Studies in Second Language Acquisition, **43** (2021), 239–267
doi:10.1017/S0272263120000376

Research Article

INITIAL PROFICIENCY AND L2 GRAMMAR DEVELOPMENT DURING SHORT-TERM IMMERSION ABROAD

CONCEPTUAL AND METHODOLOGICAL INSIGHTS

Janire Zalbidea *

Temple University

Bernard I. Issa

University of Tennessee, Knoxville

Mandy Faretta-Stutenberg

Northern Illinois University

Cristina Sanz

Georgetown University

Abstract

The first goal of this study was to examine how individual differences in initial L2 proficiency help explain L2 grammar development in oral production during short-term immersion abroad. The second goal of the study was methodological, and evaluated challenges that can result from operationalizing learners' initial L2 proficiency as pretest performance on outcome measures (as opposed to independent proficiency measures) in analyses of L2 change. L2 Spanish learners participating in summer study abroad completed an elicited imitation task and two oral production tasks. Production data were analyzed for changes in relevant grammatical complexity and accuracy

The experiment in this article earned Open Data and Open Materials badges for transparent practices. The data and materials are available at https://osf.io/nca3m/?view_only=2f238b3cec1b471797c9a89bdaa98d6f

This research was partially funded by a Professional and Scholarly Development Award from the College of Arts and Sciences at the University of Tennessee, Knoxville. We are grateful to Alfonso Hernanz, Lauren Jones, Helen Spader, Fermin Luna, and Kyrie Miranda for their assistance with various aspects of this research. We would also like to thank the SSLA editors and anonymous reviewers for their valuable feedback on earlier versions of this manuscript. All errors are our 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, PA 19122. E-mail: janire.zalbidea@temple.edu

dimensions. Results indicate that learners with higher initial L2 proficiency experience greater L2 grammar advancement from short-term immersion, and that pretest performance can be an unreliable operational estimate of initial proficiency when analyzing L2 gains. We discuss findings following cognitive accounts of SLA, and highlight methodological implications for further research in immersion contexts and beyond.

INTRODUCTION

The field of second language acquisition (SLA) has experienced growing interest in investigating second language (L2) development in study abroad (SA) contexts, with the aim of understanding how immersive, input-rich environments can foster rapid L2 achievement (e.g., Marijuan & Sanz, 2018; Tullock & Ortega, 2017). Over the past two decades, there has been a steady increase in student participation in SA, with shorter-term programs lasting 8 weeks or less leading enrollment rates among U.S. students in recent years (Institute of International Education, 2019). While the advantages of SA for certain areas of L2 development (e.g., fluency, vocabulary) are relatively uncontested, research into its contributions for developing other dimensions of learners' L2 production —particularly within the grammatical domain—is relatively scarce and has yielded divergent findings. Some studies report moderate L2 gains in certain aspects of grammar (e.g., Grey et al., 2015; Leonard & Shea, 2017; Mora & Valls-Ferrer, 2012), whereas other studies find that learners' L2 grammar develops minimally (if at all) during SA, often leading to learner frustration (e.g., DeKeyser, 2010; Isabelli-García, 2010).

To explain this persistent variability, researchers have turned to examining how individual differences may contribute to success in L2 grammar achievement. Amid the multiple individual factors relevant for SLA (e.g., age, motivation), learners' level of L2 proficiency at SA onset has featured prominently in discussions about whose grammar progresses more abroad (e.g., Collentine, 2009; DeKeyser, 2010), with diverging perspectives. Some studies suggest that more proficient learners experience more noticeable L2 gains abroad (e.g., DeKeyser, 2010; Golonka, 2006; Leonard & Shea, 2017), a finding generally attributed to their superior resources to process L2 form and meaning in communication. In contrast, other studies find greater L2 gains among less proficient learners (e.g., Baker-Smemoe et al., 2014; Llanes & Muñoz, 2009), which suggests that their less stabilized L2 knowledge may be more amenable to growth during immersion. In the present study, we investigate the contributions of initial L2 proficiency to the development of grammatical complexity and accuracy dimensions in oral L2 production during short-term SA.

Although the role of proficiency has been examined through different perspectives, an important methodological characteristic distinguishes the set of studies reporting a positive relationship between initial L2 proficiency and L2 gains abroad from those finding a negative relationship. Namely, the former administered an L2 proficiency measure that was distinct from the L2 outcome measure(s) of the study, whereas the latter employed learners' pretest performance on the L2 outcome measure(s) as a proxy for their initial L2 proficiency. Here, we sought to tease apart potential analytical and measurement-related challenges identified with the latter approach that, as we describe subsequently, are related to the statistical phenomenon of *regression toward the mean*

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

(RTM). In this study, we consider both types of proficiency operationalizations and examine how inferences drawn from the data may differ based on the chosen approach.

BACKGROUND

INITIAL PROFICIENCY AND L2 GRAMMAR DEVELOPMENT ABROAD

Despite its centrality to the field (e.g., Winke & Gass, 2018), L2 proficiency remains an opaque construct in much SLA research. Although proficiency is often equated with "L2 competence or ability … typically inferred from analyses of concrete instances of that person's L2 performance" (Thomas, 1994, p. 330), efforts have been made to theorize and model L2 proficiency. Most recently, Hulstijn (2012, 2015) has conceptualized L2 proficiency as comprising two distinct components: basic and higher language cognition. Basic language cognition is "restricted to processing oral language (listening and speaking) in utterances containing high-frequency lexical, grammatical, phonotactic and prosodic elements" (Hulstijn, 2012, p. 429), which may appear in any communicative context, irrespective of the speakers' age, literacy, or education. In contrast, higher language cognition is viewed as an extension of basic language cognition, and comprises reading and writing, as well as utterances containing more complex and lower frequency forms (Hulstijn, 2012, 2015). In this study, we examine L2 proficiency in terms of basic language cognition (as assessed by an elicited imitation task [EIT]), as this component of proficiency encompasses core linguistic subskills and knowledge considered essential to the attainment of advanced L2 oral abilities.

SA research has long been interested in investigating the role of initial L2 proficiency in L2 development, perhaps because this is an individual difference that learners can "control" by deciding when to go abroad in their L2 learning trajectory (Collentine, 2009). Initial L2 proficiency has been debated as a potential explanatory variable regarding L2 grammar development in immersion contexts, given the wide individual variability in gains observed in the empirical literature. To date, however, there is still a lack of consensus regarding the role of initial L2 proficiency in language learning while abroad. Some authors hypothesize that learners who leave for SA at higher proficiency levels experience greater L2 grammar development (e.g., DeKeyser, 2010; Lafford & Collentine, 2006; Segalowitz & Freed, 2004), as they purportedly have more cognitive resources available to simultaneously attend to form and meaning while managing L2 interaction demands in SA settings. These greater resources may also allow for increased L2 interaction during a sojourn (e.g., Freed, 1995). DeKeyser (2010) has argued that more proficient learners are also better prepared to engage in key SLA processes (e.g., monitoring) during L2 practice abroad, increasing the possibility of eventual proceduralization. From this standpoint, Lafford and Collentine (2006) consider the possibility of an initial proficiency threshold that learners may have to reach to achieve sizable L2 gains abroad.

Other authors indicate that learners can make greater L2 gains when they start their sojourn at lower proficiency levels (e.g., Brecht & Robinson, 1995; Llanes & Muñoz, 2009; Vande Berg et al., 2009), as more advanced learners' grammars may have reached a stabilization stage that stems from access to more consolidated linguistic resources (Llanes & Muñoz, 2009). By and large, these accounts suggest that lower proficiency

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

learners have "more room" for L2 development and thus "seem to make the most obvious advances" (Regan, 2003, p. 73) while abroad, at least when it comes to the outcome measures that are commonly administered in SA research (see Llanes & Muñoz, 2009).

PREVIOUS RESEARCH ON THE ROLE OF INITIAL L2 PROFICIENCY ABROAD

In spite of these proposals, the role of initial L2 proficiency (particularly for L2 grammar development) remains an open question, possibly because research has often explored this issue post hoc and employed various measures of L2 knowledge as proxies for initial L2 proficiency. In this section we summarize prior studies, organized by the relationship reported between initial L2 proficiency and L2 gains abroad. Because only one study in this domain has focused specifically on L2 grammar (Faretta-Stutenberg & Morgan-Short, 2018), we have broadened our review to consider other L2 skills (see Appendix S1 in the Online Supplementary Materials for a summary table).

Multiple studies have reported a positive relationship between initial L2 proficiency (operationalized in various ways) and L2 outcomes during SA. Golonka (2006) examined factors that could help differentiate between L2 Russian learners that experienced post-SA speaking proficiency gains on the Oral Proficiency Interview (OPI; scored based on the American Council on the Teaching of Foreign Languages [ACTFL] proficiency guidelines) and those who did not. Findings suggested that learners' pre-SA performance on the American Council of Teachers of Russian (ACTR) grammar test, as well as a series of indicators in the pre-SA OPI (e.g., self-repair rate), were higher for gainers than nongainers. In a follow-up study, Davidson (2010) also found that L2 Russian learners with greater listening proficiency and grammar knowledge in the ACTR test at SA onset experienced the greatest improvements in their OPI scores, regardless of program length. Both authors highlighted the importance of linguistic control in the target language for maximizing L2 gains abroad.

Additionally, DeKeyser (2010) found that post-SA accuracy ratings of L2 Spanish learners' oral interviews were positively predicted by their pre-SA performance on the written Modern Language Association placement test. DeKeyser concluded that "the more [learners] know, the more they can get better at using what they know through practice and add new knowledge through input and interaction" (p. 90). A more recent study by Leonard and Shea (2017), also on L2 Spanish, showed that learners with greater initial L2 vocabulary knowledge and faster processing speed at SA onset experienced greater gains in accuracy (errors per 100 words) and complexity (both lexical and grammatical), but not in fluency. Contrary to earlier research, Leonard and Shea found that initial L2 grammar knowledge was not a significant predictor. The authors suggested that having greater lexical knowledge and faster processing skills may liberate learners' attentional resources considered critical for the efficient improvement of oral L2 skills abroad.

