Studies in Second Language Acquisition, **43** (2021), 50–82
doi:10.1017/S0272263120000261

Research Article

ON THE SCOPE OF OUTPUT IN SLA

TASK MODALITY, SALIENCE, L2 GRAMMAR NOTICING, AND DEVELOPMENT

Janire Zalbidea *

Temple University

Abstract
Following calls for more modality-sensitive perspectives of SLA, this study investigated the extent to which (a) producing the second language (L2) in the oral modality impacts learner-generated noticing and L2 development of grammatical structures embedded in subsequent auditory input, and whether (b) engaging in L2 production and input processing in the written modality differentially contributes to learner noticing and L2 outcomes compared to the oral modality. Participants were beginner-level L2 Spanish learners assigned to one of three pedagogic task conditions (No-output, Speaking, Writing). Two target structures differing in their relative intrinsic salience were considered in the study. Learners' noticing behaviors were gauged using stimulated recall protocols, and L2 grammar development was measured using pre-, post-, and delayed posttests of production and written and aural acceptability judgment. Results revealed that engaging in oral output promoted greater noticing and deeper analysis of auditory input as well as more robust L2 grammar development compared to no output. However, sustained linguistic gains on the lower-salience target structure were only observed among participants who engaged in output and input processing in the written modality.

The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at: https://www.iris-database.org/iris/app/home/detail?id=york%3a937905&ref=search; https://www.iris-database.org/iris/app/home/detail?id=york%3a937906&ref=search; https://www.iris-database.org/iris/app/home/detail?id=york%3a937907&ref=search

I am most grateful to Cristina Sanz, Ronald P. Leow, Lourdes Ortega, and Andrea Révész for their valuable feedback in conducting this research. I also thank Meagan Driver, Linxi Zhang, and Aitor Martínez de la Pera for their support with various aspects of the study, and Nick Pandža for his guidance with statistical analyses. The manuscript benefitted greatly from the helpful comments of the SSLA editors and anonymous reviewers. Any remaining errors are my own.

* Correspondence concerning this article should be addressed to Janire Zalbidea, Department of Spanish and Portuguese, Temple University, Anderson Hall 413, 1114 W. Polett Walk, Philadelphia, Pennsylvania 19122. E-mail: janire.zalbidea@temple.edu

INTRODUCTION

A primary goal of psycholinguistic research in second language acquisition (SLA) is to understand how the design and implementation of different second language (L2) tasks impact their potential for advancing learner-generated SLA processes and outcomes. Although myriad task-related factors bear relevance for L2 development, all forms of L2 use can be distinguished along two fundamental properties: (a) whether they require output production or input perception only, and (b) whether they occur in the oral or written modality. Given the pervasiveness of these properties, understanding the affordances of engaging in output and using the L2 in the oral and written mediums for L2 development is critical for achieving unified accounts of SLA theory and pedagogy. Such research, beyond its conceptual and practical implications, is also scientifically important for establishing the generalizability of SLA principles across modalities (e.g., Byrnes & Manchón, 2014; Gilabert et al., 2016; Ortega, 2012).

Featuring modality from a task-based perspective, this study revisits a core tenet in SLA, namely, that producing L2 output promotes learner-generated noticing of relevant grammatical forms embedded in the input, increasing opportunities for L2 development. This "noticing function of output," as proposed by Swain (e.g., 1995, 2000, 2005), remains a central theoretical assumption in many contemporary research and pedagogic frameworks in SLA; however, the impact of output on L2 learning processes and outcomes is still not clearly understood. Earlier research has been predominantly conducted within the written modality (e.g., Izumi, 2002; Izumi & Bigelow, 2000; Uggen, 2012), and the extent to which this function of output operates in the oral modality remains to be established. Additionally, comparative modality research that examines how oral and written output-based tasks differentially contribute to L2 learning processes and outcomes is lacking. Against this background, the present study investigates the extent to which (a) producing output in the oral modality impacts learner-generated noticing and development of L2 grammar embedded in auditory input, and (b) producing output and processing input in the written modality affords L2 learning benefits over producing output and processing input in the oral modality. To provide a more comprehensive account of the contributions of output alongside modality, the study considers two target structures differing in their relative salience, a key factor posited to mediate learners' noticing behaviors (e.g., N. Ellis, 2017; Gass et al., 2017; Housen & Simoens, 2016).

BACKGROUND

OUTPUT DEMANDS, NOTICING, AND L2 GRAMMAR DEVELOPMENT

In the field of SLA, it is well assumed that noticing L2 forms in the input, as first conceptualized by Schmidt (e.g., 2001), comprises an important condition for L2 forms to be further processed in learners' working memory and, eventually, to be incorporated into their developing L2 system. Theoretical models following a limited view of L2 attention (e.g., Leow, 2015; VanPatten, 2004) hold that learners, particularly at incipient proficiencies, often disregard or only partially process the formal dimensions of the L2 (e.g., inflectional morphemes) because the most meaning-bearing dimensions of input (e.g., temporal adverbs) are prioritized in their attentional focus. Against this backdrop,

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

empirical efforts have been made to understand how pedagogic L2 tasks may be structured to best promote cognitive processes associated with successful L2 grammar development. Accumulated research indicates that external instructional interventions that aim to draw learners' attention to form (e.g., input enhancement) are not always effective because "learners often have their own internal agenda for learning" (Park, 2013, p. 75), which prompts them to engage in self-generated noticing processes. In this regard, producing L2 output has been hypothesized to serve as a powerful cognitive tool that can directly influence learners' purported internal agenda (see, e.g., Izumi, 2003; Swain, 2005). Specifically, Swain (e.g., 1995, 2000, 2005) has argued that output can have a "noticing function," such that "the activity of producing the target language may prompt [L2] learners to recognize consciously some of their linguistic problems: It may bring their attention to something they need to discover about their [L2] (possibly directing their attention to relevant input)" (Swain, 2005, p. 474).

The L2 learning benefits stemming from the noticing function of output are claimed to derive primarily from the formulation stage in Levelt's (1989) model of speech production (see Kormos, 2006, for a bilingual speech production model), where a conceptualized (nonverbal) message is converted into linguistic material (see Izumi, 2003). Formulation entails several substages of grammatical encoding, including lexical selection and integration of morphosyntactic information, giving way to surface structure. The assumption is that when learners produce or attempt to produce L2 output, they are likely to encounter some form of mismatch between the conceptualized and the formulated message (e.g., a learner may realize they lack knowledge of morphology to express future tense), as noted by Swain. This experience can, in turn, create an internal impetus for learners to search for relevant exemplars in the available L2 input on which to ground a targetlike revision. In so doing, learners are expected to pay focused attention to the formal aspects of input that are relevant to arrive at such revisions, and to notice and more deeply process any (in) consistencies between their own production and available L2 models. Repeated output opportunities can also facilitate consolidation of newly learned L2 knowledge through retrieval processes (e.g., Leow, 2015). Although Levelt's model is most widely assumed in SLA (see Kormos, 2006), more recent psycholinguistic models of speech production (e.g., usage-based, connectionist) also support the notion that production can impact the quality of newly formed linguistic representations (see Zamuner et al., 2016).

PRIOR SLA RESEARCH ON THE NOTICING FUNCTION OF OUTPUT

A number of empirical investigations, conducted predominantly in the domain of L2 writing, have put the noticing function of output to the test. Yet, despite its centrality to SLA theory and pedagogy, to date this function of output has not been conclusively supported by accumulated evidence. Indeed, out of the eight available studies summarized subsequently, approximately half have not documented substantive output effects in terms of both noticing and L2 development. This variability in earlier findings appears to stem in part from the diverse experimental operationalizations of output, the selection of target constructions, and the different quantitative and (to a lesser extent) qualitative measures used to gauge noticing behaviors.

In the seminal studies by Izumi et al. (1999) and Izumi and Bigelow (2000), training comprised two tasks: First, L2 participants wrote an essay on a topic chosen to elicit (experimental group) or not to elicit (control group) the target form (past hypothetical conditional in English). Next, they read a model essay containing target form exemplars, after being instructed to underline words that would help rewrite the essay (experimental group) or to read for comprehension (control group). For the second task, participants read a short passage, again following different instructions for underlining. Next, participants reconstructed the passage (experimental group) or responded to questions (control group). This cycle was repeated twice. Noticing was primarily operationalized as participants' underlining of the target form, and L2 improvement was assessed using written production tests and either a grammaticality judgment task (Izumi et al., 1999) or a recognition test (Izumi & Bigelow, 2000). No significant differences were found between the experimental and control groups in either noticing or L2 outcomes, although vast individual variability was reported.

Mixed evidence has emerged from a series of conceptual replications and extensions of these early studies. In Song and Suh (2008), L2 English learners who engaged in text reconstruction or picture-cued writing tasks did not evidence higher underlining rates of the past hypothetical conditional than learners who engaged in reading comprehension. Groups also improved similarly in a recognition test; nonetheless, production tests revealed an advantage for participants who produced output. A minimal role for output was also reported in Leeser (2008), where participants completed written text reconstruction tasks that required listening to (rather than reading) the passage. Participants also received a preemptive grammar review on the target form (the Spanish preterit/imperfect tense). Output effects were not observed for either noticing (operationalized as note-taking) or L2 improvement in a written production test.

In Izumi (2002), learners' note-taking behavior in a written text reconstruction task was employed as a primary measure of noticing, and the assessment was expanded to include multiple production and receptive tests. The experimental group was found to experience greater L2 gains on English relative clauses than the control group; however, the note-taking data did not reveal any output effects. In a conceptual replication of Izumi's study, Russell (2014) used underlining rate (number of target verbs underlined in input passages) and free written recall to gauge noticing of the future tense in L2 Spanish. This time, results revealed output effects in L2 outcomes and, contrary to the original study, in the noticing measures, which led Russell to suggest that the properties of the target structure may modulate the observed effects of output.

Uggen's (2012) study called further attention to the characteristics of the target form in this line of research. Uggen expanded Izumi and Bigelow's design by employing stimulated recall protocols to further tap into participants' noticing behaviors. She also considered two target forms of varying complexity (the present and past hypothetical conditional in English) and administered a delayed posttest. Essay writing and picture-cued written production tasks were employed as treatment and assessment tests, respectively. Unlike in the original study, groups received the same underlining instructions. Findings revealed no output effects for the less complex form (present conditional); yet, the benefits of output were observable in terms of noticing and L2 outcomes for the more complex form (past conditional). Results also indicated that underlining data underestimated learners' degree of noticing compared to stimulated recall data, which led Uggen to advocate for the use of qualitative measures of noticing in this paradigm.

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

A commonality of these prior studies is their focus on researching output solely within the written modality. Because written and oral tasks may offer different opportunities for focus on form, as detailed later on, Izumi and Izumi (2004) investigated whether output effects would be present in the oral modality.[1] In this case, auditory input was presented to L2 English learners at the sentence level, and participants were exposed to it both before and after output. Noticing was measured using a retrospective questionnaire. Against researchers' predictions, findings revealed an advantage for the no-output group in development of the target form (relative clauses). The authors speculated that participants might not have fully engaged in the production mechanisms that support the noticing function of output, as learners repeated input verbatim without using their own resources. Another possibility is that participants experienced high L2 processing demands posed by the oral nature of the task (e.g., Gilabert et al., 2016), which may have mitigated the expected output effects found in earlier research.

