



REPLICATION STUDY

Proficiency, language of assessment, and attention to meaning and form during L2 comprehension: Methodological considerations in L2 replication research

Paweł Szudarski^{1*}  and Sylwia Mikołajczak² 

¹University of Nottingham, Nottingham, UK and ²Adam Mickiewicz University, Poznań, Poland

*Corresponding author. E-mail: Pawel.Szudarski@nottingham.ac.uk

(Received 15 March 2021; Revised 28 March 2022; Accepted 08 April 2022)

Abstract

This study is a replication and extension of Morgan-Short et al.'s (2018) investigation into the role of attention in input processing by L1-Polish learners of L2-Spanish, with proficiency and language of assessment explored as two key methodological factors. Our aims were twofold: to investigate learners' comprehension in different conditions with their L2 proficiency controlled for, and to examine this process when learners were tested using different languages. Data from three trials (N = 136) were analyzed: Trial-English, Trial-Polish, and Trial-Spanish, where comprehension was tested in English, Polish, or Spanish, respectively. Results showed that both L2 proficiency and language of assessment significantly affected learners' performance, with their scores being lower in the *-n* morpheme condition but only when comprehension was tested in English or Spanish. We discuss these findings both theoretically and methodologically, making recommendations on designing future replication studies and improving the generalizability of L2 findings across multiple research sites.

Introduction

Replication research is typically defined as repeating specific data collection procedures from previous studies with new participants and with the overall aim of verifying original claims, improving methodological rigor and eliminating “the possible influence of artefacts or chance on findings” (Porte, 2012, p. 4). As Plonsky (2015, p. 233) stresses, “replicability is fundamental to advancing scientific knowledge,” and while replication studies are underrepresented in the field of applied linguistics (Marsden et al., 2018), recently there has been more emphasis on the benefits of reproducing findings and increasing their generalizability (Andringa & Godfroid, 2019; McManus, 2021).

One important example is Morgan-Short et al. (2018), which explored attentional mechanisms in second language (L2) learners' simultaneous processing of form and meaning during comprehension. The authors focused on the feasibility of conducting large-scale multisite replication projects, demonstrating how to produce more generalizable findings relevant to diverse learning contexts. The current investigation extends this research by addressing two key methodological aspects: the role of L2 proficiency and language of assessment in exploring L2 learners' allocation of attention during comprehension.

Attention and processing of L2 input

Attention as a concept has been studied from a range of perspectives. In L2 acquisition, researchers have explored how attention, or conscious registration of surface elements of language, facilitates learning and the uptake of information. One of the key scholars in this area is VanPatten, who proposed the Primacy of Meaning Principle (VanPatten, 2004). He argued that L2 learners were limited capacity processors whose cognitive resources were contingent on the nature of specific tasks, with attention to meaning taking priority over attention to form when L2 input is processed.

One important question explored by VanPatten (1990) concerned how attending to both form and meaning can influence learners' comprehension. Specifically, his experiment involved 202 L1-English students who listened to a passage in L2 Spanish. Four different conditions were included: listening for content (condition 1) and listening for content but also paying attention to the lexical item *inflación* (condition 2), the feminine article *la* (condition 3), and the third-person plural morpheme *-n* (condition 4). During listening, attention was operationalized as learners making check marks, each time they heard the target item. Only learners with at least 60% of checks were included in the analysis.

Comprehension was measured using a recall test administered in English (learners writing down what they recalled from the passage). While no differences were found between learners processing *inflación* versus those listening for meaning only, the *la* and *-n* experimental conditions exhibited a drop in comprehension, which meant that attention to form competed with attention to meaning. Using these results, VanPatten (1990, p. 296) concluded that simultaneous processing of L2 form and meaning was difficult, particularly for learners at early and intermediate level who might attend to form only when input is understood.

Replication efforts following VanPatten's study

Several subsequent replications were carried out, exploring both L2 reading and listening and largely confirming VanPatten's (1990) findings (for a review, see Sanz & McCormick, 2021). However, it is of note that these replications made important methodological changes to the original design. For instance, Leow et al. (2008) replaced *inflación* with the word *sol*, introduced the clitic *lo*, and crucially for the current study used a multiple-choice comprehension test (10 questions written in English). Results indicated no significant differences between the experimental conditions in terms of affecting learners' comprehension. Similarly, Morgan-Short et al. (2012) used the same design and based on their findings questioned VanPatten's Primacy of Meaning principle.