To determine the effects of initial L2 proficiency on L2 Spanish grammar gains specifically, Faretta-Stutenberg and Morgan-Short (2018) examined whether learners' pre-SA performance on an EIT, which taps into global (oral) proficiency through sentence repetition, and the *Diploma de Español como Lengua Extranjera* (DELE) test, which comprises cloze and multiple-choice tests, were related to gains in grammatical gender agreement during a semester abroad. Learners with higher EIT and DELE scores at the onset of SA experienced greater gains in article gender agreement on a grammaticality judgment task. This is, to our knowledge, the first study providing some evidence of the benefits of higher starting proficiencies—as measured by a global proficiency test like the EIT—for L2 grammar improvement abroad.

On the whole, findings from these studies point to a "more is more" account whereby greater initial L2 proficiency (as measured by pre-SA vocabulary knowledge, global oral proficiency, and discrete lexico-grammar knowledge tests) appears to set the stage for greater L2 development abroad. Nonetheless, contrasting evidence is available from multiple other studies suggesting that learners with lower initial proficiency experience greater linguistic benefits from SA than higher-proficiency learners. In an early large-scale study examining factors that predicted L2 Russian development, Brecht et al. (1995) found that gains on listening and reading proficiency tests were strongly, negatively related to participants' initial (i.e., pretest) scores in these same tests, as were gains on the OPI. Hinting that lower-level learners are more likely to experience rapid L2 improvement abroad, the authors noted that "the higher the initial level, the less the gain" (p. 46).

This "less is more" account was further supported by another large-scale project by Vande Berg et al. (2009), where learners' initial proficiency level as measured by the Simulated Oral Proficiency Interview (SOPI) was negatively related to their SOPI gains abroad. The authors determined that "learners abroad begin to plateau in their oral proficiency as captured by the SOPI" (p. 13), which they suggested is more consistent with a "ceiling" effect rather than a proficiency threshold effect. They also considered the possibility that intercultural educational differences may be contributing to the plateau effect, suggesting that rather than pursuing advanced linguistic accuracy as may be prioritized in some SA institutions, U.S. students may be "satisfied when they can speak with a certain degree of facility" (p. 14) abroad, as proposed by Engle and Engle (2004).

Baker-Smemoe et al. (2014) revisited the issue of initial L2 proficiency with L1 English learners of various target languages studying abroad in six different programs. Following Golonka (2006), learners were divided into gainers and nongainers based on their OPI change scores. Findings indicated that gainers had begun their sojourn with lower preprogram proficiencies than nongainers, leading the researchers to suggest that advancing in the ACTFL proficiency scale may be more difficult for learners who begin their sojourn at more advanced L2 levels.

Focusing on short-term immersion and native Catalan/Spanish learners of English, Llanes and Muñoz (2009) found that initial L2 proficiency, operationalized as the principal component of measures of fluency (syllables per minute) and accuracy (average errors per clauses) at SA onset, negatively predicted learners' L2 gains in fluency and accuracy. This set of results extended findings on the linguistic advantages of lower initial L2 proficiency levels abroad from earlier research using more global L2 measures (e.g., OPI) to more specific dimensions of oral ability, namely fluency and accuracy.

In sum, prior research has provided evidence that learners' L2 proficiency (operationalized through various L2 measures) at the onset of SA can affect the extent to which they benefit linguistically from immersion experiences. However, whether learners experience greater L2 improvements when they start their sojourn at lower or higher levels of L2 proficiency continues to be a matter of empirical debate. Notably, proficiency has been proposed to be an important determinant of L2 grammar development; yet, there is a paucity of research focusing on learners' L2 grammar gains in this

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

domain (cf. Faretta-Stutenberg & Morgan-Short, 2018). The present study sought to address these gaps in our understanding of the role of initial L2 proficiency abroad by considering changes in specific dimensions of learners' L2 grammar in oral production after a short-term SA experience.

METHODOLOGICAL CONSIDERATIONS IN INVESTIGATING THE ROLE OF INITIAL L2 PROFICIENCY

Although the inconsistency in current evidence on the role of initial L2 proficiency abroad might seem surprising, closer examination of the methodologies employed in prior studies reveals a pattern that, we propose, is likely contributing to these divergent findings. Specifically, two different approaches to address the role of initial proficiency emerge from previous research: Some studies have employed what we refer to as *independent* L2 proficiency measures (i.e., proficiency measures that are distinct from the L2 outcome measures used to assess L2 development abroad), whereas other studies have employed what we here term *nonindependent* measures (i.e., learners' pretest/baseline performance on the L2 outcome measure) as a proxy for their initial L2 proficiency. Interestingly, the direction of the relationship between initial proficiency and L2 gains identified appears to correspond with the type of proficiency measure employed: Studies that have employed independent measures generally report a positive relationship between initial proficiency and L2 change, whereas studies that use a nonindependent measure report a negative relationship.

Both independent and nonindependent measures comprise "concrete instances of [a] person's L2 performance" used to make inferences about their "L2 competence or ability" (Thomas, 1994, p. 330). However, as we describe subsequently, nonindependent measures of proficiency can present interpretation challenges when regressed onto or correlated with change scores (i.e., the difference in learners' performance from Time 1, or SA onset, to Time 2, or SA completion) on that same L2 measure because the relationship "between the baseline measurement and the gain is always [expected to be] negative" (Taraday & Wieczorek-Taraday, 2018, p. 1394) on account of RTM, a widespread statistical phenomenon found in repeated-measures designs. RTM implies that learners who score markedly higher or lower than the sample mean at Time 1 will tend to score closer to the sample mean at Time 2 (e.g., Kachigan, 1991; Shanks, 2017; Yu & Chen, 2015). That is, values that are more extreme (i.e., further away from the mean) at baseline, will tend to be less extreme in the follow-up measurement. This entails a negative relationship between baseline and gain values (see Campbell & Kenny, 1999). In SLA, where the focus of research is often on L2 development over time, such an effect may be interpreted as evidence that initial low scorers improved on the posttest, whereas initial high scorers did not (or at least not to the same degree as initial low scorers). However, as proposed in the literature, this effect is considered largely reflective of the magnitude of RTM in the data, rather than of actual differences among participants.

RTM occurs whenever the correlation coefficient between any two repeated measures is not equal to |1.0|. Perfect correlations between measures are very unusual in psychological and other human systems (Dormann & Griffin, 2015), which makes RTM a prevalent phenomenon. All outcome measures carry some level of error, such that what we capture comprises both the true value of the construct plus some random measurement

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

error, which is assumed to balance out at the group level but can influence single observations (e.g., Shanks, 2017; Taraday & Wieczorek-Taraday, 2018). In repeated-measures designs, such as when we assess L2 performance before and after an immersion period abroad, the true value of learners' L2 performance is expected to change, but so is the error component. RTM predicts that the more an individual's score deviates from the sample mean at Time 1, the greater "the probability that it is a result of random error at most" (Taraday & Wieczorek-Taraday, 2018, p. 1394), and that it will regress toward the sample mean at Time 2 (see Campbell & Kenny, 1999; Kachigan, 1991 for further discussion).[1] This statistical bias introduced by RTM presents challenges for SA research seeking to establish the role of initial L2 proficiency on the basis of nonindependent L2 proficiency measures only because the negative relationships that result between learners' pretest and L2 change scores may be explained on account of RTM. Arguably, additional evidence of these negative associations with an independent L2 proficiency measure would be desirable to reliably interpret findings as indicative that lower pre-SA L2 proficiency is associated with greater L2 gains.

Moreover, even when employing independent proficiency measures, SA researchers would still want to make efforts to account for potential RTM in their data, so that the unique variance in L2 gains explained by initial L2 proficiency is represented as reliably as possible even when participants differ widely in their pretest scores. One way to achieve this, as proposed in the literature, is to control for learners' initial L2 performance by incorporating participants' Time 1 scores as covariates in correlation or regression analyses (see Yu & Chen, 2015). As RTM-related error is most prevalent when outcome measures have high variability, controlling for Time 1 scores appears particularly relevant when fluctuations in L2 scores can be expected. In this study, we directly evaluate these statistical bias and control challenges concerning RTM by considering both independent and nonindependent measures of initial L2 proficiency.

Besides issues related to RTM that we mention here, multiple authors have discussed methodological concerns regarding design aspects in proficiency-related research. For instance, Tremblay (2011) has argued that research comparability can be enhanced with the use of global proficiency measures (e.g., EIT), rather than domain- (e.g., grammar) or skill-specific (e.g., reading) measures, which tend to differ in design across studies. Another aspect underscored by Freed (1995) and echoed by others (e.g., Issa & Zalbidea, 2018) is that research seeking to understand how initial L2 proficiency impacts L2 gains ought to be able to capture linguistic development among learners at differing proficiency levels. In this regard, the suitability of the OPI (the most widely used measure in prior SA research) to capture progress made by more advanced learners has been questioned (e.g., Freed, 1995). More general guidelines have also been proposed by Hulstijn (2012, 2015), including recommendations to clearly operationalize L2 proficiency, justify the choice of measure in relation to the larger study design, and establish "substantial variance in participant scores" (Hulstijn, 2015, p. 82) to avoid issues with range restriction, which can lead to attenuated relationships among variables. As described in the "Method" section, efforts were made to follow these recommendations in the present research.

With the rapid growth of SLA research in SA contexts, informing methodological practices in this domain is warranted. Here, we sought to tease apart the potential measurement-derived challenges that stem from RTM to better understand the role of initial L2 proficiency in SLA abroad. Specifically, we (a) consider possible differences in

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

the directional association of independent and nonindependent L2 proficiency measures as predictors of L2 change, and (b) examine how the predictive weight of our independent proficiency measure is impacted when initial scores are disregarded and RTM is not built into the analyses. Based on our results, we explore the possibility of reconciling prior conflicting findings by reassessing available evidence in light of RTM.

THE CURRENT STUDY

Research has shown that SA can bolster the development of oral L2 skills, yet studies focusing on learners' L2 grammar progress during immersion abroad suggest that certain students flourish while others flounder. Learners' L2 proficiency at SA onset has been proposed as a central individual factor contributing to this persistent variability in L2 grammar outcomes; however, its effects remain to be fully investigated. Toward that end, we pursued the following research question (RQ):

RQ1. To what extent do learner individual differences in initial global L2 proficiency predict changes in L2 grammatical complexity and accuracy in oral production during short-term SA?

