LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

SYSTEMATIC REVIEW ARTICLE

A Meta-Analysis and Meta-Regression of Incidental Second Language Word Learning from Spoken Input

Johanna F. de Vos,[a,b] Herbert Schriefers,[a] Michel G. Nivard,[c] and Kristin Lemhöfer[a]

[a]Radboud University, [b]International Max Planck Research School for Language Sciences, and [c]VU University Amsterdam

We meta-analyzed the effectiveness of incidental second language word learning from spoken input. Our sample contained 105 effect sizes from 32 primary studies employing meaning-focused word-learning activities with 1,964 participants with typical cognitive functioning. The random-effects meta-analysis yielded a mean effect size of $g = 1.05$, reflecting generally large vocabulary gains from spoken input in meaning-focused activities. A meta-regression with three substantive and two methodological predictors also revealed that adult participants outperformed children in terms of word learning and that interactive learning tasks were more effective than noninteractive ones. Furthermore, learning scores were higher when measured with recognition than with recall tests. Methodologically, the use of a no-input control group seemed to protect against an overestimation of learning effects, evidenced by smaller effect sizes. Finally, whether a

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. We would like to thank Dennis Schutter for his advice in the initial stages of this project and Wolfgang Viechtbauer and other members of the online Cross Validated community for their advice on the statistical analysis.

This article has been awarded an Open Data badge. All data and the analysis script are publicly accessible via the Open Science Framework at https://osf.io/92vfw. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

Correspondence concerning this article should be addressed to Johanna de Vos, Radboud University, Donders Centre for Cognition, P.O. Box 9104, 6500HE, Nijmegen, Netherlands. E-mail: johannadevos@gmail.com

DOI: 10.1111/lang.12296

pretest–posttest design was used did not influence effect sizes. All data and the analysis script are publicly available.

Keywords meta-analysis; meta-regression; incidental learning; word learning; second language acquisition; mixed-effects model; age

Introduction

Second language (L2) learners living in a country where the L2 is used are often exposed to spoken L2 input in their daily lives. Even in situations that do not explicitly revolve around word learning, such incidental exposure can still result in the acquisition of new words. In the L2 classroom as well, words can be learned incidentally when learners listen to their teacher or peers without explicitly focusing on word learning. In short, L2 learners regularly find themselves in situations where incidental word learning from spoken input is possible.

Currently, however, little insight is available about the rate at which such learning takes place. Although research on incidental L2 word learning from spoken input exists (even if much less so than on learning from written input), in this prior work, researchers have typically compared the incidental spoken condition with another condition that is in some sense enhanced, for example, with additional written input (e.g., Brown, Waring, & Donkaewbua, 2008) or with information about an upcoming vocabulary posttest (e.g., Montero Perez, Peters, & Desmet, 2018). The spoken-only condition would then function only as a baseline and not be studied in itself. Therefore, it is practically unknown how effective incidental L2 word learning from spoken input is in an absolute sense (i.e., in comparison to a no-input condition). Experts' opinions on the topic also diverge. Ellis (1999) concluded that much of incidentally learned vocabulary comes from oral input, but Schmitt (2008) was more pessimistic and concluded that the literature on this topic mostly "points to a low uptake rate from listening exposure" (p. 349).

Our study had two goals aimed at increasing our understanding of the effectiveness of incidental L2 word learning from spoken input. First, we wished to systematically quantify this effectiveness by combining all available research in one meta-analysis. The increase in power that such an analysis brings allowed us to go beyond the conclusions that one can draw on the basis of individual studies. This concerns both the generalizability of the findings and the level of power achieved. Such knowledge is relevant for teachers designing their curricula (e.g., should they include incidental learning activities with spoken input at all?), as well as for the many learners around the world who extensively rely on spoken input for their L2 acquisition.

Second, we aimed to investigate more closely how a selection of five variables affects incidental spoken L2 word learning. After all, it would be contentious to claim that there is one overall, all-embracing effect when many variables are known to influence L2 learning outcomes. At the same time, these variables have seldom been investigated within a single study of incidental L2 learning from spoken input. For example, of the 32 studies in our final sample, none compared children to adult language learners. To investigate the effect of age and other variables, we employed the technique of meta-regression, which enabled us to investigate variables that varied between studies in one analysis. We could thus expand the existing knowledge about multiple predictor variables using research that already existed.

Literature Review

Defining Incidental Learning

Various definitions of incidental learning exist. It has often been defined by what it is not: Incidental learning would be "learning without intention, while doing something else" (Ortega, 2009, p. 94). A second definition specifically applies to incidental learning in the context of experimental research. According to this definition, incidental learning is dependent on the announcement of a posttest: When learners engage in an activity without the expectation of being tested afterward, any resulting learning would be incidental (Hulstijn, 2003). A third definition is based on the nature of the activity that learners engage in: Learning can be considered incidental if it comes about as a "by-product" (Hulstijn, 2003, p. 362) of an activity that primarily revolves around meaning. This third definition was adopted for the current study. Incidental learning has often been contrasted with intentional learning, which is learning with intention, learning taking place in situations where learners know that they will be posttested, or where their activities are explicitly focused on language learning.

Mechanisms of Incidental Learning

Although it has been well established that incidental learning is much less effective than intentional learning (e.g., Hulstijn, 2003; Schmitt, 2008), this does not necessarily mean that incidental learning is ineffective in itself, which is what we aimed to investigate in the current study. After all, it is incidental learning during nontutored, everyday language use that turns learners into experienced L2 users. Multiple mechanisms have been proposed that can explain why and how incidental exposure to L2 words can result in learning.

In the first place, fast mapping might play a role. This notion, coined by Carey and Bartlett (1978), holds that children will generally try to map meaning

onto new word forms that they encounter, using logical inference. They can construct this form–meaning link with as little as one exposure (making it fast) and even when no such link is explicitly provided in the input. Fast mapping is then driven by learners' innate curiosity for word learning. Adult language learners have been found to employ fast mapping as well, both when it comes to learning the meaning of nonwords (e.g., Ramachandra, Rickenbach, Ruda, LeCureux, & Pope, 2010) and incidental L2 word learning (e.g., de Vos, Schriefers, ten Bosch, & Lemhöfer, 2018).

Second, Hulstijn (2003), citing Eysenck (1982, p. 203), argued that the processing activities that learners engage in might influence learning rates more than their intentions. Building on the notions of depth of processing (Craik & Lockhart, 1972) and elaboration (Craik & Tulving, 1975), Laufer and Hulstijn (2001) developed the Involvement Load Hypothesis. Involvement is seen as consisting of three dimensions: need, search, and evaluation. Different L2 learning activities require different amounts of these motivational (need) and cognitive (search, evaluation) constructs, and activities that require a higher involvement from the learner are expected to lead to more learning. A meta-analysis by Huang, Willson, and Eslami (2012) found support for this hypothesis: Participants who completed an output task (which supposedly was high in involvement) acquired more vocabulary than those who only read a text (which supposedly was low in involvement).

Third, learners can develop a curiosity or intention to learn words even when the activity that they engage in does not come with an announced posttest or is not explicitly focused on word learning. Thus, even in incidental learning contexts, learners can still decide to deliberately turn their attention to the input (Ortega, 2009), which can result in learning. It should be noted that this learning would only be incidental according to the two definitions from Hulstijn (2003) but not the one from Ortega (2009).

Operationalizing Incidental Learning

Because incidental learning is extremely difficult to operationalize when the learning-without-intention definition is used, we originally set out to find and analyze studies in which the learning was incidental according to the posttest-announcement and by-product definitions. However, it turned out that using posttest announcement as a criterion was problematic too.

To begin with, it was unclear whether the posttest was announced for some studies in our sample (we contacted all authors but not everyone replied). In addition, even in studies for which we knew that the posttest had not been announced, it could still have been expected by the participants. This applied,

of course, to all studies that used a pretest: When learners are tested on unknown vocabulary and are exposed to this vocabulary afterward, they probably expect a posttest. In addition, some studies used cycles of learning treatments and posttests. For example, in Winke, Gass, and Sydorenko (2010), the participants watched a video twice and completed two vocabulary tests afterward. This cycle was repeated three times. It is likely that after the first or second time, the participants knew that the vocabulary posttests were coming.

For these reasons, we based the selection of studies on Hulstijn's (2003) by-product definition only and included studies in which the word learning treatment was presented as a meaning-focused activity, such as listening to an audiobook, watching a video, or performing an interactive task with a peer. In the context of this study, we therefore speak of meaning-focused rather than incidental learning. For all included studies, we indicate in Table 2 (whenever possible) whether the posttest was announced. However, because there was much uncertainty with regard to whether the participants expected a posttest (even when it was not announced), we did not include this design feature as a variable in our analyses.

