Multi-site replication in second language acquisition research:

Attention to form during listening and reading comprehension

**Proposing authors:**

Kara Morgan-Short, University of Illinois at Chicago; Emma Marsden, University of York; Jeanne Heil, University of Maine

**Contributing authors:**

Bernard I. Issa II, University of Tennessee; Anna Mikhaylova, University of Queensland; Sylwia Mikołajczak, University of Adam Mickiewicz in Poznań; Nina Moreno, University of South Carolina; Ronald P. Leow, Georgetown University; Roumyana Slabakova, University of Southampton; Paweł Szudarski, University of Nottingham

Author Note

**Multi-site replication of:** Leow, R. P., Hsieh, H., & Moreno, N. (2008). Attention to form and meaning revisited. *Language Learning, 58*(3), 665-695. doi:10.1111/j.1467-9922.2008.00453.x

**Registered materials and protocol:** <https://osf.io/uybak>

**Registered data entry template:** [https://osf.io/jbpag](https://osf.io/jbpag%20)

**Open data:** <https://osf.io/vwytd>

**Open analysis:** [https://osf.io/nz3su](https://osf.io/nz3su%20)

**Acknowledgments:** This research was supported by a *Language Learning* Small Research Grant with some partial support also provided by the British Academy award ARP AQ160001 and by the UK Economic and Social Research Council (RES‐062‐23‐2946). Versions of this study were presented at the 2016 Second Language Research Forum and at the 2017 meeting of the European Second Language Association. We thank attendees for their helpful comments on the work. We also thank Alaidde Berenice Villanueva Aguilera, Zerbrina Valdespino-Hayden, and Charlotte Oliver, for their assistance with various aspects of this study.

**Corresponding authors:**

Kara Morgan-Short, Department of Hispanic and Italian Studies, 601 South Morgan Street (MC 315), 1706 University Hall, Chicago, IL 60304; email: karams@uic.edu

Emma Marsden, D/M/107, Department of Education, University of York, York, YO10 5DD; email: emma.marsden@york.ac.uk

Abstract

We conducted a multi-site replication study with aspects of pre-registration 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 (VanPatten, 1990; Leow, Hsieh, & Moreno, 2008). Our results are consistent with findings from the specific paradigm that we replicated (i.e., Leow et al., 2008) in that no effects 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 multi-site registered replication approach to be useful, although the registration component itself appears to be a more feasible and useful first step in order to increase the robustness and generalizability of findings in our field.

Multi-site replication in second language acquisition research:

Attention to form during listening and reading comprehension

The ability of the field of second language (L2) acquisition to arrive at robust and generalizable conclusions crucially relies on the validity and reliability of our research (Mackey & Gass, 2016; Plonsky, 2015). However, our research 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 within 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 our research.

Several recommendations for improved research practices have been posited in the L2 field (e.g., Norris, Ross, & Schoonen, 2015), including the repeated call for increased replication research (e.g., Marsden, Morgan-Short, Thompson, & Abugaber, this issue; Polio & Gass, 1997; Porte, 2012), as in other fields (for a systematic review of replication issues in the field of L2 and more generally, see Marsden et al., this issue). However, the rate of replication is extremely low in L2 research, with fewer than 1 article in 400 being a replication study (Marsden et al., this issue), 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/](https://cos.io/?_ga=2.22827646.433504577.1515600180-132528251.1513112907)) and dedicated funding for registration from the Netherlands Organisation for Scientific Research.

In the current study, we adopted some features of one such mechanism that has emerged in the field of psychology—the multi-site registered replication report approach (Simons, Holcombe, & Spellman, 2014; Simons & Holcombe, 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 the multi-site registered replication approach report and a motivation for using such an approach to examine the open question of whether attending to L2 lexical and grammatical forms when 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 multi-site registered replication report approach, report the findings, and discuss the implications of the study, both substantively in terms of L2 theory and also, to the central point of the study, in terms of the extent to which such an approach might be a viable mechanism to promote replication and robust research practices in L2 research.

Note that throughout the article, we adopt the following nomenclature in discussing different types of replications as recommended by Marsden et al. (this issue): (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 pre-defined way, and (c) ‘conceptual replications’ are replications that introduce more than one change to one or more significant variables for the purposes of extending the initial study more broadly.

**Multi-site registered replication**

In recent years, the field of psychology has engaged in several multi-site replication endeavors, including those where multiple studies have been replicated by one lab each, such as the *Estimating Reproducibility Project: Psychology* (<https://osf.io/ezcuj/>)(Open Science Collaboration, 2015), and those where multiple labs have all replicated one or more studies, such as *The* *Many Lab* replication project (<https://osf.io/89vqh/>)(Klein et al., 2014). The most systematic endeavor to promote multi-site replication has come through registered replication reports (RRRs), which were introduced in the journal *Perspectives on Psychological Science* (Simons et al., 2014; Simons & Holcombe, 2014) and are now hosted by the journal *Advances in Methods and Practices in Psychological Science*(AMPPS; <https://www.psychologicalscience.org/publications/replication>). A primary purpose of RRRs is to inform the “true size of important effects” (Simons et al., 2014, 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) RRRs are a specific, official journal article submission type, which is highly valued for academic career and funding decisions, at least for the lead authors/convenors, 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 et al., this issue, and Marsden, Morgan-Short, Trofimovich, & Ellis, this issue).

The process of submitting and publishing a RRR (<https://www.psychologicalscience.org/publications/replication>) is somewhat different from a standard article (Simons et al., 2014; Simons & Holcombe, 2014). RRRs 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 multi-site 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 and on mixed effect model analyses across sites. For the article, the discussion reflects more 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. As of the date of writing, this mechanism for publication of replications has enjoyed relative success in psychology with six published RRRs 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 RRRs.

RRRs have been argued to be beneficial to scientific inquiry because of their characteristics of preregistration, direct replication, and multi-site approach. Preregistration enables researchers to make clear distinctions to others, and to themselves, about the elements of their research that reflect a priori prediction (which entail confirmatory analyses) and postdiction (which often entail 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 since 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 multi-site approach may be particularly valuable as it allows one to isolate the signal, i.e., the effect of interest, from the noise of error, e.g., the effects of heterogeneity such as different samples, contexts, and measurement error (Simons, 2014). Also, more simply, multi-site studies usually lead to larger sample sizes than single site studies. These benefits of RRRs 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).

