Bilingualism: Language and Cognition 22 (3), 2019, 500–515 doi:10.1017/S1366728918000019

Noticing vocabulary holes aids incidental second language word learning: An experimental study*

JOHANNA F. DE VOS
Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands
HERBERT SCHRIEFERS
Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
KRISTIN LEMHÖFER
Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

(Received: February 04, 2017; final revision received: January 02, 2018; accepted: January 06, 2018; first published online 31 May 2018)

Noticing the hole (NTH) occurs when speakers want to say something, but realise they do not know the right word(s). Such awareness of lacking knowledge supposedly facilitates the acquisition of the unknown word(s) from later input (Swain, 1993). We tested this claim by experimentally inducing NTH in a second language (L2) for some participants (experimental), but not others (control). Then, in a price comparison game, all participants were exposed to spoken L2 input containing the to-be-learned words. They were unaware of taking part in an L2 study. Post-tests showed that participants who had noticed holes in their vocabulary had indeed learned more words compared to participants who had not. This held both for the experimental group as well as those participants in the control group who later reported to have noticed holes. Thus, when we become aware of vocabulary holes, the first step to improve our vocabulary is already taken.

Keywords: noticing, second language acquisition, word learning, incidental learning, mixed-effects model

Introduction

Second language (L2) learners often fail to exactly express their intended message, due to a lack of knowledge of the target language vocabulary. This is especially poignant in real-life conversations, where there is little occasion to consult a dictionary whilst speaking. Although learners are usually able to talk around their lacking word knowledge, the forced resort to circumlocution may not go unnoticed by the learners themselves.

While the awareness of being at a loss for words may be frustrating, it may well be beneficial to the second language acquisition (SLA) process. This possibility underlies one of the four hypothesised functions of output, according to Swain's Output Hypothesis (1985, 1993, 1995, 1998), namely its NOTICING FUNCTION (the other functions would be practicing, hypothesis testing, and the metalinguistic function). When learners fail to produce target language output, be it vocally or subvocally (Swain, 1995, p. 125), this "may prompt [them] to consciously recognize some of their linguistic problems" (Swain & Lapkin, 1995, p. 373). In turn, this could trigger cognitive processes involved in SLA, such as a heightened state of attention for subsequent input (Swain & Lapkin, 1995, p. 386), which may be beneficial to learning.

Swain's use of the term NOTICING differs from how it was originally used by Schmidt (1990) in his Noticing Hypothesis, which states that noticing would be a necessary condition for language learning. Schmidt (2001, p. 4) equates noticing to "awareness at a very low level of abstraction": learners' awareness of specific instances in the language input. For example, learners may notice how native speakers use a particular form in the target language (Izumi, 2013, p. 38). If learners also compare their own imperfect use of that form to the way the more proficient speaker used the form in the input, this is called NOTICING THE GAP (Schmidt & Frota, 1986). We will use the term NOTICING (THE GAP) to catch both of Schmidt's constructs in one phrase.

While noticing (the gap) concerns learners interacting with external language input, Swain's noticing function of output comes into play when learners struggle to produce language, regardless of whether the output is vocalised or not. This applies to both grammatical structures and words. In this study, we will focus on the latter. When learners become aware that an L2 target word is completely absent in their vocabulary, this is

* This research was conducted as part of a VIDI project (grant 276-89-004) awarded to Kristin Lemhöfer by the Netherlands Organisation for Scientific Research (NWO). The authors would like to thank Louis ten Bosch, Conor Dolan and Sean Roberts for their advice on the statistical analysis, and Rebecca Sachs and three anonymous reviewers for their helpful comments on earlier versions of this manuscript.

Address for correspondence:
Johanna de Vos, PO Box 9104, 6500HE Nijmegen, The Netherlands
johannadevos@gmail.com

Supplementary material can be found online at https://doi.org/10.1017/S1366728918000019

called NOTICING THE HOLE IN ONE'S INTERLANGUAGE (e.g., Doughty and Williams, 1998, p. 255). When learners struggle to produce a word they have incomplete knowledge of, it is called NOTICING THE GAP IN ONE'S ABILITY (Izumi, 2013, p. 40).

Importantly, 'noticing the gap in one's ability' is not the same as Schmidt and Frota's (1986) 'noticing the gap', because the former happens learner-internally and the latter in relation to external input. To avoid confusion in terminology, in this article we will speak of NOTICING THE HOLE (NTH) when referring to situations where learners struggle to produce output and become aware of their linguistic problem, be it because a word is completely absent in their vocabulary (a hole in one's interlanguage), or because it is only partially represented (a gap in one's ability).

How can the hypothesised facilitative effects of NTH on vocabulary learning be explained in terms of cognitive mechanisms? Imagine a learner making an unsuccessful attempt to produce a word, thereby experiencing NTH. Suppose that this learner is subsequently exposed to this word. It is hypothesised that the learner will remember the word more readily, as compared to a situation in which he/she did not notice the hole before being exposed to input. This would be an instance of the PRE-TESTING EFFECT observed in memory experiments, where an unsuccessful retrieval attempt before exposure to the relevant materials enhances learning (Grimaldi & Karpicke, 2012; Kornell, Jensen Hays & Bjork, 2009; Richland, Kornell & Kao, 2009).

Several explanations for the pre-testing effect have been offered, including the impact of unsuccessful retrieval on intentional learning behaviour (Richland et al., 2009): it could well be that failure to produce a word alters intentional learning behaviour by fostering EPISTEMIC CURIOSITY, i.e., "the desire for knowledge that motivates individuals to [...] eliminate information-gaps" (Litman, 2008, p. 1586). In turn, humans are better at learning information they are curious about (Gruber, Gelman & Ranganath, 2014; see also Kang et al., 2009). Gruber et al. (2014) name attentional processes as one potential explanation of the relationship between curiosity and learning (although they also mention it likely there are other factors too). For three retrieval-based explanations of the pre-testing effect, see Kornell et al. (2009).

Literature review: From NTH to SLA

In the present study, we experimentally manipulated NTH by confronting German learners of Dutch with their lacking L2 vocabulary knowledge. The study will be introduced in more detail in the next section. Before doing so, we present a literature review of other experimental studies concerning NTH and SLA (for two observational studies, see Hanaoka, 2007, and Hanaoka & Izumi, 2012).

Izumi, Bigelow, Fujiwara and Fearnow (1999) and Izumi and Bigelow (2000) studied the acquisition of the English past hypothetical conditional (e.g., "If Ann had traveled to Spain in '92, she would have seen the Olympics", Izumi et al., 1999, p. 426). Specifically, they investigated whether the anticipation of an output task (here: knowing that one later has to do a writing task), and the actual execution of this output task, lead to noticing and improved acquisition of the target structure.

Izumi et al. (1999, p. 423) indicate that what they call 'noticing' actually encompasses two separate processes: noticing "problems in one's interlanguage" (what we call NTH), and noticing "the relevant features in the input" (what we call noticing (the gap)). 'Noticing' was measured by letting the participants read a text containing the target structure, and asking them to underline the parts they thought were relevant to their upcoming activity. Only for the experimental group, the upcoming activity was an output task. The control group answered comprehension questions about the text. After completing their respective activities, the participants did the underlining task again.

In neither experiment did the groups differ significantly in their underlining behaviour. Thus, neither the anticipation of an output task, nor the (presumed) experience of NTH during such an output task, resulted in the learners noticing (the gap to) the target structure more often. Regarding the acquisition of the target structure, the experimental group did significantly outperform the control group in one contrast (out of many) in the 1999 study, with a large effect size of $d = 1.36$.[1] However, one should note that these studies seem to be at risk of both Type-I and Type-II errors, because no correction for multiple testing was applied, and overall sample sizes were rather small ($N = 22$ in 1999, and $N = 18$ in 2000).

A very similar study was conducted by Song and Suh (2008), using the same target structure and tasks. One additional experimental output group was added, which (supposedly) noticed holes through a picture-cued writing task that required use of the target structure. In this study, the participants in the two experimental groups did underline significantly more conditional-related items than the control participants who did not produce written output. It did not matter whether the underlining task took place before or after the output activity. Thus, in this study, anticipating and experiencing NTH in an output task increased the participants' noticing (the gap to) the target structure. It was also shown that scores on a post-test production task were higher in the experimental groups than in the control group ($d = 0.72$ and $d = 0.95$). However, differences on a recognition task were absent. The authors do not address potential reasons for

[1] All effect sizes (expressed as Cohen's *d*) mentioned in the introduction were calculated by the first author with data from the articles.