As detailed earlier, researching the role of initial L2 proficiency in SA presents a number of challenges. We noted that prior studies employing independent and nonindependent proficiency measures as predictors of L2 change have yielded contrasting findings, and we identified a number of methodological considerations regarding RTM that are particularly relevant to the latter approach. To empirically evaluate the statistical bias and control issues relating to RTM, we pursued a second question:

RQ2. Do initial global L2 proficiency (independent) and initial L2 complexity and accuracy performance (nonindependent) show different directional associations as predictors of L2 change? If so, to what extent does the predictive weight of initial global L2 proficiency change when initial L2 complexity and accuracy performance is uncontrolled in the analyses?

METHOD

STUDY ABROAD PROGRAMS

Data were collected from two short-term summer SA programs in Spain sponsored by the same U.S. public university. The programs were located in Santander and Alicante[2] and were designed for learners at intermediate and advanced curricular levels of Spanish (i.e., two semesters or at least six semesters of college-level courses), respectively. The programs were highly comparable, sharing key features regarding type of program (sheltered, U.S. faculty-led), duration (5 weeks), living arrangement (family homestay), and coursework and extracurricular activities, including aspects such as class duration (3 hours/day, 6 credit hours earned) and frequency (5 days/week), type and frequency of extracurricular activities (2–8 hours/week), and language of instruction (Spanish during classes, Spanish and English during extracurricular activities).

In the Santander program (henceforth the "Intermediate" program), learners were enrolled in an intensive Spanish language course and stayed with a host family that included at least one U.S. roommate. In the Alicante program (the "Advanced" program), learners were enrolled in two Spanish content courses and also lived with a host family, with about half of the participants staying with a U.S. roommate. In both programs, course instructors were highly functional L1 English–L2 Spanish bilinguals and homestay families were instructed to communicate only in Spanish with the students.

PARTICIPANTS

Data were collected from 35 L2 learners of Spanish (Intermediate: $n = 18$; Advanced: $n = 17$). Five participants from the Intermediate program were removed due to data unintelligibility, as a result of technical issues, or because they failed to complete one or more tasks. All participants in the final sample were native speakers of English. Besides English, seven participants reported knowledge of another native language (Urdu, Portuguese, Tagalog, Laotian, Vietnamese, Gujarati, and Hindi) and seven reported studying an additional later-learned language (American Sign Language, Latin, Turkish, or German). Participants were around 20 years of age (Intermediate: $M = 20.62$, $SD = 2.18$, n-female = 11; Advanced: $M = 20.18$, $SD = 0.88$, n-female = 15). Table 1 summarizes further participant background information.

A series of unpaired two-samples Wilcoxon tests indicated that participants in both programs were comparable in terms of their age of exposure to Spanish ($W = 79.00$, $p = .190$, $r = –.239$) and their reported motivation to learn Spanish (similarly high in both programs; $W = 85.50$, $p = .265$, $r = –.203$). As expected, participants in the Advanced program reported more formal instruction in Spanish ($W = 35.50$, $p = .002$, $r = –.575$). Participants in the Intermediate program had no prior experience studying abroad, whereas two students in the Advanced program reported participation in a previous sojourn (one spent 4 months in Costa Rica and one spent 1.5 months in Spain). Lastly, no statistical differences were found between participants in both programs regarding a series of cognitive skills, which included verbal IQ ($W = 77.50$, $p = .173$, $r = –.249$),

TABLE 1. Participant background information

	Intermediate program		Advanced program	
	M (SD)	*Mdn*	*M (SD)*	*Mdn*
Spanish language				
Age of exposure to Spanish (years old)	10.92 (5.92)	11.00	12.35 (3.26)	13.00
Years of formal instruction in Spanish	4.35 (3.14)	3.50	6.65 (2.34)	6.00
Motivation to learn Spanish[a]	5.85 (1.34)	6.00	6.47 (.62)	7.00
Cognitive skills				
Verbal IQ (KBIT-2 Verbal Knowledge)[b]	104.92 (10.66)	101.00	111.24 (12.62)	108.00
Nonverbal IQ (KBIT-2 Matrices)	102.54 (14.63)	100.00	104.82 (11.22)	103.00
Working memory (Reading span task)[c]	15.60 (7.65)	15.50	18.06 (7.13)	19.00

[a] Scores represent participants' responses to *"My motivation to learn Spanish is,"* rated using a 7-point Likert scale.

[b] All KBIT scores correspond to standardized values.

[c] Scores represent absolute span scores (max. possible = 30). Reading span data from three Intermediate participants were excluded because they did not reach the criterion on the processing measure.

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

nonverbal IQ ($W = 98.00$, $p = .613$, $r = -.092$), and working memory ($W = 66.00$, $p = .349$, $r = -.170$).

In sum, participants in the Intermediate and Advanced programs, whose data were pooled together to address the RQs of this study, took part in analogous SA programs directed by the same institution and were also comparable along a number of relevant linguistic and nonlinguistic variables.

PROCEDURE

Data for this study were collected over three sessions as part of a larger research project. Participants completed a language history questionnaire, a motivation questionnaire, and tests of IQ and working memory during the first session, which took place at the home university. The second and third sessions took place at the SA site during the first (Week 1) and last week (Week 5) of the programs, respectively. Participants completed the EIT and the oral production tasks during these on-site sessions. For both Spanish measures, materials were presented on a laptop computer and participants' responses were recorded with a high-definition audio recorder. Additionally, participants reported their L2 contact using a weekly online survey, as detailed in the following text. Extra credit and monetary compensation were awarded for participation.

MATERIALS

Participant Background: Language History, Motivation, and Cognitive Skills

Participants provided information about their background and Spanish learning experience in a language history questionnaire, and indicated their overall motivation to learn Spanish. General intelligence and working memory capacity were assessed using the Kaufman Brief Intelligence Test, Second Edition (KBIT-2; Kaufman & Kaufman, 2004) and a shortened automated reading span task (Oswald et al., 2015), respectively.

Language Contact Questionnaire

Participants in both programs completed a weekly online L2 contact questionnaire (adapted for weekly use from the posttest version of the Language Contact Profile; Freed et al., 2004). The average number of weekly L2 contact hours was comparable in the Intermediate ($M = 50.23$, $SD = 27.18$; $Mdn = 51.50$) and Advanced ($M = 42.05$, $SD = 25.96$; $Mdn = 35.33$) programs, with no statistical differences between learners in these two programs ($W = 131.00$, $p = .408$, $r = -.151$). Additionally, given our focus on oral production, we examined whether learners differed in terms of their amount of L2 speaking abroad. Participants reported similar weekly hours spent speaking with both native speakers (Intermediate: $M = 10.18$, $SD = 6.55$; $Mdn = 9.60$; Advanced: $M = 8.80$, $SD = 5.26$; $Mdn = 8.30$) and nonnative speakers of Spanish (Intermediate: $M = 5.51$, $SD = 4.96$; $Mdn = 5.33$; Advanced: $M = 4.37$, $SD = 4.50$; $Mdn = 2.50$), with no statistical differences between programs ($W = 121.50$, $p = .660$, $r = -.080$, and $W = 107.00$, $p = .900$, $r = -.023$, respectively).

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

Oral Production Tasks

L2 grammar development was assessed with two monologic oral production tasks, using the same procedure and prompts in both programs. Participants were provided with a written prompt (in English) and asked to respond to it orally in Spanish (see Appendix). In Week 1, the prompt asked about participants' initial experiences being in a Spanish language environment, as well as their goals and expectations for their sojourn. In Week 5, the prompt asked them to reflect on their goals and expectations and to discuss cultural differences and similarities.

The rationale behind the design of this task was that learners' L2 output in an open-ended oral production measure "is spontaneous and is therefore likely to be highly reflective of the learner's L2 knowledge" (Collentine, 2004, p. 233). Additionally, the task was designed to have face validity as an oral diary, such that participants would be able to express their SA experiences without being predisposed to focus on L2 grammar. The relatively unconstrained nature of the task also allowed for successful elicitation of production data from a broad range of proficiencies. Participants were asked to speak for approximately 5–8 minutes and for no longer than 15 minutes. These parameters were chosen to decrease the likelihood of excessive time pressure during L2 production.

Elicited Imitation Task

We operationalized initial global L2 proficiency as performance on an EIT (adapted from Ortega et al., 1999),[3] argued to tap into basic language cognition (e.g., Bowden, 2016; Wu & Ortega, 2013). An EIT was chosen as an independent proficiency measure because it (a) provides an objective assessment of global (oral) L2 proficiency; (b) is suitable for the study given our focus on L2 grammar in oral production; and (c) is a valid, reliable, and well-researched measure (see Bowden, 2016 for discussion of task validity). Furthermore, the use of an EIT allows for systematic comparisons among studies investigating the role of L2 proficiency, even across different target languages, fostering desirable conditions for research replicability and generalizability (Bowden, 2016).

Although only Week 1 scores are examined here, two comparable versions of the sentence repetition task were administered at Week 1 and 5 of the sojourn, with the order of administration counterbalanced across participants in both programs. Following a brief orientation to the task and practice in English, participants listened to 30 Spanish sentences, which increased in both length and complexity (from 7 to 17 syllables). After the presentation of each sentence, there was a 2-second pause, followed by a 0.5-second tone to cue the participant to repeat the sentence aloud to the best of their ability.

Coding and Scoring

Oral Production. Learners' oral production at Week 1 and 5 was coded for dimensions of grammatical complexity and accuracy that held potential to capture change across a broad proficiency spectrum.[4] Two complexity measures were calculated: mean length of analysis of speech unit (AS-unit; Foster et al., 2000) and dependent clauses per AS-unit. Mean length of AS-unit represents global complexity "that may have been achieved by any means, for example, via increased use of modification such as adjectives and adverbs"

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

TABLE 2. Sample grammatical accuracy errors

Agreement error type	Example
(1) Subject-verb	*Ayer mi familia *cociné pollo para el almuerzo* "Yesterday my family *cooked-1stp.sg. chicken for lunch"
(2) Number	*Ellos están muy *simpática* "They are very *friendly-sg."
(3) Gender	*Es *un ciudad muy bella* "It is *a-masc. very beautiful city-fem."

