EMPIRICAL STUDY

Multisite Replication in Second Language Acquisition Research: Attention to Form During Listening and Reading Comprehension

Kara Morgan-Short [iD],[a,*,†] Emma Marsden [iD],[b,*,†]
Jeanne Heil,[c,*] Bernard I. Issa II,[d] Ronald P. Leow,[e]
Anna Mikhaylova,[f] Sylwia Mikołajczak,[g] Nina Moreno,[h]
Roumyana Slabakova,[i,j] and Paweł Szudarski[k]

[a]University of Illinois at Chicago, [b]University of York, [c]University of Maine, [d]University of Tennessee-Knoxville, [e]Georgetown University, [f]The University of Queensland, [g]Adam Mickiewicz University, [h]University of South Carolina, [i]University of Southampton, [j]UiT The Arctic University of Norway, and [k]University of Nottingham, *Lead authors, †Proposing authors

We conducted a multisite replication study with aspects of preregistration in order to explore the feasibility of such an approach in second language (L2) research. To this end, we addressed open questions in a line of research that has examined whether having learners attend to form while reading or listening to a L2 passage interferes with comprehension. Our results are consistent with findings from the specific paradigm that we replicated in that no effects on comprehension were detected in analyses conducted over all sites. However, further investigation is warranted due to site-specific effects and methodological limitations. We found all aspects of the multisite registered replication approach to be useful although the registration component itself appeared to be an especially feasible and valuable first step toward increasing the robustness and generalizability of findings in our field.

Keywords replication; multisite study; preregistration; attention; second language acquisition

Introduction

The ability of the field of second language (L2) acquisition to arrive at robust and generalizable conclusions relies crucially on the validity and reliability of its research (Mackey & Gass, 2016; Plonsky, 2015). However, research in L2 acquisition may not yet represent consistent use of best practices. Several research issues, such as low statistical power, *p* hacking, and multiple researcher degrees of freedom, have been identified as problematic in other fields (e.g., Lindsay, 2015; Simmons, Nelson, & Simonsohn, 2011) and are found in the field of L2 research as well (Plonsky, 2015). Unfortunately, many of these issues are difficult to address in part because of the manner in which publication is rewarded, where "incentives for surprising, innovative results are strong" and where such incentives "may be at odds with the incentives for accurate results" (Nosek, Spies, & Motyl, 2012, p. 616). Thus, it seems imperative for mechanisms to be developed that lead to higher levels of reliability and validity in L2 research.

Several recommendations for improved research practices have been proposed in the L2 field (e.g., Norris, Ross, & Schoonen, 2015), including the

Correspondence concerning this article should be addressed to Kara Morgan-Short, Department of Hispanic and Italian Studies, 601 South Morgan Street, 1706 University Hall (MC 315), Chicago, IL 60607. E-mail: karams@uic.edu

repeated call for increased replication research (e.g., Marsden, Morgan-Short, Thompson, & Abugaber, 2018; Polio & Gass, 1997; Porte, 2012) as in other fields (for a systematic review of replication issues in L2 research and more generally, see Marsden, Morgan-Short, Thompson, & Abugaber, 2018). However, the rate of replication is extremely low in L2 research, with fewer than 1 article in 400 being a replication study (Marsden, Morgan-Short, Thompson, & Abugaber 2018), which again is at least partially due to the fact that replication is not incentivized (Nosek et al., 2012). Various mechanisms are emerging across fields, however, to incentivize and facilitate more and better-quality replication, such as the Center for Open Science (https://cos.io) and dedicated funding for registration from The Netherlands' Organization for Scientific Research (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, 2017).

In the current study, we adopted some features of one such mechanism that has emerged in the field of psychology—the multisite registered replication report approach (Simons & Holcombe, 2014; Simons, Holcombe, & Spellman, 2014)—to ascertain the feasibility and usefulness of systematically incorporating a similar approach into the field of L2 research. Below we provide a brief overview of this approach and a motivation for using it to examine the open question of whether attending to L2 lexical and grammatical forms while processing input for meaning affects learners' comprehension (Greenslade, Bouden, & Sanz, 1999; Leow, Hsieh, & Moreno, 2008; Morgan-Short, Heil, Botero-Moriarty, & Ebert, 2012; VanPatten, 1990; Wong, 2001). We then report a first attempt to emulate a multisite registered replication report approach, report the findings, and discuss the implications of the study, both substantively in terms of L2 theory and in terms of the extent to which such an approach might be a viable mechanism for promoting replication and robust research practices in L2 research—the central point of this study.

Throughout this article, we adopt the following nomenclature in discussing different types of replications as recommended by Marsden, Morgan-Short, Thompson, and Abugaber (2018): (a) *direct replications* refer to replication studies that make no intentional change to the research design of the initial study and seek to confirm the validity and reproducibility of the initial study, (b) *partial replications* are replication studies that introduce one principled change to a key variable in the initial study to test generalizability in a clearly predefined way, and (c) *conceptual replications* are replications that introduce more than one change to one or more significant variables for the purpose of extending the initial study more broadly.

Multisite Registered Replication

In recent years, the field of psychology has engaged in several multisite replication endeavors, including those where multiple studies have been replicated by one lab each, such as the Estimating Reproducibility Project: Psychology (Open Science Collaboration, 2015), and those where multiple labs have all replicated one or more studies, such as The Many Lab replication project (Klein et al., 2014). The most systematic endeavor to promote multisite replication has come through registered replication reports, which were introduced in the journal *Perspectives on Psychological Science* (Simons & Holcombe, 2014; Simons et al., 2014) and are now hosted by the journal *Advances in Methods and Practices in Psychological Science* (https://www.psychologicalscience.org/publications/replication). A primary purpose of registered replication reports is to inform the "true size of important effects" (Simons et al., p. 552) by conducting multiple direct replications of one previously published study and by analyzing the effect sizes across the replication sites. Such endeavors are incentivized because (a) registered replication reports are a specific, official journal article submission type, which is highly valued for academic career and funding decisions, at least for the lead authors/conveners, and (b) once the authors have an agreed protocol, including materials, procedures, and analyses, the editor formally accepts the article for publication regardless of the outcomes of the study, which avoids the issue of publication bias (see Marsden, Morgan-Short, Thompson, & Abugaber, 2018, and Marsden, Morgan-Short, Trofimovich, & N. Ellis, 2018).

The process of submitting and publishing a registered replication report (https://www.psychologicalscience.org/publications/ampps/rrr-guidelines) is somewhat different from a standard article (Simons & Holcombe, 2014; Simons et al., 2014). Registered replication reports begin with researchers proposing that the replication of a particular study has high replication value. If journal editors determine in consultation with reviewers that the original study merits a multisite replication, then the full protocol, including materials, procedures, and analyses, are developed such that the researchers finalize all materials and predetermine as many decision points about the procedure and analyses as possible prior to running the study and analyzing the data. This process often involves consultation with the author(s) of the initial article being replicated. The proposing authors also write a pre-data manuscript that includes the introduction, methods, and planned results sections. This manuscript is then submitted for additional review with provisional acceptance given once the researchers address reviewer and/or editor concerns. Once a project has provisional acceptance, it is registered publicly and a call is issued for other sites

to join the project and to conduct direct replications using the registered protocol. The full set of data and analyses from all sites are posted publicly, with analyses focusing on effect sizes and their 95% confidence intervals (CIs) and on mixed-effects model analyses across sites. For the article, the discussion focuses on meta-analytic issues such as measurement error and sample-size differences across sites, so that the overall effect evidenced by the project can be objectively considered. Upon publication, the author(s) of the initial study is (are) invited to contribute a brief published commentary. At the date of writing, this mechanism for publication of replications has enjoyed relative success in psychology with six published registered replication reports since 2014 (with a mean of 35.67 citations per article, as calculated from article metrics reported on each article's online journal page) and several ongoing registered replication reports.

Registered replication reports have been argued to be beneficial to scientific inquiry because of their elements of preregistration and direct replication, and because of their multisite approach. Preregistration enables researchers to make distinctions clear to others and to themselves about the elements of their research that reflect a priori prediction, which entails confirmatory analyses, and postdiction, which often entails exploratory analyses (Nosek, Ebersole, DeHaven, & Mellor, 2017). It also prevents questionable research practices such as hypothesizing after the results are known (HARKing) and *p* hacking, where decisions about data analyses, sometimes influenced by unconscious biases, may be made so that a final *p* value is less than .05 (Kerr, 1998; Lindsay, 2015). Finally, preregistration is not subject to publication bias based on whether statistically significant results were obtained or not because a provisional publication decision is made prior to data being collected and cannot be reversed based on the results. Direct replication affords the field a crucial opportunity to verify previously published findings, and doing so through a multisite approach may be particularly valuable because it allows one to isolate the signal, that is, the effect of interest, from the noise, that is, the sampling error (Simons, 2014). Also, more simply, multisite studies usually lead to larger sample sizes than single-site studies. These benefits of registered replication reports are anticipated to "lead to a better understanding of important effects . . . , and more generally advance the reproducibility and replicability" of a field of research (Simons et al., 2014, p. 554).

Although no journal-based registered replication reports exist within the field of L2 research (although *Language Learning* now accepts registered reports, which can include multisite replication studies; https://onlinelibrary.wiley.com/journal/14679922), such an infrastructure,

adapted to our field, has the potential to incentivize replication and improve certain types of research practices. However, it is particularly pertinent to ascertain the nature of any challenges for implementing such an approach that may be specific to L2 research, a field in which a range of context-specific variables is known to influence outcomes, such as proficiency and amount and nature of language experience. Thus, in order to explore the feasibility of this approach to improve the robustness and generalizability of L2 research, the first two authors of the present study led this multisite replication effort by emulating some key aspects of registered replication reports, specifically preregistering the materials and protocol for direct replications that were conducted at multiple sites with data analyses focused on reproducing previous findings and on meta-analytic effects across sites.

Attention to Form While Processing L2 Input for Meaning
Selecting Target Research for Replication

As noted above and by Marsden, Morgan-Short, Thompson, and Abugaber (2018), an initial step in any replication endeavor is to establish that an initial study has high replication value (Association for Psychological Science, 2017; Porte, 2012), which should include being influential and of continued interest in the field, being methodologically sound, and having implications for theory. Additionally, the study should not have been the subject of previous replication studies that yielded consistent results. In the case of multisite replications, all of these characteristics are probably necessary given the investment of time and resources needed for multisite replications, which may consequently be warranted only for the most central or pressing issues. However, where resources are more readily available, multisite approaches would ideally be warranted for a wider range of studies that have only some of the above characteristics.

In our case, we decided to explore a question that has been relevant to at least two important areas of inquiry in the field of L2 acquisition, that is, research about attention and awareness (e.g., Cintrón-Valentín & N. Ellis, 2016; Leow, 2001; Robinson, 1995) and research about form–meaning mapping (e.g., Doughty & Williams, 1998; R. Ellis, 2016; VanPatten, Williams, Rott, & Overstreet, 2004), with such research generally focusing on forms that are the phonemic or orthographic representations of lexical items or overt morphology. One specific area of research that informs these broad questions has examined whether attending to L2 lexical and grammatical forms while processing input for meaning affects learners' comprehension (Greenslade et al., 1999; Leow et al., 2008; Morgan-Short et al., 2012; VanPatten, 1990; Wong, 2001). This particular line of work was chosen as the medium for examining the feasibility

of registered replication reports in L2 research because this work fully warrants multisite replication. First, it is an influential line of work with the seminal study by VanPatten (1990) being one of the 10 most frequently cited articles published by the journal *Studies in Second Language Acquisition*, with 1,009 citations according to Google Scholar at the time of writing. Second, the line of research is of continued interest in the field with both theoretical and pedagogical implications, particularly for the model of input processing (VanPatten, 1996, 2015) and the pedagogical approach of processing instruction (VanPatten, 2004b, 2005). Third, researchers are interested in replicating studies in this line of research as evidenced by multiple partial and conceptual replications, which is remarkable given that the replication rate for the field of L2 research has been very low (Marsden, Morgan-Short, Thompson, & Abugaber, 2018).[1] However, the replications have not been direct replications and have not yielded fully consistent results. Finally, the validity and reliability of the research design used by Leow et al. (2008) within this line of research has been argued to be methodologically sound. Overall, the line of research addressing the extent to which attention to form while processing L2 input for meaning affects comprehension seemed to merit the efforts of a multisite replication, and we selected Leow et al. as the study to be replicated given its claimed methodological strengths.