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

the discrepancies between the outcomes of this study and the earlier studies by Izumi and colleagues.

Two issues relating to the above studies need to be discussed. Firstly, it may be that the activities in the experimental and control groups following exposure to the target structure differed in depth of processing. That is, when writing a text and thereby reproducing the target structure, this structure is likely to be processed more deeply than when answering comprehension questions. The positive relationship between depth of processing and learning has long been posited (Craik & Lockhart, 1972; Craik & Tulving, 1975) and supported, also for SLA (Laufer & Hulstijn, 2001; Leow, 2015). Therefore, potential differences between the experimental and control groups might to some extent be due to different depths of processing, rather than NTH and noticing (the gap) exclusively.

Differences in processing debt are indeed mentioned in Izumi and Izumi (2004), which is another study that used the above-described design. Unexpectedly, the researchers found that their control group improved more on the target structure than their experimental group. In the discussion, Izumi and Izumi (2004) concede that differences in processing depth may have contributed to this unexpected finding. As of yet, however, such alternative explanations cannot be empirically evaluated, because depth of processing was not measured in any of the above studies. Therefore, if researchers choose to use different treatments for the experimental and control groups, they should ideally include measurements of depth of processing to evaluate such alternative explanations.

Secondly, the adequacy of underlining as a measure of noticing (the gap) is questionable. Song and Suh (2008, p. 308) remark that this method may not be suitable "for tapping into learners' noticing and attention" and that think-aloud or stimulated recall protocols might provide a better solution. Izumi and Bigelow (2000, pp. 270–271) admit that one cannot be sure that underlining captures all items that were attended to, nor that it excludes items that were not attended to. For future studies, they recommend triangulation with other measures.

Such a triangulation was performed by Uggen (2012). Her design was very similar to Izumi and Bigelow (2000), again revolving around the past hypothetical conditional and (the anticipation of) output tasks. This time, there was an additional experimental group, which was trained and tested on the present hypothetical conditional. For the triangulation of noticing measurements, Uggen (2012) also analysed the participants' essays qualitatively, and added stimulated recall. Having finished the experimental procedure, her participants watched a video recording of the experimental session and commented on the thoughts they had had at the time. This stimulated recall measurement proved especially valuable, as it showed that in one experimental group the participants also commented on grammatical features that they had not underlined. The underlining measurement itself again was not very useful, as no differences in underlining could be detected between the two experimental groups and the control group. With regard to acquisition, the experimental group that was assigned the past hypothetical conditional showed significant improvement on this structure. The other experimental group, assigned the present hypothetical conditional, did not improve. According to Uggen (2012, p. 533), perhaps this happened because this structure was less complex and therefore less "noticeable" to the learners.

In summary, Uggen's (2012) study suggests that written output influences learners' "awareness of their linguistic limitations concerning grammar structures" (p. 506). Considering all studies discussed so far, it seems that NTH can benefit the acquisition of L2 grammatical structures, but that these structures need to be of a certain complexity. Furthermore, to measure noticing (the gap), triangulation of measurements is recommended. Uggen (2012) showed that underlining alone does not suffice.

It should be noted that so far we have only discussed studies on L2 grammar learning. The outcomes of these studies may not be directly transferrable to word learning, as grammar learning revolves around learning a rule or pattern, while vocabulary requires memorising word forms. However, the different types of noticing that were discussed above are equally relevant to grammar and word learning. After all, both types of learning can be expected to depend on a learner's attention to input (noticing (the gap)), and the learner's awareness of his/her own state of knowledge (NTH). The current study focuses on NTH. To our knowledge, there are only two studies on the effects of NTH on vocabulary learning, both of which focused on the written domain (Kwon, 2006, and Mahmoudabadi, Soleimani, Jafarigohar & Iravani, 2015). Both studies manipulated the order in which participants performed output and input tasks.

In the input task in Mahmoudabadi et al. (2015), the participants connected written words with their corresponding pictures. In the output task, the participants had to name the same pictures, but without a word list. In Kwon (2006), the input and output tasks comprised a variety of activities. The input tasks were reading a text and answering comprehension questions, looking at pictures and answering comprehension questions, and a word recognition task. The output tasks were fill-in-the-blank, answering open questions, and narrative writing. In both studies, it was assumed that the output tasks would elicit NTH (when participants failed to produce the target words). Thus, the participants in the input-before-output conditions were exposed to input containing the target vocabulary BEFORE having noticed holes, and the

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

participants in the output-before-input conditions AFTER having noticed holes.

Vocabulary post-tests were administered after the completion of all output and input tasks. Mahmoudabadi et al. (2015) found a significant facilitative effect (d = 0.98) of NTH, i.e., more word learning in the output-before-input than in the input-before-output condition. Kwon (2006) found no significant effect of NTH. Her preferred explanation (pp. 118–120) for this null result, reminiscent of Doughty's (2001) COGNITIVE WINDOW, is that the delay between the output and input tasks was too long. This may have weakened any potential effects of NTH.

Leow (1999, p. 66) has pointed out that it cannot automatically be assumed that participants will behave according to the experimental instructions or the experimenter's expectations. Accordingly, in both studies a subsample of the participants was interviewed after the treatment and post-tests. Mahmoudabadi et al. (2015) explicitly asked ten participants in the output-before-input condition (out of 43) whether they felt the need to know the words when doing the output task. All said yes. Kwon (2006) interviewed a total of ten participants (sampled from both task orders, out of 80). From the excerpts provided, it seems that at least some of the participants in the output-before-input conditions realised they did not know the words they needed during the output tasks, and became motivated to find them in the input. It is unclear whether this applies to all participants.

As in the grammar studies, in the vocabulary studies too there seems to be a confound between the NTH manipulation and opportunities for processing the input. The groups did not only differ (as intended) in whether or not the participants were expected to notice holes before exposure to input, but also in their opportunities to process that input. Only the input-before-output group could have benefitted from the retrieval of words from memory during the output task, which has been shown to facilitate vocabulary learning and retention (Barcroft, 2007). The output-before-input groups did not have this opportunity for retrieval practice, as there was nothing yet to retrieve. This difference between the conditions is conceptually distinct from, but confounded with, the NTH manipulation.

In conclusion, while we do not doubt the relevance of the above-discussed studies for L2 pedagogy, their design makes it difficult to isolate the true effect of NTH on SLA. The present study therefore employed an experimental design in which, after the NTH manipulation, exposure to input and opportunities to process that input were identical in all conditions. Still, keeping Leow (1999) in mind, we realised that we could not assume that NTH happens whenever researchers create a setting where it is expected to occur, and does not happen otherwise. Perhaps this could also explain why some of the above studies did not find significant effects of NTH. To check whether our manipulation worked as expected, we interviewed our participants regarding their NTH experience after the experiment.

The present study: NTH in incidental L2 word learning

We addressed the questions of whether NTH in spoken L2 word production facilitates the acquisition of these words from spoken input, and how well these words are retained over a short period of time. The participants were German native speakers, with Dutch being the L2. We used a task that was advertised as a price judgment task, but unbeknownst to the participants was seeded with low-frequency non-cognate Dutch words. This allowed us to investigate L2 vocabulary learning (more details will follow in the Methods section).

To induce NTH in the experimental condition, we asked the participants to name the objects in Dutch. We expected that the inability to name a given object would result in NTH. Post-experiment interviews showed that this expectation was correct. In contrast, the participants in the control condition inspected the same objects, but did not name them. This ensured that both groups were equally familiar with the materials. The expectation that this silent inspection would NOT result in NTH was also checked in post-experiment interviews. In fact, it was found that about half of the participants in the control condition had noticed holes after all. Following Leow (2000), we analysed their data separately (see Analysis). To this end, we tested more participants in the control condition, such that we could form separate groups of participants who did and who did not report experiencing NTH. In this way, we could not only assess the effect of the external, experimental induction of NTH, but also that of the spontaneous internal occurrence of NTH when it was not experimentally induced.