In sum, some evidence exists to suggest that output can positively impact learners' noticing and possibly L2 grammar development under certain conditions. However, a series of methodological limitations constrain the inferences that can be made from accumulated findings. These limitations include a lack of qualitative measures of noticing that can elucidate how learners engaged with input and an absence of delayed assessments to estimate the sustainability of output effects in long-term memory in nearly all studies (cf. Uggen, 2012). Existing research is also limited by generally low sample sizes (average $n = 13$ per group) and assessment tasks that only consider production measures (e.g., Leeser, 2008; Uggen, 2012). Furthermore, differences exist across studies in the conditions under which output is operationalized, as groups were often provided with different directions to process input (e.g., Izumi et al., 1999; Izumi & Bigelow, 2000; Russell, 2014) or additional instructional components, such as grammar reviews (e.g., Leeser, 2008), which can conceal output effects. From a conceptual perspective, a review of current evidence also reveals important gaps in our understanding of the potential of output for catalyzing L2 processing and development. With virtually all prior research dedicated to the written medium, it remains unclear whether output leads to the same purported benefits in the oral modality, and the extent to which output-based tasks differentially impact L2 noticing and learning in the oral and written modalities.

TASK MODALITY AND AFFORDANCES FOR L2 GRAMMAR DEVELOPMENT

Several voices have underscored the need to advance research addressing the role of task modality in SLA (e.g., Gilabert et al., 2016; Manchón, 2014). Such research is essential for testing the generalizability of theoretical tenets and, more broadly, for understanding the affordances of oral and written L2 practice as sites for L2 learning. Despite repeated calls for more "modality-sensitive" research agendas, as initially advocated by Harklau (2002), an overreliance on one modality over the other is apparent in many domains. Cognitive SLA research, including research on task-based language teaching (TBLT), has tended to privilege the oral modality over the written modality (see Byrnes & Manchón, 2014). Interestingly, studies on the noticing function of output have primarily centered on the written modality, as participants in several studies were students enrolled in writing courses (e.g., Izumi et al., 1999; Izumi & Bigelow, 2000; Uggen, 2012). Consequently, less is known about how output supports learner-generated noticing and learning

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

outcomes in the oral modality, despite the fact that the noticing function of output serves as a central theoretical underpinning for pedagogic frameworks predominantly researched within the oral medium (e.g., Gilabert et al., 2016; Robinson, 2003).

To understand how modality can impact learning opportunities, it is relevant to consider their differences from a psycholinguistic perspective. From a theoretical standpoint, speaking and writing are posited to follow similar production mechanisms (Cleland & Pickering, 2006). For instance, in Kellogg's (1996) model of writing, production is initiated by formulating the conceptual message and then converting it into verbal structure through translating, which involves operations such as lexical access and grammatical encoding, as in speech production (Kormos, 2006; Levelt, 1989). Subsequently, in the execution stage, the verbal structure is transcribed into text through graphomotoric movements. The final stage entails monitoring and possible revision. Despite the parallelisms between writing and speaking, some important differences exist between oral and written language processing that have implications for the L2 learning opportunities fostered by oral and written tasks (e.g., Gilabert et al., 2016; Manchón, 2014; Ortega, 2012; Williams, 2012).

Core differences between the oral and written modalities include the fact that speaking involves an online pressure component (as conceptualization occurs partly online) and that it proceeds at a considerably faster rate than writing, because the latter requires graphomotoric execution (e.g., Cleland & Pickering, 2006). Oral language is also inherently fleeting, which contrasts with the more permanent visual nature of written language. On the basis of these modality-specific characteristics, oral and written tasks are theorized to have different affordances for learners' engagement in cognitive processes known to be beneficial for L2 development, including noticing (see Gilabert et al., 2016). Specifically, the more self-regulated and visual quality of writing is predicted to facilitate retrieval from long-term memory, promote precision in grammatical encoding and monitoring, and foster the detection of linguistic problems in L2 output to a greater extent than speaking (e.g., Gilabert et al., 2016; Vasylets et al., 2017; Williams, 2012). Furthermore, as argued by Gilabert et al. (2016), in terms of L2 processing, the temporary and nonvisual nature of auditory input can pose "considerable attentional demands as new forms stay available for noticing for just a fraction of a second" (p. 123). Consequently, when learners engage in output-input comparisons in the oral medium, they "may register linguistic inconsistencies only transiently" (p. 127). Although the growing consensus in SLA seems to be that written tasks can increase opportunities for L2 form development relative to oral tasks, arguments exist to challenge this view. For instance, Kellogg (2007) has posited that the graphomotoric execution requirements as well as the activation of graphemic representations that characterizes language processing in the written modality can impose heavier demands on individuals' working memory, and hence may reduce the attentional resources left to engage in effective learner-generated focus on form.

Given these theoretical views, researchers in TBLT have begun to examine how modality impacts L2 production during task-based performance (e.g., Vasylets et al., 2017; Zalbidea, 2017) and how it influences the occurrence of focus-on-form processes associated with L2 learning (e.g., García Mayo & Azkarai, 2016). Yet, as highlighted by Gilabert et al. (2016), empirical research that jointly considers both noticing behaviors and L2 outcomes, which can thereby inform how modality contributes to both SLA processes and products, is lacking. The noticing function of output offers a fruitful conceptual basis on which to ground further modality research and theorization because the contributions of oral and written tasks can be interpreted vis-à-vis output demands. In this regard, the extent to which the noticing function of output operates in the oral modality, where learners are expected to experience higher constraints for deep L2 processing, remains to be corroborated. Comparative research is also warranted to understand how tasks involving output-input cycles in the oral modality analogize to those implemented in the written modality, where learners purportedly have greater resources available for engaging in focus on form (e.g., Gilabert et al., 2016).

THE CURRENT STUDY

Based on the gaps identified in our understanding of the roles of output alongside modality, the following research questions (RQs) and directional hypotheses were posited:

1. To what extent does engaging in oral output affect learner-generated noticing and development of L2 grammar?
2. To what extent does engaging in output and input processing affect learner-generated noticing and development of L2 grammar differentially in the oral versus the written modality?

Given the theoretical postulations previously outlined regarding the potential of L2 production, along with prior research findings, engaging in oral output was hypothesized to result in greater noticing and L2 development of grammar embedded in auditory input than not producing any output. Additionally, based on the affordances of oral and written tasks, it was hypothesized that output-input cycles in the written medium would lead to superior noticing and L2 development compared to the oral medium.

To address these research questions with attention to the methodological issues discussed earlier, output and modality were systematically controlled in this study by implementing tasks in a computerized environment. In addition, noticing behaviors were measured using stimulated recall protocols to assess potential qualitative differences in reported cognitive processes. Both immediate and longer-term learning outcomes were considered, as measured by productive and receptive assessment tasks where transfer of learning to both modalities could be explored. Lastly, to provide a more comprehensive account of the contributions of output and modality, two target structures differing in relative salience were included.

Before turning to describing the research methodology in more detail, it is worth highlighting that the present study did not aim to disentangle the independent effects of input versus output for L2 development in both modalities. Rather, by keeping input a constant, the study intended to examine the extent to which (a) output enhances input processing and increases the potential for L2 grammar development in the oral modality (by comparing a no-output [i.e., auditory input-only] condition to an output [i.e., oral output + auditory input] condition), and (b) engaging in output-input cycles in the written modality provides advantages over doing so in the oral modality (i.e., by comparing a written output + written input condition to an oral output + auditory input condition). Any modality effects are thus interpreted as arising from the combined contributions of output and input processing to SLA.

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

METHOD

TARGET STRUCTURES

Two Spanish grammatical constructions were targeted: the simple future tense and the indirect object clitic, both in the third-person singular and plural forms. The future form comprises an inflectional morpheme affixed to the infinitive verb, as shown in (1), whereas the clitic form consists of a dative pronominal element appearing in preverbal position in double object constructions, as shown in (2):

(1) Él preparará una cena
He prepare-3^{RD}.SING.FUT. a dinner
He will prepare a dinner meal

(2) Ella le da un abrazo
She him/her-3^{RD}.SING.DAT. give a hug
She gives him/her a hug

The rationale for target form selection was that, beyond core morphosyntactic differences, these structures differ in their relative intrinsic salience (i.e., perceptual or other properties inherent to linguistic forms). As this study was concerned with learners' noticing and L2 development, and more salient forms are putatively more likely to be consciously noticed and subsequently learned (e.g., Cintrón-Valentín & Ellis, 2016; Gass et al., 2017; Leow, 2015; VanPatten, 2004), considering more and less salient structures was deemed important for gaining a more nuanced understanding of the potential contributions of output and modality to SLA.

Perceptually, the future tense is argued to be more salient (i.e., easier to hear or see; Goldschneider & DeKeyser, 2001) than the clitic because it is stressed (phonetically and visually, with a written accent), it is more sonorous due to the presence of a low central vowel, which is the highest-ranked phone in Laver's (1994) sonority hierarchy, and it is not prosodically dependent on another element, as the unstressed clitic is to its adjacent verb. Importantly, the future form is also considered more salient from a functional-semantic perspective (i.e., its meaning can be more easily inferred from the input; Bulté & Housen, 2012; DeKeyser, 2005), as it transparently maps onto the absolute meaning of futurity conveyed by temporal adverbs (e.g., *tomorrow*). In contrast, the clitic bears lower semantic weight (e.g., Ortega & Long, 1997) and learners are likely to inaccurately map it onto other competing nominal coreferents in the immediate discourse (e.g., subject, direct object). Considering that, to a great extent, "the transparency of form-meaning relationships to a learner who is processing language for meaning … determines the difficulty of acquisition" (DeKeyser, 2005, p. 3) of a given target, the greater functional opacity of the clitic is deemed a key contributor to making this form a less salient, and thereby more challenging (e.g., Cintrón-Valentín & Ellis, 2016; Housen & Simoens, 2016), form than the future tense. Consistent with this account, prior instructed L2 Spanish research has shown that learners are likely to experience persistent difficulties mastering clitic structures (e.g., Ortega & Long, 1997; VanPatten, 2004), but not future tense morphology (e.g., Leeser, 2007; Russell, 2014).

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

PARTICIPANTS

Beginner-level L2 Spanish learners enrolled in introductory-level communicative language courses at a northeastern U.S. university were randomly assigned to one of three pedagogic conditions (No-output, Speaking, or Writing).[2] Table 1 summarizes language background information in the final sample ($N = 88$; $M_{age} = 19.31$, $SD = 1.90$, 52 female). Most participants reported English as their only native language ($n = 75$). Four participants indicated two native languages (Greek, Romanian, Arabic, and Patois, in addition to English) and nine reported Chinese, Urdu, Amharic, Bahasa Indonesia, Turkish, Russian, Arabic, or Japanese as their native language, along with an early age of exposure to English ($M_{age} = 5.06$, $SD = 1.57$). Kruskal–Wallis tests revealed no significant group differences for any background variable: age ($\chi^2(2) = .96$, $p = .619$), number of foreign languages ($\chi^2(2) = .25$, $p = .884$), years of Spanish education ($\chi^2(2) = 1.33$, $p = .515$), age of exposure to Spanish ($\chi^2(2) = .69$, $p = .709$), or self-rated Spanish proficiency ($\chi^2(2) = .49$, $p = .782$).