However, there are important methodological differences between these subsequent replications, rendering it difficult to make valid comparisons. Evidence for this was recently provided by Sanz and McCormick (2021), who showed how the new experimental passage employed by Leow et al. (2008) and Morgan-Short et al. (2012) was more demanding linguistically than VanPatten's (1990) original text, consequently leaving L2 participants "with little to no resources to focus on form" (Sanz & McCormick 2021: 175). Likewise, Sanz and McCormick's analysis clearly demonstrated test effects, with the format of the comprehension test (multiple-choice vs. recalling) also affecting learners' results (see "Materials" section for details).

Multisite study by Morgan-Short et al. (2018)

Considering the lack of agreement in previous studies, Morgan-Short et al. (2018) embarked on a large-scale replication research study into L2 learners' attention to form during comprehension. Employing the same design as Leow et al. (2008), the authors aimed to show how multisite projects produce results that are generalizable across different learning contexts and learner groups.

With this goal in mind, the authors recruited 631 L2-Spanish learners from seven research sites across three countries (USA, UK, and Poland) and tested them across two input modalities (reading and listening), focusing on the question whether attending to both form and meaning can interfere with L2 comprehension. In addition to recruiting more participants, statistical analyses were enhanced by including meta-analytic effect sizes, mixed-effects analyses, and Bayes factors (for details, see Morgan-Short et al., 2018).

Overall findings revealed that attending to L2 forms had no effects on learners' comprehension, with participants in the *sol* and *la* conditions performing similarly to the control meaning-only group. However, there was some evidence of lower comprehension in the *-n* condition, particularly for the higher-level Polish learners, suggesting that the bound *-n* morpheme proved most taxing. As Morgan-Short et al. (2018, p. 419) summarized, the findings were consistent but not entirely uniform across all participants, with site-specific results providing "clues to variables that could merit further examination" and consequently becoming a source of new hypotheses. The present study sought to explore such new hypotheses, targeting in particular the apparently different performance of L1-Polish learners compared to their L1-English counterparts.

Current study: Effects of L2 proficiency and language of assessment

Focusing on the validity and generalizability of replication studies (Sanz & McCormick, 2021), the current examination aimed to verify whether learners' L2 proficiency and language of assessment, two variables unexplored in Morgan-Short et al. (2018), might have affected their results. Specifically, we replicated the same design to delve more into the performance of L1-Polish learners of Spanish, examining their apparently anomalous results compared to L1-English counterparts. Our rationale then was methodologically driven: to improve the rigor of multisite studies (Moranski & Ziegler, 2021), while gaining insights into L2 learners' attentional mechanisms.

In terms of theory, the study explored the effects of L2 proficiency and language of assessment as two potential variables that affect L2 learners' allocation of attention during comprehension. The complexity of L2 comprehension has been evident in a large body of work into reading and listening, with learners' performance being affected

by, among others, L1 reading and listening abilities (Yamashita, 2002) or task effects (Brantmeier, 2005).

As regards the impact of proficiency, Vandergrift (2006) for instance found that while both L1 ability and L2 proficiency affected learners' comprehension scores, the latter had a much stronger influence. Similarly, Mecartty (2000) reported significant effects of vocabulary but not grammar on L2 listening, highlighting how specific aspects of L2 proficiency can interact with learners' performance. Specifically addressing learners' simultaneous processing of form and meaning during reading, two recent studies (Sanz & McCormick, 2021; Son et al., 2021) convincingly demonstrated the role of L2 proficiency in affecting learners' allocation of attention.

Language of assessment is another key variable to be considered, and there is a large body of research examining the role of the mother tongue in teaching and assessing foreign languages (for an overview, see Shin et al., 2020). Both the language and method of assessment can affect comprehension results, with learners' scores being a function of their abilities but also the test format (e.g., Godev et al., 2002; Gordon & Hanauer, 1995; Shohamy, 1984; Yu, 2008). For instance, Yu (2008) found that L1-Chinese learners' summarization of information in Chinese rather than in English was a better measure of their reading ability. Godev et al. (2002) reported that answering comprehension questions in learners' L1 was a better reflection of their L2-Spanish reading skills, while Joyce (2018) showed that participants scored significantly higher in L2 vocabulary when the study and testing languages were matched, with tests of learners' knowledge conducted either using L1 translations or L2 definitions. Interestingly, Brantmeier (2006) found that while the language of assessment did not affect participants' results, when learners' proficiency was considered, those with lower reading abilities performed better in their L1 versus L2. Shohamy (1984) reported similar results, showing that answering comprehension questions in the L1 was easier, particularly for lower-level learners. Taken together, this research suggests an effect of the language of assessment on learners' performance.