Initial Study and Replication Research

The initial study in this line of research—conducted by VanPatten (1990)—was theoretically grounded in questions about L2 input processing. More specifically, VanPatten asked whether learners can simultaneously "attend to both form and meaning when processing input" for comprehension, given learners' limited attentional capacity (p. 287). Three levels of L2 Spanish learners (i.e., university students from first-semester classes, from fourth-semester classes, and from third-year conversation classes) were asked to listen to a 275-word Spanish passage about inflation and were told that there would be a comprehension assessment afterwards. Before listening to the passage, participants were assigned to one of three experimental groups or to a control group. Participants in the experimental groups were asked to make a check mark on a sheet of paper each time they heard either a lexical form (*inflación*, meaning "inflation") or one of two morphosyntactic forms (*la*, a definite article meaning "the," or –*n*, the verbal morpheme indicating a third person plural subject). After the listening task, participants were asked to write down everything that they recalled from the passage in English. Their notes were then coded for the number of idea units represented. The logic behind this design was that if it were difficult for learners

to attend to the less communicatively meaningful, morphosyntactic forms while processing input for meaning, then doing so would interfere with comprehension. Thus, it was hypothesized that having to attend to the morphosyntactic forms *la* and *–n* would interfere with comprehension whereas having to attend to the lexical form *inflación* would not. The results indicated that, over all three levels, no difference was detected in comprehension between the control group and the group that attended to the lexical form but that there was a difference between the control group and the groups that attended to the morphosyntactic forms because the latter did not demonstrate having comprehended the passage as well.[2] VanPatten interpreted these results as an indication that "simultaneous processing of content and linguistic form is indeed difficult for learners" (p. 293) and suggested that, by extension, "communicatively loaded items in input received conscious attention from early stage learners" (p. 294).

The first replications of VanPatten (1990) were partial replications in that they used the same experimental design but intentionally changed one significant component of the initial study, that is, the language or the modality. Wong (2001) reproduced VanPatten's results in the aural modality with L2 English using a translated passage and the English forms *inflation* and *the.* Wong also conducted the study in the written modality, where attention arguably was not constrained to the same degree as in the aural modality, and did not reproduce VanPatten's findings. In contrast to Wong and consistent with the initial study, Greenslade et al. (1999) did find differences between conditions in L2 Spanish in the written modality. These replication studies provided some evidence to support the generalizability of VanPatten's initial findings—at least for the aural modality—that attending to morphosyntactic form interfered with comprehension. As such, the results have been used as part of the underpinning for VanPatten's input processing model (VanPatten, 1996; VanPatten & Cadierno, 1993a, 1993b) that formalized the primacy of meaning principle that "learners process input for meaning before they process it for form" (VanPatten, 2004a, p. 7).

With continued interest in the primacy of meaning principle, Leow et al. (2008) revisited the question of simultaneous attention to form and meaning in written input through a conceptual replication. Leow et al. pointed out and aimed to address certain methodological limitations of the previous studies, including the differential physical salience of the linguistic forms within the studies, the reliability and internal validity of the comprehension assessment, the operationalization of attention, and an uneven distribution of target forms in the original text. In Leow et al.'s study, second-semester L2 Spanish learners read a 358-word Spanish passage about the Aztecs that provided an equal

distribution of the target linguistic forms in the passage. Before reading the passage, participants were assigned to one of four experimental groups or to a control group: Participants in the experimental groups were asked to circle a particular form on the hard copy of the Spanish passage that they read. The forms included the lexical form *sol* (meaning "sun") or one of three grammatical forms (*la*, a definite article meaning "the"; *lo*, a direct object pronoun meaning "him" or "it"; or *–n*, the verbal morpheme indicating third person plural). In their instructions, all participants were asked to think aloud while they read the passage in order to methodologically establish that all participants were indeed processing for meaning. Participants were also told that there would be a comprehension test afterwards; however, departing from the earlier studies, the comprehension test consisted of 10 four-option multiple-choice questions. Contrary to the findings of VanPatten's (1990) and Wong's (2001) aural studies and of Greenslade et al.'s (1999) written study, Leow et al. did not find any differences in comprehension between any groups, which was consistent with the findings from Wong's written study. A partial replication of Leow et al. confirmed this finding for third-semester, university L2 Spanish learners who either did or did not think aloud while reading the passage and also showed that thinking aloud in this paradigm did not lead to reactivity effects, that is, differential performance brought on by thinking aloud (Morgan-Short et al., 2012). Thus, it was argued that when methodological limitations of the first studies were controlled, evidence was not found to support the idea that attending to grammatical forms interferes with comprehension in the written modality.[3]

The Current Study

Although this line of partial and conceptual replications stands out among research studies as being quite systematic in its investigation, limitations and open issues remain (Morgan-Short et al., 2012). First, the issue of whether attending to form affects comprehension in both aural and written modalities remains open because three of the four studies in the written modality did not find a statistically significant effect of attention to grammatical form on comprehension: No statistically significant effects were found by Leow et al. (2008), Morgan-Short et al., and Wong (2001), but statistically significant effects were reported by Greenslade et al. (1999).[4] The two studies conducted in the aural modality found that attention to grammatical form did affect comprehension, with statistically significant effects found by VanPatten (1990) and Wong. Thus, the results across the written and aural modality have been mixed. Only one study to date has used the same materials and same sample of participants

across both modalities (Wong), thus allowing a robust comparison. However, Wong's research, along with VanPatten's initial study, was characterized by the methodological limitations pointed out by Leow et al. Thus, it is still unknown whether attending to morphosyntactic form while listening to a passage would affect comprehension when these concerns are addressed.

Second, a limitation has remained partially unaddressed for the written modality. As part of the directions for the reading task, participants were directed to read the Spanish passage for comprehension and were told that, after they had read the passage, they would answer some comprehension questions without being able to refer back to the text. However, participants were not asked *not* to reread sentences, paragraphs, or the full passage. Thus, they might have gone back through sentences, paragraphs, or the entire passage to notice forms after reading for comprehension, to check comprehension after only noting the forms, or for both purposes. Some regression is natural in reading for comprehension (Rayner, 1998, 2009), but if participants specifically reread for the purpose of managing the dual nature of the experimental task, this would severely compromise the internal validity of conditions that are meant to represent simultaneous attention to both form and meaning during the process of comprehension, which is relevant to the primacy of meaning principle. Leow et al. (2008), at least partially, addressed this issue by having participants think aloud while reading and then eliminating participants who showed evidence in their think-alouds of having gone back to the passage while answering comprehension questions, a practice for which Leow et al. coined the term *backtracking*. Morgan-Short et al. (2012) also eliminated participants who showed evidence of backtracking to reread the passage for comprehension in their think-alouds. However, think-alouds do not prevent backtracking at the passage level or at more fine-grained levels, for example, at the sentence or phrase levels. Think-alouds can only reveal the presence of backtracking, and only when the behavior is verbalized. Thus, it remains an open question whether attending to form affects written comprehension when the internal validity of the experimental design is enhanced by controlling for backtracking.

A third related issue is that previous research has interpreted the results of the aural and written modality as if there were no differences between these experimental modalities other than a basic view of modality itself, that is, being in either the oral or written medium. However, different modalities involve different speeds of presentation and amounts of exposure (untimed for the written modality and timed for the aural modality) and different opportunities for backtracking (available for the written modality but not available for the aural modality) in addition to the difference in the physical medium of delivering the

input. Thus, in order to establish (a) the internal validity of the written paradigm itself and (b) the interpretation of its results in comparison to those from the aural paradigm, a written paradigm that prevents backtracking and that better matches the speed of delivery of the aural paradigm should be used. We note, however, that a highly controlled paradigm that increases the internal validity of the experimental condition may consequentially restrict its ecological validity.

In sum, we have identified the importance of understanding the constraints on attention to form during comprehension and, more specifically, of producing robust empirical evidence related to the primacy of meaning principle, given that results from previous studies have been mixed, in part because method-ological issues have limited the strength and generalizability of the conclusions that can be drawn from those results. Thus, the current multisite replication study aimed to revisit the question of whether L2 learners are able to attend to both form and meaning while processing L2 input for comprehension, the issue common to all studies in this line of research. We adopted the experimen-tal paradigm from Leow et al. (2008), which aimed to reduce methodological concerns about prior research, and carried out two partial replications of that study, with each partial replication changing one key variable. One partial repli-cation changed the modality from written to aural, and the other changed the presentation of the written modality from an untimed paper-based task to a timed computer-based task in which words were presented sequentially, one by one, as in the aural modality, in order to eliminate backtracking and to increase the comparability between the written and aural modalities. Through multisite endeavors, we then conducted direct replications of both paradigms without changes to any key variable in order to establish the replicability and general-izability of the results and to gain further insight into the size of any effects in the population.

Method

The current study replicated Leow et al. (2008) by adapting its materials (avail-able through the IRIS database at https://www.iris-database.org) to an aural and a timed written paradigm. Two partial replications were developed and run by the lead authors at a university in the United States and at universities in the United Kingdom. Subsequently, a call for participation in the multisite repli-cation effort went out via personal contacts, professional listservs, and Web sites. Included in this call was a link to a registered public Web page on the Open Science Framework that included a description of the multisite replication project, the requirements for participation in the study, and the protocol and

materials needed to execute the study (see https://osf.io/tvuer). Although several more sites responded with interest to the call, five international sites were logistically able to participate. Three sites conducted direct replications of the aural paradigm and two sites conducted direct replications of the timed written paradigm. Each site recruited participants and carried out the study independently following the protocol and material that had been registered on the Open Science Framework page. The site researchers contacted the proposing authors if they had questions about the protocol. The lead authors also developed a template for data entry and registered it via the Open Science Framework site (see https://osf.io/d5s2t), and each site deposited its data using the template (see https://osf.io/vwytd). The lead authors then developed and posted highly detailed analysis protocols (see https://osf.io/nz3su). Following these closely, each site conducted and deposited its own analyses if possible (five sites). The proposing authors carried out and deposited the analyses for two of the sites: timed written Site 2 in the United States (US) and aural Site 2 in the United Kingdom (UK). The lead researchers then conducted analyses across sites. Below is a description of the participants, materials, procedures, and analyses; readers can also consult the Open Science Framework pages and the IRIS database (https://www.iris-database.org) for the materials and analysis protocols.

Participants
Participants were recruited for the aural paradigm from four sites, including from the lead aural site in the United States (US aural Site 1) and from replicating aural sites in the United Kingdom (UK aural Site 2), the United States (US aural Site 3), and Poland (Poland aural site). For the timed written paradigm, participants were recruited from three sites, including from the lead timed written site in the United Kingdom (UK timed written Site 1) and from two replicating timed written sites in the United States (US timed written Site 2 and US timed written Site 4). All sites attempted to recruit at least 60 participants from university Spanish courses of a similar level, with at least 15 participants for each of the four conditions. This minimum a priori number was determined based on the approximate group size in Leow et al. (2008) and on the financial resources available to the lead researchers. The course level was chosen with the intention of matching the level of the participants in Leow et al., who had been enrolled in an introductory level, second-semester university Spanish language course and who also had been exposed to preterite and imperfect inflectional morphology, because these forms occurred in the comprehension passage. If participants in the equivalent to a second-semester university Spanish course had not been exposed to the preterite and imperfect, then the next course level

was targeted following Morgan-Short et al. (2012), who had tested participants enrolled in a third-semester university Spanish course.

At all sites, participants were randomly assigned to either the control condition or one of the three experimental conditions. A total of 704 Spanish L2 learners (females = 433; M_{age} = 20.48 years, SD = 6.08) across the seven sites participated in the study (see Table 1). However, three participants' data files were lost, and 15 participants were excluded from analysis for not completing the study as directed (13 for not responding to any comprehension questions, 1 for not making any mouse clicks as directed, and 1 for taking notes while reading). Finally, 55 participants from the experimental groups were excluded from the analysis presented here because they did not make at least six check marks or mouse clicks while reading or listening to the passages, which was the exclusion criterion utilized by Leow et al. (2008) and which in turn approximated the criterion used in VanPatten (1990). Thus, 631 participants across seven sites were included in the final analysis. Although efforts were made to recruit participants enrolled in university Spanish courses of a similar level, some sites evidenced higher levels of proficiency compared to others (see Table 1). Detailed condition-specific participant information and site-specific information regarding recruitment and course level information are available in Appendixes S1 and S2 in the Supporting Information online.

Materials and Procedure

The materials for this study consisted of (a) a biodata form to elicit language background information; (b) an audio file for the aural paradigm or a presentation file in either E-Prime 2.8 or Superlab 5.0 for the timed written paradigm that included a practice sentence and the comprehension passage; and (c) participant packets that included the condition-specific instructions, the comprehension test, and the proficiency test. The comprehension and proficiency tests were both paper tests in both paradigms.