As prior L2 knowledge can affect noticing behaviors (e.g., Leow, 2015), efforts were made to ensure that participants had limited working knowledge of the target structures at study onset. Participants were excluded from the statistical analyses if their pretest accuracy for the production task was above 10% (future, $n = 4$; clitic, $n = 3$), above 90% for one judgment task and at or above 80% for the other (future, $n = 10$; clitic, $n = 1$), or if they produced the target structure accurately from the beginning of the treatment task (clitic, $n = 1$). A total of 7, 3, and 9 participants in the No-output, Speaking, and Writing groups, respectively, qualified for exclusion. Thus, beyond limiting prior L2 knowledge in the final sample, these exclusions based on initial accuracy also led to increased group comparability, while still retaining sufficient data for statistical analysis in each condition. Other exclusion criteria included reporting looking up information about the target forms outside of the study (future, $n = 8$), failing to follow task instructions (future, $n = 1$, clitic, $n = 2$), and being exposed to Spanish at or before age 3 ($n = 2$). The final sample comprised 69 and 85 participants for the future and clitic form analyses, respectively (see Appendix S1 in the Online Supplementary Materials for further information on participant exclusions and distributions).

TABLE 1. Participant background information

	No-output	Speaking	Writing
	M (SD)	*M (SD)*	*M (SD)*
n[a]	30	28	30
Age	19.73 (2.64)	19.00 (1.41)	19.17 (1.32)
Number of foreign languages	1.53 (.78)	1.54 (.79)	1.40 (.56)
Years of education in Spanish	3.75 (3.06)	3.02 (3.00)	3.65 (3.32)
Age of exposure to Spanish	14.53 (4.90)	14.86 (4.30)	14.60 (4.64)
Overall self-rated Spanish proficiency[b]	3.72 (1.62)	3.56 (1.45)	4.02 (1.84)

[a] Final sample size for each condition, combining participants with data for the future form only, clitic form only, and both target forms.

[b] Averaged across ratings for four skills on a 10-point Likert scale.

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

PROCEDURE

The study followed a pre-post-delayed posttest design, with approximately 2 weeks between sessions. A summary of the testing protocol is presented in Figure 1 (all experimental materials are publicly accessible at https://www.iris-database.org).

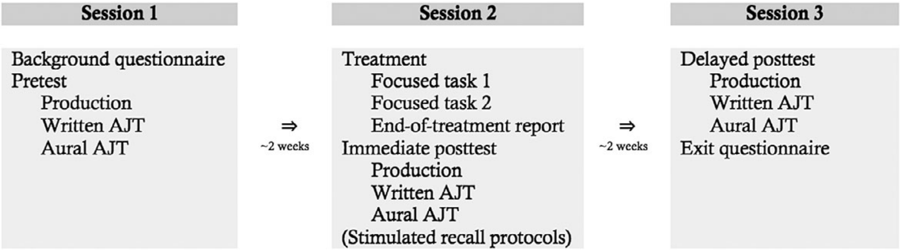
MATERIALS

Pretask

Participants worked individually on a computer. In the pretask stage, participants read a background story in English that provided the necessary contextual information for the subsequent tasks. The story explains that a famous chemist has recently discovered a substance that can provide superpowers to humans (see Figure 2). Following rumors about a planned robbery to steal the substance, the chemist is requesting the participant's help in determining which of two suspects may be behind the rumored robbery. After reviewing instructions and going through practice items, participants completed two focused tasks, acting as detectives while they learned about the suspects' routines.

Treatment

Participants were randomly assigned to one of three pedagogic conditions (see Table 2 for a summary of differences across conditions), where they completed two comparably


Session 1

Background questionnaire
Pretest
Production
Written AJT
Aural AJT

⇒ ~2 weeks

Session 2

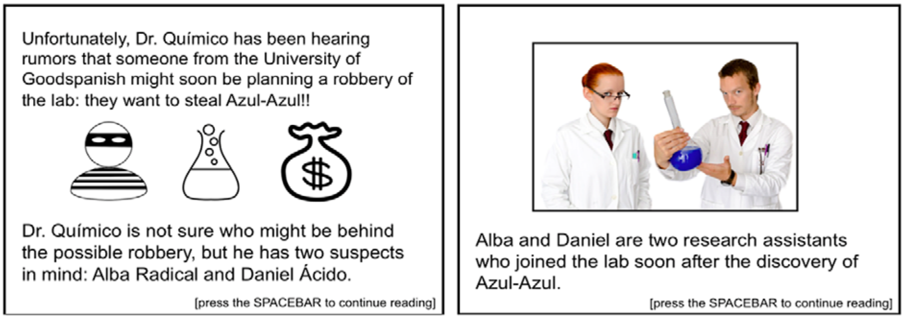
Treatment
Focused task 1
Focused task 2
End-of-treatment report
Immediate posttest
Production
Written AJT
Aural AJT
(Stimulated recall protocols)

⇒ ~2 weeks

Session 3

Delayed posttest
Production
Written AJT
Aural AJT
Exit questionnaire

FIGURE 1. Summary of testing procedure.


Unfortunately, Dr. Químico has been hearing rumors that someone from the University of Goodspanish might soon be planning a robbery of the lab: they want to steal Azul-Azul!!

Dr. Químico is not sure who might be behind the possible robbery, but he has two suspects in mind: Alba Radical and Daniel Ácido.

[press the SPACEBAR to continue reading]

Alba and Daniel are two research assistants who joined the lab soon after the discovery of Azul-Azul.

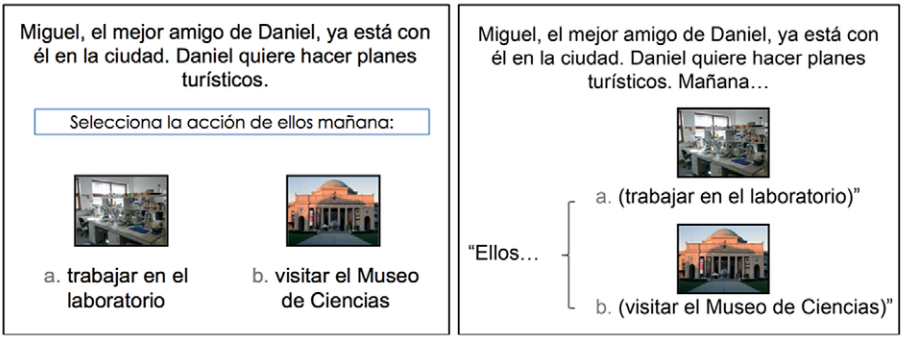
[press the SPACEBAR to continue reading]

FIGURE 2. Sample pretask slides.

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

TABLE 2. Main differences across pedagogic conditions

Group	Output		Input (feedback)	
	Yes/No	Modality	Yes/No	Modality
No-output (Control)	No	–	Yes	Auditory
Speaking (Experimental 1)	Yes	Oral	Yes	Auditory
Writing (Experimental 2)	Yes	Written	Yes	Written



Miguel, el mejor amigo de Daniel, ya está con él en la ciudad. Daniel quiere hacer planes turísticos.

Selecciona la acción de ellos mañana:

a. trabajar en el laboratorio

b. visitar el Museo de Ciencias

Miguel, el mejor amigo de Daniel, ya está con él en la ciudad. Daniel quiere hacer planes turísticos. Mañana…

"Ellos…

a. (trabajar en el laboratorio)"

b. (visitar el Museo de Ciencias)"

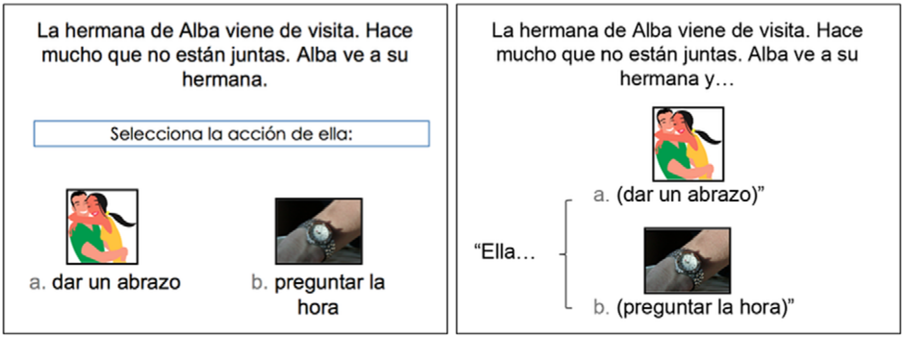
FIGURE 3. Future-focused task: Sample items in the No-output (left) and Speaking and Writing groups (right). English translation: "Miguel, Daniel's best friend, is already in the city with him. Daniel wants to do touristic activities. Select their action tomorrow: (left)/Tomorrow … They … (right) (a) work in the lab (b) visit the Science Museum."

designed focused tasks.[3] Tasks were self-paced and computerized using SuperLab software (Cedrus, San Pedro, CA). The control group completed the task without producing the L2 (No-output group), whereas the experimental groups produced L2 output in oral (Speaking group) or written form (Writing group). One task focused on the future form and introduced the male suspect's routine; the other task focused on the clitic form and introduced the female suspect's routine. Tasks were administered in a counterbalanced order.

For each task item, participants read a brief prompt about the suspect's routine and then chose which of two possible events provided a logical follow-up to the prompt (see Figures 3 and 4). In the No-output control group, participants provided their response by pressing the computer key corresponding to the chosen event. In the experimental groups, participants produced a sentence based on the chosen event by either saying it out loud into a microphone (Speaking group) or typing it into a textbox (Writing group). After each response, all participants received the same feedback in the form of targetlike model utterances containing the target structure (e.g., "Él correrá en el parque," *He will run in the park*; "Ella le da un abrazo," *She gives her a hug*), either auditorily (No-output and Speaking) or in written form (Writing), accompanied by the event picture and key (i.e., *a* or *b*). No deductive information nor explicit rules were provided.

Pedagogic conditions were thus systematically controlled in this study to allow for experimental comparison along the dimensions of interest (i.e., output and modality), such that (a) all groups completed the same conceptual task regarding event selection,

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press



La hermana de Alba viene de visita. Hace mucho que no están juntas. Alba ve a su hermana.

Selecciona la acción de ella:

a. dar un abrazo

b. preguntar la hora

La hermana de Alba viene de visita. Hace mucho que no están juntas. Alba ve a su hermana y…

"Ella…

a. (dar un abrazo)"

b. (preguntar la hora)"

FIGURE 4. Clitic-focused task: Sample items in the No-output (left) and Speaking and Writing groups (right). English translation: "Alba's sister is visiting. They have not seen each other in a long time. Alba sees her sister. Select her action: (left)/Alba sees her sister and … She … (right) (a) give a hug (b) ask for the time."

