−CO+WF], we applied the *ghlt* function of the *multcomp* package for multiple comparisons in R (Hothorn et al., 2008) to the fixed factor Treatment in the final model, and used Holm correction to adjust for multiple comparisons (Bretz et al., 2011). Nondichotomous RT distributions that deviated from normality were logarithmically transformed to bring them closer to normality. The values for VocSize and WM were standardized. For the sake of readability, RT in the average reports are the untransformed values in milliseconds.

RESULTS

WM AND VOCSIZE

Table 3 shows the descriptive statistics for the learner-related variables WM and VocSize (Cronbach's alpha = .91). Although the descriptives reveal differences in the mean scores between the groups for VocSize and WM, a one-way ANOVA indicated that these differences did not reach significance (VocSize, $F(3, 309) = 1.06$, $p = .368$, and WM, $F(3, 309) = 1.65$, $p = .178$).

TABLE 3. Means and SD for WM (maximum = 60) and VocSize (maximum = 120)

	Control *(n = 83)*		[+CO−WF] *(n = 72)*		[+CO+WF] *(n = 82)*		[−CO+WF] *(n = 76)*	
	M	SD	M	SD	M	SD	M	SD
WM	37.42	10.25	39.15	10.79	35.59	9.84	36.83	10.12
VocSize	72.08	14.90	72.60	12.75	74.72	12.75	75.26	13.47

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

TABLE 4. Immediate explicit posttests: average accuracy (in %)

	Control *(n = 83)*		[+CO−WF] *(n = 72)*		[+CO+WF] *(n = 82)*		[−CO+WF] *(n = 76)*	
	M	SE	M	SE	M	SE	M	SE
Form	30.85	1.61	54.31	1.86	92.32	0.93	96.29	0.70
Meaning	0.12	0.12	12.08	1.22	45.85	1.74	35.44	1.77
Use	28.17	1.57	40.41	1.83	59.75	1.71	61.40	1.81

TABLE 5. Delayed explicit posttests: average accuracy (in %)

	Control *(n = 79)*		[+CO−WF] *(n = 72)*		[+CO+WF] *(n = 82)*		[−CO+WF] *(n = 74)*	
	M	SE	M	SE	M	SE	M	SE
Form	46.96	1.78	72.64	1.66	92.93	0.90	91.95	1.02
Meaning	0.25	0.18	9.44	1.09	35.24	1.67	22.03	1.56
Use	25.57	1.55	42.50	1.84	57.68	1.73	57.06	1.86

Note: Four participants were administered an erroneous task and two participants were absent.

RQ 1: Explicit Measures

Form-, meaning-, and use-related learning gains were measured immediately after the treatment and two weeks later. Table 4, which presents the descriptive statistics of the immediate explicit tests, shows that the average scores were highest for *form*. The control group performed near-chance for the form recognition test and near-zero for the meaning recall test. As can be seen from Table 5, the average results on the delayed posttests show a similar pattern. However, the scores on the delayed form recognition test in the experimental group should be interpreted with caution because scores of the control group also revealed an important increase from immediate to delayed test. This seems to indicate that there was a test effect that is probably due to the multiple encounters with the target words during the immediate posttests (Rice & Tokowicz, 2019).

To check whether each of the three experimental conditions had led to significant learning gains as measured with explicit measures, we first developed a model that compared each experimental condition with the control condition. Results show that the three treatments were conducive to form-, meaning-, and use-related learning gains (see Supplement 3 for an overview) because they significantly outscored the control group (who only took the tests).

Table 6 displays the summaries of the models that were developed to compare [+CO−WF], [+CO+WF], and [−CO+WF]. The estimate levels for the fixed factor Treatment indicate that both word-focused groups systematically yielded higher scores than the context-only group. All models show the importance of the covariate VocSize for vocabulary learning. On the delayed test for *use*, female participants fared better than their male counterparts.

Pairwise comparisons between the experimental groups on the immediate posttests (Table 7) showed that for *form*, the two word-focused conditions significantly outperformed

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

TABLE 6. Accuracy on explicit immediate and delayed posttests

Posttest	Predictor	Estimate	SE	/z/	p	
Form immediate	(Intercept)	.24	.21	1.13	.26	
	[+CO+WF]	2.69	.20	11.46	< 2e-16	***
	[−CO+WF]	3.45	.25	12.76	< 2e-16	***
	VocSize	.38	.09	4.47	7.9e-06	***
Form delayed	(Intercept)	1.26	.24	5.20	2.05e-07	***
	[+CO+WF]	1.82	.20	9.33	< 2e-16	***
	[−CO+WF]	1.65	.20	8.32	< 2e-16	***
	VocSize	.40	.08	4.87	1.12e-06	***
Meaning immediate	(Intercept)	−2.54	.29	−8.85	< 2e-16	***
	[+CO+WF]	2.33	.22	10.66	< 2e-16	***
	[−CO+WF]	1.72	.22	7.81	5.96e-15	***
	VocSize	.34	.09	3.90	6.64e-05	***
Meaning delayed	(Intercept)	−2.79	.27	−10.30	< 2e-16	***
	[+CO+WF]	2.01	.20	9.81	< 2e-16	***
	[−CO+WF]	1.16	.21	5.47	4.5e-08	***
	VocSize	.29	.08	3.52	4.38e-04	***
Use immediate	(Intercept)	−.42	.18	−2.31	.02	*
	[+CO+WF]	.87	.14	6.35	2.20e-10	***
	[−CO+WF]	.92	.14	6.53	6.50e-11	***
	VocSize	.30	.06	4.96	7.06e-07	***
Use delayed	(Intercept)	−.52	.22	−2.39	.02	*
	[+CO+WF]	.69	.14	4.88	1.04e-06	***
	[−CO+WF]	.70	.15	4.77	1.83e-06	***
	VocSize	.21	.06	3.47	5.24e-04	***
	Gender(F)	.25	.13	2.00	4.6e-02	*

Note: Intercept levels represent the values of [+CO−WF]. VocSize = vocabulary size, Gender(F) = female. *$p < .05$; **$p < .01$; ***$p < .001$.

the context-only group and that [−CO+WF] significantly outperformed [+CO+WF]. This was not the case on the delayed posttests. For *meaning*, all groups significantly differed from each other, in that both word-focused groups outperformed the meaning-only group, and [+CO+WF] outperformed [−CO+WF]. For *use*, [+CO+WF] and [−CO+WF] significantly outperformed [+CO−WF]. However, no differences were observed between the two word-focused groups. Results on the delayed posttests were similar to the immediate posttests for *meaning* and *use*.