Although no journal-based RRRs exist within the field of L2 research, 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 are specific to L2 research, a field in which a range of context-specific variables are 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 the current multi-site replication effort by emulating some key aspects of RRRs, specifically pre-registering 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 when Processing L2 Input for Meaning**

As noted above and by Marsden et al. (this issue), 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 multi-site replications*,* all of these characteristics are probably necessary given the investment of time and resources needed for multi-site replications, which may consequently be warranted only for the most central or pressing issues. However, where resources are more readily available, multi-site 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 including research about attention and awareness (e.g., Cintrón-Valentín & 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 when 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 line of research was chosen as the medium for examining the feasibility of RRRs in L2 because it fully warrants multi-site replication. First, it is an influential line of work with the seminal study, VanPatten (1990), being one of the ten most highly cited articles published by the journal *Studies in Second Language Acquisition*, with 1009 citations according to Google Scholar as of the time of writing, including the citations listed in the first sentence of this paragraph. 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; VanPatten, 2015) and the pedagogical approach of processing instruction (VanPatten, 2005; VanPatten, 2004b). Third, researchers are interested in replicating studies in this line of research as there have been multiple partial and conceptual replications, which is remarkable given that the replication rate for the field of L2 research has been so low (Marsden et al., this issue).1 Note, however, that the findings from the replications have not been direct or fully consistent. 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 attending to form when processing L2 input for meaning affects comprehension seemed to merit the efforts of a multi-site replication, and we selected Leow et al. (2008) as the study to be replicated given its claimed methodological strengths. In the remainder of this section we provide (a) a review of the theoretical motivation, design and results of the VanPatten (1990) study, (b) an overview of its direct and partial replications, (c) a full review of the Leow et al. (2008) conceptual replication of VanPatten (1990) and its theoretical implications, and (d) analysis of the ongoing limitations and open issues in this line of research that motivate our replication endeavors in substantive terms.

The initial study in this line of research, 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., first-semester, fourth-semester, third-year university level) 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. Each of the experimental groups was 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 everything they recalled from the passage in English, which was 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, as 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, i.e., the language or the mode. Wong (2001) reproduced VanPatten’s results in the aural mode with English as an L2, using a translated passage and the English forms *inflation* and *the.* Wong also conducted the study in the written mode, where attention arguably was not constrained to the same degree as in the aural mode, and did not reproduce VanPatten’s findings. In contrast to Wong and consistent with VanPatten, Greenslade et al. (1999) did find differences between conditions in L2 Spanish in the written mode. Taken together with VanPatten (1990), these replication studies provided some evidence to support the generalizability of the initial findings, at least for the aural mode, that attending to morphosyntactic form interfered with comprehension. As such, the results have been used as part of the underpinning of VanPatten’s Input Processing model (VanPatten & Cadierno, 1993a; VanPatten & Cadierno, 1993b; VanPatten, 1996) 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 differential physical salience of the linguistic forms within each study, 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., second-semester L2 Spanish learners read a 358-word Spanish passage about the Aztecs that provided an even 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: Each of the experimental groups was 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 as 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 assessment afterwards, but, departing from the earlier studies, the comprehension test consisted of 10 four-option multiple choice questions. Contrary to the findings of VanPatten (1990) and Greenslade et al. (1999)’s written modality studies, and Wong’s (2001) aural modality study, Leow et al. did not find any differences in comprehension between any groups, which was consistent with the findings from Wong’s written modality. 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 confirmed that thinking aloud in this paradigm did not lead to reactivity effects, i.e., 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 mode.3

Although this line of partial and conceptual replications stands out among research 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 as three of four studies in the written mode did not find a statistically significant effect of attention to grammatical form on comprehension (no statistically significant effects found in: Leow et al., 2008; Morgan-Short et al., 2012; Wong, 2001; statistically significant effects found in: Greenslade et al., 1999).4 The two studies conducted in the aural mode found that attention to grammatical form *did* affect comprehension (statistically significant effects found in: VanPatten, 1990; Wong, 2001). Thus, the results across the written and aural mode are mixed. Note that only one study to date has used the same materials and same sample of participants across both modalities (Wong, 2001), thus allowing a robust comparison. However, Wong (2001), along with VanPatten (1990) was characterized by the methodological limitations pointed out by Leow et al. (2008). 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 are directed to read the Spanish passage for comprehension and are told that, after they read the passage, they will 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; Rayner, 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, as would be 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, which Leow et al. coined as ‘backtracking,’ when answering comprehension questions. 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, e.g., paragraph, sentence, phrase levels; they only reveal its presence, 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 experiments other than a basic view of modality itself – i.e., of being in either the oral or written medium. However, different modalities involve different speeds of presentation and amounts of exposure (untimed for the written mode and timed for the aural mode) and different opportunities for backtracking (available for the written mode but not available for the aural mode), 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 ascertaining the ‘true effect’ of the extent to which learners can comprehend L2 input while also attending to lexical items or morphosyntactic forms. A better understanding of the empirical underpinnings of the Primacy of Meaning Principle is necessary, as results from previous studies are mixed, in part because methodological issues limit the strength and generalizability of the conclusions that can be drawn from those results. Thus, the current, multi-site replication study aims to revisit the question of whether L2 learners are able to attend to both L2 form and meaning while processing L2 input for comprehension, the issue common to all studies in this line of research. We adopted the experimental 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 replication changed the mode from written to aural, and the other changed the presentation of the written mode 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 mode, in order to eliminate backtracking and to increase the comparability between the written and aural modalities. Through multi-site endeavors, we then conducted direct replications of both paradigms, without changes to any key variable, in order to establish the replicability and generalizability of the results and to gain further insight into the true magnitude of any effects.