Note. Errors are marked with an asterisk.

(Mochizuki & Ortega, 2008, p. 23). The more specific measure of dependent clauses per AS-unit was also considered relevant for our sample, as clausal complexification through subordination is posited to be a favored strategy among learners at intermediate and upper-intermediate levels (e.g., Norris & Ortega, 2009).

For accuracy, learners' morphosyntactic agreement error ratios in Spanish were coded across three domains:[5] subject-verb agreement, and plural number agreement and feminine gender agreement marked on both determiners and adjectives (see Table 2 for sample errors from the data). This complex morphology in Spanish is expected to be challenging for L2 learners, particularly those whose L1 has a relatively poor inflectional system with regard to overt morphosyntactic agreement, as is the case for English. Subject-verb and number agreement have not been shown to pose major difficulties for L2 learners beyond intermediate proficiency levels (see Foote, 2011), although lower-level learners may be less sensitive to number violations (Tokowicz & MacWhinney, 2005). In contrast, gender agreement, particularly with feminine nouns, is notable for presenting major challenges for L2 learners (even at higher proficiency levels) due to its lack of communicative value as compared to subject-verb and number agreement, among other factors (see, e.g., Alarcón, 2010). All three agreement targets are highly frequent in the input and learners were expected to have ample exposure to them abroad.

All production data were independently coded for grammatical complexity and accuracy by at least two of the researchers. Any instances of disagreement were reviewed with an additional researcher and reconciled until 100% agreement was reached on all coding. Descriptive statistics for measures of grammatical complexity and accuracy are reported in Tables 3 and 4, respectively.

Elicited Imitation Task. Participant responses to each of the 30 EIT sentences were transcribed by two independent, trained researchers. Any discrepancies were reconciled by a third researcher. Next, two independent raters were trained on the scoring protocol established by Ortega (2000), in which each sentence repetition is assigned a score of zero (no repetition, unintelligible, only one content word) to four points (perfect repetition) based on both accuracy of repetition and meaning of the utterance (see Ortega, 2000 and Bowden, 2016 for scoring guidelines). The independent ratings were compared, with any discrepancies resolved by another pair of researchers until agreement was reached on all ratings. Due to a tone playback issue with one item on one version of the EIT, responses for this item were eliminated from both versions, yielding a maximum score of 116.

Participants' EIT scores in the Intermediate program ranged from 13 to 47 ($M = 26.62$, $SD = 9.77$; 95% CI [20.71, 32.52]); in the Advanced program, scores ranged from 45 to

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

TABLE 3. Descriptive statistics for grammatical complexity

	Week 1			Week 5			
Program	*M (SD)*	Min	Max	*M (SD)*	Min	Max	*d* [95% CI]
MLU							
Intermediate	8.75 (1.44)	7.06	11.86	9.41 (1.30)	6.53	11.21	–.48 [–1.13, .17]
Advanced	9.90 (2.23)	5.44	13.43	11.39 (2.18)	7.67	15.09	–.68 [–1.02, –.32]
Grand mean	9.40 (1.99)			10.53 (2.08)			–.56 [–.84, –.27]
DC							
Intermediate	.26 (.20)	.05	.71	.26 (.20)	.00	.65	.01 [–.59, .60]
Advanced	.49 (.22)	.14	.83	.59 (.25)	.04	1.07	–.41 [–.96, .14]
Grand mean	.39 (.24)			.45 (.28)			–.21 [–.54, .13]

Note. MLU, Mean length of AS-unit; DC, Dependent clauses per AS-unit.

TABLE 4. Descriptive statistics for grammatical accuracy (error ratios)

	Week 1			Week 5			
Program	*M (SD)*	Min.	Max.	*M (SD)*	Min.	Max.	*d* [95% CI]
Subject-verb							
Intermediate	.11 (.09)	.00	.30	.12 (.11)	.00	.40	–.11 [–.96, .73]
Advanced	.05 (.04)	.00	.18	.04 (.03)	.00	.12	.11 [–.41, .62]
Grand mean	.07 (.07)			.08 (.09)			–.03 [–.50, .43]
Number							
Intermediate	.31 (.27)	.00	.71	.28 (.21)	.00	.80	.22 [–.82, 1.26]
Advanced	.07 (.08)	.00	.25	.02 (.03)	.00	.09	.84 [.07, 1.61]
Grand mean	.17 (.22)			.13 (.19)			.26 [–.24, .76]
Gender[a]							
Intermediate	.37 (.24)	.15	1.00	.23 (.14)	.04	.50	.61 [–.15, 1.07]
Advanced	.10 (.06)	.02	.20	.14 (.14)	.00	.50	–.24 [–.59, .11]
Grand mean	.22 (.21)			.18 (.15)			.21 [–.09, .52]

[a] Number of unique feminine nouns (types): at Week 1, Intermediate: *M (SD)* = 7.31 (4.23), Advanced: *M (SD)* = 12.71 (4.65); at Week 5, Intermediate: *M (SD)* = 9.77 (3.81), Advanced: *M (SD)* = 15.47 (4.40). Percentage of gender transparent (i.e., canonical) types: at Week 1, Intermediate: 57%, Advanced: 66%; at Week 5, Intermediate: 72%, Advanced: 65%.

107 ($M = 69.71$, $SD = 15.41$, 95% CI [61.78, 77.63]). Combining data from both programs yielded a desirable range of 13–107 in our full sample (see Figure 1), comparable to the distribution reported in Bowden (2016) for a sample of 37 participants. Reliability analyses computed with Week 1 EIT scores yielded a Cronbach's alpha of .98 for version A (95% CI [.96, .99]) and .97 for version B (95% CI [.95, .99]), indicating a very high level of internal consistency among this sample.

ANALYSIS

To address RQ1 on the role of initial global L2 proficiency in L2 grammar development, we computed hierarchical multiple regression models with changes in dimensions of

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

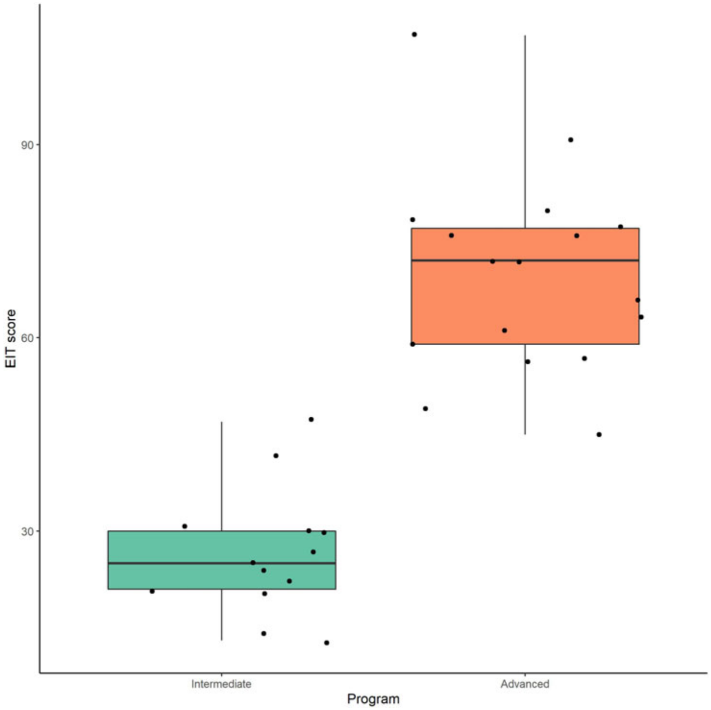

FIGURE 1. Boxplots for EIT scores.

grammatical complexity (mean length of AS-unit, MLU; dependent clauses per AS-unit, DC) and accuracy (i.e., error ratios) across targets (subject-verb, number, gender agreement) as the criterion variables. Change models were chosen to allow for direct comparisons with the prior empirical SA studies discussed here. For each dimension, we first fit a baseline model with Week 1 performance and average weekly L2 contact as control predictors because prior SA research has established that L2 contact can contribute to L2 development abroad (e.g., Collentine, 2009). Next, we fit a model that contained our control predictors and our independent variable predictor, namely initial L2 proficiency as gauged by the EIT.[6] Changes in R^2 were checked to determine the unique variance explained by initial L2 proficiency after accounting for Week 1 performance and L2 contact.

To answer the first part of RQ2, related to the directional association of initial global L2 proficiency (independent) and initial L2 performance (nonindependent) as predictors of L2 grammar change, we examined the coefficients in each of the models computed to address RQ1. To answer the second part of RQ2 about whether the predictive weight of initial global L2 proficiency differs when initial L2 complexity and accuracy are not controlled, we ran analogous hierarchical regression analyses to those run for RQ1, without incorporating Week 1 performance as a predictor.

Assumptions for multivariate regression were examined following Jeon (2015). Given our relatively modest sample size, we also employed the bootstrap estimation method (1,000 samples), which does not assume normality or homoscedasticity, to compute robust coefficients (Efron & Tibshirani, 1993). We report the bias-corrected and

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

accelerated (BCa) bootstrap confidence intervals for the EIT coefficient in the following text (other bootstrap confidence intervals can be found in Appendix S2 in the Online Supplementary Materials). For both RQs, R^2 values were interpreted based on Plonsky and Ghanbar's (2018) guidelines: ≤.20 was considered small, whereas ≥.50, large.

RESULTS

RQ1. INITIAL L2 PROFICIENCY AND L2 GRAMMAR DEVELOPMENT IN ORAL PRODUCTION ABROAD

Complexity

For measures of complexity change, the baseline models were not significant for either MLU or DC change (results from regression models displayed in Table 5).

The second models with EIT as an additional predictor were significant for MLU and DC change, accounting for 31% and 42% of total variance, respectively. In both cases, Week 1 performance emerged as a negative predictor. The increases in variance explained (ΔR^2) from Model 1 (baseline) to Model 2 were significant for both measures (MLU change: 17% increase; DC change: 26% increase), with small effect sizes. EIT was a significant positive predictor of change in both models (see Figures 2 and 3). The 95% BCa confidence intervals for the robust EIT coefficients did not cross zero in either the MLU change [.005, .050] or DC change [.002, .009] models.