RQ 2: Sensitive Measures

Table 8 to Table 13 summarize average accuracy and RT for LDT, SEMREL, and GJT. Before analyzing RT data, we inspected the data for outliers. For LDT and SEMREL, responses faster than 200 ms and responses with 2.5 SD beyond a participant's mean RT were considered outliers and removed from the dataset (immediate LDT: 2.03%;

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

TABLE 7. Pairwise comparisons for explicit posttests

Posttest	Conditions	Est.	SE	/z/	p	
Form immediate	[+CO+WF]–[+CO−WF]	2.67	.20	13.46	< 2e-16	***
	[−CO+WF]–[+CO−WF]	3.45	.25	13.76	< 2e-16	***
	[−CO+WF]–[+CO+WF]	.78	.26	2.99	.003	**
Form delayed	[+CO+WF]–[+CO−WF]	1.82	.20	9.33	< 2e-16	***
	[−CO+WF]–[+CO−WF]	1.65	.20	8.32	< 2e-16	***
	[−CO+WF]–[+CO+WF]	−.18	.22	−.80	.42	
Meaning immediate	[+CO+WF]–[+CO−WF]	2.33	.22	10.66	< 2e-16	***
	[−CO+WF]–[+CO−WF]	1.72	.22	7.81	1.2e-14	***
	[−CO+WF]–[+CO+WF]	−.60	.19	−3.22	.001	**
Meaning delayed	[+CO+WF]–[+CO−WF]	2.01	.20	9.81	< 2e-16	***
	[−CO+WF]–[+CO−WF]	1.16	.21	5.47	9.01e-08	***
	[−CO+WF]–[+CO+WF]	−.85	.18	−4.81	1.50e-06	***
Use immediate	[+CO+WF]–[+CO−WF]	.87	.14	6.35	4.41e-10	***
	[−CO+WF]–[+CO−WF]	.92	.14	6.53	1.95e-10	***
	[−CO+WF]–[+CO+WF]	.05	.14	.40	.69	
Use delayed	[+CO+WF]–[+CO−WF]	.69	.14	4.88	3.12e-06	***
	[−CO+WF]–[+CO−WF]	.70	.15	4.17	3.66e-06	***
	[−CO+WF]–[+CO+WF]	.01	.14	.07	.95	

*$p < .05$; **$p < .01$; ***$p < .001$.

TABLE 8. Immediate LDT: average accuracy (in %) and RT (ms) for critical items

	Control *(n = 82)*		[+CO−WF] *(n = 70)*		[+CO+WF] *(n = 82)*		[−CO+WF] *(n = 75)*	
	M	SE	M	SE	M	SE	M	SE
Accuracy	46.22	1.74	70.43	1.73	90.37	1.03	94.15	0.88
RT	1,497	46.19	1,188	26.01	1,048	20.68	901	14.61

Note: Results of one participant were discarded because of misunderstanding the task. One participant received an erroneous task. Results of two participants were excluded because of improper task performance.

TABLE 9. Delayed LDT: average accuracy (in %) and RT (ms) for critical items

	Control *(n = 75)*		[+CO−WF] *(n = 71)*		[+CO+WF] *(n = 82)*		[−CO+WF] *(n = 71)*	
	M	SE	M	SE	M	SE	M	SE
Accuracy	63.51	1.78	75.46	1.62	89.3 7	1.08	93.04	0.95
RT	1,068	24.20	1,002	18.54	900	11.49	849	13.16

Note: Two participants were absent, eight participants received an erroneous task and the data file of one participant was corrupted. Results of three participants were excluded because of improper task performance.

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

TABLE 10. Immediate SEMREL: average accuracy (in %) and RT (ms) for critical items

	Control *(n = 83)*		[+CO−WF] *(n = 72)*		[+CO+WF] *(n = 81)*		[−CO+WF] *(n = 75)*	
	M	SE	M	SE	M	SE	M	SE
Accuracy related	29.52	1.58	36.11	1.79	55.80	1.75	63.65	1.80
RT related	1,640	52.41	1,458	38.94	1,607	38.81	1,308	29.19

Note: One participant received an erroneous task and the data file of another participant was corrupted.

TABLE 11. Delayed SEMREL: average accuracy (in %) and RT (ms) for critical items

	Control *(n = 75)*		[+CO−WF] *(n = 71)*		[+CO+WF] *(n = 82)*		[−CO+WF] *(n = 72)*	
	M	SE	M	SE	M	SE	M	SE
Accuracy related	26.49	1.62	28.03	1.69	47.65	1.76	53.25	1.90
RT related	1,322	48.74	1,222	30.77	1,182	24.95	1,049	22.30

Note: Two participants were absent and eight participants received an erroneous task. Three participants were excluded because of improper task performance.

TABLE 12. Immediate GJT: average accuracy (in %) and RT (ms) for critical items

	Control *(n = 82)*		[+CO−WF] *(n = 72)*		[+CO+WF] *(n = 82)*		[−CO+WF] *(n = 73)*	
	M	SE	M	SE	M	SE	M	SE
Correct	57.33	1.72	66.29	1.77	72.07	1.57	79.94	1.52
RT Correct	2,213	22.41	2,088	21.46	2,101	19.59	2,010	20.13

Note: One participant was administered an erroneous task. Three participants were excluded because of improper task performance.