(b) only the experimental groups engaged in formulation and grammatical encoding through L2 production, and (c) the only difference between experimental groups was modality of output and feedback.

Each task comprised a total of 20 items (16 critical and 4 distractor items; half singular, half plural), with every four critical items followed by a distractor. Verbs and additional vocabulary were obtained from the first three chapters of participants' Spanish textbook. To control for sentential position as an additional dimension of salience, both forms were produced by participants and appeared in the feedback as the second element in the sentence, immediately after the subject.

End Report

After completing both tasks, all participants wrote a brief detective report in English where they indicated which suspect they thought was behind the planned robbery and why. This provided a larger nonlinguistic goal for task completion, which afforded an overarching focus on meaning throughout treatment (e.g., R. Ellis, 2003) (note, however, that end reports are not analyzed as part of this study).

Stimulated Recall Protocols

To tap into learners' noticing behaviors during treatment, a subset of participants ($n = 20$) took part in stimulated recall (SR) protocols at the end of Session 2.[4] SR protocols were implemented immediately after (rather than before) the assessment due to constraints in participant scheduling. During the individual interviews, participants were presented with screen-captured video and audio recordings of their performance to stimulate memory structures. The researcher stopped the recording at preset intervals for the first three to four items in each task and asked participants to discuss their reasons and thoughts regarding (a) their follow-up event choices, (b) their output (experimental groups only), and (c) the computerized feedback. Participants were also invited to share additional thoughts on their performance. Following Gass and Mackey (2017), care was taken to refrain

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

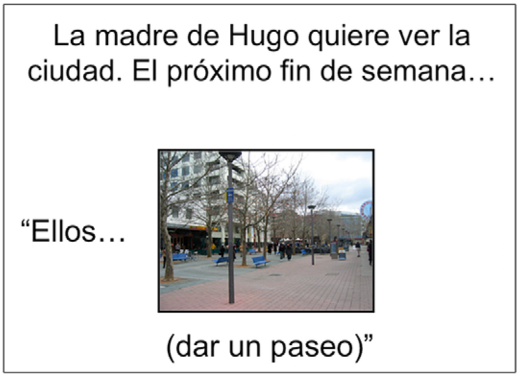
interviewees from discussing past actions in the present tense, so as to reduce the likelihood that participants would introduce intervening thoughts at the time of the interview. Delayed posttest data from learners who participated in SR protocols were excluded from the statistical analyses to avoid any confound effects.

Assessment

Participants' L2 development was assessed using production and written and aural acceptability judgment tasks. This ensured that both modalities were equally represented in the assessment and also allowed the researcher to explore transfer effects in learning (i.e., whether the No-output condition leads to improvements in output-based production tasks, or the Speaking and Writing conditions promote gains in written and aural judgment tasks, respectively).

Production. The production task required participants to read a brief prompt and form a logical follow-up sentence using the information in each slide (see Figure 5). Three versions of the task were created and their administration was counterbalanced across sessions and groups. The task comprised a total of 32 items, with 16 items focused on the future and 16 items focused on the clitic. Each of the 16 items per target form comprised 12 critical and 4 distractor items. The task was mixed-modality, in that half of the items per target form required written production (i.e., typing the sentence into a textbox), and the other half required oral production (i.e., saying the sentence out loud). Participants completed the written portion first. All items prompted learners to produce novel sentences they had not been exposed to during treatment.

Acceptability Judgment. Each acceptability judgment task (AJT) required participants to judge the grammatical acceptability of a series of utterances after reading (written AJT) or listening (aural AJT) to them. AJTs were untimed and computerized. Two versions of each task were designed, with administration counterbalanced across sessions and groups. One version was administered at pretest, the other version at posttest, and the first version



La madre de Hugo quiere ver la ciudad. El próximo fin de semana…

"Ellos…

(dar un paseo)"

FIGURE 5. Sample production task item. English translation: "Hugo's mother wants to see the city. Next weekend … They … (go for a walk)."

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

was administered again at delayed posttest. Each judgment task comprised a total of 24 sentences, with 12 items focused on the future form and 12 items focused on the clitic form. Each of the 12 items comprised 10 critical and 2 distractor sentences. Half the items were acceptable, and half were unacceptable. The AJT stimuli included one new ditransitive verb per target form that did not appear in the focused tasks, with the rest of the sentences being derived from the focused tasks by modifying pronoun numbers and changing indirect object nouns, among other procedures adopted to maximize novelty of stimuli.

CODING, SCORING, AND CONTENT ANALYSIS

Stimulated Recall Protocols

To address RQ1 and RQ2 regarding L2 noticing, stimulated recall data from 15 focal participants ($n = 5$ per condition) were transcribed and inspected for reports of cognitive processes related to the target structures (see note 4). This approach allowed for a more comprehensive assessment of noticing behaviors than a binary coding system (i.e., noticing vs. no noticing). Content analyses revealed three major levels of L2 analysis, each indicative of increasing levels of processing: (a) noticing, (b) searching, and (c) integrating.[5]

A protocol was coded as evidencing *noticing* when participants referred to or claimed having spotted the target form, its grammatical category, or its morphosyntactic context. This was the lowest level of L2 analysis found in the data and was primarily concerned with the perceptual domains of the surface form, as shown in the following sample comments: *"So like the answers were given in the infinitive form and then when it was spoken back in the feedback it was just like 'correct' instead of being infinitive. I remember noticing that after like the first time it happened"* (No-output, Participant 2, Future task); *"I noticed that they put like a* le *in front of the verb"* (Speaking, Participant 3, Clitic task).

Searching was observed when participants reported looking for patterns, comparing multiple exemplars, or observing (in)consistencies in the input. This level of analysis entailed greater cognitive effort than noticing and also dealt with the semantic aspects of the stimuli, as shown in these comments: *"That was kind of when I was like starting to realize that's like how you'd conjugate for the future because it was the same for the first one, except … the first one was singular and this one's plural"* (Speaking, Participant 1, Future task); *"I'd gotten in my head that* le *would go with* él *and* ella*, and* les *will go with* ellos*…. Eventually I figured maybe I had to look further in the sentences"* (Writing, Participant 3, Clitic task).

Lastly, learners' comments evidenced *integrating*, the deepest level of linguistic analysis, when participants hypothesized about the referential meaning of the target structure or reported engaging in semantic association processes. Integrating involved various degrees of conceptual analyses of the target form. Two types of integrating were found*: Successful* integrating occurred when participants' reported form-meaning mappings were accurate, as in this comment: "*I like started to like figure out that for future conjugations like it's* -ará *at the end or like for plural* -arán*"* (Speaking, Participant 1, Future task); *"Instead of 'I ask for help,' it's like 'I ask* you *for help' … Is that the object,* *is it? I don't know. I forget what it's called, but like it's like the action is being done to"* (Writing, Participant 2, Clitic task). *Unsuccessful* integrating was found when the reported form-meaning mappings were not targetlike: *"I was like 'Oh, that's the preterit!' and I was like 'Why is that the preterit if it's like* próximo*?' Then I was like 'Oh, wait, like maybe like the preterit and the future are like really similar in the third person plural'"* (Speaking, Participant 5, Future task); *"I put the* le *in front of the verb this time because I kind of thought that like the object of the sentence was* ayuda*, which is a singular noun"* (Speaking, Participant 3, Clitic task). A second researcher double-coded the verbal recall data (with participant group assignment blinded for coding), which resulted in an intercoder agreement of 93.33%.

Assessment: Production

Production accuracy was coded by awarding one point for each correctly produced critical item, with half a point awarded for production of target (i.e., future or clitic element) and half a point for the morphosyntactic accuracy of the structure. Accent marks were not considered. A second researcher double-coded 10% of the production data (selected from different participants at each testing session), yielding an intercoder agreement of 99.69%.[6]

Assessment: Acceptability Judgment

Participants were awarded one point for each correctly judged critical item. Internal consistency was calculated over the experimental group data in immediate and delayed posttests. Cronbach's alpha was .78 and .69 in the written AJT and .57 and .69 in the aural AJT for the future and clitic forms, respectively. Although lower than the average in L2 research (see Plonsky & Derrick, 2016), these indices are acceptable and not unanticipated in the context of this study, as lower alpha values are often observed with less proficient learner samples (Plonsky & Derrick, 2016) or when multiple participants are guessing (e.g., Indrarathne & Kormos, 2017), as expected when L2 knowledge is limited.

Focused Task Production

To gather additional insights into how the process of L2 grammar learning is influenced by modality, item-by-item L2 production accuracy in the focused tasks was also coded in the Speaking and Writing groups. This provided information on the extent to which modality impacted targetlike form incorporation in learners' output over the course of each treatment task. Participants received one point for each correctly produced critical item targeting the future or the clitic form. A second researcher double-coded 20% of the data, selected using stratified random sampling, which yielded an intercoder agreement of 99.44%.

STATISTICAL ANALYSIS

To address RQ1 and RQ2 regarding L2 grammar development, assessment data on the future and clitic forms were analyzed separately using mixed-effects modeling in R

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

(R Core Team, 2019), following a confirmatory approach. The No-output and Speaking groups and the Speaking and Writing groups were compared to address RQ1 and RQ2, respectively (note that comparisons between the No-output and Writing groups are not reported because they were not a contrast of theoretical interest). For production, linear mixed-effects models were computed with accuracy at pre-, post-, and delayed posttests, with *p*-values calculated using the *lmerTest* package (Kuznetsova et al., 2017). For acceptability judgment, binomial logistic mixed-effects models were computed with accuracy on the written and aural AJTs using the *lme4* package (Bates et al., 2015) with BOBYQA optimization. Additionally, binomial logistic mixed-effects models were computed with the experimental groups' form incorporation accuracy data.

For all assessment tasks, explanatory variables included Group (No-output, Speaking, Writing; centered on Speaking) and Time (pretest, posttest, delayed posttest; centered on posttest), which were entered into the models as fixed effects, as well as their interaction. Predictors for the form incorporation models included Group (Speaking, Writing; centered on Speaking), Item number (centered on 1), and their interaction. Random intercepts were included for participants and test versions in the production models, participants and items in the AJT models, and participants and item number in the form incorporation models. Random slopes were forward-tested with likelihood-ratio tests to arrive at the models of best fit; only random slopes that significantly improved model fit were retained (see Appendix S2 in the Online Supplementary Materials for final model terms).[7] To most clearly interpret model results (i.e., group differences at each time point and performance over time), model baselines were releveled to explore all relevant comparisons, as reported in the following text. This approach retains maximum statistical power and provides a direct test of differences without impacting the goodness-of-fit of the models to the data (e.g., Linck & Cunnings, 2015). Predicted performance is plotted based on estimations from the *effects* package (Fox & Weisberg, 2011).

RESULTS

L2 NOTICING

To address RQ1 and RQ2 regarding noticing behaviors, SR data from five focal participants in each pedagogic condition were considered. These participants were selected for content analysis because they were all native speakers of English who had begun learning Spanish after age 3 (see Appendix S3 in the Online Supplementary Materials for further background information) and none had reported looking up information about either the future or the clitic structures outside of the study. Table 3 summarizes the coding of the protocols in each group.