In short then, the present study (1) examined learners' comprehension in different attentional conditions while controlling for L2 proficiency and (2) explored whether the language of assessment affected their performance. In Morgan-Short et al. (2018), the L1-Polish learners (1) represented higher levels in Spanish and (2) were the only group of participants whose comprehension was tested in another foreign language (English), unlike their L1-English counterparts who were tested in their L1. We hypothesized that these factors might have driven the obtained results.

With these goals in mind, our study took the form of three trials based on the same design. In the first one (Trial-English), we reanalyzed data from the Polish learners in Morgan-Short et al. (2018), controlling for their proficiency. Secondly, we conducted two further replications, but rather than test Spanish comprehension in English, as was the case in Morgan-Short et al. (2018), we did it in Polish (Trial-Polish) and Spanish (Trial-Spanish). These modifications were informed by our hypothesis that the language of assessment affects the measurement of L2 comprehension. Thus, by conducting this replication study, on a theoretical level we sought to extend findings into the role of attention in processing L2 input, while on a methodological level our goal was to improve the rigor of multisite research.

Research questions

The study addressed three research questions:

RQ1: Does the proficiency of L1-Polish learners of Spanish affect their comprehension when they process L2 input for both form and meaning?

RQ2: When L2 proficiency is controlled for, does attending to form and meaning affect learners' comprehension across different conditions?

RQ3: Does language of assessment affect learners' comprehension scores?

Participants

Overall 135 L1-Polish university students of Spanish participated in three trials: Trial-English ($n = 55$), Trial-Polish ($n = 37$), and Trial-Spanish ($n = 43$). Participants were young adults ($M = 21$, $SD = 1.58$) who studied Romance languages in Poland and represented an intermediate level in Spanish as measured by the DELE test (www.dele.org). This test was used previously as a reliable measure of L2 proficiency (Seibert Hanson & Carlson, 2014) and is recognized by the Spanish Ministry of Education. Mirroring Morgan-Short et al.'s (2018) design, two levels of the grammar section of the test were used, basic and intermediate, corresponding to A1/A2 and B1 levels of the CEFR, respectively. Each level contained 12 multiple-choice questions that targeted features such as Spanish tenses or morphology. As Table 1 presents, the average proficiency score for all participants was .79 ($SD = .15$), with the majority of students scoring above 70%, which was expected given their status as students of Spanish philology (see Supplementary Materials for details on participants' proficiency within each trial). Crucially, between the three trials, there were no significant differences in participants' proficiency, $F(2, 133) = 1.96$, $p = .145$, $\eta_p^2 = .028$. Further, following Morgan-Short et al. (2018), our analysis included only those participants who made at least six check marks during listening. This threshold was introduced to exclude those students who failed to follow the research protocol and, for example, listened to the text without paying attention to the target forms.

Materials and treatment

As the design of Morgan-Short et al. (2018) was replicated, we used the same materials, which included a biodata form to elicit background information, an audio file with the experimental text, and condition-specific instructions (for details, see <http://osf.io/uybak>). The only modification was the language of the comprehension test in Trial-Polish and Trial-Spanish.

The format of the comprehension test consisted of 10 multiple-choice questions originally written by Leow et al. (2008). All questions could be answered without the need to interpret the target forms (one correct answer out of four options). However, unlike in Leow et al., where the reliability coefficient was high (.915), in our data these coefficients were lower and inconsistent (.596, -.006, and .556 in Trial-English, -Polish,

Table 1. Participant information

	No of participants (n)	Females (n)	Age M (SD)	Proficiency M (SD)	No of check marks M (SD)
Trial-English	55	48	21.13 (1.72)	.76 (.19)	9.68 (1.67)
Trial-Polish	37	33	21.24 (1.19)	.83 (.09)	7.14 (4.24)
Trial-Spanish	43	40 ^a	20.91 (1.67)	.79 (.15)	7.12 (5.04)
Total	134	121	21.09 (1.57)	.79 (.15)	7.98 (4.10)

^aOne participant self-reported as nonbinary.

and -Spanish, respectively). This aligns with Sanz and McCormick's (2021) recent study that also reported problems with this test as a measure of Spanish comprehension. As we sought to directly replicate Morgan-Short et al. (2018), we needed to employ the same test to avoid introducing any confounding variables. However, beyond the immediate context of the current study, we concur with Sanz and McCormick (2021) that future studies in this line of research should rely on other measures of comprehension.

The experimental text was taken from Leow et al. (2008). Recorded by a native-speaker of Spanish at a slightly slower pace, this text was short (3 minutes and 43 seconds) and consisted of 23 sentences, with 10 instances of each target form (*sol*, *la*, *-n*) distributed among four paragraphs.