**Methods**

The current study replicated Leow et al. (2008) by adapting its materials (available at 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 (US) and at universities in the United Kingdom (UK). Subsequently, a call for participation in the multi-site replication effort went out via personal contacts, professional listservs and websites. Included in this call was a link to a registered, public webpage on the Open Science Foundation (OSF) that included a description of the multi-site replication project, the requirements for participation in the study, and the protocol and materials needed to execute the study (see [osf.io/uybak](https://osf.io/uybak)). 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 OSF page. The site researchers contacted the lead authors if they had questions about the protocol. The lead authors also developed a template for data entry and registered it on the OSF (see [osf.io/jbpag](https://osf.io/jbpag)), and each site deposited their data using the template (see [osf.io/vwytd)](https://osf.io/vwytd/). The lead authors then developed and posted highly detailed analysis protocols (see [osf.io/nz3su](https://osf.io/nz3su/)). Following these closely, each site conducted and deposited their own analyses if possible (5 sites). For two of the sites, the lead authors carried out and deposited the analyses (for the US2-W and UK2-A sites). The lead researchers then conducted analyses across sites. Below is a description of the participants, materials, procedures and analyses, although readers can consult the OSF pages for the exact protocol and materials, as well as iris-database.org.

**Participants**

Participants were recruited for the aural paradigm from four sites, including from the lead aural site in the US (US1) and from replicating aural sites in the US (US3-A), the United Kingdom (UK2-A), and Poland (POL-A). For the timed-written paradigm, participants were recruited from three sites, including from the lead timed-written site in the UK (UK1-W) and from two replicating timed-written sites in the US (US2-W and US4-W). 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 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. (2008), who were enrolled in an introductory level, second-semester, university Spanish language course and who had been exposed to preterite and imperfect inflectional morphology, as 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. See Supplementary Information for details for each site.

At all sites, participants were randomly assigned to either the control condition or one of the three experimental conditions (see below). A total of 704Spanish (females = 433; *M* age = 20.48, *SD* age = 6.08) participants across the seven sites took part 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 6 check marks or mouse clicks while reading or listening to the passages, which was the exclusion criterion utilized by Leow et al. (2008), 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). See Supporting Information, Table SI-1, for detailed, condition-specific participant numbers and Supporting Information, Site Information, for site-specific information regarding recruitment and course level information.

INSERT TABLE 1 ABOUT HERE

**Materials and Procedures**

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 (Psychology Software Tools, Pittsburgh, PA) or Superlab 5.0 (for the timed-written paradigm; Cedrus Corporation, San Pedro, CA) 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 comprised of 23 sentences with 10 instances of each target form (*sol*, *la*, *-n*) distributed evenly among four paragraphs with no more than one target from 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 and 43 seconds long. For the timed-written paradigm, the exact same passage was used but was presented via computer with rapid serial visual presentation (RSVP, 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 RSVP 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 RSVP allows us to control ‘backtracking,’ whether it be for the purpose of spotting target forms or for comprehension. A significant amount of backtracking had been revealed through think alouds in a pilot (*N* = 21) of a paper-and-pencil version of the written paradigm. For the current study, the RSVP rate of presentation 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 ms 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 RSVP presentation of written stimuli is its ecological validity in regard to reading comprehension processes. Although some processes involved in reading are different under RSVP (Öquist & Goldstein, 2003), Juola et al. (1982) found that RSVP does not necessarily disrupt normal reading comprehension processes. More recent evidence (Ricciardi & Di Nocera, 2017) suggests that RSVP may affect reading comprehension when the rate of presentation is faster than the normal reading rate (~ 250 words per minute, wpm), but not when the rate of presentation is similar to or slower than the normal rate. The presentation rate used in the current study (97.56 wpm) was not faster than normal, and a second pilot study (*N* = 17) with the same RSVP paradigm as used in the current study yielded similar levels of comprehension as that evidenced in previous studies in this line of research (further corroborated by our main data). Thus, although the written, controlled RSVP 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 they would be given a comprehension test after hearing (or reading) the passage. 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 3rd 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, as 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. (2008) reported a reliability coefficient (Cronbach’s alpha) of 0.915 for this test (p. 681).7 Reliability was calculated again for the purposes of the current study, as it was the key dependent measure for both modalities. Given the binary nature of participants’ responses, i.e., correct or incorrect, alpha was calculated based on the KD-20 test for all participants who were included in the analyses. The alpha level for all participants across all sites was 0.197, with varying levels per site: US1-A = -0.119, US3-A = 0.153, UK2-A = 0.298, POL-A = 0.596, UK1-W = .007, US2-W = .068, US4-W = 0.008.8 Overall, these results suggest that for the current study, the set of comprehension items are not consistent with each other and thus were not measuring comprehension as a unidimensional construct. The low reliability may partially be an artifact of general low performance on the test (see Results section), as 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 participants 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 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 Spanish DELE used by Seibert Hanson and Carlson (2014, available at iris-database.org).

**Coding**

A template Excel 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, and 0 for incorrect) and each proficiency item (1 for correct and 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 and reaction time of mouse clicks to the target forms 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 the 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 immediately preceding target form (*la*, *sol*, -*n*) to the click.

**Analysis**

 The same analysis protocol was followed to conduct descriptive and (non)parametric analyses, as required by their site-specific data. All analyses reported in the results aimed to reveal whether no differences would be found between experimental and control conditions, i.e., the null hypothesis (H0), or whether differences would be found between conditions, i.e., the alternative hypothesis (H1). The first set of reported analyses examined whether the findings of Leow et al. (2008) would be reproduced at each site, as revealed though ANOVA analyses that followed Leow et al. (2008). Second, in order to gain insight into the true size of the effect of the experimental conditions compared to control, we conducted a random effects meta-analysis on effect size data across sites. The results from the ANOVA and the random effects meta-analysis analyses 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 model analysis was conducted. Finally, because conclusions from the analyses above are largely based on null hypothesis testing, which can provide evidence for the H1, i.e., that there are differences between conditions, but not for H0, i.e., that there are not differences between conditions, we report Bayes factors for differences between the experimental conditions and control, which allow us to make inferences about both H1 and H0 (Dienes, 2014). Details of each analysis are presented below in the relevant result sections.