TABLE 13. Delayed GJT: average accuracy (in %) and RT (ms) for critical items

	Control *(n = 75)*		[+CO−WF] *(n = 71)*		[+CO+WF] *(n = 82)*		[−CO+WF] *(n = 71)*	
	M	SE	M	SE	M	SE	M	SE
Correct	61.16	1.86	68.44	1.76	77.29	1.46	77.83	1.59
RT Correct	2,006	25.53	2,072	21.35	1,967	18.01	1,876	19.87

Note: Two participants were absent and eight participants were administered an erroneous task. Four Participants were excluded because of improper task performance.

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

TABLE 14. Accuracy on sensitive posttests

Posttest	Predictor	Estimate	SE	/z/	p	
LDT immediate	(Intercept)	1.09	.23	4.73	2.26e-06	***
	[+CO+WF]	1.51	.17	8.91	< 2e-16	***
	[−CO+WF]	1.97	.20	9.92	< 2e-16	***
	VocSize	.33	.08	4.21	2.56e-05	***
LDT delayed	(Intercept)	1.30	.19	7.02	2.31e-12	***
	[+CO+WF]	1.03	.15	6.71	2.00e-11	***
	[−CO+WF]	1.53	.19	8.18	9.89e-16	***
	VocSize	.33	.07	4.50	6.87e-06	***
SEMREL immediate	(Intercept)	−.63	.17	−3.68	2.35e-04	***
	[+CO+WF]	.89	.14	6.23	4.55e-10	***
	[−CO+WF]	1.24	.15	8.38	< 2e-16	***
	VocSize	.33	.06	5.22	1.75e-07	**
SEMREL delayed	(Intercept)	−1.11	.20	−5.52	3.47e-08	***
	[+CO+WF]	.95	.16	5.81	6.36e-09	***
	[−CO+WF]	1.18	.17	6.97	3.25e-12	***
	VocSize	.52	.07	7.09	1.39e-12	***
GJT immediate	(Intercept)	.74	.17	4.33	1.49e-05	*
	[+CO+WF]	.37	.14	2.52	.01	***
	[−CO+WF]	.85	.16	5.41	6.40e-08	***
	WM	.16	.06	2.53	.01	*
GJT delayed	(Intercept)	.92	.19	4.71	2.52e-06	***
	[+CO+WF]	.51	.16	3.16	.002	**
	[−CO+WF]	.55	.17	3.26	.001	**
	VocSize	.17	.07	2.27	.02	*

Note: Intercept levels represent the values of [+CO−WF]. VocSize = vocabulary size. WM = working memory. *p < .05; **p < .01; ***p < .001.

delayed LDT: 1.54%; immediate SEMREL: 1.19%; delayed SEMREL: 1.62%). For GJT, responses faster than 600 ms were considered too fast for proper processing (immediate GJT: 0.52%; delayed GJT: 1.34%). Not responding to trials or responding faster than 600 ms on more than one third of the trials was considered to reflect improper task performance.

Before turning to the results of the mixed-effects analyses, a number of aspects need to be mentioned. For SEMREL, the high accuracy level in the control group for unrelated word pairs in both the immediate and delayed posttests showed that the unrelated category probably did not produce reliable evidence of vocabulary learning. Therefore, we only analyzed the related word pairs category. Second, for GJT, outcomes on the immediate posttest showed that performance on the violated category was near-chance for all groups. Therefore, this category was not included in further analyses. Third, as for the delayed explicit form recognition posttest (cf. *supra*), accuracy scores for the delayed LDT seem to suggest that the results were affected by a test effect (Rice & Tokowicz, 2019), due to

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

TABLE 15. Pairwise comparisons for accuracy on sensitive posttests

Posttest	Conditions	Est.	*SE*	/z/	p	
LDT immediate	[+CO+WF]–[+CO−WF]	1.51	.17	8.91	< 2e-16	***
	[−CO+WF]–[+CO−WF]	1.97	.20	9.92	< 2e-16	***
	[−CO+WF]–[+CO+WF]	.46	.21	2.17	.03	*
LDT delayed	[+CO+WF]–[+CO−WF]	1.03	.15	6.71	4.01e-11	***
	[−CO+WF]–[+CO−WF]	1.53	.19	8.18	6.66e-16	***
	[−CO+WF]–[+CO+WF]	.50	.20	2.53	.01	*
SEMREL immediate	[+CO+WF]–[+CO−WF]	.89	.14	6.23	9.1e-10	***
	[−CO+WF]–[+CO−WF]	1.24	.15	8.38	< 2e-16	***
	[−CO+WF]–[+CO+WF]	.36	.14	2.53	.01	*
SEMREL delayed	[+CO+WF]–[+CO−WF]	.97	.16	5.95	5.40e-09	***
	[−CO+WF]–[+CO−WF]	1.20	.17	7.10	3.72e-12	***
	[−CO+WF]–[+CO+WF]	.23	.16	1.48	.14	
GJT immediate	[+CO+WF]–[+CO−WF]	.37	.14	2.52	.01	*
	[−CO+WF]–[+CO−WF]	.85	.16	5.41	1.92e-07	***
	[−CO+WF]–[+CO+WF]	.48	.15	3.14	.003	**
GJT delayed	[+CO+WF]–[+CO−WF]	.51	.16	3.16	.003	**
	[−CO+WF]–[+CO−WF]	.55	.17	3.26	.003	**
	[−CO+WF]–[+CO+WF]	.05	.17	.28	.78	

*$p < .05$; **$p < .01$; ***$p < .001$.

repeated encounters with the target word forms during the immediate posttests. Finally, tendencies in RT between immediate and delayed posttests for SEMREL and GJT showed faster RT in the delayed posttests. Faster RT from immediate to delayed posttests that could not be ascribed to effects of the treatment were also observed in Pellicer-Sánchez (2015) and ascribed to a practice effect from taking the same test a second time.

To check whether each of the three experimental conditions led to learning gains as measured with sensitive techniques, we first developed a model that compared the accuracy of each experimental condition with the control group (for each dependent variable). Results showed that the three treatments led to significantly higher form-, meaning-, and use-related learning gains (see Supplement 3 for an overview) than the control condition.