Having named or silently studied the pictures (i.e., after the NTH manipulation), both groups underwent the same procedure. Specifically, the participants were exposed to naturalistic L2 input from a Dutch native speaker in the form of price comparisons. The input contained, in a highly controlled way, the names of the objects previously unknown to the participants. The participants were unaware that they were expected to learn from this input and would later be tested on it.

After the exposure to input, the participants took two unannounced post-tests (immediately and after 15 minutes) to measure how many words they had learned and retained. The 15-minute interval allows us to study the earliest stages of the forgetting curve, as shown for the first time in a classic experiment by Ebbinghaus (1885/1913/2011). Ebbinghaus memorised lists of nonsense syllables (e.g., *zup*). Having studied the lists until he reached a score of 100% correct, 20

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

minutes later he could only remember 58%. This shows how rapidly newly-acquired knowledge can decay.

Post-experiment interviews confirmed that the participants were indeed unaware of the study's language learning aspect. Thus, with this task we can approach real-life incidental L2 word learning in the laboratory, while maintaining a high degree of experimental control.

Methods

Participants

The participants were 70 German students in Nijmegen, the Netherlands. Crucially, they did not know the study was targeted at German native speakers, as it was advertised as a psychological experiment about making price judgments. Non-German participants were prevented from signing up through a hidden language filter in the online participant recruitment system. Thus, the participants were fully naive regarding the language aspects of the study.

The participants were randomly assigned to the experimental and control condition. In the experimental condition, the participants tried to produce output before being exposed to the target words, and therefore noticed holes (as confirmed through interviews at the end of the experiment). We will call this condition [+O, +NTH] (see Procedure for more details on the manipulation). It should be noted that "+O" mainly reflects situations where participants TRIED to vocally produce output, but actually failed to do so. In the control condition, the participants were not required to produce output before the exposure to input, and thus were supposed not to notice holes: [−O, −NTH]. However, the post-experiment interviews revealed that almost half of the participants in the control condition had nevertheless noticed holes, as they had internally tried to name the target items. These participants were assigned to a new, third condition, which was called [−O, +NTH]. Thus, while +/- O was experimentally controlled, +/- NTH resulted from individual differences (in the original control condition only, as everyone in the experimental condition noticed holes). Testing was continued until all three conditions included a minimum of 20 participants whose data could be used.

Four participants were excluded from the analysis because they indicated during the second post-test that they had already actively known more than 25% of the target words before the experiment (see Debriefing and measures). One additional participant was excluded because he had not understood the price judgment task. The final sample thus included 65 participants (51 females), who had all been raised with German as their only native language. The participants' mean age was 22 (range 19–27); they had started learning Dutch at a mean age of 19 (range 16–24). All but one were, at the time of this study, taking higher education courses taught in Dutch, or had done so in the past. In addition to German and Dutch, all participants reported knowledge of English, and some reported knowledge of additional languages. None of the participants in the final sample guessed the purpose of the study (Dutch word learning) during debriefing.

The participants in the three conditions were compared by means of one-way independent ANOVAs on a number of dimensions that could potentially influence L2 word learning (see Table 1). Prior Dutch vocabulary size was determined with the Dutch version of the LexTALE vocabulary test (www.lextale.com; also see Lemhöfer & Broersma, 2012). To get an impression of the participants' motivation and strategy use in learning Dutch, they were asked to rate a number of statements (shown to them in German) on a 1–5 scale. We selected four of these statements for our analysis, namely: 1) "It is important to me to have a large Dutch vocabulary", 2) "The way in which something is said is not important to me, only what it means", 3) "When I hear a Dutch word I do not know, I try to learn it", and 4) "I pay attention to subtle differences between German and Dutch". All variables except LexTALE and Passive knowledge of target words were gathered through a background questionnaire that the participants completed after the experiment (see Debriefing and measures). Table 1 shows no significant differences between the groups in any of the measures (all $p > .13$).

Materials

The target words were 16 infrequent names of concrete objects that are typically unknown in L2 Dutch for German native speakers. All were non-cognates between Dutch and German, for example *garde* (German: *Schneebesen*, English: *whisk*). There were also 44 filler words that the participants should already have known. These were used to distract from the learning purpose of the study, and because we did not want to present more than one target item in a trial. The fillers were common objects (e.g., an apple). Their cognate status was not controlled. All targets and fillers were depicted through photographs. These had been found on the internet and edited in Photoshop. They were cropped to squared pictures and any words or brand logos were removed.

The complete item list can be found in the online supplementary materials (Supplementary Materials, S1). In the interest of the price judgment cover story, the words came from four semantic categories (household, clothing, tools and toys). Each category contained four target words and eleven filler words. Item selection was based on pre-test data from a similar study using the same participant population by de Vos, Schriefers, ten Bosch and Lemhöfer (2017). Of the target words selected for this study, an

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

Table 1. *Mean scores and standard deviations (in parentheses) of participant characteristics in the three conditions.*

	[+O, +NTH] $n = 21$	[−O, +NTH] $n = 20$	[−O, −NTH] $n = 24$	Test statistics
Age	23.00 (2.21)	22.25 (1.92)	22.25 (2.23)	$F(2,62) = 0.88, p = .42$
Years of learning Dutch	3.38 (2.94)	2.66 (1.57)	2.59 (1.64)	$F(2,62) = 0.91, p = .41$
Self-rated proficiency*	3.52 (0.60)	3.55 (0.69)	3.42 (0.65)	$F(2,62) = 0.27, p = .77$
Current amount of exposure to Dutch*	3.11 (0.46)	3.55 (0.87)	3.38 (0.83)	$F(2,62) = 1.80, p = .17$
Number of other languages known	2.38 (0.59)	2.35 (0.75)	2.33 (0.76)	$F(2,62) = 0.03, p = .97$
Statement 1**	3.81 (0.87)	4.25 (0.64)	3.92 (0.93)	$F(2,62) = 1.57, p = .22$
Statement 2**	2.48 (0.87)	2.25 (1.16)	2.46 (1.18)	$F(2,62) = 0.28, p = .76$
Statement 3**	3.81 (0.87)	4.10 (0.79)	3.79 (0.78)	$F(2,62) = 0.95, p = .39$
Statement 4**	3.71 (0.85)	3.90 (0.79)	3.33 (1.17)	$F(2,62) = 2.01, p = .14$
Vocabulary size (LexTALE score)***	71.0 (5.76)	69.9 (7.50)	70.5 (8.36)	$F(2,62) = 0.12, p = .89$
Passive knowledge of target words***	7.91 (8.01)	14.79 (15.20)	10.49 (8.87)	$F(2,62) = 2.10, p = .13$

Note. Variables marked with one asterisk were self-rated on a 1–5 (1 = *very low*, 5 = *very high*) scale. Variables marked with two asterisks were self-rated on a 1–5 (1 = *strongly disagree*, 5 = *strongly agree*) scale. Variables marked with three asterisks indicate a percentage. +O vs. −O refers to required output production, +NTH vs. −NTH refers to noticing the hole.

average of 1.63% (SD = 2.94, range 0–8) was known to the participants ($N = 32$) in de Vos et al. (2017); of the fillers 98.40% were known (SD = 2.18, range = 94–100) (see S1 for the raw data, Supplementary Materials).

In the current study, we did not perform a pre-test on the participants' knowledge of the target words. This would have induced NTH in the case of unknown words, which we obviously wanted to avoid in the control condition. Furthermore, the pre-test data from our earlier study showed that the target words were only known to German learners of Dutch in very rare cases, and the filler words were practically always known. Still, all participants in the current study were asked about their pre-existing knowledge of the materials at the end of the experiment (see Debriefing and measures). This allowed us to exclude already-known target words from the analysis.

Procedure

The experiment took place in a quiet laboratory room and lasted 60–75 minutes. The participants received course credit or gift vouchers for their participation. Informed consent was obtained prior to the experiment.

Manipulation

NTH was manipulated immediately before the exposure to the target words. The participants were told that the experiment concerned a price judgment task, consisting of two parts: a sorting task and a comparison task. In the sorting task, the participants were given cards with pictures of the target and filler objects, which should be sorted according to their (subjective) price. The sorting procedure was carried out separately for the objects in each of the four semantic categories. After the participants had finished sorting the first pile of cards, they were given the opportunity to inspect their sorted cards one more time. The participants had previously been instructed that in the following comparison task they would be required to make price judgments consistent with their self-made ranking.