**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. (2008), we submitted comprehension scores at each location to a one-way ANOVA with one between-subject factor (condition), including all participants in the control condition and participants in the experimental conditions who had made at least 6 checks marks to indicate hearing a target form. A general approximation of the size of the effect of condition, as measured by *eta2*, was based on the recommended interpretation of *R*2 values from Plonsky and Ghandar (submitted), i.e., small < .20 ≤ medium < .50 ≤ large. Given that both *eta*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 *eta*2 values, given the lack of field specific recommendations for *eta*2. If the ANOVA yielded a main effect of condition, post hoc analyses were conducted and consisted either of 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 the lead aural site (US1-A), comprehension scores for all the attentional conditions were around 30% mean accuracy (see Figure 1 and Supporting Information, Table SI-2). Analyses did not reveal a statistical effect of condition (*F*(3,122) = 1.052, *p* = 0.372, *η2* = 0.025). Thus, no evidence was found in support of differences in comprehension resulting from paying attention to a lexical or a grammatical form (H1).

This general pattern of results seemed to hold across the replicating aural sites (see Figure 1 and Supporting Information, Table SI-2), although mean comprehension scores were, in all conditions, descriptively higher for the UK2-A and POL-A sites. The null statistical finding from the lead site was reproduced in two of the three replicating sites: US3-A, *F*(3,208) = 0.988, *p* = 0.399, *η2* = 0.014; UK2-A, *F*(3,37) = 1.650, *p* = 0.195, *η2* = 0.118. 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 *eta*2 value: POL-A, *F*(3,51) = 3.261, *p* = 0.029, *η2* = 0.161. 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* = 0.003).

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, 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 H1 was not supported. Note, however, that one of the sites did report an effect on comprehension in relation to attending to a morphological versus a lexical form.

INSERT FIGURE 1 HERE

**Effect sizes across aural sites*.*** In order to gauge the general effect of each experimental condition compared to control across sites, we calculated Cohen’s *d* and its 95% confidence intervals (CIs) for each site and then performed a random-effects meta-analysis across sites using the metafor package in R in order to obtain a meta-analytic effect size (with the weighted mean based on variance; see Figure 2 and Table SI-3). As points of comparison, we also calculated Cohen’s *d* and its 95% CIs 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 averages for between-group comparisons, with ≥0.40<0.70 suggesting a small effect, ≥0.70<1.00 a medium effect, and ≥1.00 a large effect. Effect sizes with 95% CIs that did not cross zero were interpreted as reliable effects, providing evidence for H1 (Cumming & Finch, 2005), but the reverse was not taken to be true: 95% CIs that include zero cannot be interpreted as no effect (H0) as 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 control, 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 zero. Thus, this meta-analytic effect across the four aural sites does not provide evidence for differences in comprehension (H1) when attending to *sol* or *la* as compared to control. Note that the meta-analytic effect was consistent with the effects for each site, as the site-specific effects for these two conditions also had 95% CIs that cross zero. The meta-analytic effect size point-estimate for attending to the grammatical form –*n* compared to control, -0.38, (a) also fell within the 95% CI range from Leow et al. (2008), but was outside of the range of Morgan-Short et al. (2012), (b) approached the 0.4 value that would indicate interpretation as a small effect, but (c) had CIs that overlapped with zero. Thus, across all four sites, no evidence was provided for an effect (H1) of attending to –*n* as compared to control. This effect, however, was not consistent in each individual site. In two of the four sites (UK2-A and POL-A), 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 suggest that the results from the aural paradigm are largely consistent with Leow et al. (2008) and Morgan-Short et al. (2008) in that they do not provide evidence for H1, i.e., differences in comprehension when attending to lexical or grammatical form, which is also consistent with the overall findings from the current ANOVA analyses.

INSERT FIGURE 2 ABOUT HERE

**Timed-Written Paradigm**

Next, we examined whether the findings from Leow et al. (2008) and Morgan-Short et al. (2012) would be reproduced in a timed-written paradigm at each site. As for the aural paradigm, we submitted comprehension scores at each location to a one-way ANOVA with one between-subject factor (condition), including all participants in the control condition and participants in the experimental conditions who had made at least 6 mouse clicks to indicate seeing a target form. We interpreted *eta*2 and conducted post hoc tests following the same parameters as for the aural paradigm (see above).

For the lead timed-written site (UK1-W), comprehension scores for all the conditions were around 45% mean accuracy (see Figure 1 and Supporting Information, Table SI-2). The ANOVA for this site indicated no statistical effect of condition (*F*(3,56) = 1.243, *p* = 0.303, *η2* = 0.062). Thus, no evidence was found in support of differences in comprehension among conditions resulting from paying attention to a lexical or grammatical form (H1).

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 lead timed-written site and closer to the scores from the three US sites that administered the aural paradigm (see Figure 1 and Supporting Information, Table SI-2). The null statistical finding from the lead timed-written site was reproduced in the US4-W site (*F*(3,86) = 1.311, *p* = 0.276, *η2* = 0.044), but not in the US2-W site where an effect of condition was evidenced (*F*(3,43) = 3.480, *p* = 0.024, *η2* = 0.195), 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* = 0.016).

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 when attending to form and meaning compared to attending to meaning alone. As such, H1 is largely not supported. However, in one timed-written site (US2-W) a negative effect of attending to the lexical form *sol* was found compared to control.

**Effect sizes across timed-written sites*.*** In order to gauge the general effect of each experimental condition compared to control across the timed-written sites, we calculated Cohen’s *d* and its 95% confidence intervals (CIs) 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 SI-3).

From Figure 3, we see that the meta-analytic effect size point-estimates for attending to any form, i.e., *sol*, *la*, or –*n,* compared to control, -0.34, -0.28, -0.13 respectively, (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 small effects, and (c) had CIs that overlapped zero. Thus, these meta-analytic effects across the three timed-written sites did not lend support to H1, i.e., that attending to either a lexical or grammatical form affects comprehension compared to control. This finding was consistent for the effects for each condition for each site, with the exception of site US2-W where attending to *sol* negatively affected comprehension. Overall though, the random-effects meta-analysis did not detect evidence for effects on comprehension when attending to lexical or grammatical form (H1) in the timed-written paradigm, which is generally consistent findings from Leow et al. (2008), Morgan-Short et al. (2008), and with the ANOVA analyses reported above for this paradigm.

