AB INITIO L2 ACQUISITION OF SOCIOLINGUISTIC VARIATION IN VOCABULARY AND GRAMMAR: A PSYCHOLINGUISTIC APPROACH

By

Elizabeth Huntley

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies – Doctor of Philosophy

2024

ABSTRACT

Sociolinguistic variation (SLV) entails that language is affected by social context (i.e. register, pragmatics). Interest in the acquisition of sociolinguistic variation in a second language (L2-SLV), as a key component of communicative competence, has grown exponentially over the past thirty years (Geeslin & Long, 2014). Researchers have primarily framed L2-SLV as a skill reserved for advanced learners (Geeslin, 2018). However, many language functions at even the novice proficiency level necessitate SLV awareness (ACTFL, 2012b). This is particularly true for learners of diglossic languages such as Arabic. In this study, I expand on L2-SLV research by exploring the acquisition of SLV at the initial stages of learning in a tightly controlled experimental setting. Novice participants studied Mini-Arabii, a miniature language (e.g., Cross et al., 2020; Mueller, 2006) which mimics lexical and morphosyntactic SLV between Modern Standard Arabic (MSA) and Egyptian Colloquial Arabic (ECA). The experiment was designed to explore two fundamental questions about the acquisition of L2-SLV: 1) is learning two registers (i.e., in a curriculum that teaches SLV; "+SLV") more difficult than learning in one register (i.e., in a traditional curriculum that ignores SLV; "-SLV"); and 2) for curricula that include SLV, is it more difficult to learn through an integrated approach (where SLV registers are taught side-by-side), or through a sequential approach (where one register is taught first, and an additional register is taught second)? The training conditions I designed for my study were intended to operationalize these L2-SLV curricular approaches.

For the first comparison, +SLV vs. -SLV curriculum training, participants (novice learners of Arabic) received six exposures to the -SLV vocabulary and grammar set as well as six exposures to the +SLV vocabulary and grammar set (for +SLV, these were divided into three exposures in MSA, three in ECA). This within-participants manipulation mirrored the real-world

trade-off of choosing to cover one vs. two SLV registers within a set period of time. As for the second comparison, participants were assigned to either a +SLV-Integrated or +SLV-Sequential approach as part of their +SLV training above. Participants who received the +SLV-Integrated approach were exposed to MSA and ECA side by side on each day. Those following the +SLV-Sequential approach were exposed to MSA on the first day and ECA on the second day. Learning was measured as progress on the training tasks throughout the three-day, web-based experiment in terms of mean accuracy (both as a binary 0/1 measure and as an approximate measure through the Levenshtein Distance [Levenshtein, 1966]) and mean processing speed (log RT). Vocabulary and grammar knowledge were assessed via production and reception post-tests.

For the first comparison (+SLV vs. -SLV), participants were marginally less accurate but equally fast when they studied in two registers rather than one. As for the second comparison (+SLV-Integrated vs. +SLV-Sequential), participants also responded with comparable speed regardless of which approach they trained in. However, neither approach seemed to give participants a clear advantage in terms of accuracy. As a whole, the current study sheds nuanced light on how L2-SLV can be processed and acquired. For teachers, incorporating SLV at the initial stages of learning does not seem to come at a noticeable cost in learners' overall performance. For researchers, the results invite further investigation into why the acquisition of multiple variants affects accuracy but not processing speed, and what this means for theoretical models of L2-SLV.

Keywords: Second Language Acquisition, Sociolinguistic Variation, Psycholinguistics, Arabic

ACKNOWLEDGEMENTS

It takes a village to raise a graduate student, and I am deeply grateful for everybody who has supported me along the way.

First and foremost, I would like to thank, Dr. Aline Godfroid, for taking me on as an advisee. She is both brilliant and compassionate, and I relied heavily on her guidance to navigate my own development as a scholar and as a human. Likewise, I am incredibly grateful to the members of my dissertation committee: Drs. Kimberly Geeslin, Shawn Loewen, Ayman Mohamed, and Paula Winke. Each member challenged me to engage with the questions explored here in new and critical ways, and they have collectively strengthened both this dissertation as well as my own scholarly approach.

This dissertation would not have been possible without a substantial amount of financial and resource support. I would like to specifically thank Qatar Foundation International, the National Council of Less Commonly Taught Languages, Dr. Tressy Arts of Oxford University Press and her former mentor Dr. Jan Hoogland, the Michigan State University Dissertation Completion Fellowship, and the research and conference support from the Second Language Studies Program. Furthermore, I am deeply grateful for the statistical support I received from Ms. Sichao Wang from the Center for Statistical Training and Consulting, Dr. Bronson Hui, and my ad hoc team of friendly biostatisticians and epidemiologists (Drs. Owais Gilani, Talha Ali, and Paul Christine).

My doctoral journey took many turns, both planned and unplanned, and I am so grateful to the communities that supported and nurtured me along the way. The SSLA "Energy Family" (Drs. Susan Gass, Luke Plonsky, Hima Rawal and Ayşen Tuzcu) provided a space for gratitude and laughter in the most unexpected and beautiful ways. The SLS virtual writing groups of 2020-

2021 kept me sane and focused when the world was upside. Last but not least, my amazing family. You bring me joy every single day.

TABLE OF CONTENTS

Introduction

Human beings, as a species, are social creatures. We form tribes of all kinds to survive and to thrive. There are myriad ways in which we bring about cohesion and signal our group belonging, including through our choices in language use and expression. The field of sociolinguistics documents the ways in which language use is socially patterned. For example, speakers of English would likely identify the following–"hello", "hey", "howdy", and "sup?" – as greetings. However, they will also recognize that these variations are associated with particular social contexts–whether that is degree of formality, geographic location, or ethnic identity. This phenomenon is known as sociolinguistic variation (SLV). SLV refers to "the choices a speaker makes when selecting the forms necessary to convey a message that is appropriate in context" (Geeslin & Long, 2014, p. 3). While the implications of these choices may be readily apparent to native speakers, SLV presents unique challenges for second language (L2) learners.