TABLE 5. Explanatory models for changes in grammatical complexity

		Fit	*F*-test	*B*	*SE B*	95% CI *B*	β	*p*
ΔMLU	Model 1							
	Week 1 performance			–.28	.14	[–.56, .01]	–.37	.059
	L2 contact			–.01	.01	[–.03, .01]	–.21	.267
	Total R^2	.14	2.10					.142
	Model 2							
	Week 1 performance			–.39	.14	[–.66, –.11]	–.52**	.008
	L2 contact			–.01	.01	[–.03, .01]	–.16	.379
	EIT			.03	.01	[.01, .05]	.45*	.018
	Total R^2	.31*	3.81					.022
	ΔR^2	.17*	6.38					.018
ΔDC	Model 1							
	Week 1 performance			–.38	.18	[–.75, .001]	–.38	.050
	L2 contact			–.003	.002	[–.01, .001]	–.28	.135
	Total R^2	.16	2.62					.091
	Model 2							
	Week 1 performance			–.67	.18	[–1.03, –.30]	–.67**	.001
	L2 contact			–.002	.001	[–.005, .001]	–.23	.162
	EIT			.01	.002	[.002, .009]	.60**	.002
	Total R^2	.42**	6.31					.002
	ΔR^2	.26**	11.64					.002

Note. MLU, Mean length of AS-unit; DC, Dependent clauses per AS-unit.
* p <.05; ** p <.01.

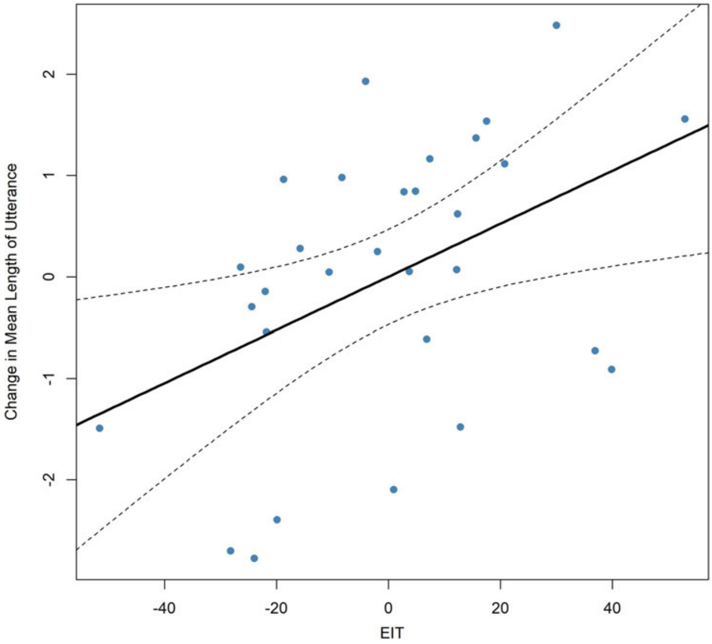

FIGURE 2. Partial effect plot: Change in MLU as a function of EIT.

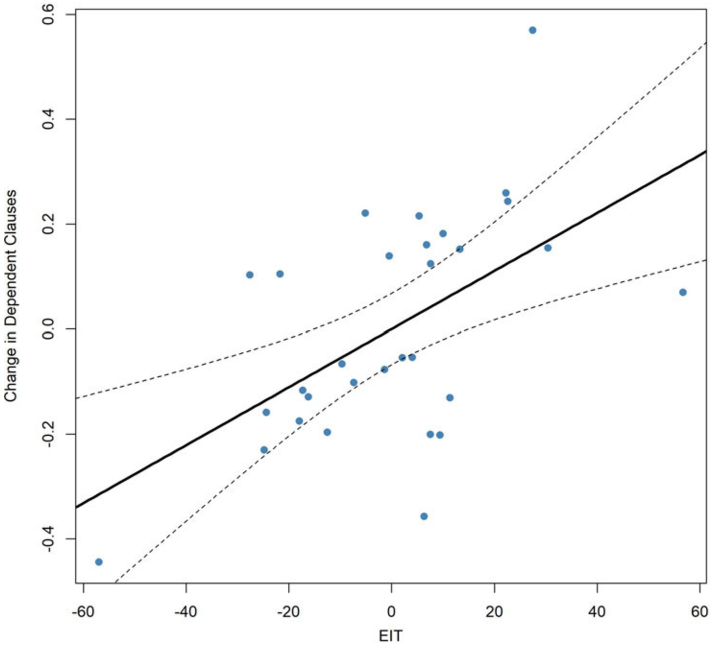

FIGURE 3. Partial effect plot: Change in DC as a function of EIT.

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

TABLE 6. Explanatory models for changes in grammatical accuracy (error ratios)

		Fit	*F*-test	*B*	*SE B*	95% CI *B*	β	*p*
ΔSubject-verb	Model 1							
	Week 1 performance			–.78	.27	[–1.35, –.22]	–.56**	.008
	L2 contact			<.001	.001	[–.002, .002]	–.01	.969
	Total R^2	.32**	6.42					.005
	Model 2							
	Week 1 Performance			–1.10	.27	[–1.65, –.55]	–.79***	<.001
	L2 contact			<.001	.001	[–.001, .001]	.02	.899
	EIT			–.002	.001	[–.003, –.001]	–.46**	.008
	Total R^2	.49**	8.22					.001
	ΔR^2	.17**	8.35					.008
ΔNumber	Model 1							
	Week 1 performance			–.87	.17	[–1.23, –.51]	–.71***	<.001
	L2 contact			–.001	.001	[–.003, .002]	–.06	.696
	Total R^2	.54***	15.45					<.001
	Model 2							
	Week 1 performance			–1.32	.14	[–1.62, –1.02]	–1.08***	<.001
	L2 contact			–.001	.001	[–.002, .001]	–.05	.601
	EIT			–.007	.001	[–.009, –.004]	–.62***	<.001
	Total R^2	.79***	32.14					<.001
	ΔR^2	.25***	30.49					<.001
ΔGender	Model 1							
	Week 1 performance			–.55	.11	[–.76, –.33]	–.73***	<.001
	L2 contact			<.001	.001	[–.002, .002]	.03	.856
	Total R^2	.52***	14.85					<.001
	Model 2							
	Week 1 performance			–.63	.13	[–.90, –.35]	–.83***	<.001
	L2 contact			<.001	.001	[–.002, .002]	.02	.892
	EIT			–.001	.001	[–.003, .001]	–.16	.367
	Total R^2	.54***	10.12					<.001
	ΔR^2	.02	.84					.367

** p<.01; *** p<.001

Accuracy

Turning to our measures of grammatical accuracy change, the baseline models were significant for all accuracy measures, with Week 1 performance emerging as a significant negative predictor in all three models (results from regression models displayed in Table 6).

The second models with EIT as an additional predictor were also significant, accounting for 49%, 79%, and 54% of total variance in subject-verb, number, and gender agreement error changes, respectively. Notably, increases in variance explained were significant for subject-verb (17%) and number (25%), corresponding to small effect sizes, but not gender (only 2%). For both the subject-verb and number models, EIT was a significant negative predictor (see Figures 4 and 5). The 95% BCa confidence intervals for the robust EIT coefficients did not cross zero in either the

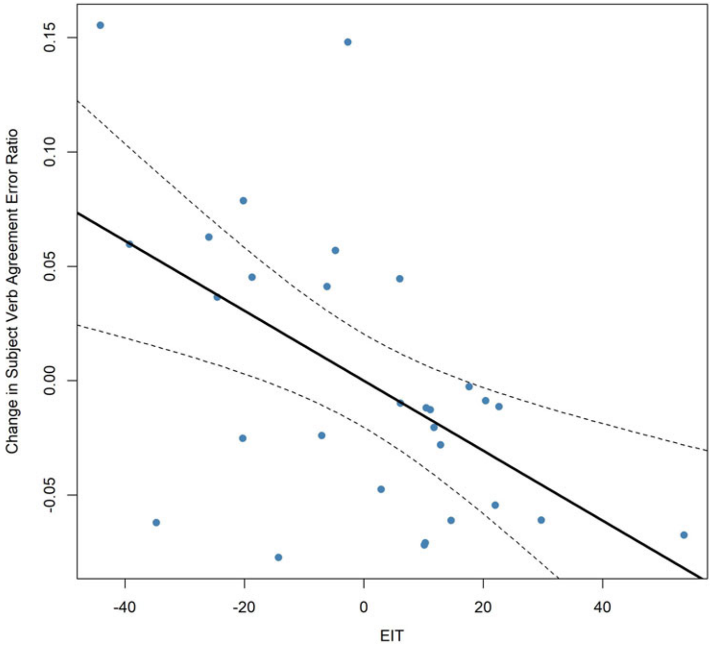

FIGURE 4. Partial effect plot: Change in Subject-verb as a function of EIT.

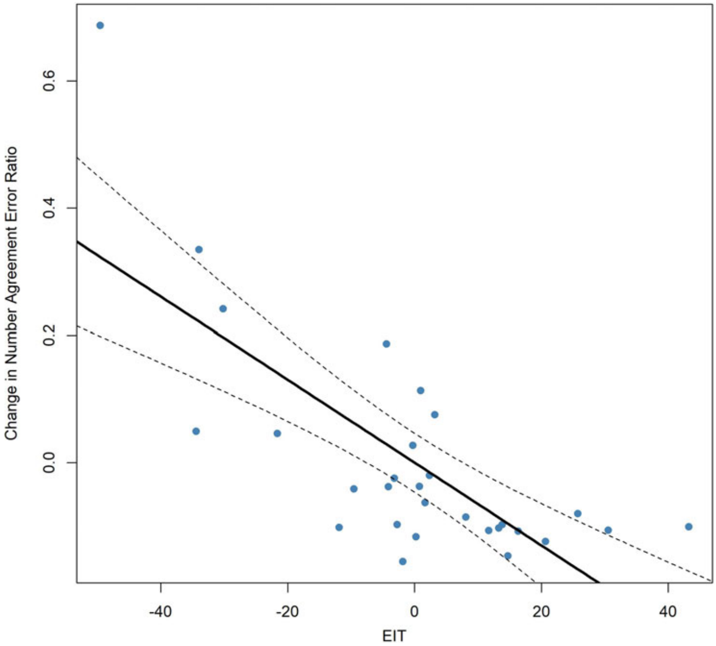

FIGURE 5. Partial effect plot: Change in Number as a function of EIT.