INSERT FIGURE 3 ABOUT HERE

**Multi-site Mixed Models 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 in R (Bates, Maechler, Bolker & Walker, 2015). The model included condition as the primary fixed effect of interest, but also included mode as a fixed effect and the interaction between condition and mode, as previous research has found different results for different modes (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 mode, random by-item intercepts and slopes for condition and mode, and random by-site intercepts nested in mode. Further specification of the random effects structure led to a failure to converge. The full model is specified below:

*Model Accuracy <-glmer(Score ~ Condition\*Mode + StdProf + (1|Subject:(Condition:Site:Mode)) + (1+Condition+Mode|Item) + (1|Site:(Mode)), data = Repldata, family = binomial(link = 'logit'), control = glmerControl(optimizer = "bobyqa"))*

An ANOVA (Type III; conducted with the ‘car’ package in R) run on the model returned significant effects of proficiency (*χ2*(1) = 61.382, *p* < 0.001) and mode (*χ2*(1) = 4.390, *p* = 0.036) as well as a marginal effect of condition (*χ2*(3) = 7.165, *p* = 0.067) that was qualified by a marginal interaction of condition by mode (*χ2*(3) = 7.417, *p* = 0.059). 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 in R) on the significant effect of mode revealed that the log odds of responding correctly to comprehension items increased by 0.36 for participants in the written mode compared to those in the aural mode (written estimated *M* = 0.34, *SE* = 0.06, 95% *CI*s = 0.26, 0.49; aural estimated *M* = 0.31, *SE* = 0.05, 95% *CI*s = 0.23, 0.41; estimate = 0.36, *SE* = 0.17, *z* = 2.095, *p* = 0.036). Follow-up Tukey tests on the relevant comparisons for the marginal condition by mode interaction did not reveal any statistically significant differences between control and experimental conditions in either mode (*p*s ≥ .296).10 These results are thus largely consistent with the pattern of findings from the ANOVA and random-effects meta-analysis in that they do not provide evidence for H1, i.e., differences in comprehension between experimental conditions and control, regardless of modality.

**Bayes Factors**

 As previously pointed out, conclusions from analyses reported above can only provide evidence for the alternative hypothesis (H1) but not for the null hypothesis (H0). Indeed, overall, evidence for H1 was not detected in the ANOVA, the meta-analytic effect size, or the mixed models results. However, whether the results support H0, i.e., no difference in comprehension between experimental and control conditions, is 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 H1 or H0, and thus can constitute evidence for H0. More specifically, a *B* value greater than 3 provides evidence for H1, whereas a *B* value less than 0.33 provides evidence for H0. Thus, in order to determine whether there was evidence for H0, i.e., that comprehension did not differ across conditions, we calculated *B* based on the mean difference between each experimental condition compared to control for the full set of data. Our calculation was based on a theory of H1 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, i.e., attention to form is predicted to negatively affect comprehension. The *SD* value, 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. Dienes (personal communication, January 30, 2018) recommends that when a maximum effect is known, then the *SD* 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 zero), the *SD* would be half of that value, i.e., 2.325. Using half of the maximum mean score from Leow et al.’s (2008) control group as the *SD* is 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 Thus, we calculated the following *B*s (with Dienes’ Bayes calculator, http://www.lifesci.sussex.ac.uk/home/Zoltan\_Dienes/inference/Bayes.htm): for *sol* compared to control: *B*H(0,2.325) = 0.05; for *la* compared to control: *B*H(0,2.325) = 0.08; for *-n* compared to control: *B*H(0,2.325) = 0.40. These results provide evidence for H0 for the *sol* and *la* conditions as compared to control, i.e., that they did not differ in regard to comprehension. For the *–n* condition compared to control, there is not sufficient evidence to make a strong conclusion as the value 0.40 falls in between 0.33 and 3, which would indicate that the result does not favor either H1 or H0. However, it is closer to 0.33, which would indicate that the data favors H0. Indeed, if we calculate *B* based on the average of the two grammatical conditions (*la* and *–n*), our result is *B*H(0,2.325) = 0.02, which is evidence for H0, suggesting that comprehension did not differ when attending to the grammatical forms as compared to control.

**Discussion**

The primary aim of the current study was to ascertain the feasibility and usefulness of incorporating a multi-site registered replication approach in the field of L2 research as a mechanism to improve the validity the field’s findings through transparent research practices and replication. In order to explore such possibilities, we conducted a multi-site replication effort that emulated some aspects of the RRR approach used in the field of psychology, e.g., pre-registration 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 modes 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 were 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. (2008) would be reproduced at each site, what the meta-analytic effect of each experimental condition was for each mode across sites, and what the overall effect of condition was across all sites regardless of mode. Below we provide a summary of the results from the multi-site replication endeavor along with the implications of the results and future directions for research. We then turn to discussion regarding the primary aim of the overall project, i.e., to assess the feasibility and usefulness of multi-site replication approaches in L2 research.

**On Attending to Form while Processing for Meaning**

Regarding whether attending to L2 form interferes with processing a text for comprehension, our results revealed the following: (a) First, the ANOVA analyses that followed Leow et al. (2008) did not evidence a statistically significant effect for comprehension for the experimental groups as compared to control, except in the US2-W site where attending to the lexical form *sol* reduced comprehension. (b) Second, the random-effects meta-analysis revealed that the meta-analytic effects for each of the experimental conditions compared to control 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. (c) Third, the mixed model also did not reveal an effect of condition on comprehension over all the sites. (d) Lastly, Bayes factors provided evidence that comprehension in two of the experimental conditions, that is, *sol* and *la*, did not differ from control. For the *–n* condition, there was not sufficient evidence to conclude that performance was similar to control, although the Bayes factor for this comparison (0.40) was closer to indicating that comprehension in the *–n* condition was similar to (≤ .33) rather than different from (≥ 3.0) comprehension in the control condition. Overall, for the population represented by the participant samples included in this multi-site replication, the results from the different analyses largely converge in that evidence is 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. (2008) and Morgan-Short et al. (2012) where attention to form was also not found to affect L2 comprehension. The results additionally extend the findings from these previous studies to an aural paradigm as well as to a timed-written paradigm where the possibility of backtracking was not possible.