The study was conducted during learners' regular classes and the same procedures were used across the three trials, with participants assigned to one of the following conditions:

Condition 1: control group

Condition 2: *sol*

Condition 3: *la*

Condition 4: *-n*

While the control group focused on comprehension only, in the remaining conditions participants listened for meaning but were also instructed to attend to the target L2 forms (*sol*, *la*, or *-n*) and, upon hearing them, make check marks on separate sheets of paper. After this listening, participants completed the comprehension test and the Spanish proficiency test.

In terms of analysis, in RQ1 we performed a series of Spearman correlations to examine the relationship between learners' Spanish proficiency and comprehension across the three trials. Establishing whether these two variables correlated with each other was deemed important theoretically and informed our subsequent analysis. Because the same design was employed across the three trials and all of them concerned the same research questions, we wanted to explore the relationship between attention to form and comprehension both cumulatively across all participants and for each of the trials separately, to take into account the impact of the language of assessment (see also Supplementary Materials for an additional analysis of correlations between learners' comprehension and number of check marks made during listening, which was Morgan-Short et al.'s [2018] way of operationalizing learners' attention to form). In RQ2, we used ANCOVAs to reanalyze data from Morgan-Short et al. (2018) and explore comprehension in relation to condition while controlling for learners' proficiency. Because in this study the Polish ($n = 55$) and British ($n = 44$) learners represented higher proficiency than their US counterparts ($n = 338$), in our analysis L2 proficiency was considered a covariate. In RQ3, to explore any potential effects of the language of assessment, ANCOVAs were used to analyze the two further sets of data where comprehension was measured in Polish (Trial-Polish) or Spanish (Trial-Spanish).

Results

Table 2 shows descriptive statistics for participants' comprehension across the three trials, followed by the results for each of the research questions.

Table 2. Descriptive statistics for listening comprehension

Condition	Trial-English (n = 55)		Trial-Polish (n = 37)		Trial-Spanish (n = 43)	
	Mean (SD)	Mean ^a (SE)	Mean (SD)	Mean ^a (SE)	Mean (SD)	Mean ^a (SE)
<i>control</i>	.41 (.24)	.32 (.04)	.48 (.16)	.48 (.05)	.62 (.24)	.59 (.05)
<i>sol</i>	.48 (.20)	.38 (.04)	.43 (.16)	.42 (.05)	.39 (.13)	.43 (.05)
<i>la</i>	.35 (.25)	.25 (.04)	.51 (.09)	.51 (.04)	.53 (.21)	.50 (.06)
<i>-n</i>	.23 (.10)	.14 (.05)	.49 (.15)	.49 (.05)	.36 (.17)	.38 (.06)
Total	.37 (.22)	.38 (.02)	.48 (.14)	.46 (.03)	.48 (.21)	.48 (.03)

^aMeans scores adjusted with proficiency treated as a covariate.

Table 3. Correlations between learners’ average comprehension and Spanish proficiency scores

	Spanish comprehension			
	Trial-English (n = 55)	Trial-Polish (n = 37)	Trial-Spanish (n = 43)	All participants (n = 135)
L2 Spanish proficiency	$\rho = .43^*$ $p = .002$	ns	$\rho = .55^*$ $p < .001$	$\rho = .39^*$ $p < .001$

*Indicates significant correlation.

Research question 1

A series of correlation tests was performed to explore the relationship between learners’ L2 proficiency and comprehension and, as explained in the preceding text, this analysis was carried out both across all participants and for each trial separately. Table 3 presents a summary of this analysis.

Results showed that the relationship between comprehension and proficiency was significant in Trial-English and Trial-Spanish, but not in Trial-Polish. These important findings highlight the role of L2 proficiency as a key factor in research exploring L2 learners’ simultaneous processing of form and meaning. Consequently, in our subsequent analysis, we employed ANCOVAs to treat proficiency as a covariate and account for its effects.

Research question 2

To examine comprehension in different experimental conditions, a one-way ANCOVA was used to reanalyze data from the L1-Polish participants in Morgan-Short et al. (2018). This analysis showed a significant effect of condition on learners’ comprehension with a large effect size, $F(3,50) = 3.71, p = .017, \eta^2 = .182$, with proficiency being a significant variable, $F(1,50) = 11.62, p = .001, \eta^2 = .189$ (see Supplementary Materials for full statistical information). The *-n* group had the lowest mean score, with post-hoc Tukey tests revealing that this condition was significantly lower than the *sol* condition ($p = .016$). All other comparisons between the conditions were nonsignificant. This analysis then demonstrated two key points: L2 proficiency mediated learners’ allocation of attention during listening, and, even with the effects of proficiency accounted for, the

-n condition negatively affected learners' comprehension compared to the lexical *sol* condition.