Meta-Analysis and Meta-Regression in L2 Research

For this study, we used the techniques of meta-analysis and meta-regression to analyze the learning outcomes of 32 studies. In general, meta-analysis allows researchers to calculate the weighted average outcome of a selection of studies. In the case of L2 word learning, such studies typically employ different tests, for example, a 10-item test requiring participants to translate words from their L2 to their first language (L1) or an 18-item L1–L2 recognition test with four answer options per item. This means that such studies cannot be directly compared in terms of the average number of words that the participants learned. To compare them, learning outcomes across studies need to be standardized by dividing the participants' gains by the standard deviation of their scores. This has been done in virtually all meta-analyses focusing on word learning (e.g., Abraham, 2008; Mackey & Goo, 2007; Montero Perez, Van den Noortgate, & Desmet, 2013). By computing these average standardized learning effects over a multitude of studies, we were thus able to address the uncertainty regarding the effectiveness of incidental L2 spoken word learning.

Furthermore, we used meta-regression to investigate how five predictor variables affect L2 incidental spoken word learning. Like ordinary multiple regression, meta-regression is used to study how well the individual independent variables predict the dependent variable. The only difference is that, as in meta-analysis, the dependent variable is not the measurement originally used in the

studies but a standardized effect size. Although meta-regression models technically are no different from other regression models, their use is still relatively rare in L2 acquisition research (for examples, however, see Goldschneider & DeKeyser, 2001, and Li, 2010).

Instead, researchers often study predictor variables by splitting their data set by the levels of their predictor(s) and calculating separate effect sizes for all these subsets (e.g., Boulton & Cobb, 2017; Mackey & Goo, 2007; Montero Perez et al., 2013). Significance can be determined by considering whether the confidence intervals of the effect sizes of the subsets overlap (Mackey & Goo, 2007) or through *Q* tests (Montero Perez et al., 2013). This has some disadvantages. In the first approach, no precise estimation of the significance level is obtained, and in both approaches, one needs to run a separate test for each contrast under investigation, which increases the chance of Type I errors if no correction is applied. Li (2010) did use meta-regression, but with software that allowed "only one independent variable to be included" (p. 350). Nowadays, better software is available. Boulton and Cobb (2017, p. 382) argued that they did not use meta-regression because that would "mainly [be] suited to continuous [predictor variables]." However, categorical predictors can easily be included in a regression model through dummy coding or other forms of contrast coding. After all, an analysis of variance model is also a regression model with categorical predictors (Field, 2009).

In the present study, five predictor variables were analyzed. Three were substantive: age of the participants, treatment, and mode of testing. In light of recent efforts to improve the quality of L2 research (e.g., Plonsky, 2013), two additional predictors focused on a methodological feature and concerned study design: whether a true control group was used and whether pretest-to-posttest gain scores were computed.

Selected Predictors of Meaning-Focused L2 Word Learning From Spoken Input

Age

The first predictor was the participants' age. While popular opinion often ascribes to the viewpoint that "younger is better" in L2 learning (Singleton & Ryan, 2004, p. 61), at the same time there is evidence that older learners might enjoy some advantages too, especially in word learning. Singleton and Ryan summarized the evidence regarding word learning as follows: There seems to be an advantage for older over younger participants in both short-term and long-term instructional studies as well as in short-term naturalistic (e.g., immersion, immigration) studies. Younger participants, however, eventually tend to

overtake older participants in long-term naturalistic studies. One explanation for these findings came from Paradis (2004) who suggested that vocabulary learning is not susceptible to a critical period for language learning (which would favor younger learners) because it relies on declarative memory. Thus, older language learners might benefit from their cognitive maturity when it comes to word learning.

In the current study, we compared the ability of L2 learners of different ages for meaning-focused L2 word learning from spoken input. As mentioned earlier, none of the studies in our sample had investigated age. There have been other studies on age effects in L2 learning, such as Snow and Hoefnagel-Höhle (1978) and Granena and Long (2013), but these did not employ an intervention in which participants were incidentally exposed to L2 spoken input. Using meta-regression, we could investigate whether there was indeed an older-is-better effect in the intervention studies included in our meta-regression.

Treatment

The second predictor was the learning treatment or intervention. A wide variety of activities can support meaning-focused learning from spoken input, such as listening to stories or audiobooks, watching videos, or interactive tasks such as solving a puzzle together. However, comparisons between such treatment types are relatively rare, especially between task-based and nontask learning activities. As with age, the technique of meta-regression is relevant for analyzing treatment effects because treatment type does not need to be manipulated within a single study. In addition, the outcomes of different studies focusing on different learning tasks have been inconclusive. For example, Ellis, Tanaka, and Yamazaki (1994) found an advantage for tasks that involved negotiation between a participant and his/her L2 conversational partner (compared to no negotiation), but Ellis and He (1999) did not find such an advantage. In this case, meta-regression can also provide a solution because combining the outcomes of multiple studies in one analysis should increase the power to detect differences between different task types.

Specifically, we compared the effectiveness of four different treatment types, which we chose because they have all been investigated regularly in primary studies. We compared audio treatments in which the input was presented auditorily only, for example, through audio books, with audiovisual treatments where the target words were also visually supported, for example, through pictures (e.g., Brown et al., 2008) or through video (e.g., Montero Perez et al., 2018). We also included two task-based treatments. These also contained audio and visual input, but in addition there was the element of a

Table 1 A schematic representation of the characteristics of the four treatments investigated

	Characteristics			
Treatment	Audio input	Visual input	Task-based	Interaction
Audio	√	✗	✗	✗
Audiovisual	√	√	✗	✗
Task/–interaction	√	√	√	✗
Task/+interaction	√	√	√	√

Note. √ = characteristic present; ✗ = characteristic absent.

meaning-focused task. Within the task treatments, we made a distinction between the presence and absence of interaction (+/– interaction) between the participant and a conversational partner. This has been commonly manipulated in task-based research (e.g., de la Fuente, 2002; Ellis & He, 1999). Thus, each of the four treatments under investigation was different from the previous one in a single aspect. This is schematically illustrated in Table 1.

Because this meta-analysis and meta-regression concerned L2 word learning from spoken input only, we excluded treatments where the spoken input was accompanied by text, such as a written transcript, glosses, L1 subtitles, or L2 captions. In the case of L2 captions, it would be unclear whether the participants learned from the spoken or written version of the input. Treatments with L1 subtitles or translations were excluded because they remove the need for participants to deduce the meaning of a new word from the context or a visual scene, which we considered an essential part of learning from spoken input. In addition, the reading process could have interfered with the listening process (for meta-analyses on the effects of subtitling, see Montero Perez et al., 2013; for meta-analytic work on glossing, see Abraham, 2008, and Yun, 2011).

Mode of Testing

From both anecdotal and scientific evidence, it is known that when learners are asked to remember previously learned words, open questions (recall) are generally more challenging than multiple-choice questions (recognition) (e.g., Donkaewbua, 2009; Montero Perez, Peters, Clarebout, & Desmet, 2014). While the first two predictors (age and treatment) presumably mainly influence the learning process itself, testing usually only takes place after the learning phase has been completed. Therefore, rather than influencing learning success, it reflects the depth at which the newly acquired word knowledge can be processed. Given the important role that testing instruments play in L2 research and in

education contexts, we chose to include mode of testing (recall vs. recognition) as our third predictor, investigating the question of whether effect-size magnitude depends on testing mode.

Methodological Predictors: Gain Scores and Control Group

Finally, we included two methodological predictors relating to study design. When designing any word learning study, one has to ensure that the vocabulary knowledge displayed in a posttest can rightfully be attributed to the treatment and not, for example, to preexisting knowledge of the target words that the participants already possessed. One solution for acknowledging preexisting knowledge is to calculate pretest-to-posttest gain scores. While this has the advantage that preexisting knowledge can be controlled for with great precision, there are also multiple disadvantages associated with this approach, especially in studies that target incidental learning.

First, the presence of a pretest might lead participants to also expect a posttest, making it questionable whether any potential learning should be considered incidental (this is why we instead concentrated on meaning-focused learning in this study). According to Schmitt (2008, p. 341), intentional vocabulary learning "almost always leads to greater and faster gains" than incidental learning. Second, as pointed out by Bisson, van Heuven, Conklin, and Tunney (2014b) as well as Nation and Webb (2011), a pretest also highlights the target words, perhaps causing learners to pay more attention to these words in later input than they would otherwise. For these reasons, studies making use of pretest–posttest designs might be expected to yield higher effect sizes than studies using nonwords or an independent control group to control for preexisting knowledge. The inclusion of a predictor in the meta-regression that represented the use of gain scores should shed more light on such potential unwanted effects in L2 word learning studies.