Table 14 displays the accuracy summaries for the models that were created to compare the experimental conditions. VocSize impacted the results for *form* and *meaning*. On the immediate GJT, working memory modulated the scores.

Pairwise comparisons for accuracy (Table 15) showed that for LDT, both [−CO+WF] and [+CO+WF] significantly outperformed [+CO−WF] on the immediate posttests. Moreover, accuracy scores in [−CO+WF] were significantly higher than in [+CO +WF]. Results for SEMREL showed that both word-focused groups significantly outperformed the meaning-only group. Additionally, the decontextualized word-focused group outperformed the contextualized word-focused group. However, this difference

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

TABLE 16. Model summaries for RT

Posttest	Predictor	Est.	SE	df	/t/	p	
LDT immediate	(Intercept)	6.73	.11	19.59	61.20	< 2e-16	***
	[+CO+WF]	−.13	.04	228.87	−3.77	2.08e-04	***
	[−CO+WF]	−.25	.04	227.37	−6.89	5.32e-11	***
	VocSize	−.04	.02	222.79	−2.75	6.53e-03	**
	Word Length	.03	.01	17.62	2.62	.02	*
LDT delayed	(Intercept)	6.64	.07	21.55	93.51	< 2e-16	***
	[+CO+WF]	−.08	.03	221.83	−2.56	.01	*
	[−CO+WF]	−.14	.03	220.90	−4.22	3.56e-05	***
	VocSize	−.05	.01	220.90	−3.65	3.27e-04	***
	Word Length	.03	.01	17.41	3.08	6.63e-03	**
SEMREL immediate	(Intercept)	7.22	.04	110.69	167.77	< 2e-16	***
	[+CO+WF]	.06	.05	239.34	1.35	.18	*
	[−CO+WF]	−.13	.05	234.33	−2.81	.005	**
	VocSize	−.04	.02	218.07	−1.77	.08	.
SEMREL delayed	(Intercept)	7.08	.04	120.39	171.43	< 2e-16	***
	[+CO+WF]	−.05	.05	232.08	−1.19	.24	
	[−CO+WF]	−.16	.05	222.84	−3.36	9.27e-04	***
	VocSize	−.05	.02	214.64	−4.05	7.23e-05	***
GJT immediate	(Intercept)	2,100.53	42.30	62.87	49.66	< 2e-16	***
	[+CO+WF]	18.72	42.27	219.32	.44	.66	
	[−CO+WF]	−74.58	43.29	217.41	−1.72	.09	.
	VocSize	−36.92	18.63	221.45	−1.98	.05	*
	WM	−29.21	17.48	217.63	−1.67	.10	.
GJT delayed	(Intercept)	2,080.96	43.74	59.50	45.57	< 2e-16	***
	[+CO+WF]	−85.77	41.90	215.14	−2.05	.04	*
	[−CO+WF]	−170.88	43.70	216.99	−3.91	1.23e-04	***
	VocSize	−76.06	18.97	216.27	−4.01	8.36e-05	***

Note: Intercept levels represent the values of [+CO−WF]. Est. = Estimate, VocSize = vocabulary size, WM = working memory.
. $p < .1$; $p < .1$; *$p < .05$; **$p < .01$; ***$p < .001$.

had disappeared in the delayed posttests. Pairwise comparisons of GJT showed that for the immediate posttests, both word-focused groups outperformed the meaning-only group. Moreover, [−CO+WF] outperformed [+CO+WF] on the immediate posttest, but not on the delayed posttest.

The summaries of the models that were developed to compare RT (Table 16) showed that VocSize, Word Length, and WM modulated RT. The positive values for VocSize indicated that more proficient learners needed less time to make responses. The positive values for Word Length suggest that the longer the word, the longer it took participants to respond. The model estimates showed that for LDT, both word-focused conditions needed less time than the context-only condition. For GJT, learners with higher WM made faster responses.

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

TABLE 17. Pairwise comparisons for RT on sensitive posttests

Posttest	Conditions	Estimate	SE	/z/	p	
LDT immediate	[+CO+WF]–[+CO−WF]	−.13	.04	−3.77	3e-04	***
	[−CO+WF]–[+CO−WF]	−.25	.04	−6.90	1.64e-11	***
	[−CO+WF]–[+CO+WF]	−.16	.03	−3.42	6e-04	***
LDT delayed	[+CO+WF]–[+CO−WF]	−.08	.03	−2.63	0.02	*
	[−CO+WF]–[+CO−WF]	−.14	.03	−4.55	1.63e-05	***
	[−CO+WF]–[+CO+WF]	−.07	.03	−2.11	0.03	*
SEMREL Immediate	[+CO+WF]–[+CO−WF]	.06	.05	1.35	.18	
	[−CO+WF]–[+CO−WF]	−.13	.05	−2.80	.01	*
	[−CO+WF]–[+CO+WF]	−.20	.04	−4.51	1.87e-05	***
SEMREL delayed	[+CO+WF]–[+CO−WF]	−.05	.05	−1.19	.23	
	[−CO+WF]–[+CO−WF]	−.16	.04	−3.36	.002	**
	[−CO+WF]–[+CO+WF]	−10	.04	−2.47	.03	*
GJT immediate	[+CO+WF]–[+CO−WF]	18.72	42.27	.44	.66	
	[−CO+WF]–[+CO−WF]	−74.58	43.29	−1.72	.17	
	[−CO+WF]–[+CO+WF]	−93.29	40.95	−2.28	.07	
GJT delayed	[+CO+WF]–[+CO−WF]	−85.77	41.90	−2.05	.08	
	[−CO+WF]–[+CO−WF]	−170.88	43.70	−3.91	2.77e-04	***
	[−CO+WF]–[+CO+WF]	−85.11	41.29	−2.06	.08	

*$p < .05$; **$p < .01$; ***$p < .001$.