Sociolinguistic awareness is clearly a key component of communicative and pragmatic competence. At the same time, it is difficult to know if and when learners will be able to handle the acquisition and processing of multiple variant structures. Introducing SLV into a curriculum can help produce learners who are able to comfortably function in L2 social environments, yet it may also slow down or even reverse traditional learning gains. This very practical conundrum has spawned a body of research in the field of Second Language Acquisition (SLA) exploring the extent to which learners are able to acquire socially variant structures. The majority of studies in this area, however, documents the acquisition of SLV for intermediate and advanced learners of Western European languages such as French and Spanish. As such, little is known about L2 acquisition of sociolinguistic variation (L2-SLV) for beginning learners, and across typologically diverse languages. Furthermore, the majority of the studies are observational rather than

experimental, meaning that it is difficult to isolate the effects of key variables of interest that may affect the acquisition of L2-SLV.

The current study seeks to fill these gaps in our understanding. It combines rigorous psycholinguistic research methods with variationist principles to answer questions that are both cognitive and pedagogical in nature. In chapter 1, I provide an overview of the key concepts that inform this study. It highlights how research in SLV, Arabic language pedagogy, and psycholinguistic theory collectively enrich and inform the research questions. In chapter 2, I outline the methodological and analytic approaches used to address these questions. I describe the features of the target language, Mini-Arabii, as well as the training and testing procedures. In chapters 3 and 4, I report on the results of research questions 1 and 2, respectively. In chapter 3, I present learning outcomes in the production and reception of vocabulary and grammar in single-register training (-SLV) versus dual-register training (+SLV), and in chapter 4, I compare the outcomes of a sequential approach (dual-register training, one learned after the other) versus integrated approach (dual-register training, registers learned side-by-side). In chapter 5, the final chapter, I narrate the results within the broader context of the cognitive and pedagogical implications discussed in Chapter 1. In doing so, through this dissertation I offer new tools to expand the L2-SLV research agenda as well as provide instructors with empirical data to help inform their curricular choices.

Chapter 1: Literature Review

Language is a fundamentally social phenomenon. While the richness of human language allows for any number of unique expressions at the individual level, on a macrosocial level different groups of people tend to use similar patterns of expression in systematic ways. This systematicity of language use among speech communities is known as sociolinguistic variation. For L2 learners, sociolinguistic variation can present an additional challenge in the acquisition process. This challenge is particularly difficult for L2 learners of a diglossic language, such as Arabic, where multiple registers of language must be acquired. The current study explores the acquisition of sociolinguistic variation in L2 Arabic.

Sociolinguistic Variation and Second Language Acquisition

The field of sociolinguistic variation (SLV) is based on the observation that members of a speech community tend to produce similar patterns of language relative to specific social contexts. These patterns of usage among different speech communities are still considered to be part of the same language "code"; i.e. African American Language (also known as African American Vernacular English) and Bostonian English are both readily identified as registers of "English." Registers themselves are not monolithic—variation can furthermore exist within a language register. Speakers of African American Language, for example, display distinct speech patterns across gender, region, profession, and social identity (King, 2020).

For native speakers of a language, variation (both between and within registers) can therefore index a variety of social phenomena ranging from social class (e.g., Labov, 1972) to geographic region (e.g., Boberg, 2000) to sexual orientation (e.g., Leap, 1995). Thus, sociolinguistic variation refers to language variation stemming from social contexts that are

external to language. This is distinct from *linguistic* variation, which refers to language variation stemming from language-internal contexts such as grammatical category affecting stress patterns (Anttila, 2006). Learning the rules of usage and interpretation of SLV can pose a challenge for L2 learners. Learners need to not only understand how native speakers use and interpret SLV, but also when and how to produce it appropriately (Geeslin, 2011; V. Regan et al., 2009).

Interest in sociolinguistic variation in a second language (L2-SLV), as a key component of communicative (Canale & Swain, 1980) and pragmatic competence (Nassif & Al Masaeed, 2020), has grown exponentially over the past thirty years (Geeslin & Long, 2014). Recent studies have explored the roles of increasingly nuanced variables such as perceptional categorization (Bedinghaus, 2015) and working memory (Zahler, 2018) on target-like acquisition of L2-SLV phenomena ranging from phonology (Pozzi & Bayley, 2020) to morphosyntax (Linford, 2016; Rehner et al., 2003). Findings in general indicate that learners who have spent time studying abroad (Linford, 2016; Pozzi & Bayley, 2020) and who have higher proficiency levels (Zahler, 2018) tend to reach higher levels of L2-SLV competence in both perception and production. For example, Bedinghaus (2015) explored the effects of learning environment (at-home versus on study abroad) on the perception of /s/-aspiration, a sociolinguistic variant in Western Andalusian Spanish. Employing a forced-choice identification task and a lexical decision task, he found that learners at baseline (before the treatment of either studying abroad or studying at-home) have more difficulty processing the aspirated dialectal variant than they do the standard variant. Moreover, learners who had spent time abroad responded to the experimental tasks more accurately and quickly than those who remained at their home university. The difference in reaction times between stimuli containing /s/-aspiration and control stimuli was hypothesized to be the processing cost for lexical access of forms containing phonological variants. Bedinghaus

thus connected phonological variation with lexical acquisition, arguing that exposure to the dialectal variant lead to faster and more accurate lexical access for words containing that variant. Such differences in lexical access are not only of theoretical importance, they also hold consequences for learners' abilities to participate in real-time communication as an index of how quickly a word can be retrieved (Godfroid, 2020).