A different approach is the use of a true control group, that is, participants who are not in any way exposed to the target words but who take the same tests as the experimental participants. In this way, researchers can again control for (group-level) preexisting knowledge, although in a less precise and individual manner than when using a pretest. In addition, researchers can control for any learning that might happen just as the result of taking tests, spontaneous fluctuations in behavior or the passing of time, and guessing. The latter is especially relevant when the L1 and the L2 are closely related. Thus, in studies without a true control group, effect sizes might be overestimated, and lower effect sizes might be found in studies with a true control group. A true control group predictor was included in the meta-regression to investigate this.

The Current Study

To summarize, this is the first meta-analysis and meta-regression to bring together all literature on meaning-focused L2 word learning from spoken input. We documented the full research process to achieve maximal transparency and reproducibility. The data and script for analysis are publicly available; in addition, the technical details that could not be included in this article due to space limitations are included in Appendixes S1–S4 in the Supporting Information online. The study addresses the following questions:

1. What is the overall effectiveness of meaning-focused exposure to spoken input in L2 word learning?
2. How strongly is this effectiveness influenced by participants' age, type of treatment, and mode of testing?
3. Are effect sizes dependent on such study design features as the use of gain scores and the use of a true control group?

Method

Search Techniques and Sources Considered

Four electronic databases were comprehensively searched for relevant studies published until and including August 2017, with no lower limit set. Three of these were subject-specific databases: PsycInfo, Linguistics and Language Behavior Abstracts, and Education Resources Information Center. These databases extensively cover the fields of psychology, linguistics, and education, and they index research on L2 learning. In addition, we inspected the ProQuest Dissertations and Theses database, a collection of four million graduate dissertations and theses from around the world.

All databases were searched for articles whose titles contained at least one of the below search terms, in combination with *vocabulary* and/or *word**. The individual search terms are shown in Example 1, separated by commas. *Acq** represents search terms related to acquisition, and *gam** refers to gaming. For instance, the first search term *incidental* was used in two searches: *incidental* AND *vocabulary* and *incidental* AND *word**. We also used search terms relating to written data (such as *subtitl**) because studies about these topics sometimes contained data that were relevant to our purposes (as explained below).

Example 1

incidental, natural*, implicit, listen*, spoken, oral, aural, task-based, interaction* AND learn*, interaction AND acq*, subtitl* AND learn*,

subtitl* AND acq*, caption* AND learn*, caption* AND acq*, gam* AND learn*, gam* AND acq*

In addition, we manually searched the reference lists of all included studies and of theoretical and review articles on incidental L2 word learning (Ellis, 1999; Gass, 1999; Huckin & Coady, 1999; Hulstijn, 2003; Restrepo Ramos, 2015; Schmitt, 2008) and inspected the online archives of the following journals (in September 2017): *Language Learning & Technology*, *System*, *Language Learning*, *Studies in Second Language Acquisition*, and *Computer Assisted Language Learning*.[1]

We screened the titles and (in case of doubt) the abstracts of all search results in the above-described databases, reference lists, and online archives. If it seemed that at least one condition in a study met the below-defined inclusion criteria, we inspected the study's method and results sections. We also included one of our own studies (de Vos, Schriefers & Lemhöfer, in press).

Inclusion Criteria and Search Outcome

The 10 inclusion criteria are listed below. Criteria 1 to 5 defined the scope of the study; Criteria 6 and 7 ensured that a study was of acceptable scientific quality; and Criteria 8 to 10 ensured that all necessary data were available:

1. The target language was a second or foreign language to the participants.
2. The target vocabulary was not explicitly taught or studied but embedded in a meaning-focused activity. The participants were not told in advance what the target vocabulary would be.
3. The participants had typical cognitive functioning.
4. The target word input (and, optionally, output) was exclusively spoken.
5. At least one dependent variable measured word knowledge.
6. It was clear to which intervention potential increases in word knowledge were attributable.
7. Preexisting word knowledge was controlled for by the use of gain scores, a true control group or a very careful selection of target items with regard to the participants' preexisting knowledge.
8. Standardized effect sizes could be calculated from the provided means and standard deviations or from raw data.
9. Information about the five predictors was available.
10. The full text of the article was available.

The screening of titles and abstracts resulted in 319 sources (e.g., articles, monographs, dissertations) that seemed relevant. Thirty of these sources (9%)

were found to meet all of the inclusion criteria and are listed under "Included studies" in Appendix S1 in the Supporting Information online. The remaining 289 sources are listed under "Excluded studies," accompanied by the reason for their exclusion.

Oswald and Plonsky (2010) have discussed the question of whether research that has not undergone peer review should be included in meta-analyses and deemed the use of both peer-reviewed and non-peer-reviewed work acceptable (pp. 91–92). Including only peer-reviewed studies has the advantage that all studies can be expected to be of an acceptable scientific quality (Burnham, 1990, cited in Oswald & Plonsky, 2010). The advantages of also including non-peer-reviewed studies include an increase in statistical power and more robust results (Oswald & Plonsky, 2010). Boulton and Cobb (2017) also included non-peer-reviewed research in their meta-analysis because they wanted to obtain a sample as comprehensive as possible, and because they considered the peer-review process to be "highly subjective" (p. 354).

Given the relative scarcity of studies that met our inclusion criteria (in combination with our number of predictors), we also chose to include non-peer-reviewed research. To maintain study quality, we implemented several methodological checks. In line with Criterion 6, posttest data were only considered if we could determine that any potential learning could be attributed to the treatment (and not to earlier posttests). Using Criterion 7, we checked whether participants' preexisting word knowledge could be accounted for.

Characteristics of the Sample

The included 30 sources contained relevant data from 32 studies (of which 24 were peer reviewed), with a total of 44 independent treatment groups that were of interest to us and eight true control groups. The total number of participants over all included groups was 1,964. The mean number of participants in the independent treatment groups was 36 ($SD = 39$, *range* = 8–187), and in the control groups it was 41 ($SD = 25$, *range* = 11–82). Ten studies were published in the 1990s, five in the 2000s, and 17 in the 2010s. The 32 primary studies are described in Table 2, with additional information provided in Appendix S1 in the Supporting Information online. The participants' proficiency in the target language covered the full spectrum, ranging from no preexisting knowledge to high proficiency. All of the studies employed custom-made vocabulary tests containing the target words that the participants had been incidentally exposed to during the intervention. Two of the studies used nonwords as targets (see Table 2). Additional information at the effect-size level (e.g., sample size,

Table 2 Basic information about the 32 included studies

Study	Age group	Gain scores	Control group	Posttest announced?	L1	L2
Al-Homoud (2008): Study 2	University	Yes	No	?	Arabic	English
Aldera & Mohsen (2013)	University	Yes	No	No	Arabic	English
Baltova (1999)	High school	Yes	No	No?	Mostly English in combination with another language	French
Birulés-Muntané & Soto-Faraco (2016)	University	No	Yes	No?	Catalan, Spanish, or Italian	English
Bisson, van Heuven, Conklin, & Tunney (2014a)	University	No	Yes	No	English	Dutch
Brown et al. (2008)	University	No	No	No	Japanese	English (but nonword targets)
de la Fuente (2002)	University	No	No	Yes	English	Spanish
de Vos et al. (in press)	University	No	No	No	German	Dutch
Donkaewbua (2009)	University	Yes	Yes	No	Thai?	English
Duquette (1993)	University	Yes	Yes	Yes?	English	French
Ellis & He (1999)	University	No	No	No	Various, mostly Asian	English
Ellis & Heimbach (1997)	Kindergarten	No	No	No	Japanese, Tagalog, Thai	English
Ellis et al. (1994): Saitama school	High school	No	No	No	Japanese	English
Ellis et al. (1994): Tokyo school	High school	No	No	No	Japanese	English
Gullberg et al. (2012): Exp. 1	University	No	No	No	Dutch	Mandarin
Gullberg et al. (2012): Exp. 2	University	No	No	No	Dutch	Mandarin
Hatami (2017)	University	No	Yes	No	Farsi	English
Hsu, Hwang, Chang, & Chang (2013)	Elementary school	Yes	No	Yes	Language(s) of Taiwan?	English

(Continued)

Table 2 Continued

Study	Age group	Gain scores	Control group	Posttest announced?	L1	L2
Karakaş & Sariçoban (2012)	University	Yes	No	No	Turkish?	English
Koolstra & Beentjes (1999)	Elementary school	No	Yes	No	Dutch?	English
Medina (1990)	Elementary school	Yes	No	No	Spanish	English
Montero Perez et al. (2018)	University	No	No	Yes and no (manipulation)	Dutch	French
Montero Perez et al. (2014)	University	Yes	No	No	Dutch	French
Nagata et al. (1999)	University	No	No	No	Japanese	English
Rodgers (2013): Study 2	University	Yes	Yes	No	Japanese	English
Sydorenko (2010)	University	No	No	No?	English (all but one, who spoke Cantonese)	Russian
Toya (1993)	University	Yes	No	Yes	Japanese	English
van Zeeland & Schmitt (2013)	University	No	No	No	Various	English (but nonword targets)
Vidal (2011)	University	Yes	Yes	No?	Spanish?	English
Winke et al. (2010)	University	No	No	No?	English (all but one, who spoke Kannada)	Spanish
Yeung et al. (2016)	Kindergarten	Yes	No	No	Cantonese	English
Yuksel & Tanriverdi (2009)	University	Yes	No	No?	Turkish?	English

Note. Detailed information about the primary studies can be found in Appendix S1 in the Supporting Information online.

treatment type, and mode of testing) is available in Appendix S2 in the Supporting Information online.