During this inspection of the cards, the treatment in the experimental and control conditions diverged. The experimental participants were asked to present their ranking to the experimenter by naming, vocally ([+O]) and in Dutch, the objects from the most expensive to the least expensive. We expected them to fail at naming the target objects, thereby experiencing NTH. If a participant did not know what a given object was called, he/she described it in Dutch. Later interviews confirmed that these participants all experienced NTH. In contrast, the control participants were asked to inspect their pile of cards in silence ([−O]), which we expected would not induce NTH. Yet, later interviews showed that this inspection did lead to NTH for about half of the control participants, who were reassigned to a newly created third group for analysis. After they had looked through their cards, the participants commenced the sorting procedure for the next category.

Comparison task

After the sorting task, all participants received naturalistic input containing the target words provided by the experimenter, the first author and a female native speaker of Dutch. The participant and experimenter were seated opposite each other, each in front of their own keyboard and computer monitor. On these monitors, two objects were displayed per trial, side by side, each picture sized

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

15x15 cm. As the objects appeared, the experimenter made a statement in Dutch about their relative price, starting with the left object (e.g., "a bed is more expensive than a fridge"). These statements were always reasonable, although not always in accordance with how the participants had previously sorted the cards. The participants were required to press one of two buttons to indicate whether or not the statement agreed with their previously established price ranking. No time limit was imposed for this response. Immediately after the response, two new objects appeared on the monitor. The objects always were visible to both the experimenter and participant.

There were four blocks (corresponding to the four semantic categories) with 40 trials per block. The order in which the semantic categories were presented was counterbalanced across the participants, and corresponded to the order of the sorting task. The position of slots for target and filler words in the trial list was fixed (see Supplementary Materials, S2), but the assignment of actual target and filler words to slots was random. Each trial contained at most one target. Each target object appeared equally often in the left or right slot. Trials containing targets were always separated by at least one trial with two fillers. Each target object (four per semantic category) was presented four times, with an inter-trial interval of four trials between the first and second, and between the third and fourth exposure. The inter-trial interval between the second and third exposure was 14 trials. The eleven fillers (per category) each appeared five or six times. Each block had a duration of approximately four minutes, and the blocks were separated by a short break.

Debriefing and measures

Following the comparison task, the participants were asked what they thought the study was about. We asked the question at this point to avoid the participants' responses becoming biased by having taken an explicit vocabulary test. After their response, they were told that the experiment was about word learning and they would therefore take a vocabulary test next. This was the first mention of the vocabulary test, which measured immediate learning gains. All objects, including the fillers, were presented successively on the computer screen, in four blocks (in the same order as before). The order of items within blocks was randomised. The participants were instructed to (try to) name the objects, and received no feedback concerning their response.[2]

After this vocabulary test, the participants filled in a questionnaire about their experience with learning Dutch and other languages. Then, they completed the Dutch version of the LexTALE vocabulary test.

Next, and about 15 minutes after the first vocabulary test, the participants were shown all the target objects (but not the fillers) once more. In this delayed vocabulary test, the participants tried again to name the objects. After each trial, the experimenter provided the correct answer, and asked whether the participant had had passive or active knowledge of the word before taking part in the experiment.

Then, the participants were interviewed to verify whether the NTH manipulation had worked as intended. Initially, we had started by asking the first participants a general question about their experience during the sorting task. However, the participants usually commented on prices rather than on NTH. We then asked them a more specific question (after a while, we stopped asking the first, unspecific question). For the experimental group, the question was: "When naming the pictures after sorting them, did you notice you were not able to produce some names?"

The control group was asked: "When looking at the pictures after sorting them, did you name the objects in silence?" "If yes, in what language?" "If in Dutch, did you notice you were not able to produce some names?" If the control group said no to the first question, the follow-up questions were not asked. We assumed that not trying to name pictures automatically meant that no NTH took place. We now consider this to be a limitation of the current study, as it would have been better to check this assumption explicitly.

Analysis

Reassignment of participants to conditions

As explained earlier, the participants in the control condition were divided into two subgroups for analysis, on the basis of the participants' self-reported experience of NTH. If participants reported that they had subvocally tried to name the objects in Dutch and noticed holes, they were assigned to the [−O, +NTH] group. This includes participants who reported using a combination of Dutch and German for subvocal naming. If participants reported they had exclusively subvocally named the objects in German (their L1) or had not named them at all, they were assigned to the [−O, −NTH] group.

Data preparation

The target words of which the participants had reported pre-existing active knowledge were excluded from the

[2] As one reviewer remarked, the immediate post-test would generate NTH in all groups, including [−O, −NTH]. This is inevitable when conducting a vocabulary test, but it confounds the retest performance of the three groups. However, we do not consider this a problem, because the hypothesised explanation of NTH's facilitative effects rests on how people process the input AFTER noticing holes, and no input was offered in between the two vocabulary tests.

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

Table 2. *A target word (*rammelaar*, English:* rattle*) and a participant's production of this word, phonetically transcribed.*

Target word	r	ɑ	m	ə	l	a:	r	
Participant's production	r	ɑ	m		l	ə	r	t
Scoring	correct	correct	correct	incorrect (deletion)	correct	incorrect (substitution)	correct	incorrect (insertion)

analysis.[3] The target words of which the participants had pre-existing passive knowledge were not excluded, because passive knowledge does not preclude word form learning for active production. To take into account any potential effects of passive knowledge on word learning, this information was included in the analysis (see Modelling).

Scoring

Learner productions were compared to target productions based on phonological similarity. To this end, we transcribed all learner productions with the DISC phonetic transcription system (Burnage, 1990), which captures every sound of Dutch in one ASCII character, including diphthongs. Details about the phonetic transcription can be found in the online supplementary materials (Supplementary Materials, S3).

Target word responses were scored at the phoneme level. This was preferred to a binary correct/incorrect score, as some word productions were partially correct (e.g., a participant saying *ramlert* to the target *rammelaar*, English: *rattle*). Instead, we counted the number of correctly and incorrectly produced phonemes. Following Levenshtein (1966), deletion, substitution and insertion of phonemes were considered incorrect. In the scoring process we employed long alignment, which lets the same phonemes appear as corresponding segments (see Heeringa, 2004). Table 2 exemplifies the scoring procedure for the *ramlert* example.

Ramlert would be counted as yielding five correct and three incorrect phonemes; the corresponding dependent variable for the statistical model for this particular production would in principle be the vector (5,3). However, the target's actual word length is 7 phonemes. Because we used a binomial probability distribution to predict the number of correct and incorrect phonemes (see Modelling), which does not allow word length to vary within words, we would adjust the final score to be (4,3). A more comprehensive explanation of this issue can be found in S3 (Supplementary Materials), but it should be noted that, for 96.4% of the responses, the length of the word produced by the participant was equal to the original word length. For the purpose of providing descriptive statistics, the original vector of correct and incorrect phonemes was also converted into a percentage. This percentage is the number of correct phonemes out of the total number of phonemes (longest alignment). In the *ramlert* example: 5 / (5+3) * 100% = 63%.

Modelling

We analysed the data using generalised two-level mixed-effects models of the binomial family with the *lme4* package (Bates, Mächler, Bolker & Walker, 2015) in *R* (R Core Team, 2013). The models were fitted by maximum likelihood estimation, using the logit link function. The vector with the number of correct and incorrect phonemes for each target word utterance was used as the dependent variable. This vector approach to the analysis of proportion data is described in Crawley (2007), and solves four problems that are associated with the alternative of using percentages as a dependent variable (Crawley, 2007, pp. 569–570).

Included as fixed effects were Condition (three levels: [+O, +NTH], [−O, +NTH], [−O, −NTH]), Testing moment (two levels: Immediate, Delayed), and their interactions. As random effects, we included random intercepts for Participant ($N = 65$) and Word ($N = 16$).