Pairwise comparisons for RT (Table 17) showed significant differences between all groups for immediate LDT. Response times in [−CO+WF] were faster than in [+CO +WF] and [+CO−WF], while responses in [+CO+WF] were faster than [+CO−WF]. For SEMREL, RT were significantly different between both word-focused groups, indicating that the contextualized word-focused group needed more time than the decontextualized word-focused group to make correct responses on meaning relatedness. The same observation was found on the delayed test. For GJT, no significant RT differences were observed on the immediate posttests. For the delayed posttest, a significant difference was observed between [−CO+WF] and [+CO−WF].

DISCUSSION

In this study, we investigated the value of combining explicit measures (i.e., pen-and-paper tasks) and sensitive measures (i.e., RT measurement) to assess the impact of three treatments: (a) contextualized input with meaning-oriented but not word-focused activities [+CO−WF], (b) contextualized input with both meaning- and word-focused activities [+CO+WF], and (c) decontextualized input with word-focused exercises [−CO +WF] on form-, meaning- and use-related aspects of 20 French target verbs.

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

RQ 1: WHAT IS THE IMPACT OF [+CO−WF], [+CO+WF], AND [−CO+WF] ON L2 VOCABULARY LEARNING AS MEASURED BY EXPLICIT MEASURES?

In answer to the first research question, our results confirm previous findings about the superiority of word-focused instruction when compared to meaning-oriented instruction (Laufer, 2017; Schmitt, 2008).

For *form* recognition, both word-focused groups outperformed the context-only group. Further, the decontextualized word-focused group recognized significantly more targets than the contextualized word-focused group. Both observations can be explained by the notion of *noticing*, that is, making the learner aware that there is something new to learn by focusing the attention on the target words (Laufer, 2020). As both word-focused conditions indeed addressed the new vocabulary in a more targeted way than the context-only condition, they fostered for a better quality of attention (Webb & Nation, 2017, p. 86), which resulted in better learning of the word forms. With regard to the decontextualized condition, presenting new words in isolation rather than as elements of a broader context involved even more deliberate noticing of the word forms (Webb & Nation, 2017, p. 87), which probably led to stronger learning. Remarkably, the control group and the meaning-only group showed an increase in the test scores from immediate to delayed posttest. This increase suggests that recurrent focused encounters with the target words during the immediate posttest battery has promoted noticing to such an extent that it eventually resulted in learning that was detected 2 weeks later. Interestingly, these additional encounters enhanced the already existing form-related knowledge in the meaning-oriented group, but did not contribute to additional learning in both word-focused groups. Moreover, results even showed a decrease of the mean score for the decontextualized group. These observations may suggest that both word-focused treatments had fully achieved their potential for word form recognition immediately after the treatment.

With respect to *meaning* recall, our results show that both word-focused conditions outperformed the meaning-only condition. Their superiority can be explained in light of the ILH (Hulstijn & Laufer, 2001), in that that the activities and the exercises in these conditions involved more engagement with the new words. As such, the word-focused activities in [+CO+WF] and the vocabulary exercises in [−CO+WF] were amenable to a high involvement load, as all three elements (*need*, *search*, and *evaluation*) were addressed. Moreover, the required use of the targets in written production on the second day of both treatments may have contributed to consolidating the initial form-meaning linkage that was established on day one. Other studies also underscored the importance of opportunities for productive knowledge development. Webb (2005), for instance, compared receptive (i.e., reading three L2 sentences with an L1 gloss) and productive (i.e., using an L2 word in a sentence) word learning. It was found that, when sufficient time was allotted, productive learning resulted in stronger receptive and productive meaning-related knowledge. Our findings are also consonant with Zou (2017) who compared the impact of vocabulary learning through cloze exercises, sentence writing, and composition writing. Both writing conditions turned out to be more efficient that the cloze condition. Interestingly, the composition-writing condition outperformed the

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

sentence-writing condition, which was explained by the fact that composition writing relied strongly on the *evaluation* component.

Additionally, with respect to the decontextualized condition, it has been argued that bilingual encoding may facilitate the initial form-meaning linkage (Schmitt, 2008; Webb & Nation, 2017) and lead to deeper memory traces (Hummel, 2010). Yet, in our study, the contextualized word-focused treatment outperformed the decontextualized word-focused treatment. This observation contradicts some earlier findings where decontextualized treatments fared best on meaning recall tests in comparison to contextualized word-focused treatments (Laufer, 2006; Llach, 2009). Our findings suggest that the news item contexts in which targets were embedded facilitated word learning, which seems compatible with contextual word learning frameworks such as the Lexical Quality Hypothesis (Bolger et al., 2008). This hypothesis holds that each encounter with an unfamiliar word results in episodic memory traces that are related to both the linguistic and nonlinguistic encoding contexts. Indeed, learners' responses on the meaning recall test demonstrated that word meanings were regularly retrieved through reactivation of the discourse contexts in which meanings were encoded, that is, the news items' content. In addition to the encoding that occurred as a consequence of text-related and word-focused activities in the contextualized word-focused condition on day one, having to use the targets in the guided output activity of day two may have strengthened the words' form-meaning link to such an extent that this treatment outclassed the decontextualized condition for meaning recall. An additional explanation for the superiority of contextualized over decontextualized instruction may be that the test format (i.e., providing the meaning) echoes better the contextualized learning condition, and hence facilitates better test performance (Nation & Webb, 2011; Webb, 2009). In the contextualized condition, ample clues for meaning establishment were indeed provided, while the decontextualized group could only rely on the L1 equivalent.