The primary studies in our sample for the most part did not aim at answering the same research questions as we did in this study. What represented the treatment condition of interest for us (i.e., meaning-focused exposure to spoken-only L2 input) was often used as a control condition for studying the effects of subtitles, captions, or glosses on incidental L2 word learning in the primary studies. This explains the low number of true control groups in our sample: The primary studies often achieved statistical control through other comparisons. For the same reason, we did not create a funnel plot of the effect sizes. Funnel plots are common in meta-analysis to indicate whether a publication bias might be present concerning a certain effect. However, because the large majority of the primary studies in our sample had not investigated the effectiveness of incidental exposure to spoken L2 input compared to a no-input condition (as we did in the current study), their publication status was not dependent on the effect size(s) that we extracted from these studies.

Age

The studies included participants in different age ranges. Because only a minority of studies reported the participants' mean age (making continuous regression impossible), we created age groups based on the type of education that participants were enrolled in. Two studies were conducted with children in kindergarten (yielding five effect sizes), three with elementary school students (10 effect sizes), three with high school students (16 effect sizes), and 24 with university students (74 effect sizes). Due to the low number of effect sizes, we grouped children in kindergarten and elementary school together for our analysis. The confound between age and education type will be addressed in the discussion.

Treatment

We distinguished four types of activities in which participants in the treatment groups engaged:

1. Audio (eight studies, 28 effect sizes): Participants listened to storybook reading, academic lectures, or audiobooks.
2. Audiovisual (18 studies, 32 effect sizes): Participants again received auditory input, but also visual support in the form of pictures or video.
3. Task/–interaction (six studies, 25 effect sizes): Participants listened to a speaker (a physically present teacher or peer), had materials that provided visual support, and engaged in a meaning-based task with these materials.

4. Task/+interaction (six studies, 20 effect sizes): The same as the previous activity, but participants also interacted with the speaker (e.g., they could ask questions). This sometimes, but not always, involved prompted production of the target words.

Appendix S2 in the Supporting Information online shows which treatments and modes of testing were used in which studies. This information is not included in Table 2 because treatment and mode of testing sometimes varied within a study; Table 2 includes study-level information only.

Mode of Testing

Testing the newly acquired word knowledge was done either by assessing recognition of words through multiple-choice questions (51 effect sizes) or their recall via open questions (54 effect sizes). Furthermore, responses could be required in the L1 or L2. Recall was always meaning based (e.g., a translation test) and recognition usually so, although in a minority of cases it was form based (e.g., lexical decision). Some studies employed one measurement type only; others employed multiple types. Similarly, some studies used one posttest, and others tested participants after various periods of retention. Thus, we could calculate 105 effect sizes from only 44 treatment groups. The dependency among these effect sizes is discussed in the Data Analysis section.

However, in the case of repeated posttesting, we only used these repeated test results if the participants could not have learned from the earlier tests (see Nation & Webb, 2011). For example, Aldera and Mohsen (2013) first administered a multiple-choice vocabulary recognition test and then a meaning translation test. It could be the case that the participants' answers to the translation test were informed by the questions and answers that they had seen previously. This would be in conflict with inclusion Criterion 6, which states that it should be clear to which intervention potential increases in word knowledge are attributable. Thus, we discarded the outcomes of recall posttests if these had been preceded by a recognition test using the same materials. Whenever this happened, it is noted in the Remarks column under "Effect sizes" in Appendix S2. Posttests could be administered in spoken or written form.

Retention Interval

Finally, it is commonly known that knowledge is gradually forgotten over time (e.g., Ebbinghaus, 1885/1913), making the retention interval between exposure and testing an important variable in L2 research (e.g., Brown et al., 2008; van Zeeland & Schmitt, 2013; Vidal, 2011). In our sample, the retention intervals ranged from immediately after the treatment (for about half of all effect sizes) to

three months (Brown et al., 2008). However, it was not possible to use retention interval as a predictor because the results in the primary studies were often not reported as a function of the retention interval. For example, Rodgers (2013) had 13 teaching sessions (generally separated by one week, but sometimes two). Episodes of a television program were shown in Sessions 3 to 12 and the posttest took place in Session 13. This means that there were at least 11 weeks between the first episode and the posttest, but only one week between the last episode and the posttest. Thus, there was not one retention interval for all items, but the scores were not reported separately for each retention interval. Similar setups were found in Al-Homoud (2008), de la Fuente (2002), Medina (1990), and Yeung, Ng, and King (2016) and in Brown et al. (2008) and Vidal (2011) with regard to the delayed posttest.

Meta-Analytic Statistics

Because we aimed to establish the effects of meaning-focused exposure to spoken input on L2 word learning, our dependent variable of interest was a learning score. We calculated learning scores using the data reported in the primary studies. A learning score always represented a contrast between test scores obtained with (or before) and without (or after) exposure to target words. Based on these contrasts, we identified four different learning score types:

1. Comparison within treatment group(s): Posttest scores to a fixed baseline.

The same level of preexisting knowledge of the target items was assumed for all participants, and their posttest scores were compared to this assumed baseline. For example, Gullberg, Roberts, and Dimroth (2012) exposed Dutch native speakers with no self-reported prior knowledge of Mandarin to a Mandarin weather forecast. Thus, any knowledge of Mandarin words demonstrated in the posttest should reflect learning as a direct result of the treatment. The posttest used in this study consisted of yes/no questions to probe word recognition. Of course, even with no existing knowledge of Mandarin, the chance of making a correct guess was 50%. Therefore, the baseline in this study was set at 50%.

2. Comparison within treatment group(s): Pretest to posttest gain scores.

Learning scores were calculated as the gain scores between a pretest and a posttest for one or more treatment groups, without using a control group. For example, Yuksel and Tanriverdi (2009) tested Turkish students' English vocabulary knowledge before and after they had watched an English movie clip.

3. Treatment group(s) compared to control group: Posttest scores only.

Learning scores were calculated by comparing the posttest scores of a treatment group (after exposure to target words) to the scores of a control group (without exposure to target words). For example, Koolstra and Beentjes (1999) tested the English vocabulary of Dutch children who had watched an English video and a comparable group of children who had not.

4. Treatment group(s) compared to control group: Pretest to posttest gain scores.

This is a combination of the learning score Types 2 and 3. The pretest-to-posttest gain scores of a treatment group were compared to the pretest-to-posttest gain scores of a control group. For example, Spanish students in the treatment group in Vidal (2011) watched a videotaped academic lecture in between taking a pretest and posttest whereas participants in the control group took only the pretest and posttest.

We used Hedges' *g* as our effect-size measure. It was calculated by multiplying Cohen's *d* by Hedges' correction factor *J* (Borenstein, 2009), which accounts for the biasing effect of small sample sizes on Cohen's *d* (Hedges & Olkin, 1985). Cohen's *d* is a standardized effect-size measure and was calculated as one of the learning scores described above, divided by the standard deviation. The calculation of all measures is described in detail in Appendix S3 in the Supporting Information online. This includes a description of the transformations that are required when combining data from studies with different designs (see Morris & DeShon, 2002). We applied such transformations to the data to ensure that the meaning of *g* was unaffected by the type of learning score it was based on. Appendix S2 in the Supporting Information online contains all 105 effect sizes and their associated characteristics (age group, treatment, etc.).

Nevertheless, one might wonder whether effect sizes that are calculated with the data of a treatment group only (learning score Types 1 and 2) are of a different magnitude than effect sizes that are calculated by comparing a treatment and control group (learning score Types 3 and 4). Similarly, it is conceivable that effect sizes that are calculated by comparing pretest to posttest gain scores (learning score Types 2 and 4) would differ from effect sizes calculated based on posttest scores only (learning score Types 1 and 3). To this end, the variables of control group inclusion (yes/no) and the use of gain scores (yes/no) were included in the analysis.

We also calculated the variance *v* associated with each effect size. This variance characterizes the distribution from which an effect size was sampled,

and therefore is different from the variance characterizing the distribution of the participants' scores that were used for computing the effect size. The effect-size variance can be used for various purposes (Morris, 2008), including to weigh effect sizes according to their inverse variance weight (Hedges & Olkin, 1985). In other words, effect sizes with larger sampling variances are weighted less. We followed this practice, which is recommended because it allows studies with more precise effect-size estimates to "contribute more to the meta-analytic average" (Oswald & Plonsky, 2010, pp. 95–96; see also Borenstein, 2009). The effect-size variance was also corrected for small sample bias by multiplying it with the squared correction factor *J* (Borenstein, 2009).