Using this model as a basis, we explored whether its fit to the data could be improved by including random slopes of Testing moment over Participant and Word, which allows for the potential scenario that not all participants or words are equally affected by the 15-minute delay. The results are reported below. We also explored some fixed effects that were not of direct interest to our research questions, but could conceivably affect word learning. These fixed effects were Passive knowledge, the interaction between Passive knowledge and Condition, and Word length (number of phonemes) (Jalbert, Neath, Bireta & Surprenant, 2011). Passive knowledge was a self-reported measurement obtained in the delayed post-test (see Debriefing and measures). We compared the different nested models using likelihood ratio tests. Alpha was set at .05. Only in case of a significant increase in model fit, in combination with a decrease in the Akaike

[3] To check the reliability of participants' self-reported previous knowledge, we compared the naming data from the participants in the [+O, +NTH] condition, who had named all objects out loud after the sorting task, to their self-reported previous knowledge. These data converged for 99.7%.

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

Table 3. *Mean percentage of correctly produced phonemes by Condition and Testing moment, and the correlation between the two testing moments for all conditions.*

	Testing moment: Immediate				Testing moment: Delayed (15 min.)				
Condition	Mean	SD	95% CI	*n*	Mean	SD	95% CI	*n*	*r*
[+O, +NTH]	28.06	11.70	22.73–33.38	21	26.03	12.07	20.54–31.52	21	0.94
[−O, +NTH]	26.25	12.68	20.31–32.18	20	23.13	13.66	16.73–29.52	20	0.92
[−O, −NTH]	16.54	10.91	11.93–21.15	24	16.52	11.22	11.78–21.26	24	0.89
Total	23.25	12.67	20.11–26.39	65	21.63	12.77	18.46–24.79	65	0.92

Note. *n* indicates the number of participants in each condition.

Table 4. *Percentage of words that were produced fully correctly, partially correctly, and fully incorrectly (by Condition and Testing moment).*

	Testing moment: Immediate			Testing moment: Delayed (15 min.)		
Condition	Correct	Partial	Incorrect	Correct	Partial	Incorrect
[+O, +NTH]	19%	15%	66%	18%	13%	69%
[−O, +NTH]	15%	18%	67%	14%	15%	71%
[−O, −NTH]	11%	9%	80%	10%	10%	80%
Total	15%	14%	71%	14%	12%	74%

Information Criterion (AIC; Akaike, 1974), were these additional effects left in the model.

Linear mixed-effects models yield beta estimates relative to the intercept, which represents one specific combination of condition levels. To perform pairwise comparisons across all condition levels, we used the *R* package *lsmeans* (Lenth, 2016). *lsmeans* uses Tukey's method for *p*-value adjustment in multiple comparisons (Tukey, 1949). As *p*-value adjustment in (generalised) linear mixed-effects models does not seem to be standard practice in the psycholinguistic literature (although it is recommended by Quené & van den Bergh, 2004), we also provide the unadjusted *p*-values.

Results[4]

Descriptive statistics

Table 3 shows the mean percentage of correctly produced phonemes over all of the words in the experiment. As was mentioned in the Scoring section, these percentages were calculated from the vectors of correct and incorrect phonemes that are used as the dependent variable in our statistical models. In Table 3, the two levels of Passive knowledge (Yes/No) were averaged over. To ease interpretation, Table 4 shows what percentage of target words were actually produced correctly, partially correctly, and incorrectly. Table 5 is similar to Table 3, but here the scores are divided by Passive knowledge (Yes/No) rather than by Testing moment (which is now averaged over).

In the following, we will report the inferential statistics that tell us whether or not the contrasts shown in these tables reached significance. Before doing so, we will report the model comparisons leading up to the final model we used to arrive at the inferential statistics.

Model comparisons

The inclusion of a random slope of Testing moment over Participant did not significantly improve model fit ($\chi^2 = 4.12$, $df = 2$, $p = .13$). Another non-significant result was found for the random slope of Testing moment over Word ($\chi^2 = 0.62$, $df = 2$, $p = .73$). Thus, these random effects were not included in the final model.

We then explored the fixed effects. Passive knowledge significantly increased model fit ($\chi^2 = 21.64$, $df = 1$, $p < .001$, AIC decreased from 7522.9 to 7503.2), as did the subsequent addition of its interaction with Condition ($\chi^2 = 34.58$, $df = 2$, $p < .001$, AIC decreased from 7503.2 to 7472.6). Word length did not improve model fit ($\chi^2 = 1.91$, $df = 1$, $p = .17$), and was again removed from the model.

Thus, the final model was specified as follows: (PhonemesCorrect, PhonemesIncorrect) ~ 1 + Condition + Testing moment + Condition:Testing moment + Passive knowledge + Condition:Passive knowledge +

[4] All data and scripts for analysis can be downloaded from http://hdl.handle.net/11633/di.dcc.DSC_2017.00028_573.

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

Table 5. *Mean percentage of correctly produced phonemes by Condition and Passive knowledge.*

	Passive knowledge: No				Passive knowledge: Yes			
Condition	Mean	SD	95% CI	*n*	Mean	SD	95% CI	*n*
[+O, +NTH]	25.47	10.39	20.74–30.20	92.09	46.88	33.86	25.36–68.39	7.91
[−O, +NTH]	22.89	14.43	16.14–29.64	85.21	28.84	29.30	12.61–45.07	14.79
[−O, −NTH]	16.65	11.98	11.59–21.71	89.51	14.74	27.62	1.00–28.47	10.49
Total	21.42	12.72	18.27–24.57	89.35	28.01	32.00	18.39–37.62	10.65

Note. *n* indicates the mean percentage of items that were passively known or unknown in each condition.

Table 6. *Estimates, standard errors,* z*-values and* p*-values of the generalised linear mixed-effects model.*

Fixed effects	Log odds (logit)	Odds ratio	Std. error	*z*-value	*p*-value
(Intercept)	−2.29	0.10	0.41	−5.64	<.001
Condition: [+O, +NTH]	0.85	2.34	0.31	2.78	.005
Condition: [−O, +NTH]	0.62	1.86	0.31	1.98	.048
Testing moment: Delayed	−0.04	0.96	0.10	−0.46	.644
Condition: [+O, +NTH] and Testing moment: Delayed	−0.10	0.90	0.13	−0.73	.466
Condition: [−O, +NTH] and Testing moment: Delayed	−0.17	0.84	0.14	−1.23	.220
Passive knowledge: Yes	−0.34	0.71	0.17	−1.99	.047
Condition: [+O, +NTH] and Passive knowledge: Yes	1.32	3.74	0.24	5.48	<.001
Condition: [−O, +NTH] and Passive knowledge: Yes	1.00	2.72	0.23	4.41	<.001
Random effects	Variance	Std. dev.			
Participant (intercept)	0.94	0.97			
Word (intercept)	1.90	1.38			

Note. The intercept represents Condition = [−O, −NTH], Testing moment = Immediate, and Passive knowledge = No.

(1|Participant) + (1|Word). In this notation, 1 represents an intercept, : represents an interaction, and | indicates random effects.

Inferential statistics: Mixed-effects model

The estimates of our model are shown in Table 6. These estimates are approximations of the binomial parameter, which here concerns the probability that a phoneme is produced correctly. The estimates are given on the logit scale, and can be back-transformed to probabilities with the formula e^x / $(1+e^x)$, where x is the logit. To obtain the logit for a specific combination of factor levels that is not the intercept, for example for [+O, +NTH] at delayed testing with no pre-existing passive knowledge, one should add the corresponding logit estimates to that of the intercept (in this example: −2.29 + 0.85 – 0.04 – 0.10 = −1.58).

The odds ratio is a measurement of effect size. With the exception of the intercept itself, these numbers show how the odds of correctly producing a phoneme change for a specific level of a factor, as compared to the level represented by intercept. For example, for participants in the [+O, +NTH] group, the odds to correctly produce a phoneme are estimated to be 2.34 times higher than for participants in the [−O, −NTH] group (at immediate testing and with no pre-existing passive knowledge, see the paragraph below).[5]

In mixed-effects models, the intercept always represents one specific combination of factor levels. Here, it represents the [−O, −NTH] group, tested immediately, and on words for which no pre-existing knowledge was reported. From Table 6, it can be seen that [−O, −NTH] under these circumstances was significantly outperformed by [+O, +NTH] (p = .005) and by [−O, +NTH] (p = .048). However, Table 6 alone does not inform us on contrasts that do not involve the intercept (for example, if we wanted to contrast [−O, +NTH] with [+O, +NTH]). Using the *lsmeans* package (Lenth, 2016), the data have been rearranged in Table 7 to show pairwise comparisons for all Condition contrasts at both testing moments. For simplicity, the levels of Passive knowledge are averaged

[5] Unfortunately, for L2 research, no standardised guidelines for the interpretation of odds ratios exist. Different guidelines that are available suggest that 1.5/2.5/4.3 (The Effect Size, n.d.), 1.5/3.5/9 (Hopkins, 2002), or 1.68/3.47/6.71 (Chen, Henian & Chen, 2010) can be considered as small/medium/large.