It is interesting to note that our conclusions about the results may 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 the US2 site, we might have concluded that attending to a lexical form seemed to have interfered with comprehension. Or, if only the effect size analysis was run for the UK2 and POL site, we might have concluded that attending to inflectional verb morphology had affected comprehension. Fortunately, because of the multi-site endeavor, site-specific results (with relatively small sample sizes) are not overly interpreted as the true effect.

The multi-site approach also provided us with unique insight into the generalizability of the finding. We see that 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. Importantly, we are also able to see possible limits on the generalizability of the findings as 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 POL site (Figure 2), one may conclude that the findings may not generalize to learners with a first language other than English, to learners who have more experience learning second languages (as most had already learned English), to learners who are not reading comprehension questions in their first language, or to learners whose Spanish language program . Note, however, that a reliable effect between the –*n* and control conditions was also evidenced at UK2, which, along with the POL 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, when attending to non-salient morphological forms, such as clitics or forms that are non-syllabic 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 multi-site study, can provide us with clues as to the factors 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 Leow et al. (2008) and Morgan-Short et al. (2012), from which the materials were adapted. The results are also consistent with the findings from the written mode in Wong (2001), who used different materials. However, the results are not consistent with the results from other studies that had used a different set of materials: VanPatten (1990) (aural mode), Greenslade (1999) (written mode), or the aural mode in Wong (2001). These studies used materials with shorter L2 passages (275 versus our 358 words) and a different method of assessing comprehension (free recall of idea units versus our multiple-choice). We also cannot rule out the possibility that the learners 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, learners’ scores on the multiple-choice test might have partially reflected random guessing, although almost all conditions at all sites performed at above-chance level (see Figure 1). In VanPatten’s (1990) study, however, very little of the recall is likely to have been based on random guessing. Thus, even though the range of percent accuracy for the current study (23% - 52%) is generally higher than the percentage of all ideas units recalled by participants in VanPatten (1990) (17% - 36%), the recall used in the VanPatten (1990) study may have better captured learners’ abilities to comprehend the passage, in part because the score is less likely to reflect guessing.12

Indeed, it seems that the comprehension test administered in this multi-site 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., with low consistency among items) or to learners 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 factors 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 RSVP presentation used for the timed-written paradigm is not ecologically valid, although, as noted above, research has shown that RSVP 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 mode. 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 (US3, UK1, and US4), participants in the –*n* condition made statistically significant 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 Supplementary Information, Table SI-1). Thus, while comprehension for these conditions was not lower than that of 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 seems to be higher rates of exclusion from the dataset for the –*n* condition (largely based on not meeting the requirement of having made six or more checks or clicks) at two sites (US 1 and US3, Supplementary Information, Table SI-1). These participants were not included in the analysis because they did not complete the attention to form task appropriately, so they do not impact the results themselves, but they still hint at a tension between attending to form and meaning. Indeed, they could be interpreted as lending broad support to the notion expressed in 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.

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, i.e., whether learners can attend to L2 forms when processing input for meaning without 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 model analyses) showed evidence of negative effects for attending to *sol* or *la*,although there was a negative, site-specific effect for *sol* at the US2 site. The meta-analytic (weighted) mean effects for these conditions were positive (0.10 and 0.19, respectively), but were not reliably above chance. Additionally, the Bayes factors for these conditions compared to control 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 learners, at least not when they were asked to indicate that they attended 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 (-0.38), and there were negative effects on comprehension in two sites (POL and UK2 as evidenced in the effect size analysis). There were also indications that participants were less successful at attending to form in the –*n* compared to the other experimental conditions (as evidenced by statistically significant lower 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 on 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 words, syllabic, and are relatively invariant, –*n* is a non-stressed, non-syllabic element at the end of a verb that can co-occur with morphemes that vary in form and meaning (e.g., -*ía****n***,–*aro****n***, –*aba****n***, –*a****n***, –*e****n*** that indicate Spanish, tense, aspect and mood).

Overall, our study provides evidence that L2 learners’ comprehension is not affected by attending to particular unbound forms, i.e., *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 attending to a bound form, i.e., -*n*, does not affect comprehension. Indeed, site-specific results both for comprehension and for how well learners were able to attend to the form suggest that a general conclusion that learners can attend to both form and meaning when focused on comprehension is not warranted, as such a conclusion may not apply uniformly to all forms and contexts. Future research that is based on current iterations of Input Processing theory (VanPatten, 2015) and other perspectives related to L2 processing (e.g., N. C. Ellis et al., 2014; N. C. Ellis & Wulff, 2015; Leow, 2015) may want to explore the boundaries of when learners can attend to form and meaning when 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 with the comprehension test. We should also note a potential lack of ecological validity in using an activity that rendered such low levels of comprehension. However, of interest are findings from Marsden and colleagues, where word and sentence-level tasks that are more ecologically valid to L2 pedagogy are processed in such a way that learners (a) successfully attend to the form of an article and the meaning of a sentence (Kasprowicz & Marsden, 2017) and (b) show the ability to learn the meaning of a word even when their attention was oriented to the meaning of a form (Marsden et al., 2013, experiment 3). Thus there seems 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 (VanPatten, 2005; VanPatten, 2004b) where explicit information about an 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 1000 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 when 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. First, regarding the dependent variable—comprehension, 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. Otherwise the validity of the test as a test of comprehension will always be challenged, and researchers will not be able to make the claim that comprehension has or has not been affected. Second, as mentioned in Leow et al. (2008), researchers need to establish that learners are engaged in both processing for form and for meaning. In the current study, we were able to control backtracking with a time-written condition where words appeared on the screen, so we can claim that learners did not 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. Learners 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, perhaps a baseline could be taken to establish the average number of regressions that each learner makes when reading, and then regressions that fall out of that normal when 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 when 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 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. Third, 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, e.g., one that is not linguistic, in order to demonstrate that the task itself is cognitively demanding and affects participants’ performance on a primary task under dual task conditions. Finally, researchers may want to better control the differences between target forms, i.e., salience, length, syllables, etc. further. For example, in order to test whether the boundedness of a morpheme makes it more difficult 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 when processing L2 input for comprehension.