Lastly, regarding *use*, both word-focused groups scored equally well and outperformed the condition that was not word-focused. Previous research had shown that decontextualized word-focused instruction and productive use were amenable to the development of grammar-related word knowledge. Webb (2007) compared learning from word pairs versus contextual learning and found that decontextualized learning of word pairs promoted the development of grammatical knowledge. Although this finding seemed counterintuitive, the learning gains for grammar were ascribed to parallel learning in other word knowledge components and the overlap with L1 meaning. Likewise, Nation (2013, p. 82) argues that the learning burden for grammatical functions is lighter when a new item roughly parallels the L1 grammatical patterns. Further, Webb's (2005) previously cited study also found that the writing task was more effective than a reading task for the development of receptive grammar-related knowledge. In our study, both the elements of an explicit focus on words and productive use may have had an additive effect for grammatical knowledge. In line with Webb's (2007) hypothesis, this effect may have been enhanced by parallel learning in other components, such as meaning, which then spilled over to grammar.

RQ 2: WHAT IS THE IMPACT OF [+CO−WF], [+CO+WF], AND [−CO+WF] ON L2 VOCABULARY LEARNING AS MEASURED BY SENSITIVE MEASURES?

In this study, sensitive measures were found to provide additional insights into the effects of the treatments with respect to speed of lexical access, the time course of meaning retrieval, and grammatical processing.

For *form* recognition, both word-focused conditions outperformed the context-only group on accuracy and RT, which is consistent with earlier research that compared meaning- and form-oriented L2 vocabulary instruction through a timed LDT (Elgort et al., 2018). Furthermore, the decontextualized word-focused group achieved higher accuracy and faster lexical access than the contextualized word-focused group. This means that decontextualized noticing of word forms through word lists supplemented with word-focused exercises not only led to more accurate word form recognition, as indicated by the explicit measures, but it also resulted in higher time-pressured accuracy scores and facilitated faster lexical access. These results show that sensitive measures not only confirm the findings of the explicit tests in terms of accuracy but they also show that word-focused instruction is conducive to faster recognition of newly learned word forms than meaning-focused instruction.

With respect to *meaning*, the RT pattern of SEMREL seems to indicate that the two word-focused groups processed the task differently. Surprisingly, the contextualized word-focused group needed significantly more time than the other groups to make correct meaning-relatedness judgments on both test moments. This seems inconsistent with the superior performance for meaning of this group on the explicit measures. A possible explanation could be that the scrutiny of episodic memory traces about the articles' content involved a processing cost reflected through longer RT. This finding suggests that sensitive measures can provide insights into how contextualized and decontextualized learning differentially affect the processing of the meaning of newly learned L2 words. Furthermore, although participants in the decontextualized treatment did not have the opportunity of scanning content-related memory traces, they yielded higher accuracy scores on SEMREL. A possible reason is that these participants may have been advantaged by the way stimuli were created. In SEMREL, most related pairs were synonyms. As participants in the decontextualized conditions were trained on exercises comprising translation and synonyms, speeded exercises echoing the learning condition may have been at their advantage. This facilitative effect of overlapping cognitive operations during initial learning and subsequent test taking has been referred to as transfer-appropriate processing (DeKeyser, 2007). Yet, the advantage of decontextualized instruction was not observed at retention, which suggests the fragility of meaning-related L2 word knowledge acquired through decontextualized learning.

For *use*, both word-focused groups outperformed the context-only group on both the immediate and delayed posttests. It was also found that the decontextualized word-focused group identified correctly built sentences with newly learned verbs more accurately than the contextualized group, but only on the immediate posttests. These findings suggest that decontextualized deliberate word-focused learning supplemented with exercises on verb structure and translation practice can have beneficial effects on the rate at which correctly built grammatical structures can be detected shortly after learning. In this context, performing an experimental task such as a time-pressured grammatical judgment

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

task may have been at the advantage of the decontextualized group. In contrast to the word-focused contextualized treatment, in which the target verbs were processed from the broader perspective of meaningful language use, the decontextualized group processed the verbs with a more narrow and more specific attentional focus on use-related properties, for example, through the example sentences in the word list, or the L1 to L2 sentence translation exercise. Additionally, as suggested by Webb (2007) and Nation (2013), the combination of L1 meaning and grammatical use during learning may have activated parallelisms with L1 structures that have resulted in more accurate and faster detection of correctly built sentences. Interestingly, this advantage was not continued at retention, which may suggest that decontextualized use-related learning loses its superiority over time. Finally, while no use-related differences were found between the two word-focused conditions for the explicit measures, the use-related sensitive measures provided two additional insights with respect to the impact of contextualized and decontextualized word-focused learning. As such, decontextualized learning is conducive to detecting correctly built sentences based on newly learnt verbs better and faster. However, this advantage was only found at the immediate level.

PEDAGOGICAL IMPLICATIONS

The insights gained from combining explicit and sensitive measures involve a number of pedagogical implications. First, both contextualized and decontextualized word-focused instruction show to be efficient for establishing form-, meaning-, and use-related receptive L2 vocabulary knowledge. More specifically, teachers might want to consider decontextualized learning to enhance knowledge related to the word form and the grammatical structures in which words are used. Additionally, focusing on both meaning and form benefits the learning of word meanings, immediately after instruction and at the long-term level. In sum, this study shows that new words are best learnt through meaningful contexts, in parallel with decontextualized techniques.

The present study also indicates the pedagogical value of prompting learners to use new L2 vocabulary either through decontextualized or meaning-oriented instruction and, in this way, echoes similar claims that have been made in the context of single word writing (Elgort et al., 2018; Webb & Piasecki, 2018), sentence writing (Webb, 2007), text writing (Zou, 2017), oral interaction (De la Fuente, 2006), collaborative output activities (Sun, 2017), and task-based language learning (Ellis, 2009).

Taken together, our study provides support for L2 vocabulary teaching approaches that advocate a balanced L2 vocabulary course design. An example is Nation's four strands approach (e.g., 2013) that states that an ideal vocabulary course consists of meaning-focused input, meaning-focused output, word-focused learning, and fluency development. Likewise, Schmitt (2008) states that "different teaching approaches may be appropriate at the different stages of acquisition" (p. 334), suggesting that an initial word-focused approach may be followed by meaning-oriented instruction, such as *linked skills* (Webb & Nation, 2017). In this approach, learners engage with the same topic across different skills and engage with new words in a receptive and productive manner. Moreover, learners can benefit from significant and repeated encounters with new words and have ample opportunities for retrieval and use. Hence, the linked skills approach caters for both incidental and deliberate learning, and is compatible with task- and

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

content-based L2 teaching approaches that advocate the implementation of word-focused instruction in the design of a task (Ellis, 2009; Van den Branden, 2016).