For both target constructions, more participants in the experimental groups than in the control group reported consciously noticing the target structures (see Appendix S4 and S5 in the Online Supplementary Materials for sample introspective comments by participant). Evidence suggesting that output raised learners' awareness of L2 knowledge gaps was observed in learners' comments, such as the following: *"I guess that also means future, like* próximo*, like 'next weekend' or something, and I was like, 'Oh! I definitely don't know what the future third person plural is!'"* (Speaking, Participant 4, Future task).

TABLE 3. SR protocols: reported levels of L2 analysis by group

			Integrating	
	Noticing	Searching	Unsuccessful	Successful
Future				
No-output	2	1	0	0
Speaking	5	5	1[a]	5
Writing	5	5	0	5
Clitic				
No-output	2	1	0	1
Speaking	5	5	3	2
Writing	5	5	2	3

Note: $n = 5$ focal participants per condition.
[a]For plural morpheme (see Table 14 on Appendix S4 for further information).

Participants in the Speaking and Writing groups also reported engaging in higher-level linguistic analysis of the target forms, namely, in searching and integrating, to a similar degree, as shown in this comment: *“I think* le *means ‘him’ or ‘her’ or… ‘for him’ or ‘for her,’ and then the* les *means ‘them’ or ‘to them,’ ‘for them,’ something like that”* (Writing, Participant 5, Clitic task). In contrast, most participants in the No-output control group reported focusing on the perceptual and content dimensions of the task, without engaging in grammatical analysis of the target structures embedded in the feedback, as can be observed in this comment: *“I guess I was just mostly looking at the picture and not totally listening to the sound as much”* (No-output, Participant 3, Clitic task).

As shown in the table, cases of successful integration of form and meaning were more common for the higher-salience future form than the lower-salience clitic form. Indeed, in the experimental groups, only one case of partially unsuccessful integration was found for the future form, as one participant mapped the plural form to past tense rather than future tense. For the clitic form, several participants reported managing multiple hypotheses about its function, and unsuccessful integration generally resulted from participants’ inaccurate mapping of the clitic to the subject or direct object noun, instead of the indirect object, as seen in this comment: *“The* le *part confused me. ‘Why there needs to be a* le*?’ … If it’s* ellas*, it’s* le*; if it’s* ellos*, it’s* les*. Maybe? … I tried to get what before taught me about the* les … *I mean, I still didn’t completely understand it”* (Writing, Participant 4, Clitic task).

L2 DEVELOPMENT: PRODUCTION

Descriptive statistics for performance on the production task (Table 4) indicate that all groups improved at posttest, with greater gains observed for the experimental groups.

At pretest, no significant differences were found between the No-output and Speaking groups (future: $b = -.01$, $SE = .06$, $p = .94$; clitic: $b = .01$, $SE = .05$, $p = .96$) or between the Speaking and Writing groups (future: $b = -.01$, $SE = .06$, $p = .86$; clitic: $b > -.001$, $SE = .05$, $p = .99$). At posttest, the Speaking group significantly outperformed the No-output group in both future and clitic production (Table 5). However, at delayed posttest, the advantage

TABLE 4. Production task: Descriptive statistics

		Future				Clitic			
		n	*M (SD)*	Min.	Max.	*n*	*M (SD)*	Min.	Max.
No-output	Pre	25	.01 (.02)	.00	.08	30	.01 (.02)	.00	.08
	Post	25	.08 (.21)	.00	.79	30	.06 (.20)	.00	1.00
	Delayed	22	.10 (.23)	.00	.75	25	.06 (.14)	.00	.54
Speaking	Pre	23	.01 (.03)	.00	.08	27	.01 (.02)	.00	.08
	Post	23	.46 (.30)	.00	.96	27	.21 (.29)	.00	.96
	Delayed	16	.27 (.29)	.00	.83	20	.10 (.24)	.00	.79
Writing	Pre	21	.00 (.00)	.00	.00	28	.01 (.02)	.00	.08
	Post	21	.57 (.32)	.00	1.00	28	.32 (.40)	.00	1.00
	Delayed	15	.16 (.31)	.00	.92	20	.14 (.26)	.00	.96

of the Speaking group was retained for the future ($b = –.20$, $SE = .07$, $p < .01$), but not the clitic form ($b = –.06$, $SE = .06$, $p = .35$). Additionally, in terms of modality, the Writing group outperformed the Speaking group in both future and clitic production at immediate posttest (Table 5), although these group differences had dissipated by the delayed posttest (future: $b = –.09$, $SE = .07$, $p = .24$; clitic: $b = .04$, $SE = .06$, $p = .56$).

Examinations of each group's trajectory over time provided additional insights. The No-output group did not evidence significant performance changes for either target structure pre-to-post (future: $b = .07$, $SE = .05$, $p = .16$; clitic: $b = .04$, $SE = .04$, $p = .32$) or pre-to-delayed (future: $b = .06$, $SE = .05$, $p = .24$; clitic: $b = .01$, $SE = .05$, $p = .86$). For the future form, both the Speaking and Writing groups improved pre-to-post ($b = .41$, $SE = .05$, $p < .001$ and $b = .55$, $SE = .05$, $p < .001$, respectively) and pre-to-delayed ($b = .25$, $SE = .06$, $p < .001$ and $b = .18$, $SE = .06$, $p < .01$, respectively). Similarly, for the clitic form, significant pre-to-post improvements were found for both the Speaking ($b = .15$, $SE = .05$, $p < .01$) and the Writing ($b = .29$, $SE = .05$, $p < .001$) groups. Yet, pre-to-delayed gains for this lower-salience form were evidenced in the Writing group ($b = .10$, $SE = .05$, $p = .046$), but not the Speaking group ($b = .07$, $SE = .05$, $p = .20$). Group differences are illustrated in Figure 6.

L2 DEVELOPMENT: WRITTEN ACCEPTABILITY JUDGMENT

Descriptive statistics for performance on the written AJT (Table 6) reveal observable improvements in written acceptability judgment accuracy for all groups.

At pretest, no significant differences were found between the No-output and Speaking groups (future: $b = .02$, $SE = .29$, $p = .96$; clitic: $b = .05$, $SE = .22$, $p = .84$), or between the experimental groups (future: $b = .48$, $SE = .31$, $p = .12$; clitic: $b = .03$, $SE = .22$, $p = .88$), as expected. At posttest, the Speaking group significantly outperformed the No-output group in accurately judging the acceptability of the future tense (Table 7), although these differences had dissipated by the delayed posttest ($b = –.36$, $SE = .33$, $p = .27$). For the clitic form, no differences were found between the Speaking and No-output groups at either immediate (Table 7) or delayed posttest ($b = –.07$, $SE = .25$, $p = .78$). Lastly, no differences were found between the Speaking and Writing groups at either the immediate

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

TABLE 5. Linear mixed-effects modeling on the production task

	Future					Clitic				
Fixed Effects	Estimate	*SE*	*t*	*p*	95% CI	Estimate	*SE*	*t*	*p*	95% CI
Intercept (S, Post)	.42	.04	9.67	<.001	[.34, .50]	.16	.04	3.99	<.001	[.08, .23]
Group (NO)	–.34	.06	–5.70	<.001	[–.46, –.23]	–.11	.05	–2.15	.03	[–.21, –.01]
Group (W)	.13	.06	2.01	.045	[.01, .25]	.14	.05	2.56	.01	[.03, .24]
Time (Pre)	–.41	.05	–7.95	<.001	[–.51, –.31]	–.16	.05	–3.40	<.001	[–.25, –.07]
Time (Del)	–.16	.06	–2.70	<.01	[–.27, –.04]	–.09	.05	–1.77	.08	[–.19, .01]
Group (NO) × Time (Pre)	.34	.07	4.75	<.001	[.20, .47]	.12	.06	1.81	.07	[–.01, .24]
Group (W) × Time (Pre)	–.14	.08	–1.85	.07	[–.28, .01]	–.13	.07	–2.06	.04	[–.26, –.01]
Group (NO) × Time (Del)	.15	.08	1.90	.06	[–.001, .30]	.06	.07	.87	.39	[–.08, .19]
Group (W) × Time (Del)	–.21	.08	–2.55	.01	[–.37, –.05]	–.10	.07	–1.45	.15	[–.24, .04]
Random Effects	Variance	*SD*				Variance	*SD*			
Intercepts ǀ Participant	.01	.11				.01	.10			
Intercepts ǀ Version	–	–				<.01	.02			
Residual	.03	.17				.03	.17			

Note: Group: NO, No-output; S, Speaking; W, Writing. Time: Pre, pretest; Post, posttest; Del, delayed posttest.

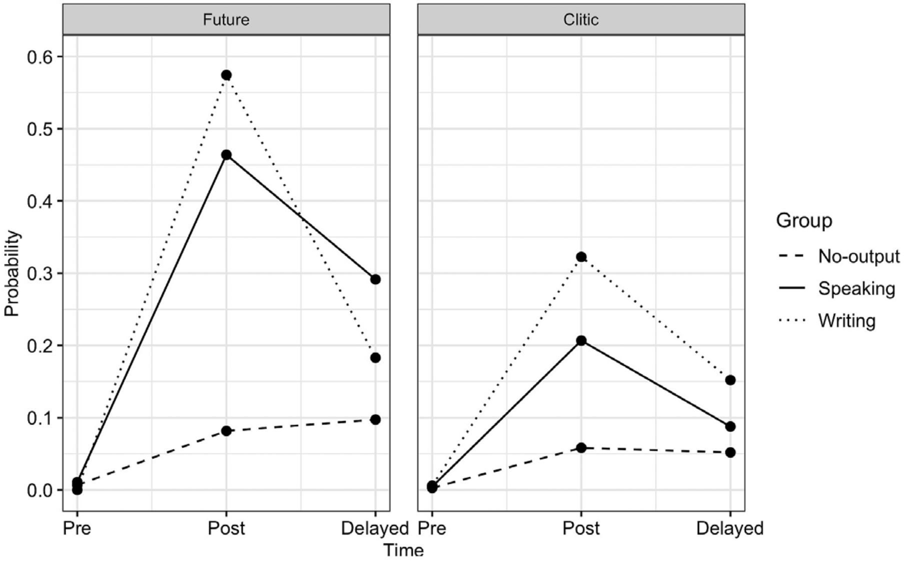


FIGURE 6. Predicted probabilities of production accuracy by condition.