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

subject-verb [–.003, –.001] or number [–.010, –.002] models (but they did for gender [–.003, .002]).

These results indicate that more proficient learners at the onset of SA experienced greater L2 grammar development as evidenced by larger changes in complexity (increase in global complexity and subordination rate) and accuracy (decrease in subject-verb and number, but not gender, agreement errors).

RQ2. THE PREDICTIVE WEIGHT OF INITIAL L2 PROFICIENCY AND INITIAL L2 PERFORMANCE

As indicated by the models computed to address RQ1, learners' Week 1 performance emerged as a significant predictor of L2 change for dimensions of grammatical complexity and accuracy, although different associations were attested than with EIT performance. Namely, increases in complexity and accuracy were greater for learners who started their sojourn with less complex and less accurate (i.e., greater errors in) L2 production, respectively. Yet, gains in both complexity and accuracy were greater for learners who had higher global L2 proficiency (i.e., higher EIT scores) at the onset of SA. The presence of these opposing relationships motivated our RQ2, which examined how the predictive weight of the EIT would be impacted when Week 1 performance was not included as a covariate. Failure to incorporate Week 1 performance was expected to impact model fit, as the negative associations between Week 1 and L2 change scores are indicative of RTM.

Before reporting results for RQ2, we further assessed the prevalence of RTM in our study using *Galton squeeze diagrams* (see Campbell & Kenny, 1999), which provide the most effective graphical representation of this statistical phenomenon (Shanks, 2017). We first computed *z*-scores for the Week 1 complexity and accuracy performance and L2 change variables. Next, we partitioned the data into quartiles based on Week 1 *z*-scores and, subsequently, calculated the mean Week 1 and L2 change values for each quartile. We then plotted these two values in a line graph. As the diagram for MLU scores in Figure 6 shows (see Appendix S3 in the Online Supplementary Materials for additional diagrams),

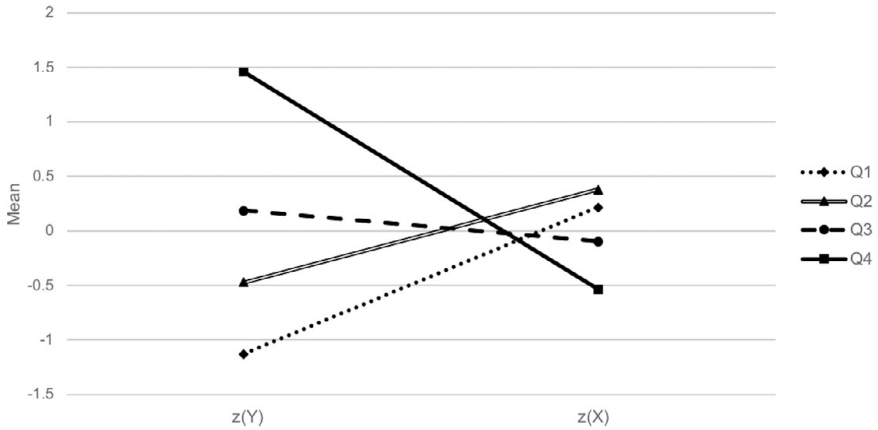

FIGURE 6. Galton squeeze diagram for MLU.

TABLE 7. Explanatory models for changes in complexity unadjusted for week 1 performance

		Fit	*F*-test	*B*	*SE B*	95% CI *B*	β	*p*
ΔMLU	Model 1							
	L2 contact			–.01	.01	[–.03, .02]	–.10	.594
	Total R^2	.01	.29					.594
	Model 2							
	L2 contact			–.002	.01	[–.02, .02]	–.04	.849
	EIT			.02	.01	[–.01, .04]	.28	.147
	Total R^2	.09	1.27					.298
	ΔR^2	.08	2.23					.147
ΔDC	Model 1							
	L2 contact			–.002	.002	[–.01, .002]	–.18	.342
	Total R^2	.03	.93					.342
	Model 2							
	L2 contact			–.001	.002	[–.004, .002]	–.12	.544
	EIT			.003	.002	[–.001, .006]	.28	.145
	Total R^2	.11	1.61					.218
	ΔR^2	.08	2.25					.145

Note. MLU, Mean length of AS-unit; DC, Dependent clauses per AS-unit.

the 25% of data points scoring highest at Week 1 (Q4) are associated with declines in MLU (i.e., negative L2 change scores), whereas the 25% of data points scoring lowest at Week 1 are associated with increases in MLU (i.e., positive L2 change scores). That is, values further away from the mean on the left axis ($z(Y) = 0$, Y = Week 1 MLU) regress to the mean on the right axis ($z(X) = 0$, X = MLU change). This pattern illustrates RTM (see Shanks, 2017; Taraday & Wieczorek-Taraday, 2018), and "implies a negative correlation between initial standing and change" (Campbell & Kenny, 1999, p. 89), as indicated in the regression models. Having established the prevalence of RTM, we turn to our analyses for RQ2.

Complexity

For complexity, the baseline models (including only L2 contact) and the second set of models (adding EIT as a predictor) were not significant for either MLU or DC change (see Table 7 for results). Contrary to what was observed for RQ1, changes in variance explained from Models 1 (baseline) to 2 were minimal and did not reach significance for either measure of complexity (MLU: 8% increase; DC: 8% increase). The 95% BCa confidence intervals for the EIT coefficient were [–.010, .043] and [<.001, .005] for the MLU and DC change models, respectively.

Accuracy

A similar pattern emerged for accuracy, as shown in Table 8, as neither the baseline nor the second set of models were significant for any dimension. Again, in contrast to RQ1, the increases in variance explained that resulted from incorporating EIT as a predictor were minimal and did not reach significance (subject-verb: 3%; number: 0.1%; gender: 11%).

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

TABLE 8. Explanatory models for changes in accuracy (error ratios) unadjusted for week 1 performance

		Fit	*F*-test	*B*	*SE B*	95% CI *B*	β	*p*
ΔSubject-verb	Model 1							
	L2 contact			–.001	.001	[–.003, <.001]	–.34	.064
	Total R^2	.12	3.71					.064
	Model 2							
	L2 contact			–.002	.001	[–.003, <.001]	–.39*	.044
	EIT			–.001	.001	[–.002, .001]	–.19	.311
	Total R^2	.15	2.39					.110
	ΔR^2	.03	1.07					.311
ΔNumber	Model 1							
	L2 contact			–.003	.002	[–.01, .001]	–.32	.088
	Total R^2	.10	3.13					.088
	Model 2							
	L2 contact			–.003	.002	[–.01, .001]	–.33	.093
	EIT			<.001	.002	[–.005, .004]	–.04	.837
	Total R^2	.11	1.53					.235
	ΔR^2	.001	.04					.837
ΔGender	Model 1							
	L2 contact			–.001	.001	[–.003, .001]	–.20	.303
	Total R^2	.04	1.10					.303
	Model 2							
	L2 contact			–.001	.001	[–.003, .002]	–.12	.534
	EIT			.002	.001	[<.001, .004]	.35	.067
	Total R^2	.15	2.42					.108
	ΔR^2	.11	3.64					.067

* $p < .05$

The 95% BCa confidence intervals for the EIT coefficient were [–.002, .001], [–.005, .004], and [<.001, .004] for the subject-verb, number, and gender agreement error ratio change models, respectively.

Results for RQ2 suggest that initial EIT scores (the independent L2 proficiency measure) and initial L2 complexity and accuracy performance (the nonindependent L2 proficiency measure) show differing associations with dimensions of L2 grammar change in our sample. Furthermore, failing to regress out learners' initial L2 grammar performance substantially impacted the predictive weight of initial global L2 proficiency that was observed in results for RQ1.

DISCUSSION

RQ1. INITIAL PROFICIENCY AND L2 GRAMMAR DEVELOPMENT ABROAD

The first research question asked about the extent to which individual differences in initial L2 proficiency explained changes in L2 grammar development in oral production abroad. Initial L2 proficiency, operationalized as EIT performance, accounted for 17–26% of the remaining variance left in changes in L2 grammatical complexity and accuracy after

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

controlling for initial performance and weekly L2 contact. Higher initial global L2 proficiency was associated with larger increases in global complexity and clausal complexity by subordination as well as greater improvements (i.e., decrease in errors) in subject-verb and number agreement accuracy, but not gender agreement accuracy.

These results align with the set of prior studies that have identified a positive role for higher initial L2 proficiency levels in promoting L2 development abroad (e.g., DeKeyser, 2010; Faretta-Stutenberg & Morgan-Short, 2018; Golonka, 2006; Leonard & Shea, 2017). In particular, our findings support and expand earlier research by Faretta-Stutenberg and Morgan-Short (2018), who also employed an EIT to gauge initial L2 proficiency. Among learners participating in semester-long programs, Faretta-Stutenberg and Morgan-Short observed that those with greater initial EIT scores experienced larger improvements in aspects of grammatical gender agreement accuracy in L2 Spanish as measured by a judgment task, but not a production task. Similarly, our study found no evidence for a positive association between higher initial L2 proficiency and gains in gender agreement production accuracy within a shorter sojourn abroad; however, initial L2 proficiency was a positive predictor of morphosyntactic production accuracy gains for other relevant targets in L2 Spanish, namely subject-verb and number agreement.