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

Table 7. *Pairwise comparisons among the estimated means for all conditions, averaged over Passive knowledge (Yes/No).*

Testing moment	Contrast	Log odds (logit)	Odds ratio	Std. error	*z*-value	Unadjusted *p*-value	Adjusted *p*-value
Immediate	[+O, +NTH] – [−O, −NTH]	1.52	4.57	0.32	4.71	<.001	<.001
	[−O, +NTH] – [−O, −NTH]	1.12	3.06	0.32	3.48	<.001	.002
	[+O, +NTH] – [−O, +NTH]	0.40	1.49	0.33	1.21	.23	.45
Delayed	[+O, +NTH] – [−O, −NTH]	1.42	4.14	0.32	4.41	<.001	<.001
	[−O, +NTH] – [−O, −NTH]	0.95	2.59	0.32	2.96	.003	.009
	[+O, +NTH] – [−O, +NTH]	0.47	1.60	0.33	1.42	.15	.33

Table 8. *Pairwise comparisons of the interaction between Condition and Testing moment.*

Contrast	Log odds (logit)	Odds ratio	Std. error	*z*-value	Unadjusted *p*-value	Adjusted *p*-value
[+O, +NTH] – [−O, −NTH]	−0.10	0.90	0.13	−0.73	.47	.75
[−O, +NTH] – [−O, −NTH]	−0.17	0.84	0.14	−1.23	.22	.44
[+O, +NTH] – [−O, +NTH]	0.07	1.07	0.13	0.54	.59	.85

over. This explains why the first two odds ratios in Table 7 (4.57 and 3.06) are not the same as those reported in Table 6 (2.34 and 1.86), which only applied to Passive knowledge = No. As can be seen, the correction of *p*-values for multiple testing does not change the significance of the findings.

The pairwise comparisons tell us that participants in the [+O, +NTH] group scored significantly higher than participants in the [−O, −NTH] group, both at immediate testing ($p < .001$) and after a 15-minute delay ($p < .001$). Both odds ratios (immediate: 4.57, delayed: 4.14) can be considered of approximately medium magnitude. More concretely, as can be calculated from Table 3, at immediate testing, the number of correctly produced phonemes was 70% higher in the [+O, +NTH] group than the [−O, −NTH] group. After 15 minutes, the [+O, +NTH] participants still produced 58% more correct phonemes as compared to the [−O, −NTH] participants.

The [−O, +NTH] participants also outperformed their peers in the [−O, −NTH] group, both at immediate testing ($p = .002$) and at delayed testing ($p = .009$). These effect sizes (immediate: 3.06, delayed: 2.59) were smaller. Still, the participants in the [−O, +NTH] group produced 59% more phonemes correctly at immediate testing, and 40% after 15 minutes, as compared to their peers in the [−O, −NTH] group. Finally, no significant difference could be detected between participants in the [+O, +NTH] and the [−O, +NTH] groups, who had both noticed holes ($p = .45$ at immediate testing, and $p = .33$ at delayed testing).

With regard to Testing moment (see Table 6 again), there was no significant decay over a period of 15 minutes time ($p = .64$). Table 8 shows that the interaction between Testing moment and Condition was not significant for any of the contrasts (all uncorrected $p > .22$).

Finally, Table 6 shows an interaction between Condition and Passive knowledge. Pre-existing passive knowledge had a negative effect on the learning rate for participants in the [−O, −NTH] group ($p = .047$). The odds ratio was 0.71, which means that these participants were 1.41 (= 1/0.71) times more likely to correctly produce a phoneme in a word they had had NO pre-existing knowledge of than a phoneme in a word they had had pre-existing knowledge of. In the participants who noticed holes, pre-existing passive knowledge had a larger and positive effect (in [+O, +NTH]: $p < .001$, odds ratio = 2.67, and in [−O, +NTH]: $p < .001$, odds ratio = 1.93; these estimates were obtained through relevelling).

Discussion

In this study, we asked whether NTH (i.e., the awareness of vocabulary holes or gaps) in spoken L2 word production facilitates the acquisition of these words from subsequent spoken input in an incidental learning environment. We created this environment by conducting the experiment outside of the classroom, and in the country where the target language was spoken. The incidental aspect of the study is also reflected by the fact that none of the 65 participants in the final sample suspected that the

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

experiment was a language learning study, as we verified in post-experiment interviews.

From two to three conditions

The original design included two conditions. In the experimental condition, the participants were required to vocally produce the target words. Because they did not actually know these target words, they failed in producing them, and thereby noticed holes in their vocabulary. Thus, 'output' in the current study does not refer to language production in the typical sense, but rather to the requirement of output. The experimental participants then were exposed to input containing the unknown vocabulary.

In the control condition, the participants studied pictures without being asked to name them and therefore were supposed not to notice holes. Then, they were exposed to the same input as the experimental group. However, about half of the participants in the control condition indicated they had subvocally tried to name (some of the) objects in (L2) Dutch. Although we did not explicitly ask them whether these subvocal naming attempts had resulted in the experience of NTH (which we consider a limitation of the current study), it does seem very likely that this was the case. In other words, these participants should have experienced what Godfroid, Housen and Boers (2010) call "learner-induced noticing" (also see Park, 2007; Williams, 1999). Given this situation, we divided the control condition into two new groups for analysis: [−O, +NTH] and [−O, −NTH]. The experimental condition was renamed [+O, +NTH].

Following Festinger's (1957) theory of cognitive dissonance, one might wonder whether the self-reported (absence of) NTH in the control participants was influenced by their post-test performance. In other words, did the participants who learned fewer words perhaps 'justify' this outcome by claiming that they had not named the objects in Dutch in the sorting task? This seems unlikely: sorting the cards took place before the participants were exposed to input, and thus bears no obvious relationship to the effort that the participants made to learn words. Indeed, during the interviews, the participants did not show any evidence of associating particular sorting strategies with particular outcomes.

We also compared the three groups on eleven variables related to word learning (see Table 1), but no significant differences were found. In the context of this study, this is a good thing: the conclusions we have drawn from our analysis should not have been biased by group-level differences in one or more of these variables. At the same time, it means that we still do not know what caused some control participants, but not others, to notice holes. The individual differences, that would explain why some people are more likely than others to experience learner-induced noticing, are something to be explored further in future research.

Effect of NTH, and underlying mechanisms

We will now consider our main research question concerning the effect of NTH on L2 word learning from spoken input. The results showed that NTH facilitates word learning, which is in line with Swain's hypothesis on the noticing function of output (1985, 1993, 1995, 1998). The effect was found both when NTH was experimentally induced by requiring the participants to produce output, and when it was not induced through required output but still internally generated by the participants. Swain (1995, p. 125) already mentioned in passing that (failure in) language production may be vocal or subvocal for the noticing function of output to have an effect. We believe to be the first to have empirically demonstrated this, through the finding that the [+O, +NTH] and [−O, +NTH] participants both outperformed the [−O, −NTH] participants. For the strength of the effect it did not matter whether vocal language production was required or was not required but happened subvocally: [+O, +NTH] and [−O, +NTH] were not significantly different from one another.

The benefit of noticing holes on L2 word learning can potentially be explained by the mechanisms that were mentioned in the introduction. These mechanisms can be summarised as learners allocating more attentional resources to the input after having become aware of their linguistic problems or vocabulary holes, and being curious as to how to resolve or fill those. Perhaps NTH functions as a type of ORIENTING, which is one of three major attentional systems proposed by Posner & Petersen (1990). This system commits attentional resources to sensory stimuli (Tomlin & Villa, 1994, p. 190). Since our explanation for the effect of NTH rests on how the participants processed the input AFTER having noticed holes, it is understandable that it did not matter whether NTH took place with or without (an attempt to) vocal output production.