LIMITATIONS AND FUTURE RESEARCH

Our research is inevitably characterized by a number of limitations. First, while sensitive measures such as RT measures are said to better represent the type of knowledge needed for fluent language use (Elgort, 2018), the sensitive measures used in the present study were receptive time-pressured tests. Consequently, the outcomes of this study are indicative of fluency of access to receptive vocabulary knowledge only. Second, as ecological validity was an asset of this study, participants were tested on sensitive measures within their school environment. However, we made sure that controlled laboratory conditions were approximated by strictly applying procedures to guarantee silence and ensure attention. A third limitation concerns some aspects of the stimuli design. In SEMREL, the control group performed equally well for unrelated pairs as the other groups, suggesting that rejecting unrelated word pairs was a default strategy whenever an unfamiliar word was presented. This hypothesis seemed confirmed by the inverse response pattern that was observed in related word pairs. A closer look at our stimuli pointed toward the importance of including a large enough filler category with mid- and low-frequency words to mask critical pairs and discourage strategical responses. Fourth, the GJT turned out to be very demanding because of the 3,250 ms cutoff. Although this methodological choice was informed by previous research and seemed acceptable in the piloting phase of the study, younger participants may suffer more quickly than adults from frustration and test fatigue, especially in test batteries where the most demanding tasks are given last. As such, the task might have been more informative if more time was provided. Lastly, as we considered only one part of speech (i.e., verbs), caution regarding the generalizability of the findings to other parts of speech is warranted. Moreover, the use of the infinitive form of the target verbs in the decontextualized word-focused treatment (both in the word lists and in most of the vocabulary exercises) may have been at the advantage of the decontextualized word-focused group for the form-related tests (form recognition and LDT) given that the stimuli used in these test were presented in the infinitive form. In addition, using the infinitive form in the meaning recall test and SEMREL may also have facilitated the identification of the target verbs in the decontextualized group.

CONCLUSION

This study assessed the value of combining explicit and sensitive measures to gauge the impact of ecologically valid L2 vocabulary instruction. On the theoretical level, results indicate that sensitive measures can complement explicit measures, in that they provide additional insights into learning effects related to lexical processing. On the pedagogical level, this study advocates a balanced approach to L2 vocabulary teaching, with opportunities for decontextualized word-focused instruction supplemented with a combination of word-focused and meaning-oriented receptive and productive activities.

https://doi.org/10.1017/S0272263120000431 Published online by Cambridge University Press

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0272263120000431.

REFERENCES

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

Bley-Vroman, R., & Masterson, D. (1989). Reaction time as a supplement to grammaticality judgements in the investigation of second language learners' competence. *University of Hawai'i Working Papers in ESL*, *8*, 207–237.

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meaning of words: An instance-based learning approach. *Discourse Processes*, *45*, 122–159. https://doi.org/10.1080/01638530701792826

Bretz, F., Hothorn, T., & Westfall, P. (2011). *Multiple comparisons using R*. Chapman & Hall.

Chen, C., & Truscott, J. (2010). The effects of repetition an L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, *31*, 693–713.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Coxhead, A. (2011). "What is the exactly word in English?": Investigating second language vocabulary use in writing. *English Australia Journal*, *27*, 3–16.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

DeKeyser, R. M. (2007). Situating the concept of practice. In R. M. DeKeyser (Ed.), *Practice in a Second language: Perspectives from applied linguistics and cognitive psychology* (pp. 1–18). Cambridge University Press.

De la Fuente, M. (2006). Classroom L2 vocabulary acquisition: Investigating the role of pedagogical tasks and form-focused instruction. *Language Teaching Research*, *10*, 263–295. https://doi.org/10.1191/1362168806lr196oa

De Neys, W., d'Ydewalle, G., Schaeken, W., & Vos, G. (2002). A Dutch, computerized, and group administrable adaptation of the operation span test. *Psychologica Belgica*, *42*, 177–190.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*, 367–413. https://doi.org/10.1111/j.1467-9922.2010.00613.x

Elgort, I. (2018). Technology-mediated second language vocabulary development: A review of trends in research methodology. *Calico Journal*, *35*, 1–29. https://doi.org/10.1558/cj.34554

Elgort, I., Candry S., Boutorwick, T., Eyckmans, J., & Brysbaert, M. (2018). Contextual word learning with form-focused and meaning-focused elaboration. *Applied Linguistics*, *39*, 464–667. https://doi.org/10.1093/applin/amw029

Ellis, R. (2009). Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics*, *19*, 221–246. https://doi.org/10.1111/j.1473-4192.2009.00231.x

Erlam, R. (2016). "I'm still not sure what a task is": Teachers designing language tasks. *Language Teaching Research*, *20*, 279–299. https://doi.org/10.1177/1362168814566087

File, K., & Adams, R. (2010). Should vocabulary instruction be integrated or isolated? *TESOL Quartlerly*, *44*, 222–249. https://doi.org/10.5054/tq.2010.219943

Frishkoff, G., Perfetti, C., & Collins-Thompson, K. (2011). Predicting robust vocabulary growth from measures of incremental learning. *Scientific Studies of Reading*, *15*, 71−91. https://doi.org/10.1080/10888438.2011.539076

Fukkink, R., Hulstijn, J. H., & Simis, A. (2005). Does training in second-language word recognition skills affect reading comprehension? An experimental study. *The Modern Language Journal*, *89*, 54–75. https://doi.org/10.1111/j.0026-7902.2005.00265.x

Godfroid, A. (2020). Sensitive measures of vocabulary knowledge and processing: Expanding nation's framework. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 433–453). Routledge.