As illustrated from the examples above, researchers have primarily conceived of L2-SLV as a skill reserved for advanced learners (Geeslin, 2018), such as those on study abroad (Geeslin & Garrett, 2018; V. Regan et al., 2009). As such, few have explored the acquisition of L2-SLV in the early stages of learning. This is despite the fact that many of the most basic language functions at even the novice level of proficiency necessitate SLV awareness (ACTFL, 2012a). Thus, one of the goals of the current study is to explore how SLV is processed and acquired in the earliest stages of learning. Furthermore, the bulk of research on L2-SLV is observational rather than experimental in nature, meaning that extralinguistic features such as learner attitudes (L. B. Schmidt, 2020) and study abroad social networks (Kennedy Terry, 2017) likely affect acquisition (The Douglas Fir Group, 2016). As noted by Geeslin (2011), however, "researchers interested in cognitive models of language must explore how it is possible for the human mind to handle the storage and production of variable structures…" (p. 462). The current study answers Geeslin's call by adopting psycholinguistic methods to isolate the effects of SLV on L2 acquisition and processing with a laboratory setting. Lastly, the majority of studies have focused on the acquisition of L2-SLV in Spanish, French, and English. Few researchers have explored the acquisition of SLV in diglossic languages such as Arabic. Thus, the current study explores the acquisition of SLV in novice L2 learners of Arabic.

Diglossia, Arabic, and L2-SLV

While variation exists in all human languages, the degree of variation in Arabic has long been considered a special case by linguists (Al-Wer et al., 2009; Ferguson, 1991; Owens, 2003; R. Schmidt, 1974). Ferguson (1959) cited Arabic as an example of linguistic diglossia, a multilingual system in which:

> …in addition to the primary dialects of the language (which may include a standard or regional standards), there is a *very divergent*, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature… which is learned largely by formal education and is used for most written and formal spoken purposes but is *not used by any section of the community for ordinary conversation*. (p. 336, emphasis added)

Ferguson conceived of diglossia as a unique language situation contrasting with bilingualism on the one hand, and "standard-with-dialect" on the other (cf. Fishman, 1967). Bilingualism, in the Fergusonian understanding, refers to the simultaneous existence of two or more codes (i.e. distinct "languages"), such as the usage of French, Arabic, and Tamazight in Morocco. "Standard-with-dialect," meanwhile, describes a multilingual situation in which multiple registers of a single code (such as General American English and African American Vernacular English) are used by a speech community. Diglossia, like "standard-with-dialect", involves the simultaneous usage of related registers. However, unlike both bilingualism and "standard-with-dialect", in diglossia there is no speech community which naturally uses the superposed prestigious register as a native language. As such, the functional boundaries of appropriate usage are somewhat more sharply delineated in diglossia than they are in bilingualism and "standard-with-dialect" multilingual situations (Mejdell, 2017).

For Arabic diglossia, the broadest contrast is typically drawn between Modern Standard Arabic (MSA) and spoken regional registers.[1] MSA, with its roots in the language of the Qur'an, functions as the highly codified prestigious register. As such, it is used in formal contexts such as literature, religious texts, newspapers, and political speeches. Conversely, the regional registers of Arabic are typically used for everyday, spoken interactions (Holes, 2004). Variation between registers (i.e., between MSA and the spoken regional colloquials), as well as within registers (between the spoken regional colloquials themselves) occurs on almost all levels of language, from phonology to lexis to morphosyntax (Brustad, 2000; Watson, 2002). Badawi (1973), for example, identified three levels of variation within the spoken register of Egypt based on the relative usage of characteristic phonemes and morphology. He associated these levels with educational background: educated spoken Arabic, semi-literate spoken Arabic, and illiterate spoken Arabic. He noted, however, that these levels represent a continuum of language patterns rather than distinct entities (Badawi, 1985). Indeed, literate native speakers typically maneuver between MSA and a colloquial register, with respect to their sociolinguistic functions, on a daily basis.

The (generally) non-overlapping functions of Arabic sociolinguistic varieties are largely related to speakers' perceptions of their relative prestige. For example, MSA is referred to in Arabic as *al-fuṣʕaː* (lit: "the most eloquent [language]"), whereas spoken colloquial registers are referred to as *al-ʕamːiːjaːt* ("the common [languages]"). While there are some examples of written colloquial Arabic (e.g., Høigilt & Mejdell, 2017), colloquial registers are typically

[1] Like sociolinguistic variation in most languages, variation can occur both between registers of a language code (i.e. geographically-based registers) as well as within registers themselves (i.e. variation by age and gender of the speaker, context of the interaction, etc.). Other examples of within-register variation that are commonly researched in Arabic linguistics are confessional (religiously-based) varieties (e.g., Blanc, 1964; Holes, 1984) and sedentary versus rural (Bedouin) varieties (e.g., Palva, 1984). See Al-Wer and Horesh (2019)for an extended overview.

considered highly unsuitable for written expression (Owens, 2013). Conversely, native speakers of Arabic, even from different regional speech communities, overwhelmingly use colloquial registers, and not MSA, for spoken communication (Bassiouney, 2009). MSA and the spoken colloquials are complete and fully functioning registers with their own unique yet interrelated lexical and grammatical systems.

The complementary nature of diglossia makes Arabic an ideal language for testing the acquisition of L2-SLV at all levels of language. At the same time, diglossia poses unique challenges for L2 learners and teachers. Arabic curricula have traditionally privileged MSA instruction at the expense of spoken registers (Ryding, 2013; Younes, 2015). The exclusion of dialects from formal instruction is due to a variety of factors, ranging from perceptions that dialects aren't appropriate for the classroom to practical issues of time and lack of curricular resources (Hashem-Aramouni, 2011). However, as the number of students interested in learning Arabic as a spoken language is increasing (Husseinali, 2006; Modern Language Association, 2016), the traditional MSA-only curriculum has been criticized for producing students who can dissect metaphors in contemporary Moroccan feminist literature but don't know how to direct a taxi driver to their desired destination (Palmer, 2007; Parkinson, 1985). Trentman (2013) explored language development as a function of L2 learners' social expectations and practice while studying abroad in Egypt. Although participants had one to two years of formal instruction prior to their arrival, Trentman concluded that "most were unable to function on a basic interactional level when they arrived in Egypt. This problem was exacerbated by the fact that they had primarily studied MSA at home, yet were expected to interact in Egyptian [colloquial] Arabic [upon arriving] abroad" (p. 560). Participants in Trentman's study had encountered Arabic SLV *sequentially*; that is, by learning in a traditional, MSA-only curriculum

first and then focusing on ECA once abroad. It seems that a sequential approach to SLV (MSA first, colloquial Arabic second) was insufficient to help students, at least initially, navigate diglossia with native speakers of Arabic.