The comprehension passage was the same passage used in Leow et al. (2008) and Morgan-Short et al. (2012), consisting of 358 words in 23 sentences with 10 instances of each target form (*sol*, *la*, –*n*) distributed evenly across four paragraphs with no more than one target form per sentence. There were also two instances of the verb *son* (the third-person plural form of the verb *ser* meaning "to be") that we did not consider as target forms in line with Leow et al. For the aural paradigm, the passage was recorded by a native speaker of Spanish at a pace that was somewhat slower than native speaker pace following VanPatten (1990). The passage was 3 minutes 43 seconds long. For the timed written paradigm, the same passage was used but was presented via computer

Table 1 Participant information by modality and site

Site/condition	Initial N	Final N	Females n	Age M (SD)	Number of languages M (SD)	Proficiency[a] M (SD)	Checks/ clicks[b] M (SD)
Aural paradigm							
US1-A	143	126	79	20.29 (4.62)	1.52 (0.76)	0.39 (0.13)	9.40 (3.61)
US3-A	241	212	142	20.67 (8.25)	1.08 (0.34)	0.38 (0.11)	9.13 (1.99)
UK2-A	43	41	34	20.73 (6.55)	1.39 (0.66)	0.72 (0.12)	10.00 (1.81)
POL-A	59	55	48	21.25 (1.70)	1.00 (0.00)	0.76 (0.19)	9.78 (1.84)
Total	486	434	303	20.62 (6.62)	1.23 (0.56)	0.47 (0.19)	9.38 (2.55)
Timed written paradigm							
UK1-W	62	60	48	18.45 (2.68)	1.37 (0.73)	0.60 (0.16)	9.24 (1.07)
US2-W	58	47	20[c]	19.17 (0.98)	1.15 (0.36)	0.40 (0.11)	9.09 (1.00)
US4-W	98	90	62	21.81 (6.09)	1.13 (0.45)	0.36 (0.11)	9.20 (1.04)
Total	218	197	130	20.16 (4.66)	1.21 (0.54)	0.44 (0.11)	9.19 (1.04)
Overall total	704	631	433	20.48 (6.08)	1.22 (0.56)	0.46 (0.19)	9.32 (2.19)

Note. US1-A = US aural Site 1; US3-A = US aural Site 3; UK2-A = UK aural Site 2; POL-A = Poland aural site; UK1-W = UK timed written Site 1; US2-W = US timed written Site 2; US4-W = US timed written Site 4. [a]A one-way analysis of variance revealed significant differences in proficiency among the sites, $F(6, 624) = 1.964$, $p < .001$, $\eta^2 = .536$. Games-Howell post hoc tests revealed that POL-A and UK2-A had higher proficiency compared to all other sites ($ps \leq .001$) but no statistical difference was detected between POL-A and UK2-A ($p = .797$); UK had higher proficiency than each of the US sites ($ps < .001$); and no differences between the US sites were detected ($ps \geq .350$). [b]For the written modality, the number of clicks represents the number of target clicks. [c]10 participants did not report gender.

with rapid serial visual presentation (Juola, Ward, & McNamara, 1982), in which one word at a time appeared sequentially in the center of the computer monitor.

The motivation for the design of the written paradigm was for it to be as comparable to the aural paradigm as possible, and rapid serial visual presentation has been claimed to emulate aural comprehension in that the participant does not control the pace of the presentation and cannot engage in previewing, regressions, or rereading (Just, Carpenter, & Woolley, 1982). This latter characteristic of rapid serial visual presentation allowed us to control backtracking. A significant amount of backtracking had been revealed through think-alouds in a pilot study ($N = 21$) of a paper-and-pencil version of the written paradigm. For the current study, the rapid serial visual presentation rate was determined by dividing the total time of the aural passage by the total number of words in the passage, giving a result of 615 milliseconds per word.[5] Thus, the two modalities were exactly matched in terms of the time that participants were exposed to the passage. One concern regarding rapid serial visual presentation of written stimuli is its ecological validity in regard to reading comprehension processes. Although some processes involved in reading are different under rapid serial visual presentation (Öquist & Goldstein, 2003), Juola et al. (1982) found that rapid serial visual presentation does not necessarily disrupt normal reading comprehension processes. More recent evidence (Ricciardi & Di Nocera, 2017) suggests that rapid serial visual presentation may affect reading comprehension when the rate of presentation is faster than the normal reading rate ($\sim$250 words per minute) but not when the rate of presentation is similar to or slower than the normal rate. The presentation rate of 97.56 words per minute used in the current study was not faster than normal, and a second pilot study ($N = 17$) with the same rapid serial visual presentation paradigm as used in the current study yielded similar levels of comprehension as those levels evidenced in previous studies in this line of research (further corroborated by our main data). Thus, although the written, controlled rapid serial visual presentation paradigm may not generalize to all reading contexts, it is arguably a valid manner of examining whether attention to form and meaning affects reading comprehension.

Participant instructions for the control and experimental conditions were based as closely as possible on those provided by Leow et al. (2008) and Morgan-Short et al. (2012). All participants were told that they would be given a comprehension test after either hearing or reading the passage according to the paradigm of their group. For the aural paradigm, participants had to listen to the passage for comprehension (control condition) or had to listen for

comprehension and make a check mark on a blank sheet of paper when they heard a target form (either the lexical form *sol*, the feminine definite article *la*, or the third person plural verb inflection *–n*).[6] The instructions for the timed written paradigm were identical except that participants were told to make a mouse click rather than a check mark when they saw a target form.

The comprehension test was the 10-item multiple-choice test used by Leow et al. (2008) and Morgan-Short et al. (2012) that asked questions in English about the passage. These questions did not specifically require the target forms to have been interpreted for meaning or function because the questions did not focus on the meaning or function of the features to be tallied. Each question was followed by four possible answers, one of which was correct. Leow et al. reported a reliability coefficient (Cronbach's alpha) of .915 for this test (p. 681).[7] Reliability was calculated again for the purposes of the current study because it was the key dependent measure for both modalities. Given the binary nature of participants' responses—correct or incorrect—alpha for internal consistency was calculated based on the Kuder-Richardson 20 formula for all participants who were included in the analyses. The alpha level for all participants across all sites was .197, with varying levels per site: US aural Site 1 = –.119, US aural Site 3 = .153, UK aural Site 2 = .298, Poland aural site = .596, UK timed written Site 1 = .007, US timed written Site 2 = .068, US timed written Site 4 = .008.[8] Overall, these results suggested that comprehension items were not consistent with each other and thus were not measuring comprehension as a unidimensional construct for the current study. The low reliability may partially be an artifact of general low performance on the test (as discussed below) because guessing is known to negatively impact reliability (Bush, 2015).

The overall procedure of the study was as follows: Participants first provided informed consent according to the institutional requirements of the specific site and then completed the background information form. Next, participants received their packets and completed a short practice task in which they made a check mark or mouse click if they heard a target word in a sentence, which was not the same as the target word in their condition. Then, participants either listened to or read the comprehension passage and made, according to their condition, check marks or mouse clicks for the listening paradigm or timed written paradigm, respectively. Participants were instructed to begin the comprehension test once they had finished reading or listening to the passage and completed the test at their own pace. Finally, participants took the proficiency test, which consisted of two sections of a version of the Diplomas de Español como Lengua Extranjera used by Seibert Hanson and Carlson (2014), available at https://www.iris-database.org.

Coding

An Excel template file facilitated systematic data entry into files specific to each site. For the aural paradigm, participants' biodata information, the number of check marks made, their responses to each comprehension item (1 for correct, 0 for incorrect), and each proficiency item (1 for correct, 0 for incorrect) were manually entered into the Excel file by each site. For the timed written paradigm, data entry was the same except that the number of mouse clicks in reaction to the target forms and the reaction time were extracted from each participant's output file by one of the lead authors. Clicks in the output file were counted if a target form preceded the click within the same sentence. We found that no target click occurred more than two words after the target word/form. If the target word/form fell at the end of a sentence, a click that occurred in the next sentence but that was no more than two words after the target form was also counted. Using this protocol, no ambiguous cases arose. Reaction times were measured from the onset of the target form (*la*, *sol*, *–n*) immediately preceding the click.

Data Analysis

The same analysis protocol was followed to calculate descriptive statistics and conduct parametric and nonparametric analyses as required by the data from each site. All reported analyses aimed to reveal whether there would be no differences between experimental and control conditions, that is, the null hypothesis (H_0), or whether differences would be found between conditions, that is, the alternative hypothesis (H_1). The first set of reported analyses examined whether the findings of Leow et al. (2008) were reproduced at each site as revealed though analyses of variance (ANOVAs) that followed those used by Leow et al. Second, in order to gain insight into the size of the effect of the experimental conditions compared to control in the population, we conducted a random-effects meta-analysis using effect-size data across sites. The results from the ANOVA and the random-effects meta-analysis are reported separately for the written and aural paradigms.[9] Third, to examine the overall effect of condition across sites and across modalities, a mixed-effects analysis was conducted. Finally, because conclusions from these analyses are largely based on null hypothesis testing, which can provide evidence for the H_1, that is, that there are differences between conditions—but not for H_0—that there are no differences between conditions, we report Bayes factors for differences between the experimental conditions and the control condition, which allowed us to make inferences about both H_1 and H_0 (Dienes, 2014).

Results

Aural Paradigm

First, we examined the extent to which the findings from Leow et al. (2008), which were broadly reproduced by Morgan-Short et al. (2012), would be reproduced in the aural paradigm at each site. Following the analysis reported by Leow et al., we submitted comprehension scores at each location to a one-way ANOVA with one between-subjects variable (condition), including all participants in the control condition and participants in the experimental conditions who had made at least six check marks to indicate hearing a target form. A general approximation of the size of the effect of condition, as measured by η^2, was based on the recommended interpretation of R^2 values from Plonsky and Ghanbar (in press), that is, small effect $< .20 \leq$ medium effect $< .50 \leq$ large effect. Given that both η^2 and R^2 represent the amount of variance explained, it seems reasonable to use the recommended interpretation of R^2 values as an approximate interpretation of η^2 values, given the lack of field-specific recommendations for η^2. When the ANOVA yielded a main effect of condition, post hoc analyses were conducted and consisted of either a Tukey test if homogeneity of variance among groups did not differ or a Games-Howell test if homogeneity of variance among conditions was shown to differ.

For US aural Site 1, the lead aural site, comprehension scores for all the attentional conditions were around 30% mean accuracy (see Figure 1 and Table S2 in Appendix S3 in the Supporting Information online). Analyses did not reveal a statistical effect of condition, $F(3, 122) = 1.052$, $p = .372$, $\eta^2 = .025$. Thus, no evidence was found in support of differences in comprehension resulting from paying attention to a lexical or a grammatical form (H_1).

This general pattern of results seemed to hold across the replicating aural sites (see Figure 1 and Table S2 in Appendix S3 in the Supporting Information online), although mean comprehension scores in all conditions were descriptively higher for UK aural Site 2 and the Poland aural site. The null statistical finding from US aural Site 1 was reproduced in two of the three replicating sites: US aural Site 3, $F(3, 208) = 0.988$, $p = .399$, $\eta^2 = .014$, and UK aural Site 2, $F(3, 37) = 1.650$, $p = .195$, $\eta^2 = .118$. In the third replicating site, however, a statistically significant effect for condition was evidenced, which accounted for a small amount of variance in the data as indicated by the η^2 value: Poland aural site, $F(3, 51) = 3.261$, $p = .029$, $\eta^2 = .161$. Post hoc Games-Howell analyses for this site indicated that comprehension was lower for the inflection –*n* condition compared to the lexical *sol* condition ($p = .003$).