Research question 3

ANCOVAs were performed to explore the effects of language of assessment on learners' comprehension. In Trial-Polish, the effects of proficiency ($F(1, 32) = .571, p = .45, = .018$) or condition ($F(3, 32) = .655, p = .59, = .058$) were nonsignificant. However, in Trial-Spanish, both proficiency ($F(1, 38) = 10.22, p = .003, = .212$) and condition ($F(3, 38) = 2.95, p = 0.45, = .189$) were significant with large effect sizes, confirming our hypothesis that the language of assessment impacts on students' comprehension results (see Supplementary Materials for details). In terms of specific experimental conditions, while post-hoc tests with Bonferroni corrections ($p < .008$) failed to reach significance, descriptively the *-n* group exhibited the lowest comprehension, indicating comprehension difficulties in this condition.

As the *-n* morpheme proved problematic, an additional two-way ANCOVA was performed to compare learners' comprehension as dependent on condition and trial. There was a significant interaction between condition and trial ($F(6, 122) = 2.75, p = .015, = .119$) and significant main effects for condition, ($F(3, 122) = 2.69, p = .049, = .062$), trial ($F(2, 122) = 4.03, p = .02, = .062$), and proficiency ($F(1, 122) = 23.81, p < .001, = .163$; see Supplementary Materials for full statistical information). Post-hoc tests pointed to clear effects of the language of assessment: overall comprehension scores for all participants were significantly higher ($p = .029$) in Trial-Spanish than in Trial-English, while for the *-n* condition specifically, learners' results in Trial-Polish were significantly higher ($p = .002$) than in Trial-English. Thus, based on these findings, it is clear then that testing these learners in English significantly affected their comprehension results.

Discussion

Focusing on the methodological rigor of L2 replication research, our study extended the findings of Morgan-Short et al. (2018) and explored whether attending to form and meaning interfered with L2 learners' comprehension. Specifically, while that study showed no overall interference across all participants, it also indicated some site-specific variance, particularly among L1-Polish learners of Spanish in the *-n* condition. These results prompted us to explore the effects of proficiency and language of assessment as two key variables that we hypothesized might affect L2 comprehension beyond the effect of condition.

Firstly, we demonstrated that comprehension was related to learners' Spanish proficiency as revealed by significant correlations in Trial-English and Trial-Spanish. In itself this result is unsurprising if we consider previous findings pointing to L2 proficiency as a strong predictor of learners' comprehension and a factor driving the dynamic transfer between L1 and L2 reading skills (e.g., Lee & Schallert, 1997; Vandergrift, 2006). However, in light of the continuing replication efforts concerning VanPatten's (1990) initial study, this evidence is important because it shows the role of L2 proficiency in mediating learners' comprehension across experimental conditions. Methodologically, this result underlines the importance of accounting for L2 proficiency effects when interpreting results (Sanz & McCormick, 2021).

In Morgan-Short et al. (2018), for instance, the negative impact of the *-n* condition was found in the higher-level Polish and British participants but not in the lower-level US learners. Similarly, a recent eye-tracking study by Son et al. (2021) reported different reading patterns in lower- and higher-level participants allocating attention to specific L2 forms, which suggests potentially more automatized reactions as a characteristic of higher proficiency learners. As speculated by Morgan-Short et al. (2012) in an earlier replication, “effects of simultaneous attention to form and meaning might only be detected when learners are beyond the lowest levels of proficiency,” with higher proficiency needed to allow more variance and show potential effects of specific experimental conditions (see Sanz & McCormick, 2021 for similar arguments).

On a theoretical level, this is in line with VanPatten’s (1990) original prediction that for beginner and intermediate learners, attending to meaning in an L2 text is sufficiently taxing, leaving them with little cognitive capacity to also attend to specific linguistic features. It appears that any potentially negative effects of attending to both form and meaning are more likely to be detected at higher levels of L2 proficiency, where we have evidence that learners have in fact understood the experimental text. Interestingly, while our findings demonstrated that learners’ L2 proficiency interacted with their listening, these correlations were significant when the comprehension test was administered in English and Spanish (.43 and .39, respectively), but not in Polish. This means that the effects of proficiency are also dependent on the way comprehension is operationalized (see the following text for discussion of the effects of language of assessment).