An increasingly popular alternative to traditional, MSA-only instruction is the "integrated approach," in which formal and spoken registers of Arabic are taught side-by-side (Al-Batal, 2018; Younes, 2015). Curricula may be integrated by presenting translation equivalents of the MSA and spoken registers (e.g., Brustad et al., 2011) or by using registers in accordance with their sociolinguistic functions such that written passages are in MSA while listening exercises are mainly in a dialect (e.g., Younes et al., 2019)[2]. The spoken registers taught in integrated textbooks are simplified versions of a variety of educated spoken Arabic associated with a major urban center, such as Cairo or Damascus (Brustad et al., 2011, p. xxi; Younes et al., 2019, p. xxi) (Brustad et al., 2011, p. xxi; Younes et al., 2019, p. xxi). In the Al-Kitaab integrated textbook series, for example, register is represented through different colored text: vocabulary and grammar points in Egyptian Colloquial Arabic are depicted in green, Levantine Colloquial Arabic in purple, and MSA in blue. Any forms that are shared between the varieties are presented in black (Brustad et al., 2011, p. xvii). In the 'Arabiyyat al-Naas integrated textbook series, the register of a form is not overtly identified apart from in the summary chapter of each unit in a section known as the "Sociolinguistic Corner" (Younes et al., 2019, p. 87). The decision to not draw attention to the differences between registers is intentional, in order to present Arabic

[2] Typically, curricula which include SLV follow the model of "MSA + 1" in which students gain productive and receptive skills in two registers: MSA and a spoken colloquial register (Vanpee, Forthcoming). An alternative to this model is known as "multidialectal approaches" (Al Masaeed, 2022; Trentman & S'hiri, 2020) In a multidialectal approach, in addition to gaining productive and receptive skills in MSA and a primary spoken register, students are exposed to spoken registers from other regions of the Arabic-speaking world. The goal is to develop receptive skills in multiple spoken registers, akin to how native speakers of Arabic can understand spoken registers other than their own. Although multidialectal approaches are not as common as the "MSA + 1" approach, they were the subject of a 2022 AAAL colloquium titled "Translingual Approaches in World Language Education: Perspectives from Arabic Learning Contexts" (Al Masaeed et al., 2022).

to students as "one communication system, not two" in which the varieties fulfill complementary functional roles (Younes, 2018, p. 25).

The integration of spoken dialects into standard L2 Arabic curricula is an extremely divisive issue within the Arabic teaching community (Younes & Huntley, 2019). Alhawary (2013), in a position paper, argued that the integrated approach "increase[s] the learning burden for the learner who would be faced with too much input to comprehend and proceduralize." This statement has been partially challenged by research documenting that L2 learners can indeed demonstrate appropriate SLV usage (Nassif & Al Masaeed, 2020; Soliman, 2014), and that an awareness of the linguistic features of dialects assists in overall comprehension of texts in an unfamiliar dialect (Trentman & S'hiri, 2020). The bulk of research on acquisition of L2-SLV in Arabic, however, reports on learners who have studied Arabic for a minimum of two years. Thus, little is known about the acquisition of Arabic L2-SLV at the beginning stages of acquisition. One exception is Huntley (forthcoming), who investigated *ab initio* vocabulary acquisition in MSA and a dialect. In a lab-based study, I operationalized the traditional curriculum (MSA-only) and the integrated curriculum (MSA + a dialect [Egyptian Colloquial Arabic]) as two distinct training conditions. The results of the study did not support Alhawary's (2013) claim, at least for lexis, as no significant differences were detected between training conditions apart from a slight processing speed benefit of ECA words. As noted previously, however, SLV in Arabic is not limited to vocabulary alone. Thus, the current study builds off of Huntley (forthcoming) by exploring how L2 learners can acquire both lexical and grammatical SLV. Furthermore, it compares learning by operationalizing the curricular approaches described above as training conditions.

Theoretical frameworks for L2-SLV

To date, there are no major theoretical frameworks which directly address L2-SLV from a cognitive perspective. Typically, researchers interested in studying the acquisition of L2-SLV have turned towards analytic frameworks, such as variationist approaches, to understand the phenomena of interest (Bayley & Tarone, 2012). Variationist approaches allow researchers to create probabilistic models for the frequency of SLV features in as a function of social and linguistic constraints. By comparing the distribution of these features in L1 and L2 datasets, researchers following a variationist approach can draw conclusions about the degree to which independent variables of interest (such as proficiency or attitude) affect the frequency of use (such as an allophonic variant) in these populations (e.g., Kennedy Terry, 2017; B. Regan, 2022). While such approaches offer a rich perspective on how and why L2 speakers develop knowledge of SLV, they are primarily applied as a granular descriptive tool to the output of more advanced learners. As such, they do not directly lead to a theoretical framework for considering more broadly how the phenomenon of SLV, that is, the existence of multiple variants whose usage is socially dictated, is acquired by L2 learners. However, this does not mean that there are no frameworks which can speak to issues of quantity, frequency, and ordering—issues inherent to understanding how multiple variants can be acquired. The following section draws on elements from paired associate learning, the Declarative-Procedural Model, Skill Acquisition Theory, and Usage-Based approaches to provide a theoretical framework for understanding the L2-SLV acquisition process.

Frequency-Based Approaches to Learning: More is Less?