Overall, results for the aural paradigm across all four sites reproduced Leow et al.'s (2008) and Morgan-Short et al.'s (2012) findings for the written paradigm
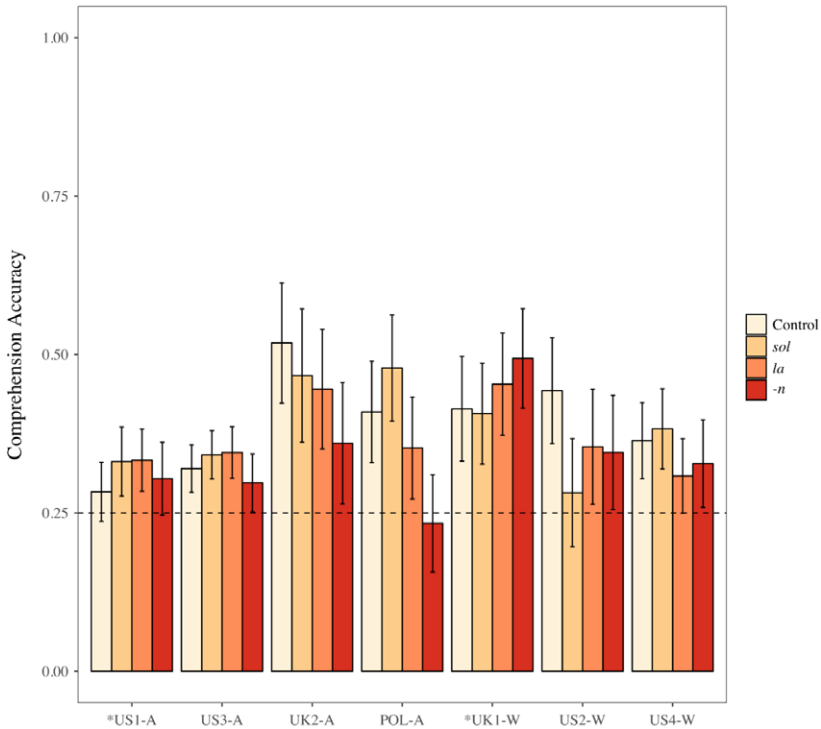
Figure 1 Mean comprehension accuracy scores for each site by condition. Error bars represent 95% confidence intervals. The dashed line represents chance level performance, that is, 0.25 accuracy. US1-A = US aural Site 1; US3-A = US aural Site 3; UK2-A = UK aural Site 2; POL-A = Poland aural site; UK1-W = UK timed written Site 1; US2-W = US timed written Site 2; US4-W = US timed written Site 4 (*lead site). [Color figure can be viewed at wileyonlinelibrary.com]

in that differences in comprehension were not detected for participants who attended to form and meaning as compared to participants who attended to meaning alone. Thus, H_1 was not supported. However, one of the sites did report an effect on comprehension in relation to attending to a morphological versus a lexical form.

Effect Sizes Across Aural Sites

To measure the general effect of each experimental condition compared to the control condition across sites, we calculated Cohen's *d* and its 95% CI for each site and then performed a random-effects meta-analysis across sites using the metafor package (Viechtbauer, 2010) in R (R Core Team, 2016) to

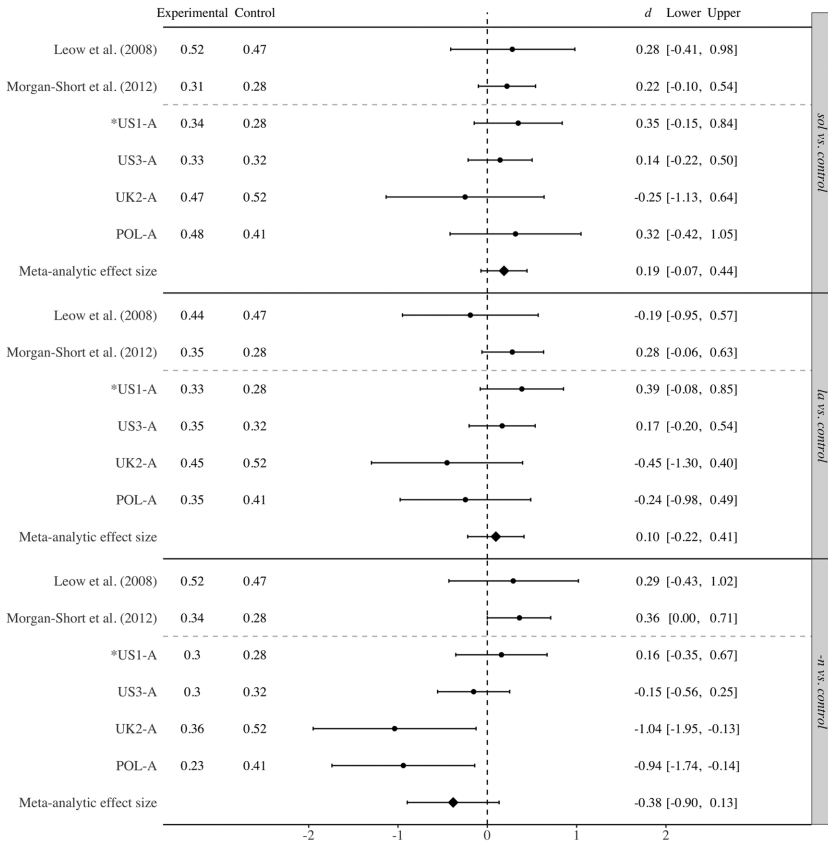|	Experimental	Control		d	Lower	Upper	
Leow et al. (2008)	0.52	0.47		0.28	[-0.41,	0.98]	
Morgan-Short et al. (2012)	0.31	0.28		0.22	[-0.10,	0.54]	
*US1-A	0.34	0.28		0.35	[-0.15,	0.84]	
US3-A	0.33	0.32		0.14	[-0.22,	0.50]	
UK2-A	0.47	0.52		-0.25	[-1.13,	0.64]	
POL-A	0.48	0.41		0.32	[-0.42,	1.05]	
Meta-analytic effect size				0.19	[-0.07,	0.44]	
Leow et al. (2008)	0.44	0.47		-0.19	[-0.95,	0.57]	
Morgan-Short et al. (2012)	0.35	0.28		0.28	[-0.06,	0.63]	
*US1-A	0.33	0.28		0.39	[-0.08,	0.85]	
US3-A	0.35	0.32		0.17	[-0.20,	0.54]	
UK2-A	0.45	0.52		-0.45	[-1.30,	0.40]	
POL-A	0.35	0.41		-0.24	[-0.98,	0.49]	
Meta-analytic effect size				0.10	[-0.22,	0.41]	
Leow et al. (2008)	0.52	0.47		0.29	[-0.43,	1.02]	
Morgan-Short et al. (2012)	0.34	0.28		0.36	[0.00,	0.71]	
*US1-A	0.3	0.28		0.16	[-0.35,	0.67]	
US3-A	0.3	0.32		-0.15	[-0.56,	0.25]	
UK2-A	0.36	0.52		-1.04	[-1.95,	-0.13]	
POL-A	0.23	0.41		-0.94	[-1.74,	-0.14]	
Meta-analytic effect size				-0.38	[-0.90,	0.13]	

Figure 2 Forest plot of random-effects meta-analysis across the aural sites for the effect on comprehension of each experimental condition compared to the control condition. For each comparison, the figure reports mean accuracy and plots raw effect sizes with 95% confidence intervals by site. The overall meta-analytic (weighted mean) effect for each comparison is also plotted with its 95% confidence intervals. US1-A = US aural Site 1; US3-A = US aural Site 3; UK2-A = UK aural Site 2; POL-A = Poland aural site (*lead site).

obtain a meta-analytic effect size with the weighted mean based on variance (see Figure 2 and Table S3 in Appendix S3 in the Supporting Information online). As points of comparison, we also calculated Cohen's *d* and its 95% CI for Leow et al. (2008) and Morgan-Short et al. (2012), who had not reported effect sizes. Effect sizes were interpreted in the light of the original studies and Plonsky and Oswald's (2014) field-specific recommendations for

between-group comparisons, with $0.40 \leq d < 0.70$ suggesting a small effect, $0.70 \leq d < 1.00$ a medium effect, and $d \geq 1.00$ a large effect. Effect sizes with 95% CIs that did not cross 0 were interpreted as reliable effects, providing evidence for H_1 (Cumming & Finch, 2005), but the reverse was not taken to be true, that is, 95% CIs that include 0 cannot be interpreted as no effect (H_0) because they also include a range of values that could be interpreted as an effect.

From Figure 2, we see that the meta-analytic effect-size point estimates for attending to the lexical form *sol* and for attending to the grammatical form *la* compared to the control condition, 0.19 and 0.10 respectively, (a) fell within the 95% CIs from Leow et al. (2008) and Morgan-Short et al. (2012), (b) were close to zero and did not approach the 0.40 value that would be interpreted as a small effect, and (c) had CIs that overlapped with zero. Thus, this meta-analytic effect across the four aural sites does not provide evidence for differences in comprehension (H_1) while attending to *sol* or *la* compared to the control condition. The meta-analytic effect was consistent with the effects for each site because the site-specific effects for these two conditions also had 95% CIs that included zero. The meta-analytic effect-size point estimate for attending to the grammatical form *–n* compared to the control condition, – 0.38, (a) also fell within the 95% CI range from Leow et al. but was outside of the range of Morgan-Short et al., (b) approached the 0.40 value that would indicate interpretation as a small effect, but (c) had a CI that overlapped with zero. Thus, across all four sites, no evidence was provided for an effect (H_1) of attending to *–n* compared to the control condition. This effect, however, was not consistent in each individual site. In two of the four sites (UK aural Site 2 and Poland aural site), the 95% CIs did not cross zero, suggesting that for these site-specific samples, attending to *–n* negatively affected comprehension compared to control, with effects on the border between medium and large. Overall, though, the random-effects meta-analysis suggested that the results from the aural paradigm were largely consistent with those reported by Leow et al. and Morgan-Short et al. in that they did not provide evidence for H_1, that is, they did not provide evidence that there were differences in participants' comprehension while attending to lexical or grammatical form, which was also consistent with the overall findings from the current study's ANOVA analyses.

Timed Written Paradigm
Next, we examined whether the findings from Leow et al. (2008) and Morgan-Short et al. (2012) were reproduced at each site implementing the timed written paradigm. As for the aural paradigm, we submitted comprehension scores

at each location to a one-way ANOVA with one between-subject variable (condition), including all participants in the control condition and participants in the experimental conditions who had made at least six mouse clicks to indicate seeing a target form. We interpreted η^2 and conducted post hoc tests following the same parameters as for the aural paradigm.

For UK timed written Site 1, the lead timed written site, comprehension scores for all the conditions were around 45% mean accuracy (see Figure 1 and Table S2 in Appendix S3 in the Supporting Information online). The ANOVA for this site indicated no statistical effect of condition, $F(3, 56) = 1.243$, $p = .303$, $\eta^2 = .062$. Thus, no evidence was found in support of differences in comprehension among conditions resulting from paying attention to a lexical or grammatical form (H_1).

Results from the two timed written replicating sites differed in their consistency with results from the lead site. Both replicating sites had mean comprehension scores around 35%, which is descriptively lower than the value reported at the lead timed written site and closer to the scores from the three US sites that administered the aural paradigm (see Figure 1 and Table S2 in Appendix S3 in the Supporting Information online). The null statistical finding from UK timed written Site 1 was reproduced in US timed written Site 4, $F(3, 86) = 1.311$, $p = .276$, $\eta^2 = .044$, but not in US timed written Site 2, where an effect of condition was evidenced, $F(3, 43) = 3.480$, $p = .024$, $\eta^2 = .195$, which accounted for a small amount of variance in the data. Post hoc Tukey analyses indicated lower comprehension for the lexical *sol* condition compared to the control condition ($p = .016$).

In sum, results from two of the three timed written sites reproduced the earlier findings of the untimed written paradigms of Leow et al. (2008) and Morgan-Short et al. (2012) in that differences in comprehension were not detected while participants were attending to form and meaning compared to those attending to meaning alone. As such, H_1 is largely not supported. However, in US timed written Site 2, a negative effect of participants attending to the lexical form *sol* was found compared to the control participants' performance.