Secondly, our study extended previous findings by showing that even with L2 proficiency accounted for, there were differences in learners’ comprehension across the experimental conditions. Namely, we found that the *-n* group achieved the lowest comprehension score and was significantly lower than the *sol* group, indicating a trade-off in learners’ comprehension of meaning and simultaneous processing of specific L2 forms. This is similar to the results of Morgan-Short et al. (2018) and Son et al. (2021), where the *-n* morpheme also proved demanding and interfered with learners’ comprehension. It is worth adding that the latter study contained an additional experimental condition with the grammatical item *lo* which also required more cognitive efforts than processing the word *sol*, leading the authors to the conclusion about the differential effects of purely grammatical (e.g., *-n*) versus lexical or lexico-grammatical forms (e.g., *sol*) on learners’ scores. This growing body of research suggests then that for L2 learners, when they process input for both meaning and form, the type of forms they attend to is an important variable that determines differences in attention allocation. As explained by Morgan-Short et al. (2018), the form *-n* as an unstressed and bound morpheme at the end of Spanish verbs differs in linguistic terms from an independent word such as *sol* and consequently is more likely to be cognitively demanding.

Thirdly, our analysis produced a number of key methodological insights related to the effects of the language of assessment. Specifically, we found that when L1-Polish learners of Spanish were tested in the L1 (Trial-Polish), their results were comparable across the experimental conditions, irrespective of what target forms they attended to. However, with the comprehension test administered in Spanish (Trial-Spanish), the results resembled those when the test was performed in English (Trial-English), with the *-n* condition leading to significantly lower scores. This means that depending on how comprehension was operationalized and measured, there emerged different patterns in learners’ performance.

These results thus confirm our hypothesis about the impact of the language of assessment and support earlier research which indicated differences in L2 learners’

comprehension depending on whether they were tested in their L1 versus L2 (e.g., Godev et al.; Shohamy, 1984; Yu, 2008). They also provide new methodological insights into the findings of Morgan-Short et al. (2018), particularly regarding the performance of the L1-Polish participants. Unlike their British or American counterparts, these learners were the only group in that study whose understanding of L2-Spanish was demonstrated in another foreign language (English) rather than in their L1 (Polish). Our study shows that this constituted additional difficulty for these learners, which in turn led to lower comprehension scores. Indeed, when the average scores of all Polish participants were considered together regardless of condition, the results in Trial-English were significantly lower than in Trial-Spanish. Interestingly, this difference was nonsignificant when Trial-Polish was compared with Trial-Spanish, suggesting potentially that matching the study and testing languages (Joyce 2018) may be the optimal solution, particularly for multisite designs that involve learners with different levels of language learning experience and multilingualism. Based on our results, we argue that future multisite studies should consider the language of assessment to be used across multiple research sites, considering context-specific policies and equity issues for all participants. For heritage speakers based in the United States, for instance, testing them through their L1 might have very different theoretical and practical implications than testing L1-Polish learners of Spanish through Spanish, one of several foreign languages they may speak. In fact, in Poland testing L2 learners through the target language (be it English, German, or Spanish) is generally accepted, and informal feedback from our participants suggested that, as students of Spanish, they had expected to be tested in this language as this was the practice they had been used to doing.

Returning to the main question about the role of attention in L2 learners' processing of input, it is important to discuss the meaning of our findings with respect to the effects of specific experimental conditions. If we assume that answering comprehension questions in one's L1 is easier, then the conditions for the L1-Polish participants in Trial-Polish resembled those in Morgan-Short et al. 2018 (L1-English participants tested in English), and in both cases the results converged: learners' comprehension was not adversely affected by the experimental conditions. However, different conclusions were drawn based on the results of Trial-Spanish, where the *-n* condition clearly interfered with learners' comprehension. Showcasing the benefits of replication research, on the one hand, these results also highlight the fundamental role of methodological decisions in affecting the validity and generalizability of L2 findings on the other (see Sanz & McCormick, 2021 for similar arguments about different operationalizations of L2 comprehension).

Methodological implications

While this study was a single-site examination of L1-Polish learners of Spanish at university level, its starting point was the intention to extend the multisite findings of Morgan-Short et al. (2018) and underline the importance of methodological considerations in replication research. By employing the same design and procedures across three trials, we were able to provide a fuller account of site-specific effects and consequently gain a better understanding of attentional mechanisms when L2 learners pay simultaneous attention to form and meaning. More broadly, however, our findings also brought to the fore a number of important methodological points that should inform the design of future replication studies, particularly those involving multisite designs (Moranski & Ziegler, 2021).

Firstly, with Spanish proficiency being a significant factor in the operationalization of comprehension as a construct, we provided strong evidence for the importance of this variable in mediating L2 learners' allocation of attention during listening, particularly when comprehension is measured in learners' different languages. Therefore, this stresses the need for future studies in this line of research to control for the effects of L2 proficiency. As emphasized by Sanz and McCormick (2021: 165), the key to the discussion is "the possibility that with low levels of comprehension, the potential effects of a secondary task (i.e. attending to form) are hidden" (see also Son et al., 2021 for recent findings showing differences in reading behaviors between higher- and lower-level learners).