At its most elemental level, integrating L2-SLV into a curriculum entails that learners must assimilate not just one but multiple forms or rule systems for a single concept. As a "numbers game," this is an ostensibly more difficult[3] task–the number of forms to be learned within a limited set of exposures has been doubled. In other words, the ratio of type frequency (examples of a form) to token frequency (instances in which that form appears) increases in the L2 input, making the learning of each more type difficult (Ellis, 2009; Ellis & Collins, 2009).

Paired associate learning is a classic learning paradigm which may speak to this additional level of difficulty (Arndt, 2012). Paired associate learning is based on repetition – the more times a form and meaning are associated, the more likely they are to be connected in the mind of the learner (Mandler & Huttenlocher, 1956; Webb, 2007). From the "numbers game" perspective, adding more forms to the conceptual meaning should require more repetitions to attain the same level of connection, and is thus likely more difficult. Recall from the previous section that Huntley (forthcoming) probed this assumption by comparing learning outcomes for associating one versus two Arabic allophonic variant forms to a single meaning. In the analysis, I failed to find significant differences for accuracy and processing speed between words learned in one register (a traditional curriculum) vs. two, apart from a slight processing speed advantage in form recognition of Egyptian Colloquial Arabic items. Thus, for lexical items, it remains unclear whether or not associating additional forms with a singular concept, the crux of L2-SLV, is universally more "difficult".

Although paired associate learning is most typically used in the domain of laboratory-based L2 vocabulary research, the challenge of mapping multiple structures to a meaning could

[3] The term "ostensibly" is used here because "difficulty", as a measurable construct, is not well-defined in SLA (see Housen & Simoens, 2016, for a discussion)

also be extended to grammar knowledge. Again, the basic assumption would hold that requiring the association of more rules to express a single meaning would be more difficult. While this assumption has not been directly tested, it is in keeping with rule-based theories of grammar learning such as the Declarative-Procedural Model (Ullman, 2001b, 2001a, 2004) or Skill Acquisition Theory (DeKeyser, 1997, 2017). Within the Declarative-Procedural Model, adding an additional layer of rule to be learned (such as checking for "social context" before deciding which set of verb endings to apply) would be more cumbersome. Over time and with sufficient practice, this extra step in morphological processing could be proceduralized (as posited by Skill Acquisition Theory), but there are no shortcuts in the acquisition process. Again, more content in the input would require more practice for mastery.

In sum, there are no models that directly speak to the phenomenon of L2-SLV. Since L2-SLV requires that multiple forms be associated with concepts depending on social context, the notion of frequency—both token frequency and type frequency—can offer insights into this potential learning burden. In essence, if language learning is about associating a linguistic structure with a concept, then pairing two structures with a concept is ostensibly more difficult and will lead to worse outcomes. Of course, not all agree that second language acquisition is strictly a "numbers game" where more forms results in less learning. The next section discusses Usage-Based approaches, which offer an alternative model to type versus token frequency for how multiple forms and concepts might be cognitively associated in mutually supportive ways.

Usage Based Approaches

As discussed previously, there are no theoretical approaches which directly speak to the phenomenon of L2-SLV. However, Geeslin and Long (2014) point to several cognitive

approaches which may be applicable to the task of mapping multiple forms to a singular meaning. Of particular relevance to the current study are Usage-Based approaches. Akin to variationist approaches, Usage-Based approaches (UBA) posit a probabilistic model of language acquisition: the amount of input and the creation of associations within it shape acquisition. While the specific, testable hypotheses typically utilized by UBA are beyond the scope of the current study, many key UBA constructs can be borrowed to understand the rationale and methodology adopted here to address L2-SLV. The relevant constructs are discussed below:

Constructions and Associative Language Learning in UBA.

In UBA, language is made up of constructions. Constructions are, in essence form-function-meaning units. Constructions exist at nearly[4] every level of language, from simple bound morphemes (such as adding *-er* to the end of verbs like *paint* and *run* creates active participle nouns like *painter* and *runner*) to larger syntactic frames (such as "noun-verb-noun-noun" is the construction for ditransitive verbs in phrases like "she gave him a high-five" or "he baked his mother a cake"). The mapping of associations between form, meaning, and function in the learner's mind is created through frequency of association in the input. In fact, UBA posit that learning is a fundamentally associative process: each time a form is paired with its meaning in the input, their mutual association is strengthened in the learner's mind. Thus, frequency of exposure plays a paramount role in establishing these associations.

[4] At the time of writing, I could not find evidence of this being extended to the level of phonemes, but presumably association of input would extend to phoneme-grapheme correspondence. This association is relevant in the context of allophonic SLV (e.g., Bedinghaus, 2015; Pozzi & Bayley, 2020; B. Regan, 2022). The effect of SLV on phoneme-grapheme correspondence has been explored in the context the of phonemic awareness for native speakers of Arabic. A series of studies by Saiegh-Haddad and colleagues suggests that SLV impacts the development of L1 reading skills (Saiegh-Haddad, 2003, 2004; Saiegh-Haddad & Geva, 2008; Saiegh-Haddad & Taha, 2017; Schiff & Saiegh-Haddad, 2018).

Within an SLV context, register variants would be considered an example of constructions. As such, register would thus strongly dictate when variant constructions co-occur. This could be realized in the form of a social tagging, where "register" would be part of the form-meaning-function unit (K. L. Geeslin, personal communication, September 30, 2022). There is some evidence of this association with L1 speakers of Arabic. For example, Ibrahim and Aharon-Peretz conducted an auditory lexical decision task study for L1 Arabic speakers in Israel (who also speak Hebrew). They found that found spoken colloquial Arabic had a stronger priming effect than either Modern Standard Arabic or Hebrew on the recognition of spoken colloquial Arabic items[5] (2005). For L2 learners, studying variants in context in an integrated approach (where registers are learned side by side) would help build up the associative strength of these social tags. Conversely, a sequential approach would mean that learners would not develop (whether explicitly or implicitly) a sensitivity towards register as part of the form-meaning-function unit until later in their studies. They would have to go back and re-apply this "social tag" to previously learned words. Whether or not this process is actually more difficult than learning in an integrated approach is an empirical question the current study seeks to answer.