**On Multi-site Registered Replications**

 The principle objective of the current study was to ascertain the feasibility and usefulness of incorporating a multi-site registered replication report approach (Simons et al., 2014; Simons & Holcombe, 2014) in the field of L2 research. Many aspects of this RRR mechanism were adopted by the current study. First, seven different sites collected data independently for the study. Second, some aspects of the study were publically registered prior to data collection and analysis. Although the materials and protocols for the partial replications (US1-A and UK1-W) were not registered, they were for the direct replications. Also, a template was registered for data entry at each site, and the detailed analysis plan was publically 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 creates full transparency of the research findings and helped to discourage *p*-hacking.

Note that the current study differed from the full registered report aspect of RRRs in that they 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 work flow. Thus, given (a) that registration should not require additional work or resources and (b) that infrastructure exists to support it, e.g., IRIS 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 benefitted 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 US1-A and UK1-W sites. These adjustments may 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, as a registered report article type is now offered through *Language Learning* (Marsden et al., this issue). We recommend that Registered Reports be adopted more widely as one (of many) mechanisms to promote replication and robust research practices (see Marsden et al. this issue).

 In regard to the multi-site aspect of the study, the clear benefits of increasing the external validity of a finding and providing insight into the true size of an effect, 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, e.g., to pay research assistants for the purpose of coordination or even to help 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 Report (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 multi-site 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 et al., this issue), and the negative effects of this could be even more harmful in a high profile, multi-site 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, multi-site 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, e.g., L1, L2 experience, educational context, socio-linguistic, -economic, -cultural and -affective factors. These L2 specific challenges are in addition to the more ‘normal’ challenges of multi-site work, such as individual differences among participants’ age, cognitive abilities, or educational attainment, and also on top of practical issues, such as the availability of software, hardware and incentivization to engage participants. However, multi-site approaches also provide an opportunity to measure variables that otherwise might not be able to be explored.

In sum, although multi-site endeavors should be pursued, both for replication and initial research (for an example of an initial multi-site 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 multi-site research for some studies. Researchers interested in initiating such initiatives may choose to take advantage of resources such as the ‘Study Swap’ on the OSF site, where researchers look for and offer themselves as multi-site collaborators, and the ‘Call for Replication collaborators’ on the IRIS site.

**Additional Reflections**

We would finally like to provide a more reflective, introspective discussion about issues in conducting this multi-site replication study given the findings of the synthesis of self-labelled replication in the L2 field (Marsden et al., this issue). As stated previously, the first two authors of this article began the project with the intention of examining the feasibility of incorporating multi-site registered replication reports into our field. With this goal in mind, we (that is, Morgan-Short and Marsden, throughout this ‘Additional Reflections’ section) searched for a research question, paradigm, and materials that were suitable for our purpose. In addition to the motivations for choosing the paradigm that we reported above, we were also concerned with the availability of materials. We had both conducted previous studies with the current paradigm and thus could access 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 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 or our research also made us vulnerable to the very concern raised by Marsden et al. (this issue) about author-overlap between initial and replication studies, i.e., that we were making ourselves susceptible to the risk of researcher bias, which could engender questionable research practices. We took measures to counter these, such as pre-registration of the protocol and analysis, by employing a multi-site approach, and by 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 datasets, in order 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 Spanish DELE (Seibert Hanson & Carlson, 2014) in large part because it was available on IRIS and freely available. 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 was comprised of 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 cross-modal 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 only able to do so in three sites. Our reflection here relates to the factors that influence our fields’ decisions about what to replicate. Resource requirements are definitely one consideration, even though Marsden et al. (this issue) 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 initial results of our results at a conference, we were challenged about whether we should expect others to join us in order to replicate null results. Further illustrating this concern, Marsden et al. (this issue) found that replications of studies with null findings were extremly 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 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 multi-site 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 (IPA) before data is collected, researchers who receive IPA 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 a researcher does not receive IPA, 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 infrastructure (such as IRIS, the OSF, Registered Reports) can inform decision-making and facilitate a more collaborative ethic in the field.

**Conclusions**

 We conducted a multi-site replication study with aspects of pre-registration in order to explore the feasibility and usefulness of a multi-site 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 an L2 passage would interfere with comprehension of that passage, results from the current study indicated that (a) an effect of attention to form on comprehension was not detected in by-site ANOVA analyses that followed previous research, except for one 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, the comprehension for two of the experimental conditions was similar to that of control. Thus, overall, for the population represented by the participant samples included in this multi-site 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 difficulty attending to the bound, morphosyntactic form as evidenced by effect size analyses and also by effects related to how well learners 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 when processing input for meaning makes it difficult for learners to comprehend.

In regard to the feasibility and usefulness of incorporating a multi-site registered replication report approach (Simons et al., 2014; Simons & Holcombe, 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 as compared to our previous research endeavors. The Registered Reports 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 work flow 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. In regard to the multi-site 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 for the field of L2 research than similar endeavors in social and cognitive psychology. Overall though, both mechanisms should be adopted by our field to some degree in order to increase the robustness and generalizability of findings in our field.

**Notes**

1 Note, however, that only one replication of VanPatten (1990) self-labeled as a replication. See Marsden et al. (submitted) 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 Note that Leow et al. (2008) and Morgan-Short et al. (2008) additionally addressed issues of depth of processing based on the 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 as the current study did not administer think alouds and thus cannot address questions of depth of processing.

4 Note that 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 they could not conclude that there were no differences, as 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 replications studies preceding Leow et al.’s (2008) 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 this calculation on the data and found an alpha level of 0.153 across all participants and 0.213 for participants in the non-think aloud group, which consisted of participants who were not asked to think aloud during the task as in the current study.

8 Note that, descriptively, the reliability coefficients tend to be higher for sites that evidenced higher levels of proficiency, and were also generally higher for the aural mode than for the written mode.