TABLE 6. Written AJT: Descriptive statistics

		Future				Clitic			
Group	Time	*n*	*M (SD)*	Min.	Max.	*n*	*M (SD)*	Min.	Max.
No-output	Pre	25	.55 (.20)	.20	1.00	30	.52 (.15)	.20	.80
	Post	25	.60 (.27)	.10	1.00	30	.54 (.18)	.30	1.00
	Delayed	22	.65 (.23)	.30	1.00	25	.56 (.17)	.30	.90
Speaking	Pre	23	.55 (.13)	.30	.80	27	.51 (.14)	.20	.80
	Post	23	.74 (.22)	.30	1.00	27	.59 (.21)	.20	.90
	Delayed	16	.70 (.19)	.40	1.00	20	.57 (.17)	.40	.90
Writing	Pre	21	.64 (.18)	.40	1.00	28	.51 (.16)	.20	.90
	Post	21	.77 (.21)	.50	1.00	28	.66 (.20)	.40	1.00
	Delayed	15	.69 (.19)	.40	1.00	20	.62 (.20)	.30	1.00

(Table 7) or the delayed posttest for the future ($b = .05$, $SE = .36$, $p = .89$) or the clitic form ($b = .23$, $SE = .26$, $p = .38$).

Analyses of each group's trajectory revealed further insights. For the future form, the No-output group showed no significant pre-to-post changes ($b = .23$, $SE = .19$, $p = .24$), but it did pre-to-delayed ($b = .46$, $SE = .20$, $p = .03$). Both the Speaking and Writing groups improved pre-to-posttest ($b = .94$, $SE = .21$, $p < .001$, and $b = .72$, $SE = .23$, $p < .01$, respectively), although pre-to-delayed changes were significant in the Speaking group ($b = .83$, $SE = .24$, $p < .001$), but not the Writing group ($b = .40$, $SE = .25$, $p = .11$). Different achievement patterns were found for the clitic form. In this case, the No-output group did not significantly improve pre-to-post ($b = .14$, $SE = .18$, $p = .45$) or pre-to-delayed ($b = .24$, $SE = .19$, $p = .22$). Conversely, pre-to-post gains were significant in both the

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

TABLE 7. Logistic mixed-effects modeling on the written AJT

	Future					Clitic				
Fixed Effects	Estimate	*SE*	Odds	*p*	95% CI	Estimate	*SE*	Odds	*p*	95% CI
Intercept (S, Post)	1.18	.24	3.26	<.001	[.71, 1.67]	.46	.27	1.59	.08	[–.07, 1.00]
Group (NO)	–.70	.30	.50	.02	[–1.30, –.10]	–.24	.22	.78	.27	[–.68, .19]
Group (W)	.26	.33	1.30	.43	[–.39, .92]	.34	.23	1.41	.14	[–.11, .80]
Time (Pre)	–.94	.21	.39	<.001	[–1.36, –.53]	–.43	.19	.65	.03	[–.81, –.04]
Time (Del)	–.11	.25	.90	.65	[–.59, .38]	–.07	.21	.93	.73	[–.50, .35]
Group (NO) × Time (Pre)	.71	.29	2.04	.01	[.15, 1.28]	.29	.27	1.33	.28	[–.24, .81]
Group (W) × Time (Pre)	.22	.32	1.24	.49	[–.41, .84]	–.31	.27	.73	.26	[–.85, .23]
Group (NO) × Time (Del)	.34	.32	1.41	.29	[–.29, .97]	.17	.29	1.19	.55	[–.40, .75]
Group (W) × Time (Del)	–.21	.36	.81	.56	[–.92, .50]	–.11	.31	.90	.72	[–.71, .49]
Random Effects	Variance	*SD*				Variance	*SD*			
Intercepts | Participant	.56	.75				.19	.44			
Intercepts | Item	.20	.45				.90	.95			

Note: Group: NO, No-output; S, Speaking; W, Writing. Time: Pre, pretest; Post, posttest; Del, delayed posttest.

TABLE 8. Aural AJT: Descriptive statistics

		Future				Clitic			
Group	Time	*n*	*M (SD)*	Min.	Max.	*n*	*M (SD)*	Min.	Max.
No-output	Pre	25	.55 (.15)	.20	.80	30	.52 (.15)	.30	.80
	Post	25	.55 (.16)	.20	.80	30	.58 (.14)	.30	.89
	Delayed	22	.61 (.18)	.30	.90	25	.57 (.17)	.30	1.00
Speaking	Pre	23	.54 (.12)	.30	.80	27	.53 (.10)	.40	.70
	Post	23	.69 (.18)	.40	1.00	27	.65 (.16)	.40	.90
	Delayed	16	.65 (.16)	.40	.90	20	.54 (.13)	.30	.70
Writing	Pre	21	.55 (.19)	.20	.80	28	.53 (.13)	.30	.80
	Post	21	.69 (.22)	.30	1.00	28	.65 (.21)	.30	1.00
	Delayed	15	.60 (.17)	.30	.90	20	.65 (.21)	.40	1.00

Speaking ($b = .43$, $SE = .19$, $p = .03$) and Writing groups ($b = .73$, $SE = .19$, $p < .001$). However, while the Writing group showed significant pre-to-delayed changes for this lower-salience form ($b = .55$, $SE = .21$, $p < .01$), the Speaking group did not ($b = .35$, $SE = .21$, $p = .096$), akin to what was observed in the production task.

L2 DEVELOPMENT: AURAL ACCEPTABILITY JUDGMENT

Descriptive statistics for performance on the aural AJT (Table 8) indicate improvements in aural acceptability judgment accuracy across all groups, with greater accuracy among experimental groups.

At pretest, no significant differences were found between the No-output and Speaking groups (future: $b = .08$, $SE = .22$, $p = .74$; clitic: $b = –.07$, $SE = .21$, $p = .74$) or between the Speaking and Writing groups (future: $b = .07$, $SE = .23$, $p = .76$; clitic: $b = –.01$, $SE = .22$, $p = .97$). At posttest, the Speaking group significantly outperformed the No-output group in accurately judging the acceptability of utterances with the future tense (Table 9), although these differences had dissipated by the delayed posttest ($b = –.24$, $SE = .26$, $p = .35$), as in the written AJT. For the indirect object clitic, no significant differences were found between the Speaking and No-output groups at posttest (Table 9) or delayed posttest ($b = .10$, $SE = .24$, $p = .69$). Furthermore, the Speaking and Writing groups were not significantly different at posttest for the future or the clitic (Table 9). However, at delayed posttest, although no differences were found for the future ($b = –.19$, $SE = .28$, $p = .49$), the Writing group significantly outperformed the Speaking group in judging auditory utterances featuring the clitic ($b = .55$, $SE = .25$, $p = .03$).

Additional differences were observed in each group's performance over time. For the future tense, the No-output group did not evidence significant pre-to-post ($b = –.01$, $SE = .19$, $p = .98$) nor pre-to-delayed changes ($b = .27$, $SE = .20$, $p = .18$). In contrast, the Speaking group significantly improved pre-to-post ($b = .74$, $SE = .21$, $p < .001$), as did the Writing group ($b = .66$, $SE = .22$, $p < .01$). Significant pre-to-delayed changes were found in the Speaking group ($b = .59$, $SE = .23$, $p = .01$), but not the Writing group ($b = .32$, $SE = .23$, $p = .17$). For the clitic form, the No-output group did not show any pre-to-post ($b = .32$, $SE = .19$, $p = .09$) or pre-to-delayed changes ($b = .26$, $SE = .19$, $p = .18$). Conversely, significant pre-to-post gains were found in the Speaking group ($b = .60$, $SE = .20$, $p < .01$)

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

TABLE 9. Logistic mixed-effects modeling on the aural AJT

	Future					Clitic				
Fixed Effects	Estimate	*SE*	Odds	*p*	95% CI	Estimate	*SE*	Odds	*p*	95 % CI
Intercept (S, Post)	.89	.22	2.43	<.001	[.45, 1.33]	.76	.28	2.15	<.01	[.20, 1.33]
Group (NO)	–.67	.23	.51	<.01	[–1.13, –.22]	–.35	.22	.70	.10	[–.78, .07]
Group (W)	–.01	.25	.99	.97	[–.50, .48]	.06	.22	1.06	.78	[–.38, .50]
Time (Pre)	–.74	.21	.48	<.001	[–1.15, –.34]	–.60	.20	.55	<.01	[–.99, –.21]
Time (Del)	–.15	.23	.86	.52	[–.61, .31]	–.51	.22	.60	.02	[–.94, –.08]
Group (NO) × Time (Pre)	.74	.28	2.10	<.01	[.19, 1.30]	.28	.27	1.33	.30	[–.25, .82]
Group (W) × Time (Pre)	.08	.30	1.08	.79	[–.51, .67]	–.07	.28	.93	.81	[–.62, .48]
Group (NO) × Time (Del)	.43	.31	1.53	.17	[–.18, 1.03]	.45	.29	1.57	.13	[–.13, 1.03]
Group (W) × Time (Del)	–.19	.34	.83	.58	[–.85, .48]	.48	.31	1.62	.12	[–.12, 1.10]
Random Effects	Variance	*SD*				Variance	*SD*			
Intercepts ǀ Participant	.15	.38				.13	.36			
Intercepts ǀ Item	.39	.63				1.07	1.03			

Note: Group: NO, No-output; S, Speaking; W, Writing. Time: Pre, pretest; Post, posttest; Del, delayed posttest.

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

TABLE 10. Summary of findings: Evidence of immediate and sustained L2 development by pedagogic condition

	Immediate (pre-post)			Sustained (pre-delayed)		
	No-output	Speaking	Writing	No-output	Speaking	Writing
Future (higher salience)	–	✓	✓	✓	✓	✓
Clitic (lower salience)	–	✓	✓	–	–	✓

Note: A checkmark indicates significant differences from pretest in at least one L2 outcome measure.

and the Writing group ($b = .67$, $SE = .20$, $p < .001$). As in the written AJT, the Writing group significantly improved pre-to-delayed ($b = .65$, $SE = .22$, $p < .01$), but the Speaking group did not ($b = .10$, $SE = .21$, $p = .65$). An overview of the main findings concerning evidence of L2 development in each group is presented in Table 10. While this table broadly summarizes learning effects within each pedagogic condition, it is important to keep in mind that differences in group trajectories—particularly between experimental conditions—were not large enough to drive interactions for every outcome measure, as further discussed in the following text.

FORM INCORPORATION: FOCUSED TASK PERFORMANCE

For the future form, targetlike form incorporation rates were similar in the Speaking ($M = .47$, $SD = .29$, Min = .00, Max = .87) and Writing ($M = .52$, $SD = .24$, Min = .00, Max = .94) groups. Only a main effect for Item Number was observed (Table 11), which indicates that both groups progressed at comparable rates over the course of the task. However, for the clitic form, the Writing group ($M = .41$, $SD = .25$, Min = .00, Max = .88) evidenced more targetlike form incorporation in their output than the Speaking group ($M = .25$, $SD = .20$, Min = .00, Max = .81). In this case, a Group × Item Number interaction was found (Table 10), which corresponds to a marginal effect of Item Number for the Speaking group (i.e., this group improved just marginally over the course of the task) and a significant effect of Item Number for the Writing group (i.e., the Writing group improved significantly as the task progressed). These group differences are illustrated in Figure 7.