The Role of Prior Knowledge.

UBA also acknowledge the effect of prior knowledge plays on determining what to parse in the input. For example, if a learner's L1 is considered to be relatively simple in terms of morphology, then the learner will likely be less attuned to morphological markings in the L2 (e.g., Ellis & Sagarra, 2011; Nassif, 2019; Sagarra & Ellis, 2013). This concept, referred to as

[5] Furthermore, Hebrew and MSA had similar priming effects. The authors argue that, as such, MSA is more akin to an L2 than a first language (see also Ibrahim, 2009).

"learned attention," has been explored in studies on competing temporal cues. Findings show that L1 speakers of English, a morphologically "impoverished" language, tend to look for temporality in lexical items (i.e., adverbs of time) rather than verb markers (i.e., past tense). Their sensitivity toward morphology is thus "blocked" by their prior language experience (Ellis & Sagarra, 2010).

Note that all of these studies were designed to explore the role of L1 knowledge on L2 acquisition. UBA thus provide room to hypothesize how the prior establishment of associative networks may facilitate (or mitigate) the subsequent development of new associations in the learner's mind. As discussed previously, incorporating SLV into the L2 curriculum means that learners must map multiple surface variants to the same construction. Part of the fundamental empirical question explored in this dissertation is whether it is better to first establish a solid "base" from which to build on (an additive model, as in the sequential approach), or to build up the associations between these form-meaning-function units simultaneously as part of a complete language system (a symbiotic model, as in the integrated approach). UBA would suggest that prior knowledge could potentially block the addition of new information in the mind of the learner. For example, if the surface form "sinjaab" has already been associated with the concept of "squirrel", it may be more difficult to later add on the allophonic variant "singaab" than if the two forms are conceptually associated from the beginning. The same line of reasoning could extend to morphological variants as well.

In sum, UBA provide a powerful framework for testing specific hypotheses about SLA. While UBA are not typically used in SLV lines of research, their underlying constructs on how language is learned (constructions as form-meaning-function units, associative learning and social tagging, the role of prior knowledge) can be extended to SLV to provide insights on how

variation can be acquired by L2 learners. These different approaches to acquiring SLV are operationalized in the current study through the creation of Mini-Arabii, as discussed in the following section.

Artificial Languages as a Tool in SLA Research

Artificial languages (ALs) are language-like systems designed to mimic, to varying degrees, elements of natural languages (Morgan-Short, 2020). They have been called "'test tube' models of natural language" (Morgan-Short et al., 2012, p. 3), in that ALs allow researchers to explore the underlying processes and products of second language acquisition in tightly controlled settings. A further benefit of ALs is that learners are able to reach high levels of accuracy in relatively short periods of time (Tagarelli et al., 2019). Neurolinguistic evidence from electrophysiological measures (ERP) and functional magnetic resonance imaging (fMRI) suggests that learners of ALs can exhibit native-like brain processing patterns, adding a degree of ecological validity to the use of ALs in SLA research (Friederici et al., 2002; Petersson, 2004).

ALs have primarily been adopted in experiments at the nexus of psychology, linguistics, and neurocognition, such as for exploring the relationship between implicit and explicit learning and knowledge. Researchers can adopt or modify an AL to target their own particular areas of inquiry. For example, Walker et al. (2020) were interested in how adult learners can simultaneously acquire elements of vocabulary and grammar through exposure alone, as is often the case in immersion settings. They developed a semi-artificial language composed of 16 pseudowords (nouns, verbs, adjectives, and case markers) that approximated Japanese syntax (see also Rebuschat et al., 2021). Most importantly, participants' exposure to the target language combined training and testing in the same experimental blocks. As such, the researchers were

able to track implicit-statistical learning over the course of the experiment. The authors found that participants were indeed able to simultaneously acquire the grammar and vocabulary of their semi-artificial language through exposure alone.

A subset of ALs relevant to the current study are "miniature languages." Miniature languages represent trimmed-down versions of a natural language consisting of a limited set of novel, meaningful forms and structures. Mueller (2006, p. 235) describes mini-languages as "L2 in a nutshell," in that researchers can probe the acquisition of higher order language (i.e. sentence-level processing and production) with a greater degree of ecological validity than what artificial languages can offer. Cross et al. (2020) developed Mini-Pinyin, to explore how elements of Mandarin syntax can be acquired in an implicit-statistical learning paradigm. Like Mandarin, Mini-Pinyin word order can be sequence-based or dependency-based through the use of coverbs and classifiers. Through a grammaticality judgment task, Cross et al. found that L2 learners can indeed develop grammatical sensitivity to the rules of Mandarin-syntax through exposure alone. The authors argued that research on typologically diverse miniature languages such as Mini-Pinyin is crucial for expanding the generalizability of empirically-driven theories in Second Language Acquisition.

The current study builds on the push to research the acquisition of typologically diverse languages by developing the miniature language of Mini-Arabii. Mini-Arabii is a miniature language system based on simplified Arabic, both MSA and the spoken colloquial registers. The vocabulary and grammar of Mini-Arabii provide a tool for studying the effects of lexical and grammatical variation on acquisition within an experimental context. As such, Mini-Arabii can be used to explore how SLV in lexis and morphosyntax is acquired and processed by learners. Furthermore, the variant elements can be flexibly distributed to operationalize real-world

curricular approaches to SLV. As such, the current study will offer both theoretical and pedagogical implications for how the acquisition of L2-SLV.