Interrater Reliability

The information presented in Appendixes S1 and S2 in the Supporting Information online was extracted from the primary studies by two raters: the first author of this article and three graduate students in linguistics or psychology, who divided the work. For the quantitative data, such as means and standard deviations, the interrater agreement was 90%. For the qualitative data, including the five predictor variables, the interrater agreement was 96%. Following Boulton and Cobb (2017), the interrater agreement had been calculated by considering the number of discrepancies relative to the total number of cells. The first and second raters resolved all discrepancies together through discussion and by rereading the primary studies.

Data Analysis

Research Question 1 focused on the overall effectiveness of meaning-focused exposure to spoken input for L2 word learning. To address this issue, we explored several random-effects meta-analytic models with the metafor package (version 1.9-9; Viechtbauer, 2010) in R (version 3.3.1; R Development Core Team, 2016). While the 32 studies in our sample all met the inclusion criteria, they still represented a wide variety of study designs, L1–L2 combinations, materials, and so on. Therefore, it was expected that there would be heterogeneity among the effects that were estimated in each study. In other words, it was likely that the variation in effect sizes would exceed the variation that would have been expected due to random variables alone, such as participant sampling.

Such expected between-study heterogeneity can be statistically accounted for by including random intercepts at the study level. In our case, this means that the meta-analytic model estimated a unique effect for each of the 32 studies, rather than assuming that the 32 studies all estimated the exact same effect. In

a similar vein, the true effects could be imagined to vary across the 105 effect-size samples (even if some of them came from the same study), for instance, as a function of treatment or testing instrument. To accommodate this, random intercepts were also introduced at the sample level (see Konstantopoulos, 2011).

We investigated whether the inclusion of random effects indeed significantly improved model fit by comparing the models using likelihood ratio tests. To this end, we first ran a null model with neither fixed nor random effects. The null model was compared to a model with only random intercepts at the study level. This latter model was then compared to a model with random intercepts at both the study and the sample levels. The best-fitting of these models, hereafter referred to as Model 1, was used to answer Research Question 1.

Research Questions 2 and 3 asked how effect-size estimates for meaning-focused L2 spoken word learning are influenced by five predictors: (a) age group: kindergarten/elementary school versus high school versus university, (b) treatment type: audio versus audiovisual versus task/–interaction versus task/+interaction, (c) mode of testing: recognition test versus recall test, (d) use of gain scores: yes versus no, and (e) use of a true control group: yes versus no. These five predictors were added as fixed effects to Model 1 (a random-effects meta-analytic model), yielding Model 2 (a mixed-effects meta-regression model). Parameters for both Model 1 and Model 2 were estimated with the restricted maximum likelihood procedure, which takes into account the number of fixed-effects parameters that are estimated.

Because Model 2 was a regression model, we investigated whether there were any effect sizes that exerted a disproportionate influence on the estimation of the model parameters using Cook's distance (D_i, Cook, 1977). An early rule of thumb was that a Cook's distance larger than 1 should be considered reason for concern (Cook & Weisberg, 1982, cited in Field, 2009). A simulation by McDonald (2002) has shown that 0.85 may be a more appropriate guideline. Therefore, three effect sizes with $D_i > 0.85$ were excluded from the data set, and Model 2 was rerun on this reduced data set.

Finally, an important assumption in standard meta-analytic models is that the effect-size estimates are independent (Hedges, Tipton, & Johnson, 2010). Because we calculated 105 effect sizes from 44 treatment and eight control groups, the sampling errors of the effect sizes were not always independent. The traditional approach in case of dependency is to compute the weighted average of all the effect sizes coming from the same treatment group (Borenstein, Hedges, Higgins, & Rothstein, 2009). This solves the issue of dependency, but at the same time information is lost. This would be especially problematic in the meta-regression, for example, if a study used both recognition and recall tests.

An alternative recommended by Hedges et al. for dealing with dependency is to explicitly model the correlations among the effect-size estimates coming from the same treatment group.

However, as also pointed out by Hedges et al. (2010), the correlations between the effect-size estimates needed to implement this strategy are often not available. In these cases, it is accepted that correlations from comparable studies be used to estimate the missing values (Borenstein, 2009). Reassuringly, Ishak, Platt, Joseph, and Hanley (2008) found through simulation studies that "the results of multivariate meta-regressions were relatively insensitive to incorrect values of within-study correlations" (cited in Hedges et al., 2010, p. 45). We will also report a sensitivity analysis in which we investigated how robust the outcomes of our meta-analysis and meta-regression were when different correlational values were assumed (the methodological details are explained in Appendix S4 in the Supporting Information online).

To take the correlation between the sampling errors of the effect sizes into account, we constructed the whole variance–covariance matrix of the sampling errors (see Appendix S3 in the Supporting Information online for information on the variance calculations). The covariances were calculated according to the standard definition of covariance: $\text{covariance}_{a,\,b} = \text{correlation}_{a,\,b} \times SD_a \times SD_b$. The full covariance matrix can be found in Appendix S2 in the Supporting Information online (under "Covariance matrix"); the formulas in the cells show which correlation was used or assumed. Alpha was set at .05.

Results

Research Question 1

Model comparisons showed that the model with random intercepts at the study level fit the data significantly better than the null model, $\chi^2(1) = 461.74$, $p < .0001$ (Akaike's information criterion [AIC] dropped from 1,214.60 to 754.87). In turn, the model with random intercepts both at the study and sample level (i.e., the effect-size level) fit the data significantly better than the model with random intercepts at only the study level, $\chi^2(1) = 536.03$, $p < .0001$ (AIC dropped from 754.87 to 220.84). This means that the true effect sizes were both heterogeneous between and within studies, as we had expected. This best-fitting model was used for further analysis and is hereafter called Model 1.

Model 1 yielded a weighted average effect-size estimate of $g = 1.05$, 95% confidence interval (CI) [0.81, 1.28], $SE = 0.12$, $z = 8.77$, $p < .0001$, based on the effect sizes of learning gains obtained in the individual studies. Over all studies, this learning gain was significantly larger than 0. Thus, L2 learners experience a significant increase in their vocabulary knowledge after

Table 3 Results from Model 2 after the exclusion of three influential cases

					95% CI	
Fixed effects	*β*	*SE*	*z*	*p*	LL	UL
Intercept	−0.00	0.49	−0.01	.99	−0.96	0.95
Age: high school	0.74	0.59	1.27	.21	−0.41	1.89
Age: university	0.92	0.41	2.26	.02	0.12	1.72
Treatment: audiovisual	0.07	0.16	0.43	.67	−0.25	0.39
Treatment: task/–interaction	0.10	0.44	0.22	.83	−0.76	0.95
Treatment: task/+interaction	0.73	0.45	1.61	.11	−0.16	1.61
Testing: recognition	0.42	0.09	4.42	< .0001	0.23	0.60
Gain scores: yes	0.03	0.32	0.11	.91	−0.59	0.66
Control group: yes	−0.47	0.23	−2.03	.04	−0.93	−0.02
Random effects	Variance	*SD*				
Intercept (study)	0.49	0.70				
Intercept (sample)	0.05	0.23				

Note. $k = 102$. CI = confidence interval; LL = lower limit; UL = upper limit.

meaning-focused exposure to spoken L2 input. Variance at the study level was estimated by the model as 0.31, while variance at the sample level was estimated as 0.21. Profile likelihood plots of these variance components, included in Appendix S4 (Figure A) in the Supporting Information online, showed that we could be confident in these variance estimates. The intraclass correlation was $0.31/(0.31 + 0.21) = 0.59$. This represents a fair correlation (Cicchetti, 1994) between effect sizes coming from the same study, which provided further justification for the inclusion of random intercepts at the study level.

Research Questions 2 and 3

For reasons of space, the outcomes for Model 2 as computed on the full data set are given in Table A in Appendix S4 in the Supporting Information online. In Table 3, we present the model outcomes after three cases with Cook's distance values of 2.52, 2.47, and 1.04 were excluded from the data set (see Data Analysis).[2] The beta estimates show the estimated increase or decrease in effect sizes (in standard deviation units) compared to the predictor level that was represented by the intercept (age = kindergarten/elementary school; treatment = audio; testing = recall test; gain scores = no; control group = no).