Effect Sizes Across Timed Written Sites

To measure the general effect of each experimental condition compared to the control condition across the timed written sites, we calculated Cohen's *d* and its 95% CI for each site and performed a random-effects meta-analysis across sites, using the same methods and approach to interpretation as for the aural data (see Figure 3 and Table S3 in Appendix S3 in the Supporting Information online).
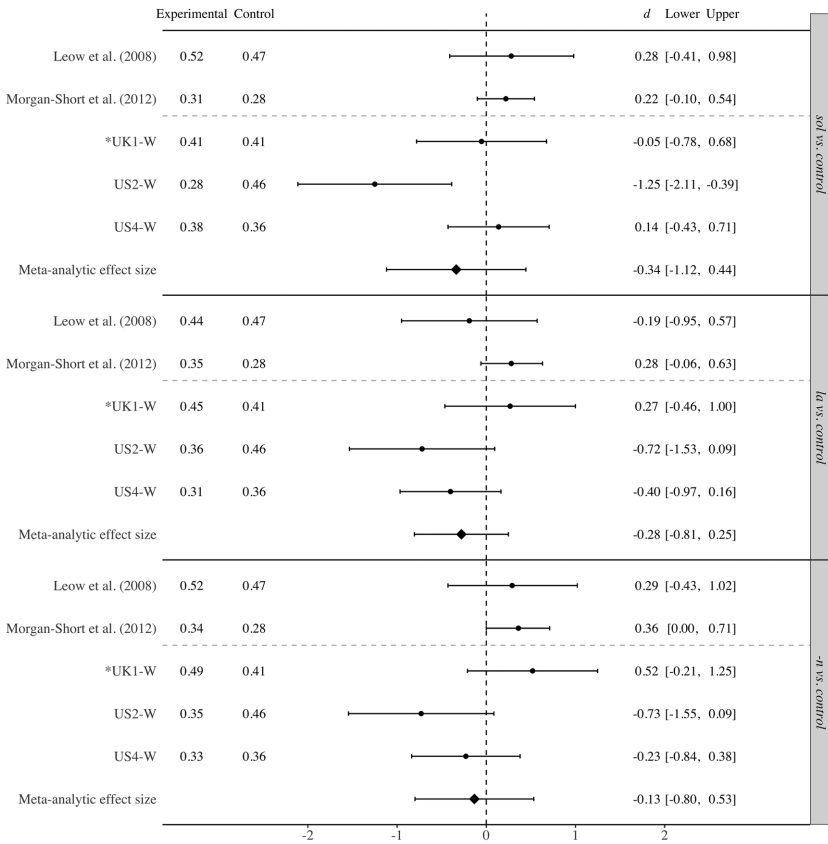
	Experimental	Control		d	Lower	Upper
Leow et al. (2008)	0.52	0.47		0.28	[-0.41,	0.98]
Morgan-Short et al. (2012)	0.31	0.28		0.22	[-0.10,	0.54]
*UK1-W	0.41	0.41		-0.05	[-0.78,	0.68]
US2-W	0.28	0.46		-1.25	[-2.11,	-0.39]
US4-W	0.38	0.36		0.14	[-0.43,	0.71]
Meta-analytic effect size				-0.34	[-1.12,	0.44]
Leow et al. (2008)	0.44	0.47		-0.19	[-0.95,	0.57]
Morgan-Short et al. (2012)	0.35	0.28		0.28	[-0.06,	0.63]
*UK1-W	0.45	0.41		0.27	[-0.46,	1.00]
US2-W	0.36	0.46		-0.72	[-1.53,	0.09]
US4-W	0.31	0.36		-0.40	[-0.97,	0.16]
Meta-analytic effect size				-0.28	[-0.81,	0.25]
Leow et al. (2008)	0.52	0.47		0.29	[-0.43,	1.02]
Morgan-Short et al. (2012)	0.34	0.28		0.36	[0.00,	0.71]
*UK1-W	0.49	0.41		0.52	[-0.21,	1.25]
US2-W	0.35	0.46		-0.73	[-1.55,	0.09]
US4-W	0.33	0.36		-0.23	[-0.84,	0.38]
Meta-analytic effect size				-0.13	[-0.80,	0.53]

sol vs. control / *la vs. control* / *-n vs. control*

Figure 3 Forest plot of random-effects meta-analysis across the timed written sites for the effect on comprehension of each experimental condition compared to the control condition. For each comparison, the figure reports mean accuracy and plots raw effect sizes with 95% confidence intervals by site. The overall meta-analytic (weighted mean) effect for each comparison is also plotted with its 95% confidence intervals. UK1-W = UK timed written Site 1; US2-W = US timed written Site 2; US4-W = US timed written Site 4 (*lead site).

Figure 3 shows that the meta-analytic effect-size point estimates for participants attending to any form, that is, *sol*, *la*, or *–n*, compared to those in the control condition, were –0.34, –0.28, and –0.13, respectively, and (a) fell within the 95% CI range from Leow et al. (2008) but were below the range of Morgan-Short et al. (2012), (b) varied in size but did not reach the 0.40 value to be interpreted as a small effect, and (c) had CIs that overlapped with zero. Thus, these meta-analytic effects across the three timed written sites did not

lend support to H_1. This finding was consistent for the effects for each condition for each site with the exception of US timed written Site 2, where attending to *sol* negatively affected comprehension. Overall though, the random-effects meta-analysis did not detect evidence for effects on comprehension while participants attended to lexical or grammatical form in the timed written paradigm, which is generally consistent with findings from Leow et al. and Morgan-Short et al. and with the ANOVAs reported above for this paradigm.

Multisite Mixed-Effects Analyses

To consider the effect of condition on comprehension accuracy across site and modality, the data from each site were entered into a mixed-effects logistic model using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R. The model included condition as the primary fixed effect of interest but also included modality as a fixed effect and the interaction between condition and modality because previous research has found different results for different modalities (e.g., Wong, 2001). The model also included proficiency, whose values were standardized and centered at zero, as a control variable. The maximal random-effects structure supported by the data was used (Barr, Levy, Scheepers, & Tily, 2013) and included random by-subjects intercepts nested in condition, site, and modality; random by-item intercepts and slopes for condition and modality; and random by-site intercepts nested in modality. Further specification of the random-effects structure led to a failure to converge. The full model is specified below:

> Model Accurac ← glmer(Score ~ Condition*Modality + StdProf + (1|Subject:(Condition:Site:Modality)) + (1+Condition+ Modality|Item) + (1|Site:(Modality)), data = Repldata, family = binomial(link = 'logit'), control = glmerControl(optimizer = "bobyqa"))

An ANOVA (Type III) conducted with the car package (Fox & Weisberg, 2011) in R and run on the model returned significant effects of proficiency, $\chi^2(1) = 61.382$, $p < .001$, and modality, $\chi^2(1) = 4.390$, $p = .036$, as well as a marginal effect of condition, $\chi^2(3) = 7.165$, $p = .067$, that was qualified by a marginal interaction of condition by modality, $\chi^2(3) = 7.417$, $p = .059$. The effect of proficiency indicated that participants with higher levels of proficiency scored more accurately on the comprehension test. Follow-up Tukey tests conducted with the multcomp package (Horthorn, Bretz, & Westfall, 2008) in R on the significant effect of modality revealed that the log odds of responding correctly to comprehension items increased by 0.36 for

participants in the written modality compared to those in the aural modality (written estimated $M = 0.34$, $SE = 0.06$, 95% CI [0.26, 0.49]; aural estimated $M = 0.31$, $SE = 0.05$, 95% CI [0.23, 0.41]; estimate $= 0.36$, $SE = 0.17$, $z = 2.095$, $p = .036$). Follow-up Tukey tests on the relevant comparisons for the marginal condition by modality interaction did not reveal any statistically significant differences between control and experimental conditions in either modality (ps $\geq$ .296).[10] These results were thus largely consistent with the pattern of findings from the ANOVA and random-effects meta-analysis in that they did not provide evidence for H_1, that is, they did not provide evidence that there were differences in comprehension between experimental conditions and control, regardless of modality.

Bayes Factors

As previously pointed out, conclusions from analyses reported above can provide evidence only for H_1, not for H_0. Indeed, overall, evidence for H_1 was not detected in the ANOVA, the meta-analytic effect size, or the mixed-effects model results. However, whether the results supported H_0, that is, no difference in comprehension between experimental and control conditions, remained an open question. A Bayesian approach (Dienes, 2014) can provide insight into this question because Bayes factors (B) indicate whether a result is more likely to occur under H_1 or H_0 and thus can constitute evidence for H_0. More specifically, a B value greater than 3 provides evidence for H_1 whereas a B value lower than 0.33 provides evidence for H_0. Thus, in order to determine whether there was evidence for H_0, that is, that comprehension did not differ across conditions, we calculated B based on the mean difference between each experimental condition compared to the control condition for the full set of data.

Our calculation was based on a theory of H_1 with a half-normal distribution and a SD of 2.325 (Dienes, 2014). The half-normal distribution indicates that small effects are more likely than large effects and that effects are predicted to be in one direction by the theory, that is, attention to form is predicted to negatively affect comprehension. The standard deviation, which represents the plausible predicted difference, was determined based on the mean comprehension score of 4.65 from Leow et al.'s (2008) control group. Z. Dienes (personal communication, January 30, 2018) has recommended that, when a maximum effect is known, then the standard deviation should be equivalent to half of the maximum effect. Given that comprehension could not be negatively affected by more than 4.65 (as the minimum comprehension score cannot go lower than 0), the standard deviation would be half of that value, or 2.325. Using half of the

maximum mean score from Leow et al.'s control group as the standard deviation was also, interestingly, convergent with the average size of the reduction in comprehension (51%) across the VanPatten (1990), Greenslade et al. (1999), and Wong (2001) studies when a statistically significant effect was found.[11]

We calculated the Bs for the following contrasts of conditions using Dienes' Bayes calculator (http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm): *sol* versus control, $B_{H(0, 2.325)} = 0.05$; *la* versus control, $B_{H(0, 2.325)} = 0.08$; *–n* versus control, $B_{H(0, 2.325)} = 0.40$. These results provided evidence for H_0 for the *sol* and *la* conditions compared to the control condition, that is, that they did not differ in regard to comprehension. For the *–n* condition compared to the control condition, there was not sufficient evidence to make a strong conclusion because the B value fell between 0.33 and 3, which indicated that the result did not favor either H_1 or H_0. However, it was closer to 0.33, which indicated that the data more closely favored H_0. Indeed, calculating B based on the average of the two grammatical conditions (*la* and *–n*) yielded $B_{H(0, 2.325)} = 0.02$, which was evidence for H_0, suggesting that comprehension did not differ for participants attending to the grammatical forms compared to control participants.

Discussion

The primary aim of the current study was to ascertain the feasibility and usefulness of incorporating a multisite registered replication approach in the field of L2 research as a mechanism for improving the validity of the field's findings through transparent research practices and replication. To explore such possibilities, we conducted a multisite replication study that emulated some aspects of the registered replication reports approach used in the field of psychology, for example, preregistration of the materials and protocol for direct replications that were conducted at multiple sites. More specifically, we first conducted two partial replications of Leow et al. (2008) in the aural and written modalities to examine whether attending to form while listening to or reading a L2 passage for meaning in timed conditions would interfere with comprehension of that passage. After the partial replications had been run, their procedures and materials were registered and replicated directly by multiple sites. Analyses were then conducted to examine whether the results from Leow et al. were reproduced at each site, what the meta-analytic effect of each experimental condition was for each modality across sites, and what the overall effect of condition was across all sites regardless of modality. We provide a summary of the results from this multisite replication endeavor along

with the implications of the results and future directions for research. We then discuss the primary aim of the overall project, that is, assessing the feasibility and usefulness of multisite replication approaches in L2 research.

Attending to Form While Processing for Meaning
Regarding whether attending to L2 form interferes with the processing of a text for comprehension, our results revealed the following:

- The ANOVAs that followed Leow et al. (2008) did not show a statistically significant effect for comprehension for the experimental groups compared to the control groups, except in US timed written Site 2, where attending to the lexical form *sol* reduced comprehension.
- The random-effects meta-analysis revealed that the meta-analytic effects for each of the experimental conditions compared to the control condition were not reliable and did not reach the level to be considered small effects. In addition, they mostly fell within the 95% CIs of Leow et al.'s effect sizes.
- The mixed-effects model also did not reveal an effect of condition on comprehension over all the sites.
- Bayes factors provided evidence that comprehension in two of the experimental conditions, that is, *sol* and *la*, did not differ from those in the control condition. For the *–n* condition, there was not sufficient evidence to conclude that performance was similar to the control condition although the Bayes factor for this comparison (0.40) was closer to indicating that comprehension in the *–n* condition was similar to (i.e., $B \leq 0.33$) rather than different from (i.e., $B \geq 3.0$) comprehension in the control condition.

Overall, for the population represented by the participant samples included in this multisite replication, the results from the different analyses largely converged in that evidence was not provided for an effect of attending to lexical or grammatical form on L2 comprehension. These results are consistent with the findings from Leow et al. and Morgan-Short et al. (2012), where attention to form was also not found to affect L2 comprehension. The results extend the findings from these previous studies to an aural paradigm as well as to a timed written paradigm where backtracking was not possible.

It is interesting to note that our conclusions about the results might have been different if the study had been conducted at just one of these sites or had been based on just one analytic approach. For example, if the written study had only been run at US Site 2, we might have concluded that attending to a lexical form seemed to have interfered with comprehension. Or if only the effect-size analysis had been run for UK Site 2 or for the Poland site, we might

have concluded that attending to inflectional verb morphology had affected comprehension. Fortunately, because of the multisite endeavor, site-specific results (with relatively small sample sizes) were not overly interpreted as the population effect.