Our findings are in line with cognitive accounts postulating a positive role for initial L2 proficiency abroad (e.g., DeKeyser, 2010; Lafford & Collentine, 2006; Segalowitz & Freed, 2004). These accounts predict that learners who start their sojourn at higher levels of L2 proficiency experience greater linguistic development abroad, particularly as it pertains to L2 grammar. At least in part, this is thought to result from the fact that more proficient learners have greater attentional resources to allocate to both the formal and functional dimensions of L2 grammar during communication abroad, compared to less proficient learners, whose primary focus may be on meaning (Lafford & Collentine, 2006). More proficient learners also have greater proceduralized knowledge (DeKeyser, 2010) and are better “able to access it rapidly to communicate in spontaneous, real-time contexts” (Loewen & Sato, 2017, p. 4). As a result, they may be more likely to make progress toward strengthening their L2 grammar during SA, experiencing morphosyntactic maturation as signaled by larger increases in L2 complexity and accuracy by the end of an intensive short-term sojourn. In addition to greater language control, more proficient learners may also show enhanced cognitive dispositions associated with L2 learning aptitude, as suggested by research outside of SA contexts (e.g., Li, 2016; Sparks et al., 2012), which could further support their rapid L2 grammar achievement during short stays abroad.

Although the present study did not set out to directly examine proficiency threshold effects in SA, our data can provide some preliminary insights in this regard, at least with respect to the aspects of L2 grammar investigated here. Indeed, the relationships observed with learners’ EIT performance in this study do not appear to suggest a clear minimum global proficiency threshold that learners must reach prior to starting their sojourn to experience gains (e.g., Collentine, 2009; Lafford & Collentine, 2006), nor do they seem consistent with a plateau effect in L2 development for learners with higher starting proficiencies (e.g., Vande Berg et al., 2009). Rather, findings point to a general linear association between initial L2 proficiency and changes in grammatical complexity and accuracy in oral production, without pronounced fluctuations observed in the direction of this relationship at any particular point in the relatively wide proficiency scale represented

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

by our sample. Further research with larger sample sizes, wide proficiency ranges, and a variety of L2 outcome measures is needed to solidify our understanding of proficiency thresholds in L2 development in SA contexts.

In all, findings from this study are not consistent with the notion that learners at higher proficiency levels have no room for L2 improvement or make “less obvious” gains (Regan, 2003). Rather, results suggest that learners with greater basic language cognition at the onset of their sojourn are better positioned to make rapid L2 grammar gains in oral production during an intensive, short stay abroad. An exception to this generalization was observed for accuracy gains in grammatical gender agreement, for which initial L2 proficiency did not emerge as a significant explanatory variable. As noted earlier, gender agreement is known to be particularly challenging for L2 learners even at more advanced levels, particularly when their L1 does not have grammatical gender (e.g., Alarcón, 2010; Grey et al., 2015). We speculate that factors other than initial L2 proficiency (e.g., type of L2 instruction) may be more important in explaining gender agreement accuracy improvements in unstructured L2 oral production tasks such as the one employed in this study.

The rapid development of oral skills is a major focus of SA programs and popularly believed to be a guaranteed outcome. Over the years, empirical research has problematized this notion, revealing that not all L2 domains progress at the same rate during immersion abroad (see Marijuan & Sanz, 2018). Here, we show that initial L2 proficiency predicts changes in several dimensions of L2 grammar in oral production. Our results suggest that students and program administrators can expect aspects of L2 grammar development to be positively related to initial L2 proficiency, and that such development should be considered on a fine-grained scale with regard to oral production, given the nuanced nature of linguistic improvement in short-term SA. Specifically, a more advanced learner can expect to refine their oral L2 skills, with subtle but noteworthy increases in grammatical complexity (increased use of subordinate clauses and longer utterances) and accuracy (more precise morphosyntactic formulation with respect to subject-verb and number agreement), during a brief stay abroad.

RQ2. THE PREDICTIVE WEIGHT OF INITIAL L2 PROFICIENCY AND INITIAL L2 PERFORMANCE

The second research question asked about the directional associations of independent and nonindependent L2 proficiency measures as predictors of L2 grammar gains, and how the predictive weight of initial global L2 proficiency is impacted when initial L2 performance is uncontrolled. The models computed for RQ1 revealed opposite relationships between each type of proficiency measure and L2 change: Whereas higher EIT scores (independent measure) were positively associated with complexity and accuracy development, higher initial L2 complexity and accuracy performance (nonindependent measure) were negatively related to change in these same dimensions. Additionally, when initial performance on outcome measures was not incorporated as a covariate (i.e., when RTM was not built into the analyses), the predictive weight of the EIT observed in the L2 change models was substantially impacted.

The contrasting relationships with L2 change attested here for the two types of proficiency measures may appear counterintuitive under the assumption that both

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

learners' pretest performance and their scores on an independent L2 proficiency test can be reliable estimates of initial L2 proficiency when regressed on L2 gain scores. These seemingly contradictory effects, however, can be reconciled if the negative relationship between nonindependent (i.e., pretest performance) measures and L2 change is taken to manifest RTM. As described earlier, RTM is a common statistical phenomenon whereby an inverse (i.e., negative) relationship is found between baseline scores and the magnitude of change for a given metric (e.g., Campbell & Kenny, 1999; Taraday & Wieczorek-Taraday, 2018; Yu & Chen, 2015). Given the prevalence of RTM in repeated-measures designs, its presence is certainly not unexpected in our sample. Indeed, results from the unadjusted models that did not regress out initial L2 performance scores highlight the pull of RTM across our dataset (Yu & Chen, 2015). Absence of initial L2 complexity and accuracy performance as a covariate resulted in a substantial underestimation of the role of initial global L2 proficiency (as measured by the EIT) in predicting L2 grammar gains abroad.

These findings underscore the importance of (a) employing L2 proficiency measures that are independent from the L2 outcome measures in research seeking to understand the impact of initial L2 proficiency on SLA abroad, and (b) accounting for RTM in the analyses (e.g., by considering learners' Time 1 scores; Yu & Chen, 2015). Although Time 1 performance measures may provide an intuitive and convenient view into learners' starting L2 abilities, their reliability as operational estimates of proficiency in this paradigm may be challenged. Namely, if results indicate a relationship with L2 change scores, interpretation is contested because RTM may stand as sufficient to account for that effect. Thus, going forward, we raise caution against interpreting differences in learners' L2 gains abroad as arising from differences in their Time 1 scores (e.g., Campbell & Kenny, 1999) and advocate for the use of independent global L2 proficiency measures, in line with Tremblay (2011) and Hulstijn (2012, 2015). Independent measures such as the EIT can also promote desirable conditions for research replicability and generalizability (Bowden, 2016), as noted earlier.

Based on our results and the known pervasiveness of RTM across repeated-measures designs, we propose that the contradictory findings from earlier research on the role of initial L2 proficiency may be reconciled by considering which type of proficiency estimate (i.e., independent or nonindependent) was employed as a predictor of L2 gains in each study. Because earlier evidence that lower proficiency learners experience greater L2 development abroad has thus far been reported in studies employing nonindependent measures, it is possible that this evidence may be reevaluated in light of RTM effects. Consequently, we may tentatively conclude that the bulk of findings to date provides broader support for the "more is more" account, such that learners who are more proficient at the onset of immersion appear to experience greater L2 development abroad, including grammar development.

It is also worth noting that the methodological implications derived from this study are relevant for SLA research more broadly, such that findings caution against research practices where unintentionally biased inferences may be drawn by the underrecognized effects of RTM. One potential scenario that can be affected by RTM artifacts is the creation of post hoc subgroups to explore L2 development based on learners' pretest performance (see Shanks, 2017). In this scenario, the effect of the independent variable on L2 change scores could be overestimated among subgroups identified as "low-scorers" at

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

pretest because their initial scores would be more likely to improve based solely on RTM. In sum, examining the effect of the independent variable on one or more subsets of a sample based on their lower (or higher) pretest scores may lead to unwarranted conclusions that disregard the bias introduced by RTM.

Although RTM has not been a matter of much methodological discussion in the field of SLA to date, it has been substantially considered across several other disciplines where, by and large, it is viewed as a psychometric obstacle in making inferences about performance change on the sole basis of baseline scores (e.g., Campbell & Kenny, 1999; Yu & Chen, 2015). Here, we have sought to illustrate the relevance of assessing potential RTM effects when examining the role of initial L2 proficiency in L2 development during SA with the broader goal of contributing to recent calls for increased methodological awareness within SLA (e.g., Marsden & Plonsky, 2018). To address RTM, we plotted our data into Galton squeeze diagrams (Campbell & Kenny, 1999) and followed Yu and Chen's (2015) guidelines to regress out initial values on change scores. While the most conventional approach to address our first research question would entail using posttest scores as criterion variables, change score models were employed because they ensured direct methodological comparability with the previous empirical SA studies discussed here and they also allowed us to easily highlight RTM in our data. However, we note that debates persist around the use of change score models in this type of paradigm and that additional methods are available to account for regression artifacts (see e.g., Campbell & Kenny, 1999; Yu & Chen, 2015).

CONCLUSIONS AND LIMITATIONS

Findings from this study suggest that the persistent variability found in SA research for L2 grammar outcomes, compared to other aspects of learners' L2 (e.g., fluency, vocabulary), may be in part explained by learners' linguistic abilities at program onset. From a theoretical perspective, results support cognitive accounts postulating that more proficient learners are better positioned to experience rapid growth in L2 grammar dimensions of oral production during short-term SA. Our conclusions also have pedagogical relevance, as they can help manage the expectations of various SA stakeholders, including administrators and student participants, regarding the rate of noticeable progress in grammar-related aspects of oral L2 production during short-term programs. Lastly, this study brings to light relevant methodological implications for advancing research on the role of initial L2 proficiency in SA contexts, and calls attention to the largely underrecognized effects of RTM in L2 research.

Findings from this study should be interpreted considering its limitations. One aspect to note is that the scope of the study is limited to a finite set of L2 grammar dimensions. This, in large part, is due to the variability in L2 speech production arising from the open-ended nature of the task and the broad range of proficiencies in our sample. Although the advancement of oral skills continues to be a major focus of SA research, future studies should also consider other L2 outcome measures (e.g., L2 processing behaviors) to better understand how L2 grammar develops for different learners. Additionally, we focused solely on short-term study abroad programs here, which constitute the most popular option for SA participants in the United States (Institute of International Education, 2019). Nonetheless, given that the role of learner individual differences can interact with the intensity of L2 exposure (e.g., Muñoz, 2012), future research may wish to examine how the predictive weight of initial L2 proficiency on L2 grammar development is impacted by program duration. A further limitation is that, although data were collected from two programs, the study comprised a relatively moderate sample size. Indeed, SA programs tend to enroll relatively small cohorts, which can pose persistent sampling challenges, particularly when researchers seek to establish suitably wide ranges for their variables. Although we note that a larger sample would have been desirable here, bootstrapping proved useful for determining the robustness of initial L2 proficiency as a predictor of L2 grammar development in the study.