For age, no significant difference could be detected between participants in kindergarten/elementary school (the level represented by the intercept) and high

school. However, there was a significant 0.92 increase in effect-size magnitude for participants in university compared to kindergarten/elementary school. The estimated effect-size difference between participants in university and those in high school was 0.92 – 0.74 = 0.18. We changed the order of variable levels in the model (this is called relevelling) to obtain test statistics for this contrast, which showed that the difference was nonsignificant, $\beta = 0.18$, 95% CI [–0.78, 1.13], $SE = 0.49$, $z = 0.36$, $p = .72$.

Treatment had four levels. As we did with the age variable, each of these treatment levels in turn was made the intercept. One contrast was significant, namely task/–interaction versus task/+interaction, $\beta = 0.63$, 95% CI [0.26, 1.00], $SE = 0.19$, $z = 3.37$, $p < .001$. Thus, participants learned more words when the learning task that they engaged in involved interaction with a conversational partner than when it did not. All other contrasts were nonsignificant (all ps $> .11$).

Using a recognition test significantly increased effect sizes with 0.42 standard deviation units relative to using a recall test. No difference was found between studies that controlled for previous knowledge through gain scores and studies that did not use gain scores. On the other hand, studies that used a control group that received no target word input yielded significantly lower effect sizes than studies that did not use a control group, with a difference of 0.47 standard deviation units.

Table 3 also shows that variance at the study level was estimated as 0.49, while variance at the sample level was estimated as 0.05. Profile likelihood plots again showed that we could be confident in these parameter estimates (see Figure B in Appendix S4 in the Supporting Information online). For Model 2, the intraclass correlation was estimated as 0.49/(0.49 + 0.05) = 0.91. This represent a very high correlation between effect sizes coming from the same study. In other words, almost all of the variance was between studies, not within.

An inspection of Table A in Appendix S4 in the Supporting Information online revealed that the outcomes of Model 2 carried out using the reduced and unreduced data sets were very similar. The more salient differences are that the age effect was slightly more pronounced in the reduced data set and that the control group effect was nonsignificant in the unreduced data set. However, the p value for the latter effect only increased from .04 to .06, so this difference was not very substantial.

Sensitivity Analysis

For studies that did not report the correlation coefficient(s) needed for our analyses, we had borrowed a correlation coefficient from the most similar study

that we could find in our sample. To investigate how sensitive our outcomes were to variation in the magnitude of these correlations, we carried out a sensitivity analysis, which is reported in Appendix S4 in the Supporting Information online. We observed only very small changes in the magnitude of the estimated predictor effects and their associated p values, both for Model 1 and Model 2. The direction of all effects was preserved. This indicates that our data were robust to variations in correlation.

Discussion

Meta-Analysis

This study set out to quantify the overall effectiveness of meaning-focused exposure to spoken input for L2 word learning. We found an estimated average Hedges' g of 1.05. This means that participants on average improved their vocabulary knowledge by 1.05 standard deviations after meaning-focused exposure to spoken L2 input. Because the effect size is expressed in terms of standard deviation units, it is not possible to state in an absolute sense the number of learned words to which such an effect size would correspond.

Plonsky and Oswald (2014) present guidelines for the interpretation of standardized effect sizes in the context of L2 research. For between-group designs, they regard $d = 0.4$ as small, $d = 0.7$ as medium, and $d = 1.0$ as large (and g is equal to d except that it is corrected for small-sample bias). The studies included in our meta-analysis used both between-group and within-group designs. However, the above guidelines still apply to our results because we had already taken intragroup correlations into account when calculating effect sizes for within-group designs. Thus, the estimated effect size of $g = 1.05$ can be considered a large effect.

Finding this large effect could be considered surprising in light of the general pessimism that surrounds the effectiveness of learning from listening (e.g., Schmitt, 2008). However, while the linguistic input in all of the studies in our sample was exclusively spoken, three-quarters of these studies provided additional support for learning in the form of pictures, video, or learning tasks. Another explanation for the large effect size might be that the participants in some studies were aware of the upcoming posttest and therefore perhaps paid more attention to the target words. This may have resulted in larger effect sizes than would be found in studies in which the learning was purely incidental.

To put our finding in perspective, the outcome of the one other meta-analysis (that we know of) in which absolute L2 word learning gains were studied should be considered. Mackey and Goo (2007) studied improvements in vocabulary and grammar before and after an interactive treatment. Averaging over these

two domains, they found an effect size of $d = 1.09$. This is very close to our estimated effect size of $g = 1.05$. It is currently impossible to say how absolute learning from spoken input compares, for example, to learning from written input, because we believe no such meta-analysis has been conducted yet—this would be an important avenue for future research.

In conclusion, meaning-focused treatments for L2 word learning may be more effective than has previously been thought. Still, it is possible that the magnitude of our overall effect size was mostly attributable to specific subsections in our data set, such as the studies using task-based learning. The meta-regression provides more insight into this.

Meta-Regression

We investigated whether the effectiveness of meaning-focused exposure to spoken L2 words is predicted by three substantive and two methodological variables.

Age

A positive age effect was found: Participants in university significantly outperformed children in kindergarten and elementary school. Effect sizes for high school participants also were higher than for younger children, but this difference did not reach significance (potentially due to small sample sizes for both age groups). Multiple explanations for the superiority of university students over kindergarten and elementary school children are conceivable. To begin with, older learners have more experience in language learning. For example, they might know more strategies to derive word meaning from context. In addition, they possess a higher degree of cognitive control (Craik & Bialystok, 2006), making it easier for them to focus on a task. There are also potential explanations based on confounds between age group and other variables.

First, the older groups on average had more years of experience with the L2, and, relatedly, typically seemed to have been more proficient (see Appendix S1 in the Supporting Information online). In turn, learners with higher proficiency levels are known to learn vocabulary faster (Vidal, 2011). If more primary studies with adult learners are conducted in which the participants have no prior experience with the target L2 (such as Gullberg et al., 2012), a future meta-regression could circumvent this issue. In any case, age, proficiency levels, and years of experience with the target language should be clearly reported in all primary studies (this was not the case in our sample) so that future meta-analysts can better control for these variables.

Second, the adult participants perhaps also had a higher motivation to learn words. With regard to the classroom studies, the adults had chosen to enroll in language classes. For the children, language study simply was part of the school's curriculum. With regard to the laboratory studies, presumably the adults had volunteered to participate whereas the children would have been signed up by their school or parents. If more primary studies are conducted with adult participants other than self-selected language learners or if a study's language learning aspect is hidden from the participants (such as in de Vos et al., 2018, in press), this confound could be alleviated.

Third, different average intelligence quotient levels can be expected between the general school population and university students. Because "foreign language aptitude partially overlaps with traditional intelligence" (Ortega, 2009, p. 165), this may also have influenced the results, implying an urgent need for L2 acquisition researchers to include other adult learner populations in their studies. For now, we conclude that the combined variable of age and educational context favors university students over child learners in L2 meaning-focused spoken word learning.

Treatment

Regarding the incidental learning treatment, the effect-size estimates increased in magnitude as expected: task/+interaction > task/–interaction > audiovisual > audio. Nevertheless, only the difference between task/+interaction and task/–interaction was significant, with a small/medium effect size. In other words, for L2 spoken word learning, it is beneficial if there is an element of interaction to a learning task. This conforms to earlier literature demonstrating positive effects of interaction (e.g., Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka, 2006; Mackey & Goo, 2007). For future research, it would be interesting to make a distinction within the task/+interaction category between interactive tasks with and without prompted target word production. This has already been done in a few primary studies such as de la Fuente (2002) and Ellis and He (1999). Their findings pointed to a superior role of output. However, the sample sizes in our meta-regression were too small to allow such an investigation.

According to our results, audiovisual treatments had no significant added value over audio-only treatments for L2 word learning. Another observation is that only the difference between task/–interaction and task/+interaction was significant, while the estimated effect of audio (and audiovisual) treatments versus learning task/+interaction was actually larger. This may seem curious, but it is likely a consequence of the fact that task/–interaction versus task/+interaction was often manipulated within studies whereas no studies contrasted audio or

audiovisual treatments with treatments involving learning tasks. Thus, the former contrast could be estimated with more precision and more easily achieve significance. Given this consideration, combined with the finding that the effect of the audio versus task/+interaction contrast was actually estimated to be large (albeit nonsignificant), we are hesitant to confidently conclude that there would be no difference between interactive treatments involving learning tasks on the one hand and audio and audiovisual treatments on the other. More primary studies are needed from which learning scores for one or more of these treatment types can be extracted so that this question can be reconsidered with more statistical power.

Mode of Testing

Testing recognition of the newly learned words compared to recall was found to lead to higher effect sizes. Thus, the ability to recognize a word among several alternatives is achieved more easily than the ability to retrieve the word freely from memory. This outcome confirms Nation's (2001) receptive–productive distinction, which is one of the two dimensions in his model of knowing a word. The other dimension consists of nine subtypes of word knowledge, divided over three domains (form, meaning, and use), but for reasons of statistical power these domains and subtypes were not investigated in the current meta-regression. Future studies aiming to evaluate this proposed second dimension of word knowledge can draw from a rich base of primary studies that have already investigated such questions (e.g., Hatami, 2017; van Zeeland & Schmitt, 2013; Winke et al., 2010).