Suggested direction for future research: Mediation analysis

The mechanisms discussed in the above paragraph could be empirically investigated in future studies using mediation analysis (see Imai, Keele, Tingley & Yamamoto, 2011; MacKinnon, Fairchild & Fritz, 2007). Finding empirical support for such hypothesised pathways would mean a great step forward in our understanding of exactly HOW the positive relationship between NTH and L2 word learning comes about. It is almost certain that at least one further variable must be involved (and potentially more). After all, the realisation of a vocabulary

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

hole in itself does not fill up that hole with the right word form. Rather, an explanation based on mediation through a third and fourth variable was already given: experiencing NTH could make learners curious about the word forms missing in their vocabulary, which in turn could lead them to allocate more attention to the input, leading to more word learning.

Thus, we propose the following chain of processes: NTH → curiosity → attention → word learning (while recognising that this chain is not necessarily exhaustive, and that alternative chains could exist as well). In order to investigate this chain, a future study should also measure curiosity and the amount of attention paid to the target vocabulary during the comparison task. Attention might be measured using eye tracking (e.g., Godfroid, Boers & Housen, 2013). Curiosity could potentially be measured in a stimulated recall procedure (Gass & Mackey, 2000) after the task is finished. If participants were questioned on their curiosity about learning words before or during exposure to input, this would likely trigger NTH in participants assigned to conditions in which no NTH should take place.

In the current study, the incidental finding that some participants in the original control condition had noticed holes enabled us to make some additional comparisons between the groups that we had not initially foreseen. For studying the noticing function of output, this was very interesting. If one wanted to conduct a mediation analysis, however, it would be necessary to have access to a manipulation of NTH that works predictably for all participants. Specifically, participants in a control group should NOT experience NTH. One potential solution for the current set-up could be to leave out the sorting task for the control group. Then, the control participants would not experience NTH before being exposed to input. This would have the disadvantage, however, that participants in the [+NTH] group would already be more familiar with the materials at the start of the comparison task.

Alternatively, mediation analysis can also be applied to studies in which participants are assigned to conditions based on their self-reported experience of NTH, as we did in the current study. However, a prerequisite is that we would need to know the factor(s) that lead some learners but not others in the [−O] condition to experience NTH (Imai et al., 2011). The factors we included in Table 1 did NOT explain this difference, so further exploration would be required.

A disadvantage of applying mediation analysis to a study using non-random assignment is that one cannot be sure whether a significant third variable actually is a mediator variable, rather than a confounding variable. In the latter case, the third variable would both cause learners to experience NTH on the one hand, and on the other hand to be more curious or to allocate more attention to language. The question of mediation versus confounding can be resolved if a predictable method for manipulating NTH is found: only if the third variable is a mediator and not a confound, a relationship between the independent variable (NTH) and the mediator should become visible upon manipulating the independent variable.

Effect of testing moment

Another question of this study was at what rate newly-acquired L2 word knowledge is again forgotten. We found no significant decrease in scores over a period of 15 minutes (although a trend towards decay was visible). Thus, it seems that Ebbinghaus's (1885/1913/2011) nonsense syllables were forgotten sooner (he only remembered 58% after 20 minutes) than the L2 vocabulary in this experiment. Of course, learning a list of nonsense syllables is not the same as learning meaningful L2 names of real objects. Potentially, the current participants had a higher motivation to remember the vocabulary they had just learned, or benefited from the connection that could be made between the word forms and their object referents.

Perhaps due to the short delay of 15 minutes, there was no significant interaction effect between Condition and Testing moment either: word knowledge did not decay at different rates depending on the condition. Thus, the differences between the conditions that were observed at immediate testing persisted 15 minutes later. Readers interested in the retention of word knowledge over longer periods of time are referred to de Vos et al. (2017). That study did show a significant decline in word knowledge in tests after both 20 minutes and six months following exposure. However, that study was different from the current study in several aspects. In conclusion, the retention of incidentally acquired L2 word knowledge over short periods of time seems to depend on the task in which this knowledge was acquired.

Effect of passive knowledge

Because we worked with natural language items, there was the possibility that the participants would already have (some) knowledge of the target words before commencing the experiment (even though we had pre-tested all our items on a similar participant group, see Materials). This was checked through self-report at the end of the experiment. Words that a participant already had actively known before taking part were removed from the analysis. Words of which only passive knowledge was reported were included in the analysis, and we investigated whether such pre-existing passive knowledge was related to learning success on the word level.

Participants who had noticed holes (with or without required output) achieved significantly higher learning scores on those words they had already had passive

https://doi.org/10.1017/S1366728918000019 Published online by Cambridge University Press

knowledge of. For the participants who had not noticed holes, the relationship was the other way around: they achieved significantly LOWER learning scores on words they had already passively known before. While this initially may seem surprising, an explanation is conceivable.

The participants had not been told that they would be tested on object names in a picture-naming post-test. Thus, when they were exposed to the target words in the comparison task, they probably were not consciously preparing themselves for such a task. Since it is known that people generally pay more attention to novel stimuli (e.g., Horstmann & Herwig, 2016; Johnston, Hawley, Plewe, Elliott & Dewitt, 1990), it is likely that the participants paid more attention to the target words they had never heard before, and, as a result, better acquired their word forms. This could explain the (weak) negative effect of pre-existing passive knowledge on word learning for the participants who had not noticed holes.

Why would this not apply to the participants who HAD noticed holes in the sorting task? Their passive knowledge was of no use in the moment when they had to retrieve the names of the target objects from memory. Thus, these participants experienced NTH for all the target objects they could not name, regardless of whether or not they knew their names passively. This also means that they presumably became curious about all of these names, again regardless of passive knowledge status. Then, in the price comparison task, the participants probably paid extra attention to all the objects they were unable to name before. Upon hearing these objects' names, the participants' already existing knowledge of these names was reactivated, and there was less new information to be learned. This could explain the positive relationship between pre-existing passive knowledge and word learning in the participants who had noticed holes, and why the directionality of the relationship differed between participants who had and had not noticed holes.

Conclusion

This study showed that noticing holes in one's vocabulary facilitates subsequent incidental L2 word learning from spoken input. Participants who reported awareness of not being able to produce certain words acquired more of these words from later input, as compared to participants who did not report such awareness. It did not matter whether this awareness had been experimentally induced by requiring the participants to vocally produce output (and fail), or whether it was learner-generated and resulted from subvocal (failure in) output production. The current study does not yet allow us to also draw conclusions about the cognitive mechanisms that explain the increase in word learning rates following the experience of NTH. Therefore, we suggest that future researchers use mediation analysis to explore the mechanisms that underlie the effects of Swain's noticing function of output.

In addition to these theoretical insights, there are two practical lessons to be drawn from this study. Firstly, when it comes to studying NTH (and presumably other forms of noticing too), even under identical treatment conditions participants can differ in their actual NTH experience. This means that NTH will always need to be monitored, rather than just assumed to be present or absent. Secondly, although for word learning it did not matter whether NTH was induced by pushing the learners to produce output or was learner-generated, only the pushed-output treatment generated NTH for all participants in the first place. Thus, if language teachers wanted their students to experience NTH, pushing them to produce output seems worthwhile.

In conclusion, when learners become aware of their vocabulary holes, the first step in filling these holes is already taken. The fact that these results were found in a setting that did not explicitly encourage participants to learn words is very promising. Conceivably, in classroom contexts focused on language learning, effects of NTH might be even more pronounced. This should be investigated in future studies: such knowledge would be very relevant to both language teachers and learners.

Supplementary material

To view supplementary material for this article, please visit http://hdl.handle.net/11633/di.dcc.DSC_2017.00028_573 and https://doi.org/10.1017/S1366728918000019

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* pp. 716–723. doi:10.1109/tac.1974.1100705

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57,* pp. 35–56. doi:10.1111/j.1467-9922.2007.00398.x

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

Burnage, G. (1990). *CELEX: A Guide for Users*. Nijmegen: SSN. Downloaded from http://wwwlands2.let.ru.nl/members/software/celex_intro.pdf, 8 September 2016.