The Current Study

The aim of the current study is to explore L2-SLV acquisition through a psycholinguistic lens using Arabic, a diglossic language. The goal is to create a miniature language, Mini-Arabii, to investigate how SLV is processed and acquired at the lexical and morphosyntactic levels. Variation is operationalized as learning words and phrases in either one register (-SLV) or two registers (+SLV). Each register has its own set of grammar and vocabulary. In my experiment, I operationalize three different curricular approaches to SLV, as follows:

1) -SLV (in which vocabulary and grammar following a single set of rules from one register is studied, representing a traditional curriculum which does not introduce SLV), and

2) +SLV-Integrated (in which two registers, each following their own set of rules, are taught side-by-side, representing the introduction of SLV early on in L2 instruction), and

3) +SLV-Sequential (in which one register is taught first, then a second register is taught second, representing a traditional no-variation curriculum followed by study abroad).

In keeping with the operationalization of each curricular choice and its associated trade-offs (studying one vs. two registers within the same time constraints), participants received six exposures to -SLV vocabulary and grammar, and six exposures to +SLV vocabulary and grammar (three in each register).The comparison of -SLV vs. +SLV curricula is a within-subjects manipulation, where each participant serves as their own control. In contrast, the comparison of +SLV approaches (Integrated vs. Sequential) is a between-subjects manipulation. Half of the participants are assigned to +SLV-Integrated, and the other half are assigned to

+SLV-Sequential. The two registers in the +SLV conditions will henceforth be referred to as "MSA" and "ECA" (referencing "Modern Standard Arabic" and "Egyptian Colloquial Arabic" respectively).

Acquisition were measured throughout the experiment in terms of accuracy and processing speed (reaction time). The experiment lasted three days, which included two days of training and a third day of posttests. The conditions across the two experiments were designed to answer the following research questions:

RQ1) How does learning one register (-SLV) compare to learning two registers (+SLV)?

Hypothesis 1: Based on frequency-based learning approaches, learning one register (-SLV) will lead to superior outcomes compared to learning two registers (+SLV). However, based on the findings from Huntley (forthcoming), the degree of superiority may vary according to the outcome measure (reaction time versus accuracy [binary or approximate]).

RQ2) To compare different approaches to +SLV learning, what is the effect of training condition (Integrated vs. Sequential) on register knowledge both individually (MSA or ECA) and collectively (combined)?

Hypothesis 2: Based on Usage-Based approaches, learning in an Integrated condition compared to in a Sequential condition will lead to superior outcomes on MSA and ECA register knowledge. However, based on the findings from Huntley (forthcoming), the degree of superiority will vary according to the outcome measure (reaction time versus accuracy [binary or approximate]).

Chapter 2: Methodology

Participants

Participants were recruited from a large Midwestern public university. First, a call for participation was sent out via the Office of the Registrar to all undergraduates who listed the United States as their home country. This residency restriction was chosen to target L1 speakers of English. From this initial call, 1,396 people filled out the eligibility survey. The eligibility survey screened for participants who possessed the following criteria: 1) normal or corrected-to-normal vision, 2) normal or corrected-to-normal hearing, 3) between the ages of 18-35; 4) right-handedness, 5) no prior exposure to Arabic, Turkish, Hindi / Urdu, or Farsi, 6) experience studying a language with grammatical gender (such as French or Spanish), 7) self-identify as L1 English speakers, and 8) currently enrolled in or have completed an undergraduate degree. 126 eligible participants were invited to join the study, and 80 completed the study. Of those, four were screened out for poor data quality, resulting in a total of 76 participants (38 female, 33 male, 5 non-binary). The average participant age was 21.14 years old (SD: 2.58 years). 13 participants reported an additional language as an L1; these languages were Bengali, Bosnian, French, German (n = 2), Gujarati (n = 3), Igbo, Sinhalese, and Spanish (n = 3). The languages with grammatical gender that participants reported studying included Czech (n = 1), French (n = 12), German (n = 15), Gujarati (n = 1), Italian (n = 2), Latin (n = 1), Marathi (n = 1), Russian (n = 3), Slovak (n = 1), and Spanish (n = 58). Participants reported a wide range of explicit learning experiences of these languages, from studying a few days a week on Duolingo to having lived abroad. In all, these participants were selected because they represent the typical student who embarks on L2 Arabic study in the Global North, and who are thus affected by the curricular choices operationalized in this research.

Materials

For the purpose of this study, I developed Mini-Arabii, a miniature language that mimics key elements of SLV in Arabic. All linguistic features of Mini-Arabii, ranging from phonology to morphosyntax, were derived from Arabic. They were modified for the purposes of the experiment, in an effort to balance the experimental rigor and ecological validity of the study. The following sections detail the methodological choices made when constructing Mini-Arabii.

Phonology and Orthography in Mini-Arabii

Several linguistic features of Mini-Arabii were simplified to match North American English (see Table 54 in appendix A). This step was taken to avoid confounding the effects of sociolinguistic variation with the effects of unfamiliar phonology (Cook et al., 2016; Hayes-Harb & Masuda, 2008) and orthography (Mathieu, 2016; Showalter & Hayes-Harb, 2013). In terms of phonology, geminated consonants and long vowels were shortened and pharyngeal emphatic phonemes were de-pharyngealized (/ðˤ/ → / ð/); /tˤ/ → /t/; /dˤ/ → /d/. As for the phoneme /ʕ/: when appearing at the end of a syllable (i.e. CV<u>C</u>), it was removed; when appearing at the beginning of a syllable (i.e. <u>C</u>VC): it was replaced with the voiced palatal glide /j/. The changes for the phoneme /ʕ/ reflect how English L1 speakers unfamiliar with Arabic phonology often interpret it initially. In terms of orthography, the grapheme combination ⟨th⟩ represents /θ/, whereas ⟨dh⟩ represents /ð/. The single grapheme ⟨q⟩, representing /q/, was retained as it appears in borrowed words such as “Qur’an” and “Iraq”. Finally, the glottal stop /ʔ/ is represented as ⟨’⟩. All sentences in Mini-Arabii were presented in transliteration.