The multisite approach also provided a unique insight into the generalizability of the finding. The finding was fairly consistent among L2 learners who were taking Spanish language classes of similar levels at seven universities in three countries. Thus, the finding may generalize to different learners in different universities and in different countries. Important also are the possible limits on the generalizability of the findings because the results were largely consistent but not entirely uniform among the different sites. That is, from site-specific results that deviate from the overall results, we can begin to formulate evidence-based hypotheses for further studies that could investigate these apparent anomalies that sit within more robust, broader trends. For example, given the reliable effects between the –n and control conditions in the Poland site (Figure 2), one may conclude that the findings may not generalize (a) to learners with a first language (L1) other than English, (b) to learners who have more experience learning L2s (as most had already learned English), or (c) to learners who are not reading comprehension questions in their L1. However, a reliable effect between the –n and control conditions was also evidenced at UK Site 2, which (along with the Poland site) had a higher level of proficiency compared to all other sites (see Table 1). Thus, the overall null effect on comprehension for condition may not generalize to learners who are at higher levels of proficiency while they are attending to nonsalient morphological forms such as clitics or forms that are nonsyllabic or verb final. Although it is not possible to draw these conclusions from the current study itself, anomalous findings in the context of a large multisite study can provide clues to variables that could merit further examination.

Another aspect of generalizability to be considered is whether similar effects would be found when using different materials. Our overall results are consistent with those reported by Leow et al. (2008) and Morgan-Short et al. (2012), from whose studies the materials were adapted. The results are also consistent with the findings for the written modality in Wong (2001), who used different materials. However, the results are not consistent with the results from other studies that used a different set of materials: the aural modality in VanPatten (1990) and in Wong and the written modality in Greenslade et al. (1999). These studies used materials with shorter L2 passages (275 vs. our 358 words) and a different method of assessing comprehension (free recall of idea units vs. our multiple-choice test). We also cannot rule out the possibility that the participants

in the current study might have shown an effect of attending to form if their comprehension had been assessed with a free recall comprehension assessment. For the current study, participants' scores on the multiple-choice test might have partially reflected random guessing although almost all conditions at all sites performed at an above-chance level (see Figure 1). In VanPatten's study, however, very little of the recall was likely to have been based on random guessing. Thus, even though the range of percent accuracy for the current study (23–52%) was generally higher than the percentage of all ideas units recalled by participants in VanPatten (17–36%), the recall used in the VanPatten study may have better captured participants' abilities to comprehend the passage in part because the score was less likely to reflect guessing.[12]

Indeed, it seems that the comprehension test administered in this multisite endeavor was not an ideal assessment of comprehension. First, the reliability of the comprehension test was quite low. This could have been due to the test itself (i.e., low item consistency) or to participants not having understood enough to demonstrate comprehension reliably on the test. These issues are difficult to tease apart, but we note that participants showed generally low levels of comprehension—as indicated by accuracy in the control groups, whose means ranged from 28% to 52% accuracy, with chance level at 25%. Apart from the low reliability of the comprehension test, the generally low level of comprehension may have made it difficult to detect effects of different conditions because, if comprehension was near floor, there might not have been room for it to be negatively affected and/or there might not have been sufficient variance in the data to show any effect of variables that might explain variance. We return to suggestions to resolve these issues below.

Two additional limitations of the current study should be acknowledged before we draw general conclusions from the experiment. First, one may argue that the rapid serial visual presentation used for the timed written paradigm is not ecologically valid although, as noted above, research has shown that rapid serial visual presentation does not necessarily disrupt normal reading comprehension processes (Juola et al., 1982), especially if the rate of presentation is not fast (Ricciardi & Di Nocera, 2017). This was corroborated by our findings that comprehension was the same or higher in the written modality. Second, although only a few site-specific effects were observed in the comprehension data, there were effects found in other regards. For example, in three sites— US Site 3, UK Site 1, and US Site 4—participants in the –*n* condition made statistically significantly fewer checks or clicks than participants in the other conditions, even though all participants made at least six checks or clicks as required to be included in the analysis (see Table S1 in Appendix S2 in the

Supporting Information online). Thus, while comprehension for these conditions was not lower than that for the other conditions within each site, there might have been some trade-off between attending to form (as operationalized by making checks or clicks) and comprehension. Finally, descriptively there seemed to be higher rates of exclusion from the data set for the –*n* condition largely based on not meeting the requirement of having made six or more checks or clicks at two sites—US Site 1 and US Site 3 (see Table S1 in Appendix S2 in the Supporting Information online). These participants were not included in the analysis because they did not complete the attention to form task appropriately, so they did not impact the results themselves, but the results still hint at a tension between learners attending to both form and meaning. Indeed, they could be interpreted as lending broad support to the notion expressed in the primacy of content words principle (VanPatten, 2015), that "learners process content words in the input before anything else" (p. 115), though this would need to be corroborated with precise operationalizations of processing for form and meaning.

Theoretical Implications and Suggestions for Future Work
Considering the results across different sites and analyses along with the noted limitations of the study, what conclusions can be drawn in regard to the original theoretical question that motivated this line of research, that is, whether learners can attend to L2 forms while processing input for meaning without their comprehension being affected? We argue that, for the current paradigm, the results provide consistent evidence that a task of attending to an unbound form, whether it be a more communicatively meaningful lexical form or an arguably less communicatively meaningful grammatical form, does not interfere with comprehension during both listening and reading, at least when levels of comprehension are not very high. Neither of the across-site analyses (i.e., the meta-analytic effect or mixed-effects model analyses) showed evidence of negative effects for attending to *sol* or *la*, although there was a negative, site-specific effect for *sol* at US Site 2. The meta-analytic (weighted) mean effects for these conditions were positive ($d = 0.10$ and 0.19, respectively) but were not reliably above chance. Additionally, the Bayes factors for these conditions compared to the control conditions suggest that comprehension in these conditions was similar to that of the control group. For these conditions, attending to form did not seem to make comprehension any more difficult for participants, at least not when they were asked to indicate that they attended to the forms by making a check mark or mouse click.

Extending such a conclusion to the bound morpheme –*n* condition, however, may need to be tempered. Although neither of the across-site analyses evidenced a negative effect for this condition, the overall meta-analytic (weighted) mean effect was negative ($d = -0.38$), and there were negative effects on comprehension in two sites—the Poland site and UK Site 2 as evidenced in the effect-size analysis. There were also indications that participants were less successful at attending to form in the –*n* condition compared to the other experimental conditions (as evidenced by statistically significantly fewer numbers of check marks or clicks and descriptively more participants who did not make the minimum number of checks or clicks for inclusion). The reasons for the site-specific effects in the –*n* condition are difficult to know as the –*n* form differs in many ways from the forms *sol* and *la*: Whereas *sol* and *la* are syllabic, full words that are relatively invariant, –*n* is an unstressed, nonsyllabic element at the end of a verb that can co-occur with morphemes that vary in form and meaning (e.g., –*ían*, –*aron*, –*aban*, –*an*, –*en*) and that indicate Spanish tense, aspect, and mood.

Overall, our study provides evidence that L2 learners' comprehension is not affected by their attending to particular unbound forms, that is, *sol* and *la*, within a context where learners are reading or listening to a relatively short, controlled L2 passage. However, our study is not able to provide positive evidence that learners attending to the bound form –*n* does not affect their comprehension. Indeed, site-specific results both for comprehension and for how well participants were able to attend to the form suggested that a general conclusion that learners can attend to both form and meaning while focused on comprehension is not warranted because such a conclusion may not apply uniformly to all forms and contexts. Future research that is based on current iterations of the input processing theory (VanPatten, 2015) and other perspectives related to L2 processing (e.g., N. Ellis et al., 2014; N. Ellis & Wulff, 2015; Leow, 2015) may want to explore the boundaries of when learners can attend to form and meaning while processing input for comprehension (see Marsden, Williams, & Liu, 2013, for such a study).

We are similarly tentative in drawing any implications for pedagogy due to the concerns mentioned above. In terms of informing L2 instruction for learners like those in this study, it might be tempting to draw on our finding that offline comprehension was not generally affected by a requirement to also allocate attention to lexical or grammatical forms in the input and suggest that classroom activities that require learners to attend to (e.g., underline or circle) specific items in the input may not adversely affect overall comprehension. Such a conclusion might be premature given our concerns about the comprehension test. We should also note a potential lack of ecological validity in using an

activity that resulted in such low levels of comprehension. However, of interest are findings from Marsden and colleagues' studies, where word- and sentence-level tasks that are more ecologically valid to L2 pedagogy were processed in such a way that participants (a) successfully attended to the form of an article and the meaning of a sentence (Kasprowicz & Marsden, 2017) and (b) showed their ability to learn the meaning of a word even while their attention was oriented to the meaning of a form (Marsden et al., 2013, Experiment 3). Thus, there seem to be conditions where learners can successfully attend to both L2 form and meaning, and perhaps one of the roles of L2 instruction is to create these conditions. This is precisely the purpose of processing instruction (Van-Patten, 2004b, 2005), where explicit information about a L2 form is provided and then input is structured so that learners can attend to the form and process its meaning.

   Given the more than 1,000 citations of VanPatten (1990), it is clear that the field of L2 acquisition has a strong interest in understanding the conditions in which learners can or cannot attend to form and meaning while processing L2 input for comprehension. However, even after analyzing data from 631 L2 learners with a paradigm that incorporated methodological improvements over previous paradigms, we still do not have clear answers to all pertinent theoretical questions. In order to address this critical question in a more robust manner, we believe that future experiments will need to incorporate the following methodological recommendations into their research design.

- Regarding the dependent comprehension variable, studies should fully pilot their measure of comprehension in order to establish (a) that scores will not be around floor or ceiling level and (b) that the reliability of the comprehension test is acceptable. More specifically, we would not recommend that the current comprehension test be used unless it is established that its reliability is acceptable for a particular population. Without a reliable test of comprehension, researchers will not be able to make valid claims regarding whether comprehension has or has not been affected.
- As mentioned in Leow et al. (2008), researchers need to establish that participants are engaged in both processing for form and for meaning. In the current study, we were able to control backtracking with a timed written condition where words appeared on the screen, so we can claim that participants do not first process the input for meaning and then go back and find the forms, but we still cannot claim that they consistently attended to meaning or successfully attended to form, especially for the –n group. Perhaps an ecologically valid manner of creating such conditions would be through an

eye-tracking paradigm where sentences could be presented one at a time. Participants would be asked to read the sentence normally, to make a mouse click if they noticed a target form, and not to go back and reread the sentence for meaning or to find a form. The number of check marks would be taken as evidence of attention to form and an evaluation of the eye-tracking data could be taken as evidence of attention to meaning. For example, a baseline could be taken to establish the average number of regressions that each participant makes while reading, and then regressions that fall outside that normal range while a participant is reading experimental stimuli could be used to eliminate such trials. For a listening paradigm, as suggested by a reviewer, participants could also be asked to make mouse clicks while listening and the timing of the mouse clicks could be recorded and aligned with the timing to the aural passage so that researchers could reasonably establish that the clicks were in response to the target forms. This would provide more confidence in the internal validity of learners attending to form. However, it is not clear how it could be established that learners attended to meaning and did not just listen for the forms except for above-chance performance on a subsequent comprehension test.

- Future research may want to establish that the dual task aspect of the experiment is sufficiently challenging to the participant such that it does require cognitive resources. Perhaps the task could be tested first with a different paradigm, for example, one that is not linguistic, to demonstrate that the task itself is cognitively demanding and affects participants' performance on a primary task under dual-task conditions.

- Researchers may want to better control the differences between target forms, that is, salience, length, syllables, and the like. For example, in order to test whether the boundedness of a morpheme makes it more difficult for participants to attend to both form and meaning, researchers could examine the effect of attending to direct object pronouns in Spanish, which have the same form whether they are bound or unbound.

Overall, these and/or other methodological improvements should be established by researchers moving forward on the issue of attention to form and meaning while processing L2 input for comprehension.