Chen, H., Cohen, P., & Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics – Simulation and Computation, 39,* pp. 860–864. doi:10.1080/03610911003650383

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal*

Learning and Verbal Behavior, 11, pp. 671–684. doi:10.1016/s0022-5371(72)80001-x

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104,* pp. 268–294. doi:10.1037/0096-3445.104.3.268

Crawley, M. J. (2007). *The R Book*. Chichester: John Wiley & Sons. doi:10.1002/9781118448908

de Vos, J. F., Schriefers, H., ten Bosch, L., & Lemhöfer, K. (2017). Incidental spoken L2 word learning and retention: An experimental study. Manuscript submitted for publication.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and Second Language Instruction,* pp. 206–257. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139524780.010

Doughty, C., & Williams, J. (1998). Pedagogical choices in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on Form in Classroom Second Language Acquisition,* pp. 197–262. Cambridge, UK: Cambridge University Press.

Ebbinghaus, H. (1885, translated 1913, reprinted 2011): *Memory: A Contribution to Experimental Psychology*. Eastford, CT: Martino Fine Books. doi:10.1037/10011-000

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Gass, S. M., & Mackey, A. (2000). *Stimulated Recall Methodology in Second Language Research*. Mahwah, NJ: Lawrence Erlbaum Associates. doi:10.4324/9781410606006

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye tracking. *Studies in Second Language Acquisition, 35,* pp. 483–517. doi:doi.org/10.1017/S0272263113000119

Godfroid, A., Housen, A., & Boers, F. (2010). A procedure for testing the Noticing Hypothesis in the context of vocabulary acquisition. In M. Pütz & L. Sicola (Eds.), *Inside the Learner's Mind: Cognitive Processing and Second Language Acquisition,* pp. 169–197. Amsterdam/Philadelphia: John Benjamins. doi:10.1075/celcr.13.14god

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40,* pp. 505–513. doi:10.3758/s13421-011-0174-0

Gruber, M. J., Gelman, B. D., & Ranganath, C. (2014). States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. *Neuron, 84,* pp. 486–496. doi:10.1016/j.neuron.2014.08.060

Hanaoka, O. (2007). Output, noticing, and learning: An investigation into the role of spontaneous attention to form in a four-stage writing task. *Language Teaching Research, 11,* pp. 459–479. doi:10.1177/1362168807080963

Hanaoka, O., & Izumi, S. (2012). Noticing and uptake: Addressing pre-articulated covert problems in L2 writing. *Journal of Second Language Writing, 21,* pp. 332–347. doi:10.1016/j.jslw.2012.09.008

Heeringa, W. (2004). Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. dissertation, University of Groningen.

Hopkins, W. G. (2002). A New View of Statistics. Downloaded from http://www.sportsci.org/resource/stats/effectmag.html, 29 June 2017.

Horstmann, G., & Herwig, A. (2016). Novelty biases attention and gaze in a surprise trial. *Attention, Perception, & Psychophysics, 78,* 69–77. doi:10.3758/s13414-015-0995-1

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review, 105,* pp. 765–789. doi:10.1017/s0003055411000414

Izumi, S. (2013). Noticing and L2 development: Theoretical, empirical, and pedagogical issues. In J. M. Bergsleithner, S. Nagem Frota & J. K. Yoshioka (Eds.), *Noticing and Second Language Acquisition: Studies in Honor of Richard Schmidt,* pp. 37–50. Honolulu, HI: National Foreign Language Resource Center.

Izumi, S., & Bigelow, M. (2000). Does output promote noticing and second language acquisition? *TESOL Quarterly, 34,* pp. 239–278. doi:10.2307/3587952

Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the Output Hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition, 21,* pp. 421–452. doi:10.1017/s0272263199003034

Izumi, Y., & Izumi, S. (2004). Investigating the effects of oral output on the learning or relative clauses in English: Issues in the psycholinguistic requirements for effective output tasks. *The Canadian Modern Language Review, 60,* pp. 587–609. doi:10.3138/cmlr.60.5.587

Jalbert, A., Neath, I., Bireta, T. J., & Surprenant, A. M. (2011). When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 338–353. doi:10.3758/s13421-011-0094-z

Johnston, W. A., Hawley, K. J., Plewe, S. H., Elliott, J. M., & Dewitt, M. J. (1990). Attention capture by novel stimuli. *Journal of Experimental Psychology: General, 119,* pp. 397–411. doi:10.21236/ada221394

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T., & Camerer, C. F. (2009). The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory. *Psychological Science, 20,* pp. 963–974. doi:10.2139/ssrn.1308286

Kornell, N., Jensen Hays, M., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* pp. 989–998. doi:10.1037/a0015729

Kwon, S. H. (2006). Roles of output and task design on second language vocabulary acquisition. Ph.D. dissertation, University of Florida.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22,* pp. 1–26. doi:10.1093/applin/22.1.1

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44,* pp. 325–343. doi:10.3758/s13428-011-0146-0

Lenth, R. V. (2016). Least-squares means: The *R* package *lsmeans. Journal of Statistical Software, 69,* pp. 1–33. doi:10.18637/jss.v069.i01

Leow, R. P. (1999). The role of attention in second/foreign language classroom research: Methodological issues. In *Papers from the 2nd Hispanic Linguistics Symposium*, pp. 60–71.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior. Aware versus unaware learners. *Studies in Second Language Acquisition, 22,* pp. 557–584. doi:10.1017/s0272263100004046

Leow, R. P. (2015). *Explicit Learning in the L2 Classroom: A Student-Centered approach*. New York, NY: Routledge. doi:10.4324/9781315887074

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory, 10,* pp. 707–710.

Litman, J. A. (2008). Interest and deprivation factors of epistemic curiosity. *Personality and Individual Differences, 44,* pp. 1585–1595. doi:10.1016/j.paid.2008.01.014

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58,* pp. 593–614. doi:10.1146/annurev.psych.58.110405.085542

Mahmoudabadi, Z., Soleimani, H., Jafarigohar, M., & Iravani, H. (2015). The effect of sequence of output tasks on noticing vocabulary items and developing vocabulary knowledge of Iranian EFL learners. *International Journal of Asian Social Science, 5,* pp. 18–30. doi:10.18488/journal.1/2015.5.1/1.1.18.30

Park, E. S. (2007). *Learner-Generated Noticing of L2 Input: An Exploratory Study* (Doctoral dissertation, Teachers College, Columbia University, US).

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13,* 25–42. doi:10.1146/annurev.ne.13.030190.000325

Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43,* pp. 103–121. doi:10.1016/j.specom.2004.02.004

R Core Team (2013). R: A language and environment for statistical computing.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15,* pp. 243–257. doi:10.1037/a0016496

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* pp. 129–158. doi:10.1093/applin/11.2.129

Schmidt, R. (2001). *Attention*. In P. Robinson (Ed.), *Cognition and Second Language Instruction,* pp. 3–32. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139524780.003

Schmidt, R., & Frota, S. N. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to Learn: Conversation in Second Language Acquisition,* pp. 237–326. Rowley, MA: Newbury House.

Song, M.-J., & Suh, B.-R. (2008). The effects of output task types on noticing and learning of the English past counterfactual conditional. *System, 36,* pp. 295–312. doi:10.1016/j.system.2007.09.006

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in Second Language Acquisition,* pp. 235–253. Rowley, MA: Newbury House.

Swain, M. (1993). The Output Hypothesis: Just speaking and writing aren't enough. *The Canadian Modern Language Review, 50,* pp. 158–164.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principles and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson,* pp. 125–144. Oxford: Oxford University Press.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on Form in Classroom Second Language Acquisition,* pp. 64–81. Cambridge, UK: Cambridge University Press.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16,* 371–391. doi:10.1093/applin/16.3.371

The Effect Size (n.d.). Downloaded from psych.unl.edu/psycrs/971/meta/effect_sizes.ppt, 29 June 2017.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition, 16,* 183–203. doi:10.1017/s0272263100012870

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5,* pp. 99–114. doi:10.2307/3001913

Uggen, M. S. (2012). Reinvestigating the noticing function of output. *Language Learning, 62,* pp. 506–540. doi:10.1111/j.1467-9922.2012.00693.x

Williams, J. (1999). Learner-generated attention to form. *Language Learning, 49,* pp. 583–625. doi:10.1111/0023-8333.00103.