Secondly, our study confirmed the fundamental role of measurement in L2 research by pointing to reliability issues with the multiple-choice test. This format was first employed by Leow et al. (2008) to replace VanPatten's (1990) original recall assessment, but as Sanz and McCormick (2021) convincingly show, the former is a coarser measure which offers a poorer representation of the construct of comprehension, particularly when learners' understanding is low. The inconsistent performance of this test across different studies (Morgan-Short et al., 2018; Sanz & McCormick, 2021; Son et al., 2021; current study) is a clear limitation that should be addressed in future research. Unless the comprehension of the experimental text is measured reliably, any attempts at examining learners' simultaneous attention to form and meaning run the risk of hiding potential variation resulting from the effects of conditions.

Lastly, in addition to the format of assessment, our study also revealed the effects of language of assessment, with the language in which comprehension was measured affecting learners' results. These are particularly relevant findings if the field is to rely more on multisite studies that involve participants representing different learning contexts or levels of linguistic competence (Andringa & Godfroid, 2020). As Moranski and Ziegler (2021: 224) observe, in such studies researchers are likely to face the challenge of working in contexts that may differ along many methodological dimensions such as participants' levels, proficiency measures or site-specific requirements of language programs. Thinking of increasing the robustness of L2 findings (Plonsky, 2015), future studies should select and agree upon such methodologies that avoid the introduction of confounding variables and ensure the comparability of results across sets of site-specific data (e.g., testing students from all research sites in the same language or ensuring participants' similar proficiency levels in the target language). Finally, power analysis and extensive piloting of materials to be employed across multiple research sites are an effective way of optimizing and standardizing research designs and procedures aimed at replication and reproducibility efforts (for an example, see Peters et al., *under review*).

Conclusions and future research

Building on Morgan-Short et al. (2018), the aims of the current study were twofold: firstly to extend the existing findings into the role of attention in L2 learners' processing of input and, secondly, to explore the effects of L2 proficiency and language of assessment on learners' comprehension as key methodological considerations in replication L2 research. With the effects of L2-Spanish proficiency accounted for, our analysis showed that learners' listening performance was significantly lower in the *-n* group compared to the *sol* experimental conditions, but only when comprehension was tested in English and Spanish rather than in Polish (learners' L1). As such, this

study provides valuable insights into the complexity of studying L2 comprehension across different learning contexts, both at a theoretical and practical level. Importantly, our findings demonstrated the benefits of replication studies in consolidating findings and explaining the impact of unexpected factors (McManus, 2021), while underlining the importance of key methodological decisions that need to be considered in planning and conducting multisite L2 research (Moranski & Ziegler, 2021).

It should be acknowledged that the study has several limitations. Firstly, given the low reliability of the multiple-choice comprehension test, future research should explore other ways of measuring L2 comprehension such as cloze tests (Cheng, 2004) or aural formats (Kim & Godfroid, 2019). Secondly, in our study we followed the design of Leow et al. (2008) and Morgan-Short et al. (2018) and operationalized attention as learners' check marks made during the listening process each time a given target form was heard. As shown by Son et al., (2021), online methodologies such as eye tracking are able to provide concurrent data that are more fine-grained operationalizations of attention at different levels and depth of processing. Thirdly, given the impact of proficiency as a key variable in L2 learners' simultaneous processing of form and meaning, future studies should control for the level of learners' target language both across and within research sites that participate in multisite designs. Finally, our findings should be replicated with more participants who represent diverse L2 populations, including multilingual learners or learners of varied L2 proficiency, whose different levels of metalinguistic awareness might affect their performance (e.g., Han & Peeverly, 2007).

Acknowledgments. We are grateful to Kara Morgan-Short and Emma Marsden for their leadership and support during this project. We would also like to thank the editor and three anonymous SSLA reviewers for their valuable comments. A special thank you is due to Beatriz González-Fernández and Renata Seredyńska for reading earlier versions of our manuscript.

Supplementary Materials. To view supplementary material for this article, please visit <http://doi.org/10.1017/S0272263122000171>.