DISCUSSION

RESEARCH QUESTION 1: OUTPUT DEMANDS, SALIENCE, AND L2 GRAMMAR DEVELOPMENT

The first research question asked whether engaging in oral output impacts learner-generated noticing and development of L2 grammar. Based on the theoretical postulations on the role of production, it was hypothesized that the Speaking group would evidence greater noticing and L2 development than the No-output group. As predicted, participants who produced output engaged in deeper levels of L2 analysis and experienced significant and more sustained L2 learning outcomes following task-based practice. The advantage

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

TABLE 11. Logistic mixed-effects modeling on targetlike form incorporation in the focused tasks

	Future					Clitic				
Fixed Effects	Estimate	*SE*	Odds	*p*	95% CI	Estimate	*SE*	Odds	*p*	95% CI
Intercept (S, Item 1)	–1.82	.35	.16	<.001	[–2.64, –1.18]	–1.68	.32	.19	<.001	[–2.35, –1.06]
Group (W)	–.07	.49	.93	.88	[–1.05, .96]	.20	.44	1.22	.65	[–.68, 1.08]
Item Number	.19	.03	1.20	<.001	[.12, .26]	.04	.02	1.04	.07	[–.003, .08]
Group (W) × Item Number	.06	.05	1.06	.23	[–.04, .17]	.07	.03	1.08	.01	[.02, .13]
Random Effects	Variance	*SD*	Correlation			Variance	*SD*			
Intercepts | Participant	.89	.94				1.21	1.10			
Item Number | Participant	.01	.09	.77			–	–			

Note: Group: S, Speaking; W, Writing.

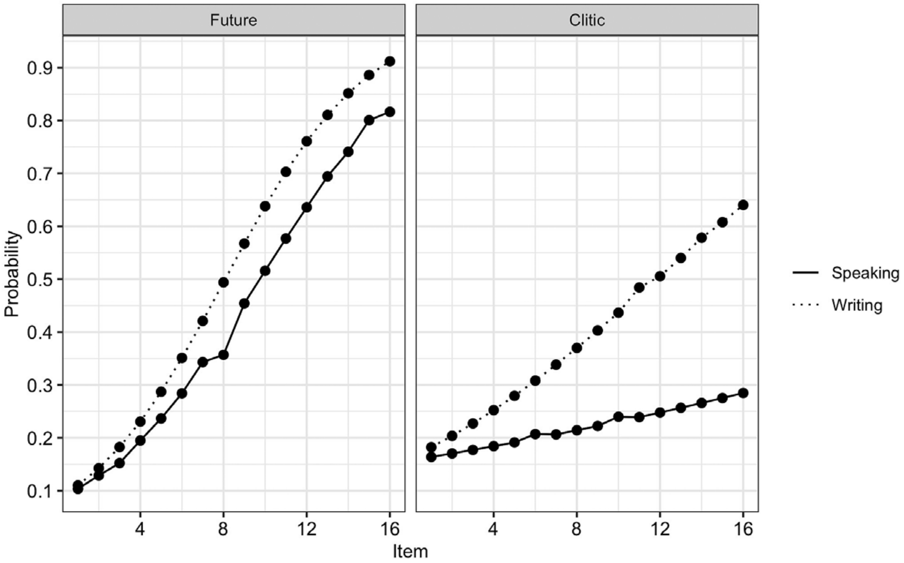


FIGURE 7. Predicted probabilities of targetlike form incorporation by condition.

of the Speaking group over the No-output group was most apparent for gains on the more salient structure, namely, the future form.

These findings do not align with Izumi and Izumi (2004), the only prior study investigating the applicability of the noticing function of output in the oral modality, where a detrimental role for output was found in learners' noticing behaviors and L2 improvement. The sizeable facilitatory effects for output observed in the present study point at the distinct operationalization of output between studies, rather than the increased processing demands imposed by the oral modality of the output and input components, as the most plausible account for this divergence. Indeed, as ventured by Izumi and Izumi, participants in the experimental group were not required to rely on their own linguistic resources for L2 production. As a result, it is likely that the type of output that they engaged in differed widely from the current study, in that it did not sufficiently activate the formulation processes purported to underlie the noticing function of output (e.g., Izumi, 2003; Swain, 2005).

Results from the present study critically extend the benefits of written output-input cycles identified in previous research (e.g., Izumi, 2002; Russell, 2014; Uggen, 2012) to the oral modality. Using SR protocols to gauge noticing, Uggen (2012) found that the properties of the target form constrained the role of output, such that effects were observable only for the more complex target structure, which was argued to be more salient to learners. In contrast, in the current study, engaging in output procedures led to deeper processing and L2 development of both more and less challenging structures differing in relative salience, which further attests to the potential of output for supporting robust learner-generated focus-on-form processes in the oral medium. Methodological differences aside, results from the current research are in keeping with Uggen's observation that greater target structure salience is associated with increased learner noticing

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

and L2 accuracy. Specifically, although immediate output effects were found here for both the future and clitic forms, the longer-term facilitatory effects of oral output in terms of enduring L2 gains in production and acceptability judgment were evidenced for the more salient future form only. Further discussion on the relevance of target form salience alongside modality is provided in the next section.

Findings for the first research question are consistent with and extend Swain's (e.g., 2005) theoretical proposal that output promotes psycholinguistic processes beneficial for SLA, particularly when accompanied by relevant input (see also Izumi, 2003; Leow, 2015). Together, qualitative and quantitative results suggest that, as expected, the requirement to produce L2 speech directed participants' attention to the status of their morphosyntactic knowledge concerning the future and clitic forms, which in turn promoted an internal impetus to focally attend to pertinent input (i.e., target structure models) in the auditory feedback. Such input, as suggested by learners' verbal reports, appeared to be processed more deeply and with higher cognitive effort in the Speaking group relative to the No-output group. The greater level of linguistic analysis promoted by output arguably facilitated further processing of intake and, ultimately, incorporation of newly developed L2 grammar knowledge, as indicated by the Speaking participants' improved L2 performance across tasks.

In sum, analyses of verbal reports and L2 outcome measures provide unequivocal evidence to support that the noticing function of output is operative in the oral modality, where learners are expected to experience heightened temporal demands for L2 output and input processing. More specifically, findings suggest that this beneficial function of output can extend in scope to, beyond noticing (e.g., Swain, 1995, 2000, 2005), promote deeper, more integrative processing of form and meaning in relevant L2 input (e.g., Izumi, 2003; Leow, 2015), giving way to robust L2 outcomes that are observable in both auditory and written measures. These findings lend credence to a central role of output in SLA that had been largely assumed, but not experimentally corroborated to date.

RESEARCH QUESTION 2: MODALITY EFFECTS, SALIENCE, AND L2 GRAMMAR DEVELOPMENT

The second research question asked whether output and input processing in the written modality, compared to the oral modality, would differentially impact noticing and development of L2 grammar. Results provide some support for the hypothesis that the Writing group would experience greater noticing and linguistic development than the Speaking group. Participants in both modality groups reported engaging in deep levels of L2 analysis to a similar degree and experienced robust L2 development of the higher-salience future form. However, for the lower-salience clitic form, only the Writing group appeared to evidence longer-term linguistic gains (although group differences were limited to immediate production and sustained aural acceptability judgment accuracy). Writing also led to more targetlike incorporation of this less salient form throughout the focused task.

A plausible explanation for the broad similarities observed between the Speaking and Writing conditions, which contrast with the more extensive differences described earlier between the No-output and Speaking conditions, lies at the common denominator between conditions, namely, output. Given that both experimental groups met the

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

production requirements for the noticing function of output to unfold (e.g., both engaged in grammatical encoding processes under the same conceptual demands, see Izumi, 2003; Swain, 2005), it stands to reason that focused attention and further processing of relevant input, as well as subsequent linguistic development, were facilitated to a considerable degree in both modalities. Although these findings point to the cross-modal applicability of output-input cycles for enhancing learner-generated L2 grammar noticing behaviors and outcomes in the context of focused tasks, some modality effects were apparent, particularly on the lower-salience form, which deserve due consideration.

The finding that the Writing condition evidenced greater immediate production as well as sustained aural acceptability judgment accuracy on the clitic form indicates that output-based focused tasks in the written modality provided some superior affordances for the development of the less salient, more challenging target construction of the study relative to the oral modality. Analysis of participants' form incorporation during the focused task offers additional insights in this regard, revealing critical modality effects in the process—not just the outcome—of L2 development. Indeed, participants in the Writing group showed superior rates for targetlike incorporation of the clitic, an advantage that increased as the task progressed, which suggests that learners were able to establish more accurate form-meaning connections in this modality by the end of treatment (e.g., N. Ellis, 2004; VanPatten, 2004). These incorporation differences are consonant with the stability of both productive and receptive gains for the clitic form found also in the Writing condition, as form incorporation has been associated with more elaborate input processing as well as syntacticization and long-term development of L2 forms (e.g., Robinson, 2003; Schmidt, 2001). Despite these benefits observed for Writing over Speaking with regard to the lower-salience clitic form, it is worth noting that modality effects did not systematically arise across all the assessment tasks of the study, as was the case for output effects in RQ1.

Concerning learners' noticing behaviors, analyses of participants' recall comments did not reveal pronounced differences between the Writing and Speaking groups in terms of reported cognitive processes engaged during treatment. This, once again, appears consistent with the notion that the noticing function of output is operative in both modalities. With regard to the lower-salience clitic form, it is also possible that both Writing and Speaking allowed learners to reach deep levels of L2 analysis, but that modality differences existed in the time course of L2 processing. For instance, based on the form incorporation data, it may be reasonable to assume that form-meaning integration processes were elaborated to a greater extent over the course of the task in the Writing condition. Regardless, all inferences made from the recall data should be interpreted with caution given the limited number of interviewed participants and the timing of SR protocol implementation in this study, as further discussed in the following text.

From a theoretical standpoint, the advantages observed in the Writing group can be explained with reference to the idiosyncrasies of each modality as sites for language learning. Specifically, findings support the account that the slower pace of writing and the nontransient visual quality of feedback in the written condition maximized learners' opportunities for monitoring and promoted enhanced rehearsal of relevant exemplars in memory following output (e.g., Gilabert et al., 2016; Williams, 2012). This, in turn, may have allowed learners to draw comparisons and test hypotheses about the clitic form more rapidly and effectively within the same output and input parameters. Arguably, these affordances of the written modality for engaging in learner-generated focus-on-form

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

processes become most relevant when the target structure is less salient, particularly when its function cannot be easily inferred from the input, as is the case with the clitic. Indeed, as observed in the recall comments, participants experienced challenges linking the indirect object clitic with its target coreferent in both modalities, in large part because there were alternative candidates, such as the subject and the direct object, that could potentially serve this role (e.g., DeKeyser, 2005; N. Ellis, 2017). Access to multiple stable, visual cues in the written input, which participants in the Writing group could focally attend to and jointly compare at their own pace and discretion, appears to have allowed them to reach targetlike associations sooner relative to the Speaking group, and to rehearse and strengthen those associations over the course of the task, plausibly promoting knowledge consolidation (e.g., N. Ellis, 2004; Leow, 2015).