Multisite Registered Replications
Registered Replication Reports in L2 Research
The principal objective of the current study was to ascertain the feasibility and usefulness of incorporating a multisite registered replication report approach

(Simons & Holcombe, 2014; Simons et al., 2014) in L2 research. Many aspects of this registered replication report mechanism were adopted by the current study. First, seven different sites collected data independently for the study. Second, some aspects of the study were publicly registered prior to data collection and analysis. Although the materials and protocols for the partial replications—US aural Site 1 and UK timed written Site 1—were not registered, they were registered for the direct replications. Also, a template was registered for data entry at each site, and the detailed analysis plan was publicly posted but not registered. The fact that the materials, procedures, and the data entry sheet for the direct replications were registered decreased multiple researcher degrees of freedom, prevented intended deviations from the direct replication, and lessened the likelihood of unintended changes. Thus, researchers should have been largely impeded from affecting the results of the study during the course of the study, even unintentionally. Also, the fact that the analysis protocol, data, and results were posted publicly created full transparency of the research findings and helped to discourage *p* hacking.

The current study differed from the full registered report aspect of registered replication reports in that these require that the motivation, materials, procedures, and analysis be fully peer reviewed, approved, and then registered before any data collection occurs. Although our materials were not peer reviewed or approved, our experience with registration allows us to comment on the feasibility of a registered report approach. Based on our experience over the course of the full research project, planning the methods, procedures, and analyses in order to register them before they were carried out did not incur additional work or resources but rather required a significant shift in the order of our workflow. Thus, given (a) that registration should not require additional work or resources and (b) that infrastructure exists to support it—for example, the IRIS database and the Open Science Framework—we argue that a full implementation of a registered report mechanism is feasible and has considerable benefits, especially for replication studies of important initial studies that have high levels of internal validity and reliability.

Indeed, we believe that our study would have benefited from following a complete registered report approach had such a mechanism been available. As noted above, registered reports through journals involve peer review before data collection. With peer review of the materials and protocol, we might have made adjustments to the design, materials, and protocols developed for the partial replications of Leow et al. (2008) carried out at the lead sites—US aural Site 1 and UK timed written Site 1. These adjustments might have increased our ability to interpret the results in regard to their theoretical and pedagogical

implications. Fortunately, future L2 research can now benefit from peer review prior to data collection along with other benefits of registered reports such as publication decisions that are unbiased by statistical significance because a registered report article type is now offered through *Language Learning* (Marsden, Morgan-Short, Trofimovich, & N. Ellis, 2018). We recommend that registered reports be adopted more widely as one of many mechanisms to promote replication and robust research practices (see Marsden, Morgan-Short, Thompson, & Abugaber, 2018).

With respect to the multisite aspect of the study, the clear benefits of increasing the external validity of a finding and providing insight into the size of an effect in the population, within the constraints of the materials and procedures chosen, makes such an approach highly recommendable. However, there are also clear challenges to this approach. Such a large endeavor requires time to coordinate the research among the various sites as well as financial resources, for example, paying research assistants to coordinate the sessions or even helping to pay for sites to purchase software that is required to conduct the study. Indeed, the current study was supported by the then *Language Learning* Small Research grant program, and the funds were used for research assistant support and for participant compensation at some sites. Whereas the registered replication reports now hosted by the *Journal of Advances in Methods and Practices in Psychological Science* started with a fund of $250,000 (https://www.psychologicalscience.org/publications/replication#FUND), no such funding exists for the field of L2 research.

Other aspects of multisite endeavors may be challenging for our field. The potential for real or perceived bullying in replication research has been noted in psychology (see Bohannon, 2014, also discussed in Marsden, Morgan-Short, Thompson, & Abugaber, 2018), and the negative effects of this could be even more harmful in a high-profile, multisite replication effort, given that multiple researchers may be perceived as going against one researcher or research team. However, we hope that with the careful establishment of infrastructure and an enhanced collaborative and synthetic ethic in the research community, such risks should be minimized. Also, multisite endeavors may be particularly difficult to pursue in context-sensitive fields such as L2 research where many context-dependent variables are known to be at play, for example, L1 or L2 experience; educational context; and sociolinguistic, economic, cultural, and affective variables. These L2-specific challenges are in addition to the more normal challenges of multisite work, such as individual differences among participants' age, cognitive abilities, or educational attainment, and are also on top of practical issues, such as the availability of software, hardware, and

incentivization to engage participants. However, multisite approaches also provide an opportunity to measure variables that perhaps otherwise could not be explored.

In sum, although multisite endeavors should be pursued, both for replication and initial research (for an example of an initial multisite study, see VanPatten, Collopy, Price, Borst, & Qualin, 2013), such an approach may be difficult to adopt systematically as a field. However, individual researchers may choose to engage in multisite research for some studies. Researchers interested in initiating such initiatives may choose to take advantage of resources such as the Study Swap on the Open Science Framework site, where researchers look for and offer themselves as multisite collaborators, and the Calls for Replication Collaborators on the IRIS site.

Additional Reflections

We would like to provide a more reflective, introspective discussion about issues in conducting this multisite replication study given the findings of the synthesis of self-labeled replication in L2 research (Marsden, Morgan-Short, Thompson, & Abugaber, 2018). As stated previously, the proposing authors of this article began the project with the intention of examining the feasibility of incorporating multisite registered replication reports into our field. With this goal in mind, we searched for a research question, paradigm, and materials that were suitable for our purpose. In addition to considering the replication value of previous research, we were also concerned with the availability of materials. We had both conducted previous studies with the current paradigm and thus had access to the materials and were familiar with them. In an ideal world, the availability of the materials would not be a major consideration in choosing a paradigm to replicate. Having access to the materials entailed advantages such as reducing the number of researcher degrees of freedom and minimizing the heterogeneity among studies that would have necessarily resulted from recreating materials from a study for which we did not have the materials. However, access to materials that we had already used in our research also made us vulnerable to the very concern raised by Marsden, Morgan-Short, Thompson, and Abugaber about author overlap between initial and replication studies, that is, that we were making ourselves susceptible to the risk of researcher bias, which could engender questionable research practices. We took steps to counter these, such as preregistration of the protocol and analysis, employing a multisite approach, and making all the data and analyses openly available. But the fact remains that we were restricted in our choices due to the lack of wide availability of full sets of materials with full protocols, score sheets, analysis protocols, and even previous

data sets to combine and compare our analyses (Marsden, Mackey, & Plonsky, 2016).

An additional issue of note is related to our choice of the Spanish proficiency test. We opted to use the Diplomas de Español como Lengua Extranjera (Seibert Hanson & Carlson, 2014) in large part because it was freely available via IRIS. This illustrates the point that the open availability of materials, vital for ascertaining parity across multiple sites, is very helpful. The availability of particular materials, however, may lead to the use of such materials over others, for better or worse. This may engender the undesirable situation that openness results in an overuse of certain materials. Thus, we need to work as a field to make all materials available, not just some. Equal visibility of all materials will reduce the potentially harmful effects of choosing materials just because they are available.

Finally, in deciding which research paradigm to replicate, we considered another study that was also closely related to the theoretical issues of form–meaning connections and attention in L2 acquisition (Marsden et al., 2013). However, this other paradigm required specialized software and comprised three experiments. Thus, we elected to go with a study that we believed would be more feasible for multiple researchers at different institutions. We were also hesitant about replicating Marsden et al. because of its null results in terms of crossmodal priming. In the end, we chose to replicate Leow et al. (2008), which also had null results but was situated in a line of research where statistically significant results had been evidenced. We also ended up using specialized software although we were able to do so only in three sites. Our reflection here relates to the variables that influence our field's decisions about what to replicate. Resource requirements are definitely one consideration, even though Marsden, Morgan-Short, Thompson, and Abugaber (2018) did find some replication research with considerable demands on resources. Another arguably more important issue, though, relates to the extent to which researchers will undertake replications of studies with null findings. Indeed, when presenting our initial results at a conference, we were challenged about whether we should expect others to join us in replicating null results. Further illustrating this concern, Marsden, Morgan-Short, Thompson, and Abugaber found that replications of studies with null findings were extremely rare.

Overall, although there were significant risks in moving forward with the paradigm that we chose, we also had assurance that we could conduct the study because of the financial support from a *Language Learning* Small Research grant (which entailed the potential for a publication as the journal retained first rights to publication for awardees). We also note that the award was made largely

in recognition of the primary purpose being to investigate the multisite replication approach itself, thus giving us an additional incentive to invest the effort. Without these assurances, proceeding with the study would have carried more risks. Similarly, registered reports can provide an important form of assurance and confidence for researchers to engage in motivated and methodologically sound research endeavors that may otherwise be deemed overly risky. Because the theoretical motivation, research design, and materials are fully reviewed and given in-principle acceptance before data are collected, researchers who receive in-principle acceptance know that their study will be published regardless of the statistical significance of the results, so long as they follow the approved protocol. Even if researchers do not receive in-principle acceptance, they will have received valuable feedback before having run their study.

By reflecting on our decisions in this way, we open potentially sensitive, though likely widespread, concerns to collective scrutiny. However, we believe that such transparency (about motivations, materials, analyses, and data) along with open science infrastructure (such as IRIS, the Open Science Framework, and registered reports in *Language Learning*) can inform decision making and facilitate a more collaborative ethic in the field.

Conclusion

We conducted a multisite replication study with aspects of preregistration in order to explore the feasibility and usefulness of a multisite registered replication approach in the field of L2 acquisition. In doing so, we addressed ongoing questions about attention to L2 form and meaning. In regard to the question about whether attending to form while listening to or reading a L2 passage would interfere with comprehension of that passage, results from the current study indicated (a) that an effect of attention to form on comprehension was not detected in by-site ANOVA analyses that followed previous research except at one site where attending to the lexical form *sol* led to reduced comprehension, a potential anomaly that we cannot account for; (b) that the random-effects meta-analytic effect size for each experimental condition compared to control was not reliable or of meaningful magnitude; (c) that no effect of condition was evidenced when examining the data across all sites and modalities; and (d) that across all sites, text comprehension for two of the experimental conditions was similar to that for the control condition. Thus, overall, for the population represented by the participant samples included in this multisite replication, the results from the different analyses largely converge and provide evidence that attending to at least some lexical or grammatical forms does not seem to affect L2 listening or time-controlled reading comprehension, at least for

unbound forms when comprehension is relatively low. Importantly, though, there was some indication of participants having difficulty attending to the bound, morphosyntactic form as evidenced by effect size analyses and also by effects related to how well participants were able to engage in the task to attend to this form. These conclusions, however, must be considered in light of the limitations of the experiment, particularly the low reliability of the comprehension test. Because of this and other limitations of this paradigm, we recommend that new paradigms be developed to further investigate questions of whether attention to L2 form while processing input for meaning makes comprehension difficult.

Regarding the feasibility and usefulness of incorporating a multisite registered replication report approach (Simons & Holcombe, 2014; Simons et al., 2014) into the field of L2 research, we found that the registered report aspect would be a feasible mechanism for our field even though it required a shift in the workflow of our project compared to our previous research endeavors. The registered report mechanism now available through *Language Learning* will require additional time to register the design, materials, and planned analyses of a research project and also to undergo peer review prior to data collection. However, we argue that this adjustment in the workflow has multiple benefits, including peer review that is unbiased by results, feedback on design prior to data collection, and increased transparency of research practices more generally. With respect to the multisite aspects of the endeavor, our experience suggested that this approach can be accomplished with appropriate resources. However, it might pose more of a challenge in terms of its feasibility in L2 research than similar endeavors in social and cognitive psychology. Overall though, both mechanisms—registered replications and multisite collaborations—should be adopted by our field to some degree to increase the robustness and generalizability of findings in our field.

<div align="right">Final revised version accepted 23 March 2018</div>

Notes

1 However, only one replication of VanPatten (1990) self-labeled as a replication; see Marsden, Morgan-Short, Thompson, and Abugaber (2018) for implications of the failure to self-label as a replication.

2 This pattern held for the first-semester and third-year learners and was slightly different for the fourth-semester learners whose results followed the general pattern except that there was not a difference between the experimental group that paid attention to the grammatical form –*n* and the control group.

3 Leow et al. (2008) and Morgan-Short et al. (2012) additionally addressed issues of depth of processing based on think-aloud protocol data. Leow et al. found overall low levels of processing and suggested that attentional resources may not have been tapped in such a way that they would interfere with comprehension. Morgan-Short et al. found variation in the levels of processing and a positive correlation between more in-depth processing and increased comprehension, which would not be predicted by the primacy of meaning principle. These results are not fully considered here because the current study did not administer think-alouds and thus cannot address questions of depth of processing.

4 Conclusions based on studies that did not find statistically significant effects are also somewhat limited in that they relied exclusively on null hypothesis testing, meaning that they could only conclude that no differences were detected but that they could not conclude that there were no differences because the null result could be due to an actual lack of differences, low power, or high degrees of variance in the data (Dienes, 2014).