Gain Scores

Finally, we investigated two predictors related to the way in which effect sizes were calculated. No effect for the use of pretest-to-posttest gain scores could be detected. This is not in accordance with our expectation that studies using a pretest–posttest design would yield higher effect sizes because the participants already knew what target words to look for and were likely expecting a posttest. Two explanations are conceivable.

First, in some of the studies, explicit efforts had been made to minimize the impact of the pretest. For example, Nagata, Aline, and Ellis (1999) conducted the pretest three months before the treatment "to ensure that the subjects did not pay focused attention to the lexical items when they performed the task" (p. 140). Montero Perez et al. (2014) told their participants "that such tests are typically administered at the beginning of the academic year" (p. 126). Such approaches are commendable and can help to minimize unwanted pretest effects.

Second, a large number of the studies that did not use gain scores to control for preexisting knowledge worked from the assumption that all the target words were unknown to the participants. Although in many cases this assumption seemed justified (e.g., in the studies employing nonwords), in some other cases the question of whether this assumption was valid remains. For example, de la Fuente (2002) used words from indigenous languages that are used in Latin American Spanish. It is conceivable that some of her participants, who were students of Spanish at a university in the United States, might have encountered these words while traveling. If in some of the non-gain-scores studies the participants' preexisting knowledge has been underestimated (thus resulting in larger effect sizes), this may have obscured the comparison between studies that did and did not use gain scores.

The above arguments can explain why, in our sample, there was no effect of gain scores. Of course, this does not mean that future researchers can disregard potential influences of pretest use. If researchers want to use a pretest, they should always make efforts to hide the purpose of the pretest and/or its relationship to a following word learning treatment as well as clearly report when and under what circumstances a pretest was conducted.

Control Group

Studies in which treatment groups were compared to true control groups with no exposure to the target words yielded smaller effect sizes than studies that did not contain a control group. The effect was small, but significant. This is in line with our hypothesis, discussed previously. A potential explanation mentioned there was guessing: Participants might (partly) guess words rather than properly learn them, especially when target words are cognates between the L1 and L2. This would then lead to an overestimation of learning effects if no control group (that can also engage in guessing) is used.

Another explanation could be that participants could have learned from the tests themselves, although we tried to exclude this option as much as possible by only including data from repeated posttests where a test effect seemed unlikely. Regardless of the exact explanation, the finding of a significant control group effect shows the need for control group inclusion in studies that aim to evaluate the effectiveness of word learning interventions. Nation and Webb (2011) argue that control groups can also be useful in determining (and correcting for) unwanted side effects of pretest use. There is still a lot of room for improvement because, in our sample, only a quarter of the studies included a true control group.

Limitations and Recommendations for Future Research

As meta-analyses and meta-regressions are a product of the primary studies that they are based on, we were dependent on the information reported in the primary studies to run our analyses. Unfortunately, some relevant studies could not be included because means or standard deviations were not reported (see Appendix S1 in the Supporting Information online). We urge researchers to always report the means and standard deviations for all treatment and control groups included in their study, a seemingly simple thing to do. Guidelines for reporting quantitative results can be found in Norris, Plonsky, Ross, and Schoonen (2015).

Also not often reported in primary studies were correlations between repeated tests on the same participants. Such correlations are needed in meta-analyses and meta-regressions to calculate effect sizes and variances for repeated-measures designs as well as to construct the variance–covariance matrix that is needed to control for dependency between effect sizes. Therefore, we recommend that future researchers report correlations both between pretest and posttest(s) and between repeated posttests (or provide the raw data from which these can be calculated). Although our sensitivity analysis showed that our outcomes were not much influenced by borrowing correlations from other studies, it would be better if future meta-analyses could be as precise as possible.

We had originally intended this study to be concerned with incidental learning. However, it was often unclear whether the participants expected to be posttested on vocabulary, and we therefore resorted to meaning-focused learning. To enable future researchers to properly study incidental learning (and perhaps contrast it with intentional learning), it is important that all authors of primary studies report whether the posttest was announced to the participants. Ideally, after finishing the experiment, they would also interview the participants about whether they were expecting a posttest. If researchers wish to study incidental learning and have the posttest come as a surprise, they should try to prevent situations where participants can guess that a posttest might be administered.

In the meta-regression, we focused on three substantive variables (age, treatment, and mode of testing) that are central to L2 word learning. Many other variables of potentially high importance could not be included. For example, one variable that can influence L2 word learning is exposure frequency (e.g., Brown et al., 2008; van Zeeland & Schmitt, 2013; although exposure effects are not always found, as in Gullberg et al., 2012). However, it was not possible to control frequency of exposure in this study (or make it a predictor), because

this variable was either not controlled in some of the primary studies (e.g., de la Fuente, 2002; Donkaewbua, 2009) or no information about exposure frequency was provided (e.g., Koolstra & Beentjes, 1999; Medina, 1990). Similarly, there was no possibility of incorporating retention interval as a variable in the current study. Proficiency is another important variable that could not be included for various reasons, one of them being the sample size of our data set. In addition, proficiency was measured with different instruments (or was not measured at all) across studies, and (within studies) the participants sometimes were of different proficiency levels. We recommend that future primary researchers control and report exposure frequency, retention interval, and proficiency and that future meta-analysts include these variables in their designs.

Conclusion

Until now, no consensus has existed on the effectiveness of word learning from meaning-focused exposure to spoken L2 input. Our meta-analysis showed that this type of learning is very well possible, on average yielding a large effect. Whether this finding also applies to studies in which the learning was strictly incidental is still to be seen. For this to be investigated, primary studies should first become more transparent in reporting the exact pretest and posttest instructions that the participants were given and ideally verify posttest expectancy through interviews.

In our meta-regression, we detected significant effects for age (higher scores for older participants), treatment (higher scores for learning tasks with interaction than without), and mode of testing (higher scores on recognition tests than on recall tests). Studies using a true control group yielded lower (probably more realistic) effect sizes. All of these novel insights could be extracted from already existing research, which shows the great potential that the technique of meta-regression has for furthering our knowledge in any given domain of language learning.

Final revised version accepted 13 March 2018

Notes

1 All issues of *Language Learning & Technology* and *System* were inspected. We found no new studies that fit the inclusion criteria. Therefore, we only inspected the issues of *Language Learning*, *Studies in Second Language Acquisition*, and *Computer Assisted Language Learning* published after 2010. Again, this yielded no results that had not already been found in the database search.

2 These were both effect sizes from Yeung et al. (2016) and the recall effect size from Brown et al. (2008), respectively.

References

References marked with an asterisk indicate studies included in the meta-analysis/meta-regression.

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, *21*, 199–226. https://doi.org/10.1080/09588220802090246

*Al-Homoud, F. (2008). *Vocabulary acquisition via extensive input*. (Unpublished doctoral dissertation). University of Nottingham, Nottingham, UK.

*Aldera, A. S., & Mohsen, M. A. (2013). Annotations in captioned animation: Effects on vocabulary learning and listening skills. *Computers and Education*, *68*, 60–75. https://doi.org/10.1016/j.compedu.2013.04.018

*Baltova, I. (1999). *The effect of subtitles and staged video input on the learning and retention of content and vocabulary in a second language*. (Unpublished doctoral dissertation). University of Toronto, Toronto, ON.

*Birulés-Muntané, J., & Soto-Faraco, S. (2016). Watching subtitled films can help learning foreign languages. *PLOS ONE*, *11*, 1–10. https://doi.org/10.1371/journal.pone.0158409

*Bisson, M.-J., van Heuven, W. J., Conklin, K., & Tunney, R. J. (2014a). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics*, *35*, 399–418. https://doi.org/10.1017/S0142716412000434

Bisson, M.-J., van Heuven, W. J., Conklin, K., & Tunney, R. J. (2014b). The role of repeated exposure to multimodal input in incidental acquisition of foreign language vocabulary. *Language Learning*, *64*, 855–877. https://doi.org/10.1111/lang.12085

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.). *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–235). New York: Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. https://doi.org/10.1002/9780470743386

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *67*, 348–393. https://doi.org/10.1111/lang.12224

*Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*, 136–163.

Burnham, J. C. (1990). The evolution of editorial peer review. *Journal of the American Medical Association*, *263*, 1323–1329. https://doi.org/10.1001/jama.1990.03440100023003

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Proceedings of the Stanford Child Language Conference*, *15*, 17–29.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, *19*, 15–18. https://doi.org/10.2307/1268249

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.