9 Parallel analyses were conducted for the reaction time data that was output by the software programs, i.e., E-Prime and SuperLab, in the aural mode. However, we do not report or discuss these analyses in the article as there are several constraints on a valid interpretation of that 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 are not matched. Whereas sol and la are words, syllabic, and are relatively invariant, –n is a non-stressed, non-syllabic 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 that indicate Spanish, tense, aspect and mood). Also, there may be variability within the -n condition as it was found on the end of verbs of different lengths, whereas for sol and la the length was always shorter, constant, and pre-defined. Because the study was not designed as a reaction time study, these factors 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, as there is not any reaction time data for the condition. A second potential issue is that, although the timing specifics of the software programs used to present the passage is not expected to differ meaningfully, the timing specifics of different hardware configurations may have impacted the reaction time output (Stahl, 2006). Future multi-site 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 OSF analysis page (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 is not a contrast of theoretical interest.

11 The percentage reduction of comprehension for the grammatical conditions that differed from control 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) For Greenslade et al. (1999), comprehension reduced by 43% for the la condition and by 39% for the –n condition. (c) For Wong (2001), comprehension reduced by 77 % for the the aural condition. Thus the average reduction in comprehension is 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 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*(3), 255-278. 10.1016/j.jml.2012.11.001

Bohannon, J. (2014). Replication effort provokes praise—and ‘bullying’ charges. *Science, 344*(6186), 788-789. 10.1126/science.344.6186.788

Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education, 40*(2), 218-231. 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. 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*(2), 170.

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

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

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*(03), 547-579.

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). NY: Routledge.

Ellis, R. (2016). Focus on form: A critical review. *Language Teaching Research, 20*(3), 405-428. 10.1177/1362168816628627

Greenslade, T., Bouden, L., & Sanz, C. (1999). Attending to form and content in processing L2 reading texts. *Spanish Applied Linguistics, 3*(1), 65-90.

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*(2), 208-227. 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*(2), 228.

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,* , amw051. 10.1093/applin/amw051

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217. 10.1207/s15327957pspr0203\_4

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š, Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology, 45*(3), 142-152. 10.1027/1864-9335/a000178

Leow, R. P. (2001). Attention, awareness, and foreign language behavior. *Language Learning, 51*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). NY: Routledge.

Leow, R. P., Hsieh, H., & Moreno, N. (2008). Attention to form and meaning revisited. *Language Learning, 58*(3), 665-695. 10.1111/j.1467-9922.2008.00453.x

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*(12), 1827-1832. 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. *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1-21) Taylor and Francis.10.4324/9780203489666

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (this issue). Self-labelled replication in second language research: Narrative and systematic reviews, and recommendations for the field.

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*(4), 619-654. 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*(04), 659-685. 10.1017/S027226311200037X

Norris, J. M., Ross, S. J., & Schoonen, R. (2015). Improving second language quantitative research. *Language Learning, 65*(S1), 1-8. 10.1111/lang.12110

Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. (2017). The preregistration revolution. Retrieved from 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*(6), 615-631. 10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 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*(4), 539-558. 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*(1), 408-411. 10.3758/s13428-015-0577-0

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

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

Polio, C., & Gass, S. M. (1997). Replication and reporting. *Studies in Second Language Acquisition, 19*(04), 499-508.

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

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422. 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*(8), 1457-1506. 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. 10.1016/j.chb.2016.12.047

Robinson, P. (1995). Attention, memory, and the “noticing” hypothesis. *Language Learning, 45*(2), 283-331.

Seibert Hanson, A., & Carlson, M. (2014). The roles of first language and proficiency in L2 processing of spanish clitics: Global effects. *Language Learning, 64*(2), 310-342. 10.1111/lang.12050

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science, 22*(11), 1359-1366. 10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76-80. 10.1177/1745691613514755

Simons, D. J., & Holcombe, A. O. (2014). Registered replication reports: A new article type at perspectives on psychological science. . *Observer, 27*(3)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*(5), 552-555. 10.1177/1745691614543974

Stahl, C. (2006). Software for generating psychological experiments.. *Experimental Psychology, 53*(3), 218-232. 10.1027/1618-3169.53.3.218

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*(4), 1713-1717. 10.3758/s13428-015-0677-x

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., 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*(2), 506-527.

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: Lawrence Erlbaum Associates.

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

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

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. *Modern Language Journal, 77*(1)10.2307/329557

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

VanPatten, B. (1990). Attending to form and content in the input. *Studies in Second Language Acquisition, 12*(3), 287-301.

Wong, W. (2001). Modality and attention to meaning and form in the input. *Studies in Second Language Acquisition, 23*(3), 345-368.

|  |
| --- |
| Table 1*Participant Information by Modality and Site* |
| Site/Condition | Initial *N* | Final*N* | Gender*N* female | Age*M* (*SD*) | Num Lang*M* (*SD*) | Profa*M* (*SD*) | NumChecks/Clicksb*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 (011) | 9.13(1.99) |
| UK2-A | 43 | 41 | 34 | 20.73 (6.55) | 1.39 (0.66) | 0.72 (.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 | 20c | 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) |
| *Notes.* aA one-way ANOVA revealed significant differences in proficiency among the sites (*F*(6,624) = 1.964, *p* < .001,  = 0.536). Games-Howell post hoc tests revealed that POL-A and UK2-A had higher proficiency compared to all other sites (*p*s ≤ .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 (*p*s < .001); and no differences between the US sites were detected (*p*s ≥ .350). bFor the written mode, the number of clicks represents the number of target clicks. c10 participants did not report gender. |



*Figure 1.* Mean comprehension accuracy scores for each site by condition are reported. Error bars represent 95% confidence intervals. The dashed line represents chance level performance, i.e., 0.25 accuracy, and ‘\*’ indicates lead sites.



*Figure 2.* Forest plot of random-effect 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 and reports the raw effect sizes with 95% CIs by site. The ‘\*’ indicates lead sites. The overall meta-analytic (weighted mean) effect for each comparison is also plotted with 95% CIs.

* Figure 3.* Forest plot of random-effect 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 and reports the raw effect sizes with 95% CIs by site. The ‘\*’ indicates lead sites. The overall meta-analytic (weighted mean) effect for each comparison is also plotted with 95% CIs.