Godfroid, A., Loewen, S., Jung, S., Park, J. H, Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition*, *37*, 269–297. https://doi.org/10.1017/S0272263114000850

González-Fernández, B., & Schmitt, N. (2019). Word knowledge: Exploring the relationship and order of acquisition of vocabulary knowledge components. *Applied Linguistics*, *41*, 481–505. https://doi.org/10.1093/applin/amy057

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*, 346–363. https://doi.org/10.1002/bimj.200810425

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*, 539–558. https://doi.org/10.1111/0023-8333.00164

Hummel, K. (2010). Translation and short-term L2 vocabulary retention: Hindrance or help? *Language Teaching Research*, *14*, 61–74. https://doi.org/10.1177/1362168809346497

Ionin, T., & Zyzik E. (2014). Judgment and interpretation tasks in second language research. *Annual Review of Applied Linguistics*, *34*, 37–64. https://doi.org/10.1017/S0267190514000026

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13.

Laufer, B. (2006). Comparing focus on form and focus on forms in second-language vocabulary learning. *The Canadian Modern Language Review/La revue canadienne des langues vivantes*, *63*, 149–166. https://doi.org/10.3138/cmlr.63.1.149.

Laufer, B. (2017). The three "I"s of second language vocabulary learning: Input, instruction, involvement. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning. Volume III* (pp. 343–354). Routledge.

Laufer, B. (2020). Evaluating exercises for learning vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 351–368). Routledge.

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

Laufer, B., & Rozovski-Roitblat, B. (2015). Retention of new words: Quantity of encounters, quality of task, and degree of knowledge. *Language Teaching Research*, *19*, 687–711. https://doi.org/10.1177/1362168814559797

Lee, S., & Pulido, D. (2017). The impact of topic interest, L2 proficiency, and gender on EFL incidental vocabulary acquisition through reading. *Language Teaching Research*, *21*, 118–135. https://doi.org/10.1177/1362168816637381

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207. https://doi.org/10.1111/lang.12117

Littlewood, W. (2014). Communication-oriented language teaching: Where are we now? Where do we go from here? *Language Teaching*, *47*, 349–362. https://doi.org/10.1017/S0261444812000134

Llach, M. P. A. (2009). The effect of reading only, reading and comprehension, and sentence writing in lexical learning in a foreign language: Some preliminary results. *Revista Española de Lingüística Aplicada*, *22*, 9–33.

Loewen, S. (2015). *Introduction to instructed second language acquisition*. Routledge.

Lonsdale, D., & Le Bras, Y. (2009). *A frequency dictionary of French core vocabulary for learners*. Routledge.

Mason, B., & Krashen, S. (2010). A reader response to File and Adams's "The reality, robustness, and possible superiority of incidental vocabulary acquisition." *TESOL Quarterly*, *44*, 790–793. https://doi.org/10.5054/tq.2010.238721

Nation, I. S. P. (2013). *Learning vocabulary in another language*. 2nd ed. Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Heinle.

Noreillie, A. S. (2019). *It's all about words*. Three empirical studies into the role of lexical knowledge and use in French listening and speaking tasks (Unpublished doctoral dissertation). KU Leuven, Leuven.

Pellicer-Sánchez, A. (2015). Developing automaticity and speed of lexical access: The effects of incidental and explicit teaching approaches. *Journal of Spanish Language Teaching*, *2*, 126–139. https://doi.org/10.1080/23247797.2015.1104029

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary learning from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, *22*, 31–55.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes-No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*, 489–509. https://doi.org/10.1177/0265532212438053

Peters, E. (2012). The differential effects of two vocabulary instruction methods on EFL word learning: A study into task effectiveness. *International Review of Applied Linguistics*, *50*, 213–238. https://doi.org/10.1515/iral-2012-0009.

Peters, E., Hulstijn, J., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German vocabulary through reading: The effect of three enhancement techniques compared. *Language Learning*, *59*, 113–151. https://doi.org/10.1111/j.1467-9922.2009.00502.x

Peters, E., Velghe, T., & Van Rompaey, T. (2019). The development of an English and French vocabulary test. *Internation Journal of Applied Linguistics*, *170*, 53–78. https://doi.org/10.1075/itl.17029.pet

R Core Team (2018) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Rice, C. A., & Tokowicz, N. (2019). A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition*, *42*, 439–470. https://doi.org/10.1017/S0272263119000500

Rott, S. (2004). A comparison of output interventions and un-enhanced reading conditions on vocabulary acquisition and text comprehension. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, *61*, 169–202. https://doi.org/10.3138/cmlr.61.2.169

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*, 329–363. https://doi.org/10.1177/1362168808089921

Schmitt, N. (2010). *Researching vocabulary. A vocabulary research manual*. Palgrave Macmillan.

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*, 261–274. https://doi.org/10.1017/S0261444819000053

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*, 484–503. https://doi.org/10.1017/S0261444812000018

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

Sun, C. H. (2017). The value of picture-book reading-based collaborative output activities for vocabulary retention. *Language Teaching Research*, *21*, 96–117. https://doi.org/10.1177/1362168816655364

Van den Branden, K. (2016). The role of teachers in task-based language education. *Annual Review of Applied Linguistics*, *36*, 164–181. https://doi.org/10.1017/S0267190515000070

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*, 33–52. https://doi.org/10.1017/S0272263105050023

Webb, S. (2007). Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research*, *11*, 63–81. https://doi.org/10.1177/1362168806072463

Webb, S. (2009). The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal*, *40*, 360–376. https://www.jstor.org/stable/44486803

Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., & Piasecki, A. (2018). Re-examining the effects of word writing on vocabulary learning. *International Journal of Applied Linguistics*, *169*, 72–94. https://doi.org/10.1075/itl.00007.web

Zou, D. (2017). Vocabulary acquisition through cloze exercises, sentence-writing and composition-writing: Extending the evaluation component of the involvement load hypothesis. *Language Teaching Research*, *21*, 54–75. https://doi.org/10.1177/136216881665241