Following this account, study results contribute to the growing body of evidence motivating the integration of task modality into current taxonomies of task characteristics, as advocated by several voices (e.g., Vasylets et al., 2017; Zalbidea, 2017), given its potential to influence not only focus-on-form processes (e.g., greater targetlike form incorporation) but also longer-term L2 development. More generally, findings also call attention to the utility of employing a combined process- and product-oriented perspective in L2 research seeking to unearth the role of modality, and point to the relevance of investigating and conceptualizing modality as a task feature that may hinder or boost the L2 learning opportunities that are already afforded by the cognitive demands of the task (e.g., output and input processing requirements) (e.g., Gilabert et al., 2016).

LIMITATIONS AND CONCLUSION

Before discussing the broader implications of this research, some limitations should be acknowledged. First, although SR protocols were deemed suitable for this study given their compatibility with different modalities, introspective data should be carefully considered given the issues of veridicality and reliability due to potential memory decay from the time elapsed since treatment performance (e.g., Leow, 2015). In the present study, SR protocols were implemented immediately after the posttest assessment because of constraints in participant scheduling. Although none of the participants reported difficulties recalling their treatment performance, nor referred to the assessment tasks during the interviews, a shorter delay would have certainly been desirable (see Gass & Mackey, 2017). Another aspect to consider is that only a subset of the sample could take part in the SR interviews due to logistic limitations, and hence the extent to which the observed behaviors are representative of the whole group cannot be ascertained. Thus, while the amount of introspective data considered in this study is in keeping with other studies in instructed L2 research (e.g., Cerezo et al., 2016), further investigation is needed to strengthen our understanding of output and modality effects on learners' cognitive processes. Additionally, although it was important to consider both more and less challenging structures in the present study, this also required restricting the number of exemplars per form included in the materials to minimize the risk of participant fatigue from prolonged testing. The operationalization of form salience adopted here is also limited to the extent that forms varied along multiple relative dimensions that may have differentially contributed to the degree of learning difficulty experienced by beginner-level participants. More research on the contributions of target structure salience that also considers L2 learners' construal of different salience dimensions is warranted (e.g., Gass et al., 2017).

Despite these limitations, this study has provided novel empirical support for a cornerstone function of output which, to date, had been largely assumed but remained uncorroborated. In addition to expanding the theoretical scope of output as a central construct in L2 development, findings from this research also contribute to advancing more modality-sensitive perspectives of SLA (e.g., Byrnes & Manchón, 2014; Harklau, 2002). The different facilitative effects attested for both oral and written tasks also hold relevance for L2 pedagogy, and add to the growing body of research problematizing the distinctive opportunities that each modality builds for L2 learning. Future studies may examine the contributions of output and modality for L2 development among various learner populations, such as those differing in age and familiarity with the L2. For instance, younger L2 learners with lower command of reading and writing in the target language may benefit differently from the affordances of oral and written tasks.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0272263120000261.

NOTES

[1] Other research, although not specifically focused on the "noticing function of output," has explored the value of output for supporting SLA processes within the oral modality (e.g., Philp & Iwashita, 2013).

[2] A total of 75 participants were enrolled in first-semester Spanish courses and 13 (No-output: $n = 4$; Speaking: $n = 5$; Writing: $n = 4$) were enrolled in second-semester courses. Data were collected over two semesters.

[3] Focused tasks elicit the use of a particular target form (e.g., future form) to accomplish a nonlinguistic goal requiring a focus on meaning (e.g., selecting the most plausible event) (see R. Ellis, 2003).

[4] Although 15 participants ($n = 5$ per group) were initially scheduled for the SR interviews, the total number of interviewed participants reached 20 (5, 6, and 9 in the No-output, Speaking, and Writing groups, respectively), as some participants in the initial sample were excluded. One participant in the Speaking group was a nonnative English speaker, and three participants in the Writing group were exposed to Spanish before age 3 ($n = 1$) or had reported looking up the future form outside of the study ($n = 2$). Hence, their data were not included for consistency and comparability reasons. To keep the number of focal participants equal across groups (i.e., $n = 5$), as initially planned, SR data from the last participant who completed the interview protocol in the Writing condition were not considered. Two SR participants in the No-output condition scored above the 90%–80% pretest AJT accuracy criterion for the future form (see Appendix S4 in the Online Supplementary Materials); while these participants were excluded from the statistical analyses, they were retained for the SR analyses for the same consistency reason and because further data collection was no longer viable.

[5] Given the nonconcurrency of learners' SR verbalizations in this study, the cognitive processes identified here share similarities but do not fully overlap with those found in previous research (see Appendix S6 in the Online Supplementary Materials for further discussion).

[6] Files from nine different consecutive participant numbers were double-coded in the pretest, posttest, and delayed posttest production datasets to increase representativeness (this procedure resulted in 44%, 19%, 37% of the data belonging to the No-output, Speaking, and Writing groups, respectively, following random group assignment).

[7] Task version and item number were excluded from the random effects structure of the future tense production model and the clitic treatment model, respectively, because they explained no significant variance. Whether or not these random effects were included in the models had no bearing on the significance or

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

interpretation of results. Additionally, the order of focused task administration was excluded from the random effects structure of all models after testing models with participant nested within administration order and observing that it explained no variance for any subset of the data. Production data were initially analyzed applying logistic modeling on trial-level data; however, this led to convergence problems on account of participants' ~0% accuracy at pretest.

REFERENCES

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software*, *67*, 1–48.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy, and fluency in SLA* (pp. 21–46). John Benjamins.

Byrnes, H., & Manchón, R. M. (2014). Task-based language learning: Insights from and for L2 writing—An introduction. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning—Insights from and for L2 writing* (pp. 1–23). John Benjamins.

Cerezo, L., Caras, A., & Leow, R. P. (2016). The effectiveness of guided induction versus deductive instruction on the development of complex Spanish gustar structures: An analysis of learning outcomes and processes. *Studies in Second Language Acquisition*, *38*, 265–291.

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.

Cleland, A. A., & Pickering, M. J. (2006). Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language*, *54*, 185–198.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning*, *55*, 1–25.

Ellis, N. C. (2004). The processes of second language acquisition. In B. VanPatten, J. Williams, S. Rott, & M. Overstreet (Eds.), *Form-meaning connections in second language acquisition* (pp. 49–76). Erlbaum.

Ellis, N. C. (2017). Salience. In M. Hundt, S. Mollin, & S. Pfenninger (Eds.), *The changing English language: Psycholinguistic perspectives* (pp. 71–92). Cambridge University Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Sage.

García Mayo, M. P. , & Azkarai, A. (2016). EFL task-based interaction: Does task modality impact on language-related episodes? In M. Sato & S. Ballinger (Eds.), *Peer interaction and second language learning: Pedagogical potential and research agenda* (pp. 241–266). John Benjamins.

Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research* (2nd ed.). Routledge.

Gass, S. M., Spinner, P., & Behney, J. (2017). *Salience in second language acquisition*. Routledge.

Gilabert, R., Manchón, R., & Vasylets, L. (2016). Mode in theoretical and empirical TBLT research: Advancing research agendas. *Annual Review of Applied Linguistics*, *36*, 117–135.

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

Harklau, L. (2002). The role of writing in classroom second language acquisition. *Journal of Second Language Writing*, *11*, 329–350.

Housen, A., & Simoens, H. (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition*, *38*, 163–175.

Indrarathne, B., & Kormos, J. (2017). Attentional processing of input in explicit and implicit conditions: An eye-tracking study. *Studies in Second Language Acquisition*, *39*, 401–430.

Izumi, S. (2002). Output, input enhancement, and the noticing hypothesis. *Studies in Second Language Acquisition*, *24*, 541–577.

Izumi, S. (2003). Comprehension and production processes in second language learning: In search of the psycholinguistic rationale of the output hypothesis. *Applied Linguistics*, *24*, 168–196.

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

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

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

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

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 57–71). Lawrence Erlbaum Associates.

Kellogg, R. T. (2007). Are written and spoken recall of text equivalent? *American Journal of Psychology*, *120*, 415–428.

Kormos, J. (2006). *Speech production and second language acquisition*. Erlbaum.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). *lmerTest* package: Tests in linear mixed-effects models. *Journal of Statistical Software*, *82*, 1–26.

Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.

Leeser, M. J. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning*, *57*, 229–270.

Leeser, M. J. (2008). Pushed output, noticing, and development of past tense morphology in content-based instruction. *Canadian Modern Language Review*, *65*, 195–220.

Leow, R. P. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. Routledge.

Levelt, W. (1989). *Speaking: From intention to articulation*. MIT Press.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207.

Manchón, R. M. (2014). The internal dimension of tasks: The interaction between task factors and learner factors in bringing about learning through writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning—Insights from and for L2 writing* (pp. 27–53). John Benjamins.

Ortega, L. (2012). Epilogue: Exploring L2 writing–SLA interfaces. *Journal of Second Language Writing*, *21*, 404–415.

Ortega, L., & Long, M. H. (1997). The effects of models and recasts on the acquisition of object topicalization and adverb placement in L2 Spanish. *Spanish Applied Linguistics*, *1*, 65–86.

Park, E. S. (2013). Learner-generated noticing behavior by novice learners: Tracing the effects of learners' L1 on their emerging L2. *Applied Linguistics*, *34*, 74–98.

Philp, J., & Iwashita, N. (2013). Talking, tuning in and noticing: Exploring the benefits of output in task-based peer interaction. *Language Awareness*, *22*, 353–370.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, *100*, 538–553.

R Core Team. (2019). *R: A language and environment for statistical computing* (version 3.3.1.). R Foundation for Statistical Computing.

Robinson, P. (2003). The cognition hypothesis, task design, and adult task-based language learning. *Second Language Studies*, *21*, 45–105.

Russell, V. (2014). A closer look at the output hypothesis: The effect of pushed output on noticing and inductive learning of the Spanish future tense. *Foreign Language Annals*, *47*, 25–47.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press.

Song, M. J., & Suh, B. R. (2008). The effects of output task types on noticing and learning of the English past counterfactual conditional. *System*, *36*, 295–312.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principles and practice in applied linguistics* (pp. 125–144). Oxford University Press.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 97–114). Oxford University Press.

Swain, M. (2005). The output hypothesis: theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–484). Lawrence Erlbaum Associates.

Uggen, M. S. (2012). Reinvestigating the noticing function of output. *Language Learning*, *62*, 506–540.

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press

VanPatten, B. (2004). Input and output in establishing form-meaning connections. In B. VanPatten, J. Williams, S. Rott, & M. Overstreet (Eds.), *Form-meaning connections in second language acquisition* (pp. 29–47). Erlbaum.

Vasylets, O., Gilabert, R., & Manchon, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning*, *67*, 394–430.

Williams, J. (2012). The potential role(s) of writing in second language development. *Journal of Second Language Writing*, *21*, 321–331.

Zalbidea, J. (2017). “One task fits all”? The roles of task complexity, modality, and working memory capacity in L2 performance. *Modern Language Journal*, *101*, 335–352.

Zamuner, T. S., Morin-Lessard, E., Strahm, S., & Page, M. A. (2016). Spoken word recognition of novel words, either produced or only heard during learning. *Journal of Memory and Language*, *89*, 55–67.

https://doi.org/10.1017/S0272263120000261 Published online by Cambridge University Press