References

- Andringa, S., & Godfroid, A. (2019). Call for participation. *Language Learning*, 69, 5–10.
- Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, 40, 134–142.
- Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension in Spanish. *Modern Language Journal*, 89, 37–53.
- Brantmeier, C. (2006). The effects of language of assessment and L2 reading performance on advanced readers' recall. *The Reading Matrix*, 6, 1–17.
- Cheng, H. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals*, 37, 544–553.
- Godev, C. B., Martinez-Gibson, E. A., & Toris, C. C. M. (2002). Foreign language reading comprehension test: L1 versus L2 in open-ended questions. *Foreign Language Annals*, 35, 202–221.
- Gordon, C. M., & Hanauer, D. (1995). The interaction between task and meaning construction in EFL reading comprehension tests. *TESOL Quarterly*, 29, 299–324.
- Han, Z.-H., & Peeverly, S. (2007). Input processing: A study of ab initio learners with multilingual backgrounds. *International Journal of Multilingualism*, 4, 17–37.
- Joyce, P. (2018). L2 vocabulary learning and testing: The use of L1 translation versus L2 definition. *Language Learning Journal*, 43, 217–227.
- Kim, K. M., & Godfroid, A. (2019). Should we listen or read? Modality effects in implicit and explicit knowledge. *Modern Language Journal*, 103, 648–664.

- Lee, J.-W., & Schallert, D. L. (1997). The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: A test of the Threshold Hypothesis in an EFL context. *TESOL Quarterly*, 31, 713–739.
- Leow, R. P., Hsieh, H., & Moreno, N. (2008). Attention to form and meaning revisited. *Language Learning*, 58, 665–695.
- Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, 68, 321–391.
- McManus, K. (2021). Are replication studies infrequent because of negative attitudes? Insights from a survey of attitudes and practices in second language research. *Studies in Second Language Acquisition*. Advance online publication. <https://doi.org/10.1017/S0272263121000838>
- Mecarty, F. (2000). Lexical and grammatical knowledge in reading and listening comprehension by foreign language learners of Spanish. *Applied Language Learning*, 11, 323–348.
- Moranski, K., & Ziegler, N. (2021). A case for multisite second language acquisition research: Challenges, risks, and rewards. *Language Learning*, 71, 204–242.
- Morgan-Short, K., Heil, J., Botero-Moriarty, A., & Ebert, S. (2012). Allocation of attention to second language form and meaning. *Studies in Second Language Acquisition*, 34, 659–685.
- Morgan-Short, K., Marsden E., Heil, J., Issa, B. I., Mikhaylova, A., Mikołajczak, S., Moreno, N., Leow, R. P., Slabakova, R., & Szudarski, P. (2018). Multi-site replication in SLA research: Attention to form during listening and reading comprehension. *Language Learning*, 68, 392–437.
- Peters, E., Puimege, E. & Szudarski, P. (under review). Repetition and incidental learning of multiword units: A conceptual replication of Webb, Newton, and Chang (2013). Stage 1 Registered Report with In-Principle Acceptance. *Language Learning*.
- Plonsky, L. (2015). Quantitative considerations for improving replicability in CALL and applied linguistics. *CALICO Journal*, 32, 232–244.
- Porte, G. K. (Ed.) (2012). *Replication research in applied linguistics*. Cambridge University Press.
- Sanz, C., & McCormick, T. J. (2021). VanPatten (1990)'s long and winding story and the nature of replication studies. In M. J. Leeser, G. D. Keating, and W. Wong (Eds.) *Research on second language processing and processing instruction: Studies in honor of Bill VanPatten*. (pp. 153–181). John Benjamins.
- Seibert Hanson, A., & Carlson, M. (2014). The roles of first language and proficiency in L2 processing of Spanish clitics: Global effects. *Language Learning*, 64, 310–342.
- Shin, J.-Y., Dixon, L. Q., & Choi, Y. (2020). An updated review on the use of L1 in foreign language classrooms. *Journal of Multilingual and Multicultural Development*, 41, 406–419.
- Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing*, 1, 147–170.
- Son, M., Lee, J., & Godfroid, A. (2021). Attention to form and meaning revisited: Insights from eye tracking. *Studies in Second Language Acquisition*. Advance online publication. <https://doi.org/10.1017/S0272263121000565>
- Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency. *Modern Language Journal*, 90, 6–18.
- VanPatten, B. (1990). Attending to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition*, 12, 287–301.
- VanPatten, B. (Ed.) (2004). *Processing instruction: Theory, research, and commentary*. Lawrence Erlbaum Associates.
- Yamashita, J. (2002). Mutual compensation between L1 reading ability and L2 language proficiency in L2 reading comprehension. *Journal of Research in Reading*, 25, 81–95.
- Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing*, 25, 521–551.

Cite this article: Szudarski, P. and Mikołajczak, S. (2022). Proficiency, language of assessment, and attention to meaning and form during L2 comprehension: Methodological considerations in L2 replication research. *Studies in Second Language Acquisition*, 1–13. <https://doi.org/10.1017/S0272263122000171>