Craik, F. I. M., & Bialystok, E. (2006). Cognition through the lifespan: Mechanisms of change. *Trends in Cognitive Sciences*, *10*, 131–138. https://doi.org/10.1016/j.tics.2006.01.007

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

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

*de la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition*, *24*, 81–112. https://doi.org/10.1017/S0272263102001043

*de Vos, J. F., Schriefers, H., & Lemhöfer, K. (in press). Noticing vocabulary holes aids incidental second language word learning: An experimental study. *Bilingualism: Language and Cognition*.

de Vos, J. F., Schriefers, H., ten Bosch, L., & Lemhöfer, K. (2018). *High learning rates for interactive L2 vocabulary acquisition in a lab-based immersion setting*. Manuscript submitted for publication.

*Donkaewbua, S. (2009). *Vocabulary learning through listening in another language*. Saarbrücken, Germany: Lambert Academic Publishing.

*Duquette, L. (1993). *L'étude de l'apprentissage du vocabulaire en contexte par l'écoute d'un dialogue scenarisé en français langue seconde* [The study of vocabulary learning in context by listening to a dialogue in scenario form in French as a second language] (Report No. CIRAL-B-187). Quebec, Canada: International Center for Research on Language Planning.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Teachers College Press. https://doi.org/10.1037/10011-000

Ellis, R. (1999). Factors in the incidental acquisition of second language vocabulary from oral input. In R. Ellis (Ed.), *Learning a second language through interaction* (pp. 35–61). Amsterdam: John Benjamins. https://doi.org/10.1075/sibil.17.06ell

*Ellis, R., & He, X. (1999). The roles of modified input and output in the incidental acquisition of word meanings. *Studies in Second Language Acquisition*, *21*, 285–301. https://doi.org/10.1017/s0272263199002077

*Ellis, R., & Heimbach, R. (1997). Bugs and birds: Children's acquisition of second language vocabulary through interaction. *System*, *25*, 247–259. https://doi.org/10.1016/s0346-251x(97)00012-2

*Ellis, R., Tanaka, Y., & Yamazaki, A. (1994). Classroom interaction, comprehension, and the acquisition of L2 word meanings. *Language Learning*, *44*, 449–491. https://doi.org/10.1111/j.1467-1770.1994.tb01114.x

Eysenck, M. W. (1982). Incidental learning and orienting tasks. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 197–228). New York: Academic Press.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Gass, S. (1999). Discussion: Incidental vocabulary learning. *Studies in Second Language Acquisition*, *21*, 319–333.

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning*, *51*, 1–50. https://doi.org/10.1111/1467-9922.00147

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*, 311–343. https://doi.org/10.1177/0267658312461497

*Gullberg, M., Roberts, L., & Dimroth, C. (2012). What word-level knowledge can adult learners acquire after minimal exposure to a new language? *International Review of Applied Linguistics in Language Teaching*, *50*, 239–276.

*Hatami, S. (2017). The differential impact of reading and listening on L2 incidental acquisition of different dimensions of word knowledge. *Reading in a Foreign Language*, *29*, 61–85.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Cambridge, MA: Academic Press. https://doi.org/10.1016/b978-0-08-057065-5.50001-4

Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65. https://doi.org/10.1002/jrsm.5

*Hsu, C.-K., Hwang, G.-J., Chang, Y.-T., & Chang, C.-K. (2013). Effects of video caption modes on English listening comprehension and vocabulary acquisition using handheld devices. *Educational Technology & Society*, *16*, 403–414.

Huang, S., Willson, V., & Eslami, Z. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. *The Modern Language Journal*, *96*, 544–557. https://doi.org/10.1111/j.1540-4781.2012.01394.x

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language: A review. *Studies in Second Language Acquisition*, *21*, 181–193. https://doi.org/10.1017/s0272263199002028

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Oxford, UK: Blackwell. https://doi.org/10.1002/9780470756492.ch12

Ishak, K. J., Platt, R. W., Joseph, L., & Hanley, J. A. (2008). Impact of approximating or ignoring within-study covariances in multivariate meta-analyses. *Statistics in Medicine*, *27*, 670–686. https://doi.org/10.1002/sim.2913

*Karakaş, A., & Sariçoban, A. (2012). The impact of watching subtitled animated cartoons on incidental vocabulary learning of ELT students. *Teaching English with Technology*, *12*, 3–15.

Keck, C., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.13

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*, 61–76. https://doi.org/10.1002/jrsm.35

*Koolstra, C. M., & Beentjes, J. W. (1999). Children's vocabulary acquisition in a foreign language through watching subtitled television programs at home. *Educational Technology Research and Development*, *47*, 51–60. https://doi.org/10.1007/BF02299476

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

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*, 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 407–452). Oxford, UK: Oxford University Press.

McDonald, B. (2002). A teaching note on Cook's distance—A guideline. *Research Letters in the Information and Mathematical Sciences*, *3*, 127–128.

*Medina, S. L. (1990). The effects of music upon second language vocabulary acquisition. *The Annual Meeting of the Teachers of English to Speakers of Other Languages*. Retrieved from ERIC database (ED 352–834).

*Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology*, *18*, 118–141.

*Montero Perez, M., Peters, E., & Desmet, P. (2018). Vocabulary learning through viewing video: The effect of two enhancement techniques. *Computer Assisted Language Learning*, *31*, 1–26. https://doi.org/10.1080/09588221.2017.1375960

Montero Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, *41*, 720–739. https://doi.org/101016/j.system.2013.07.013

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*, 364–386. https://doi.org/10.1177/1094428106291059

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105–125. https://doi.org/10.1037//1082-989X.7.1.105

*Nagata, H., Aline, D., & Ellis, R. (1999). Modified input, language aptitude and the acquisition of word meanings. In R. Ellis (Ed.), *Learning a second language through interaction* (pp. 133–149). Amsterdam: John Benjamins. https://doi.org/10.1075/sibil.17.09nag

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

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

Norris, J. M., Plonsky, L., Ross, S. J., & Schoonen, R. (2015). Guidelines for reporting quantitative methods and results in primary research. *Language Learning*, *65*, 470–476. https://doi.org/10.1111/lang.12104

Ortega, L. (2009). *Understanding second language acquisition* (1st ed.). London: Hodder Education. https://doi.org/10.4324/9780203777282

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110. https://doi.org/10.1017/S0267190510000115

Paradis, M. (2004). *A neurolinguistic theory of bilingualism*. Amsterdam: John Benjamins. https://doi.org/10.1075/sibil.18

Plonsky, L. (2013). Study quality in SLA. An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. https://doi.org/10.1111/lang.12079

R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org

Ramachandra, V., Rickenbach, B., Ruda, M., LeCureux, B., & Pope, M. (2010). Fast mapping in healthy young adults: The influence of metamemory. *Journal of Psycholinguistic Research*, *39*, 213–224. https://doi.org/10.1007/s10936-009-9133-3

Restrepo Ramos, F. D. (2015). Incidental vocabulary learning in second language acquisition: A literature review. *PROFILE Issues in Teachers' Professional Development*, *17*, 157–166. https://doi.org/10.15446/profile.v17n1.43957

*Rodgers, M. P. H. (2013). *English language learning through viewing television: An investigation of comprehension, incidental vocabulary acquisition, lexical coverage, attitudes, and captions*. (Unpublished doctoral dissertation). Victoria University of Wellington, Wellington, New Zealand.

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

Singleton, D., & Ryan, L. (2004). *Language acquisition: The age factor* (2nd ed.). Clevedon, UK: Multilingual Matters.

Snow, C. E., & Hoefnagel-Höhle, M. (1978). The critical period for language acquisition: Evidence from second language learning. *Child Development*, *49*, 1114–1128. https://doi.org/10.2307/1128751

*Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning & Technology*, *14*, 50–73.

*Toya, M. (1993). Form of explanation in modification of listening input in L2 vocabulary learning. *Occasional Paper #23*, Honolulu: Department of English as a Second Language, University of Hawai'i. http://hdl.handle.net/10125/38658

*van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, *41*, 609–624. https://doi.org/10.1016/j.system.2013.07.012

*Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*, 219–258. https://doi.org/10.1111/j.1467-9922.2010.00593.x

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

*Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, *14*, 65–86.

*Yeung, S. S., Ng, M., & King, R. B. (2016). English vocabulary instruction through storybook reading for Chinese EFL kindergarteners: Comparing rich, embedded, and incidental approaches. *Asian EFL Journal*, *18*, 89–112.

*Yuksel, D., & Tanriverdi, B. (2009). Effects of watching captioned movie clip on vocabulary development of EFL learners. *Turkish Online Journal of Educational Technology*, *8*, 48–54.

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, *24*, 39–58. https://doi.org/10.1080/09588221.2010.523285

Supporting Information

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

Appendix S1. Study Characteristics.
Appendix S2. All Effect Sizes.
Appendix S3. Calculating Meta-Analytic Statistics.
Appendix S4. Supplements to the Analyses.