Despite these limitations, this study has provided novel insights into the role of learners' initial L2 proficiency in L2 grammar development in intensive, short-term SA, both from a conceptual and a methodological standpoint. To further build upon our understanding of the role of initial proficiency in L2 development during SA, additional research across different L2 domains and SA program types is necessary. Toward this end, multisite research, where we plan to combine data from multiple SA programs to secure larger and more representative learner samples, is currently underway.

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/S0272263120000376.

NOTES

[1] RTM effects can be expected when two repeated measures are not perfectly correlated, regardless of the type of measurement scale or the distributional properties of the data (see Shanks, 2017).

[2] In Alicante, located in the Autonomous Community of Valencia, Catalan/Valencian is co-official with Spanish. We examined learners' production for mentions of *catalán* or *valenciano* to consider how Alicante's multilingual context may have impacted student experiences (see Tullock & Ortega, 2017). Only one participant mentioned *valenciano* during Week 1, expressing interest in learning a bit of it during their stay. However, no student reported exposure to Catalan/Valencian in the L2 contact surveys. Host families were instructed to speak only in Spanish with students, and classes were held entirely in Spanish. Furthermore, a faculty program leader reported that students' observations about Catalan/Valencian were limited to noticing this language on street signage. Thus, it appears that participants' contact with Catalan/Valencian was minimal.

[3] Faretta-Stutenberg and Morgan-Short (2018) replaced pronouns and verb endings containing the second-person plural form *vosotros* with *ustedes*, and created an equivalent version of the EIT to allow for pre- and posttest administration.

[4] Production was coded for the first 45 AS-units (the maximum produced in the Intermediate program).

[5] As advised by a reviewer, we also considered examining perfective and imperfective aspect production; however, the limited number of past tense verbs produced by participants precluded this type of analysis.

[6] EIT performance was significantly correlated with performance in all dimensions of grammatical complexity and accuracy at Week 1, in the expected directions (see Appendix S4 in the Online Supplementary Materials).

REFERENCES

Alarcón, I. V. (2010). Gender assignment and agreement in L2 Spanish: The effects of morphological marking, animacy, and gender. *Studies in Hispanic and Lusophone Linguistics*, *3*, 267–300.

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Variables affecting L2 gains during study abroad. *Foreign Language Annals*, *47*, 464–486.

Bowden, H. W. (2016). Assessing second-language oral proficiency for research: The Spanish Elicited Imitation Task. *Studies in Second Language Acquisition*, *38*, 647–675.

Brecht, R. D., Davidson, D. E., & Ginsberg, R. B. (1995). Predictors of foreign language gain during study abroad. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 37-66). John Benjamins.

Brecht, R. D., & Robinson, J. L. (1995). On the value of formal instruction in study abroad. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 318–334). John Benjamins.

Campbell, D. T., & Kenny, D. A. (1999). A primer on regression artifacts. Guilford.

Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition*, *26*, 227–248.

Collentine, J. (2009). Study abroad research: Findings, implications, and future directions. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 218–233). Wiley-Blackwell.

Davidson, D. E. (2010). Study abroad: When, how long, and with what results? New data from the Russian front. *Foreign Language Annals*, *43*, 6–26.

DeKeyser, R. M. (2010). Monitoring processes in Spanish as a second language during a study abroad program. *Foreign Language Annals*, *43*, 80–92.

Dormann, C., & Griffin, M. (2015). Optimal time lags in panel studies. *Psychological Methods*, *20*, 489–505.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman and Hall.

Engle, L., & Engle, J. (2004). Assessing language acquisition and intercultural sensitivity development in relation to study abroad program design. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *10*, 219–236.

Faretta-Stutenberg, M., & Morgan-Short, K. (2018). Contributions of initial proficiency and language use to second-language development during study abroad: Behavioral and event-related potential evidence. In C. Sanz & A. Morales-Front (Eds.), *The Routledge handbook of study abroad research and practice* (pp. 421–435). Routledge.

Foote, R. (2011). Integrated knowledge of agreement in early and late English-Spanish bilinguals. *Applied Psycholinguistics*, *32*, 187–220.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*, 354–375.

Freed, B. F. (1995). Language learning and study abroad. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 3–33). John Benjamins.

Freed, B. F., Dewey, D. P., Segalowitz, N., & Halter, R. (2004). The language contact profile. *Studies in Second Language Acquisition*, *26*, 349–356.

Golonka, E. M. (2006). Predictors revised: Linguistic knowledge and metalinguistic awareness in second language gain in Russian. *Modern Language Journal*, *90*, 496–505.

Grey, S., Cox, J. G., Serafini, E. J., & Sanz, C. (2015). The role of individual differences in the study abroad context: Cognitive capacity and language development during short-term intensive language exposure. *Modern Language Journal*, *99*, 137–157.

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, *15*, 422–433.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins.

Institute of International Education (2019). *Open Doors Data Summary*. https://www.iie.org/Research-and-Insights/Open-Doors/Data/US-Study-Abroad/Duration-of-Study-Abroad

Isabelli-García, C. (2010). Acquisition of Spanish gender agreement in two learning contexts: Study abroad and at home. *Foreign Language Annals*, *43*, 289–303.

Issa, B., & Zalbidea, J. (2018). Proficiency levels in study abroad: Is there an optimal time for sojourning? In C. Sanz & A. Morales-Front (Eds.), *The Routledge handbook of study abroad research and practice* (pp. 453–463). Routledge.

Jeon, E. H. (2015). Multiple regression. In L. Plonsky (Ed.), *Advancing qualitative methods in second language research* (pp. 131–158). Routledge.

Kachigan, S. K. (1991). *Multivariate statistical analysis: A conceptual introduction* (2nd ed.). Radius Press.

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test, Second Edition (KBIT-2): Manual*. NCS Pearson.

Lafford, B., & Collentine, J. (2006). The effect of study abroad and classroom contexts on the acquisition of Spanish as a second language: From research to application. In R. Salaberry & B. Lafford (Eds.), *Spanish second language acquisition: From research to application* (pp. 103–126). Georgetown University Press.

Leonard, K. R., & Shea, C. E. (2017). L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors. *Modern Language Journal*, *101*, 179–193.

Li, S. (2016). The construct validity of language aptitude: A meta-analysis. *Studies in Second Language Acquisition*, *38*, 801–842.

Llanes, A., & Muñoz, C. (2009). A short stay abroad: Does it make a difference? *System*, *37*, 353–365.

Loewen, S., & Sato, M. (2017). *Instructed Second Language Acquisition (ISLA): An overview*. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 1–12). Routledge.

Marijuan, S., & Sanz, C. (2018). Expanding boundaries: Current and new directions in study abroad research and practice. *Foreign Language Annals*, *51*, 185–204.

Marsden, E., & Plonsky, L. (2018). Data, open science, and methodological reform in second language acquisition research. In A. Gudmestad & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 219–228). John Benjamins.

Mochizuki, N., & Ortega, L. (2008). Balancing communication and grammar in beginning-level foreign language classrooms: A study of guided planning and relativization. *Language Teaching Research*, *12*, 11–37.

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, *46*, 610–641.

Muñoz, C. (2012). *Intensive exposure experiences in second language learning*. Multilingual Matters.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*, 555–578.

Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners* [Unpublished doctoral dissertation]. University of Hawai'i at Manoa.

Ortega, L., Iwashita, N., Rabie, S., & Norris, J. M. (1999). *A multilanguage comparison of measures of syntactic complexity* [Funded project]. University of Hawai'i, National Foreign Language Resource Center.

Oswald, F. O., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, *47*, 1343–1355.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 *values.* Modern Language Journal*, 102, 713–731.*

Regan, V. (2003). L'acquisition de la variation native par les apprenants L2 pendant un séjour dans la communauté linguistique native et dans la sale de classe: Le cas des apprenants irlandais du Français langue seconde. *Journal of Educational Thought*, *37*, 283–301.

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*, 173–199.

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, *24*, 752–775.

Sparks, R. L., Patton, J., & Ganschow, L. (2012). Profiles of more and less successful L2 learners: A cluster analysis study. *Learning and Individual Differences*, *22*, 463–472.

Taraday, M., & Wieczorek-Taraday, A. (2018). Regression toward the mean. In B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation* (pp. 1393–1395). Sage.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–336.

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition*, *27*, 173–204.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: “Clozing” the gap. *Studies in Second Language Acquisition*, *33*, 339–372.

Tullock, B., & Ortega, L. (2017). Fluency and multilingualism in study abroad: Lessons from a scoping review. *System*, *71*, 7–21.

Vande Berg, M., Connor-Linton, J., & Paige, R. M. (2009). The Georgetown consortium project: Interventions for student learning abroad. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *18*, 1–75.

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press

Winke, P., & Gass, S. M. (2018). Individual differences in advanced proficiency. In P. A. Malovrh & A. G. Benati (Eds.), *The handbook of advanced proficiency in second language acquisition* (pp. 157–178). Routledge.

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*, 680–704.

Yu, R., & Chen, L. (2015). The need to control for regression to the mean in social psychology studies. *Frontiers in Psychology*, *5*, 1574.

APPENDIX

ORAL PRODUCTION TASK PROMPTS

First week: First impressions

What were your first impressions of Santander/Alicante (the people, the city, your classes)? Can you recall anything particularly interesting or surprising that you did not know or had not expected? How has being in a Spanish language environment influenced your interactions with the people around you? What are your goals and expectations for the upcoming weeks?

Last week: Reflections

Reflect on the goals and expectations you had about studying abroad in terms of classes, language, cultural, your own personal growth/change. How have these goals been realized and/or how did they change during your stay? Did you note particular differences/similarities between the Spanish culture and your own culture (e.g., regarding the family you stayed with, the type of house you stayed in, the food you ate)?

Note: Prompts adapted from the BarSA Project (e.g., Grey et al., 2015) materials.

https://doi.org/10.1017/S0272263120000376 Published online by Cambridge University Press