5 A presentation rate by syllables was also considered but was deemed very difficult to read by the researchers because of the variation in word duration based on the different number of syllables among the words.

6 A fourth experimental condition (the masculine pronoun clitic *lo*) from Leow et al. (2008) and Morgan-Short et al. (2012) was not included in the current study. This fourth condition did not appear in VanPatten (1990) nor the replication studies preceding Leow et al.'s conceptual replication and was not used in the current study to facilitate securing at least 15 participants per condition.

7 Morgan-Short et al. (2012) did not report a measure of reliability, but the first author of that article performed a reliability analysis on the data and found alpha levels of .153 across all participants and .213 for participants in the non-think-aloud group, which consisted of participants who were not asked to think aloud during the task as was the case in the current study.

8 Descriptively, the reliability coefficients tended to be higher for sites that reported higher levels of proficiency and were also generally higher for the aural modality than for the written modality.

9 Parallel analyses were conducted for the reaction time data obtained from E-Prime and SuperLab in the aural modality. However, we do not report or discuss these analyses in this article because there are several constraints on a valid interpretation of these data. First, the study was not specifically designed for the purpose of examining reaction time data. For example, the target forms in the experimental conditions were not matched. Whereas *sol* and *la* are syllabic, full words that are relatively invariant, *–n* is an unstressed, nonsyllabic element at the end of a verb that can co-occur with morphemes that vary in form and meaning (e.g., *–ían*, *–aron*, *–aban*, *–an*, *–en*) and that indicate Spanish tense, aspect, and mood. Also, there may be variability within the *–n* condition because it was found on the end of verbs of different lengths whereas for *sol* and *la* the length was

always shorter, constant, and predefined. Because the study was not designed as a reaction time study, these variables are inherently confounded with condition. Additionally, even if a valid interpretation of differences among experimental conditions could be made, they would still not be parallel to that of the accuracy data because there were no reaction time data for the control condition. A second potential issue is that, although the timing specifics of the software programs used to present the passage are not expected to differ meaningfully, the timing specifics of different hardware configurations may have impacted the reaction time output (Stahl, 2006). Future multisite studies collecting reaction time or online data will need to consider validating the timing of different software and hardware systems (Plant, 2016), although for paradigms that do not require precise millisecond timing, the benefits of collecting data across larger samples may outweigh disadvantages in timing variability (van Steenbergen & Bocanegra, 2016). Given these issues, we are not confident in being able to offer a valid interpretation of the reaction time data, but we do provide access to the data and results on our public Open Science Framework analysis page (https://osf.io/nz3su) in the folder for Written Data Analysis under Files.

10 The marginal interaction of condition and modality appeared to be driven by a significant difference between the written control condition and the aural *–n* condition, which was not a contrast of theoretical interest.

11 The percentage reduction of comprehension for the grammatical conditions that differed from the control conditions in previous studies are as follows: (a) VanPatten (1990), comprehension reduced by 42% for the *la* condition and by 58% for the *–n* condition; (b) Greenslade et al. (1999), comprehension reduced by 43% for the *la* condition and by 39% for the *–n* condition; and (c) Wong (2001), comprehension reduced by 77 % for the aural condition. Thus, the average reduction in comprehension was 52%.

12 A disadvantage of the recall assessment, as pointed out in Leow et al. (2008), is the inability to account for the relatively large amount of variance and individual approaches to the recall process.

References

Association for Psychological Science. (2017). *Registered replication reports at AMPPS: Instructions for authors*. Retrieved June 23, 2017, from https://www.psychologicalscience.org/publications/ampps/rrr-guidelines

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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

Bohannon, J. (2014). Replication effort provokes praise—and "bullying" charges. *Science*, *344*, 788–789. https://doi.org/10.1126/science.344.6186.788

Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education*, *40*, 218–231. https://doi.org/10.1080/02602938.2014.902192

Cintrón-Valentín, M. C., & Ellis, N. C. (2016). Salience in second language acquisition: Physical form, learner attention, and instructional focus. *Frontiers in Psychology*, *7*, 1284. https://doi.org/10.3389/fpsyg.2016.01284

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*, 170–180.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Doughty, C., & Williams, J. (1998). *Focus on form in classroom second language acquisition*. New York: Cambridge University Press.

E-Prime (Version 2.8) [Computer software]. Pittsburgh, PA: Psychology Software Tools.

Ellis, N. C., Hafeez, K., Martin, K. I., Chen, L., Boland, J., & Sagarra, N. (2014). An eye-tracking study of learned attention in second language acquisition. *Applied Psycholinguistics*, *35*, 547–579. https://doi.org/10.1017/S0142716412000501

Ellis, N. C., & Wulff, S. (2015). Usage-based approaches to SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 75–93). New York: Routledge.

Ellis, R. (2016). Focus on form: A critical review. *Language Teaching Research*, *20*, 405–428. https://doi.org/10.1177/1362168816628627

Fox, J., & Weisberg, S. (2011). *An {R} Companion to Applied Regression, Second Edition*. Thousand Oaks, CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Greenslade, T., Bouden, L., & Sanz, C. (1999). Attending to form and content in processing L2 reading texts. *Spanish Applied Linguistics*, *3*, 65–90.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal 50*(3), 346–363.

Juola, J. F., Ward, N. J., & McNamara, T. (1982). Visual search and reading of rapid serial presentations of letter strings, words, and text. *Journal of Experimental Psychology: General*, *111*, 208–227. https://doi.org/10.1037/0096-3445.111.2.208

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*, 228–238.

Kasprowicz, R., & Marsden, E. (2017). Towards ecological validity in research into input-based practice: Form spotting can be as beneficial as form-meaning practice. *Applied Linguistics*. Published online February 13, 2017. https://doi.org/10.1093/applin/amw051

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š, Bernstein, M. J., et al. (2014). Investigating variation in replicability. *Social Psychology*, *45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Leow, R. P. (2001). Attention, awareness, and foreign language behavior. *Language Learning*, *51*, 113–155. https://doi.org/10.1111/j.1467-1770.2001.tb00016.x

Leow, R. P. (2015). Toward a model of the L2 learning process in instructed SLA. In R. P. Leow (Ed.), *Explicit learning in the L2 classroom: A student-centered approach* (pp. 236–250). New York: Routledge.

Leow, R. P., Hsieh, H., & Moreno, N. (2008). Attention to form and meaning revisited. *Language Learning*, *58*, 665–695. https://doi.org/10.1111/j.1467-9922.2008.00453.x

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*, 1827–1832. https://doi.org/10.1177/0956797615616374

Mackey, A., & Gass, S. M. (2016). *Second language research: Methodology and design* (2nd ed.). New York: Routledge.

Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). New York: Routledge. https://doi.org/10.4324/9780203489666

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*. https://doi.org/10.1111/lang.12286

Marsden, E., Morgan-Short, K., Trofimovich, P., & Ellis, N. C. (2018). Editorial: Introducing Registered Reports at *Language Learning*: Promoting transparency, replication, and a synthetic ethic in the language sciences. *Language Learning*, *68*. https://doi.org/10.1111/lang.12284

Marsden, E., Williams, J., & Liu, X. (2013). Learning novel morphology: The role of meaning and orientation of attention at initial exposure. *Studies in Second Language Acquisition*, *35*, 619–654. https://doi.org/10.1017/S0272263113000296

Morgan-Short, K., Heil, J., Botero-Moriarty, A., & Ebert, S. (2012). Allocation of attention to second language form and meaning. *Studies in Second Language Acquisition*, *34*, 659–685. https://doi.org/10.1017/S027226311200037X

Nederlandse Organisatie voor Wetenschappelijk Onderzoek. (2017). *Replication studies*. Retrieved November 1, 2017, from https://www.nwo.nl/en/research-and-results/programmes/replication+studies

Norris, J. M., Ross, S. J., & Schoonen, R. (2015). Improving second language quantitative research. *Language Learning*, *65*, 1–8. https://doi.org/10.1111/lang.12110

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2017). *The preregistration revolution*. Retrieved January 3, 2018, from https://osf.io/2dxu5

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. https://doi.org/10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*. https://doi.org/aac4716. 10.1126/science.aac4716

Öquist, G., & Goldstein, M. (2003). Towards an improved readability on mobile devices: Evaluating adaptive rapid serial visual presentation. *Interacting with Computers*, *15*, 539–558. https://doi.org/10.1016/S0953-5438(03)00039-0

Plant, R. R. (2016). A reminder on millisecond timing accuracy and potential replication failure in computer-based psychology experiments: An open letter. *Behavior Research Methods*, *48*, 408–411. https://doi.org/10.3758/s13428-015-0577-0

Plonsky, L. (2015). *Advancing quantitative methods in second language research*. New York: Routledge.

Plonsky, L., & Ghanbar, H. (in press). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *The Modern Language Journal*.

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Polio, C., & Gass, S. M. (1997). Replication and reporting. *Studies in Second Language Acquisition*, *19*, 499–508.

Porte, G. (2012). *Replication research in applied linguistics*. New York: Cambridge University Press.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, *62*, 1457–1506. https://doi.org/10.1080/17470210902816461

Ricciardi, O., & Di Nocera, F. (2017). Not so fast: A reply to Benedetto et al. (2015). *Computers in Human Behavior*, *69*, 381–385. https://doi.org/10.1016/j.chb.2016.12.047

Robinson, P. (1995). Attention, memory, and the "noticing" hypothesis. *Language Learning*, *45*, 283–331.

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

Seibert Hanson, A., & Carlson, M. (2014). The roles of first language and proficiency in L2 processing of Spanish clitics: Global effects. *Language Learning*, *64*, 310–342. https://doi.org/10.1111/lang.12050

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80. https://doi.org/10.1177/1745691613514755

Simons, D. J., & Holcombe, A. O. (2014). Registered replication reports: A new article type at perspectives on psychological science. *Observer*, *27*. Retrieved June 22, 2017, from http://goo.gl/YMlCqv

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*, 552–555. https://doi.org/10.1177/1745691614543974

Stahl, C. (2006). Software for generating psychological experiments. *Experimental Psychology*, *53*, 218–232. https://doi.org/10.1027/1618-3169.53.3.218

Superlab (Version 5.0) [Computer software]. San Pedro, CA: Cedrus Corporation.

VanPatten, B. (1990). Attending to form and content in the input. *Studies in Second Language Acquisition*, *12*, 287–301.

VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Norwood, NJ: Ablex.

VanPatten, B. (2004a). Input processing in second language acquisition. In B. VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* (pp. 5–31). Mahwah, NJ: Erlbaum.

VanPatten, B. (Ed.). (2004b). *Processing instruction: Theory, research, and commentary*. Mahwah, NJ: Erlbaum.

VanPatten, B. (2005). Processing instruction. In C. Sanz (Ed.), *Mind and context in adult second language acquisition: Methods, theory, and practice* (pp. 267–281). Washington, DC: Georgetown University Press.

VanPatten, B. (2015). Input processing in adult second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 113–134). Mahwah, NJ: Erlbaum.

VanPatten, B., & Cadierno, T. (1993a). Explicit instruction and input processing. *Studies in Second Language Acquisition*, *15*, 225–241.

VanPatten, B., & Cadierno, T. (1993b). Input processing and second language acquisition: A role for instruction. *The Modern Language Journal*, *77*, 45–57. https://doi.org/10.2307/329557

VanPatten, B., Collopy, E., Price, J. E., Borst, S., & Qualin, A. (2013). Explicit information, grammatical sensitivity, and the first-noun principle: A cross-linguistic study in processing instruction. *The Modern Language Journal*, *97*, 506–527. https://doi.org/10.1111/j.1540-4781.2013.12007.x

VanPatten, B., Williams, J., Rott, S., & Overstreet, M. (Eds.). (2004). *Form-meaning connections in second language acquisition*. Mahwah, NJ: Erlbaum.

van Steenbergen, H., & Bocanegra, B. R. (2016). Promises and pitfalls of web-based experimentation in the advance of replicable psychological science: A reply to Plant

(2015). *Behavior Research Methods*, *48*, 1713–1717. https://doi.org/10.3758/
   s13428-015-0677-x

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package.
   Journal of Statistical Software, *36*(3), 1–48. http://www.jstatsoft.org/v36/i03/

Wong, W. (2001). Modality and attention to meaning and form in the input. *Studies in
   Second Language Acquisition*, *23*, 345–368.

Supporting Information

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

Appendix S1. Site-Specific Data Collection Information.
Appendix S2. Detailed Condition-Specific Participant Information.
Appendix S3. Additional Data.