Vocabulary Items in Mini-Arabii

The vocabulary items in Mini-Arabii can be divided into two types according to the level at which SLV occurs: items which vary phonetically (nouns), and items which vary morphosyntactically (verbs).[6] The phonetic variation which occurs between nouns was based on four major phonetic changes which occur between MSA and ECA (Khalil, 2020; Nydell, 1993; Watson, 2002). These changes, selected due to their saliency and frequency, are as follows:

1) /d͡ʒ/ → /g/ (alveo-palatal fricative to velar plosive)

   ex. / mu**d͡ʒ**rima/ → /mu**g**rima/

2) /q/ → /ʔ/ (uvular plosive to glottal plosive)

   ex. /ħari:**q/** → /ħari:**ʔ/**

3) interdental → alveolar (θ/ → /t/; /ð/ → /d/; /ðˤ/ → /dˤ/)

   ex. **/ð**ura/ → **/d**ura/

4) interdental → sibilant (/θ/ → /s/; /ð/ → /z/; /ðˤ/ → /zˤ/)

   ex. /bi**ð**ra/→ /bi**z**ra/

First, all potential target nouns which reflect each of these four phonetic changes in one of three positions (word-initial, word-medial, and word-final) were systematically identified. Oxford University Press provided non-commercial access to a spreadsheet containing all entries in the Oxford Arabic Dictionary (Arts, 2014), organized by headword, part of speech, root form, and verb form. Next, root forms were split into separate columns using the stringi package (Gagolewski, 2022) in R which, unlike more commonly used packages such as stringr (Wickham, 2022), has the ability to parse right-to-left writing systems. Finally, root forms were filtered to meet the desired combination of letter-location combinations described above.

[6] While there are sociolinguistically variant lexical items in Arabic that also receive inflection, such items will not be included in the present study to avoid confounding the influence of phonetic variation and grammatical variation.

From the group of potential nouns identified, two concrete nouns (one masculine, one feminine) for each of the change–by-position combinations were selected. Thus, in total, 24 nouns were chosen for the study (see Table 55 in appendix A). Half of the nouns ended in *-a*, making them grammatically feminine within the rules of Mini-Arabii, and half were grammatically masculine with no marked ending (balance of grammatical gender will be relevant to the grammar topics, discussed below). The concreteness ratings for the target nouns ranged from 4.42 – 5 (out of 5 maximum, M = 4.82, SD = 0.16) as determined by the Brysbaert, Warriner, and Kuperman database (2014). The length of the target nouns ranged from 1-3 syllables (M = 2.13 SD = 0.54). Finally, within each set of grammatically masculine and feminine nouns, half (three) could be plausible subjects and half could be plausible objects.

In addition to the 24 nouns, four verbs were chosen for the study (see Grammar Structures section below).[7] The 24 nouns and four verbs were divided into two even sets of items, each consisting of 12 nouns and two verbs (14 items total). In keeping with a counterbalanced within-subjects experimental design, participants learned one set of vocabulary in two registers (in the +SLV conditions, whether Integrated or Sequential) and one set in only one register (in the -SLV condition), as explained in greater detail below.

Grammar Structures in Mini-Arabii

The grammar structures chosen for this study were 1) negated present-tense subject-verb agreement, and 2) negated past-tense subject-verb agreement. These structures were inflected for two persons: third person singular feminine and third person singular masculine. Past- and

[7] There are fewer verbs in the target stimuli compared to nouns. This is because the relative learning burden for acquiring a single verb and its rule-based conjugation system is likely greater than for acquiring a single noun. The two learning targets draw on potentially different learning systems (rule-based versus item-based learning).

present-tense subject-verb agreement were chosen because they entail roughly equivalent rules of inflection (the addition of either prefixes or suffixes). This approximate equivalence makes past- and present-tense a suitable grammatical pair for a within-subjects, counterbalanced design. Negation was chosen because the negation particles can vary between register, representing SLV morphosyntactic variation. The Mini-Arabii verb inflections and negation particles were adapted from Arabic and modified for the purposes of the experiment. Any deviation from Arabic was carefully and systematically implemented to design counterbalanced stimuli.

The underlying rules of subject-verb agreement in Mini-Arabii were as follows: nouns which ended in an *-a* (i.e. "nid͡ʒma" / "star") were grammatically feminine; all other unmarked nouns were grammatically masculine (i.e. "ħal:aq" / "barber"). In the present tense, feminine nouns were conjugated by adding the prefix *t-* to the uninflected verb, whereas masculine nouns were conjugated by adding the prefix *y-*. In the past tense, feminine nouns were conjugated by adding the suffix *-at* to the uninflected verb, whereas masculine nouns required no additional affix to the verb stem.

For negation, participants learned that one tense follows the same rules of negation (the particle *lam* is placed in front of the verb) regardless of register, and another tense follows separate negation rules for each register (in MSA the particle *la* was placed in front of the verb, in ECA the particle *ma* was placed in front of the verb and the suffix *-sh* was added to the end of the inflected verb; see Figure 1). The tense which followed the same rules of negation between registers operationalized learning in a traditional curriculum, where no morphosyntactic variation occurs. The tense in which negation varied between registers represented learning in a curriculum that incorporates SLV (either Integrated or Sequential). The assignment of varying and non-varying tenses was counterbalanced: half of the participants (Group A) learned that

there is variation in the past tense, but not the present; the other half (Group B) learned the opposite: that there is variation in the past tense, but not the present (see Figure 1).[8]

Figure 1

Illustration of counterbalanced training conditions operationalizing variation and no-variation curricula

Counterbalanced Training		**Present Tense** *(s/he doesn't eat)*		**Past Tense** *(s/he didn't eat)*	
		No variation		Variation	
				MSA	ECA
	A	lam t-akal lam y-akal		la akal-at la akal-Ø	ma akal-at-sh ma akal-Ø-sh
		Variation		No variation	
		MSA	ECA		
	B	la t-akal la y-akal	ma t-akal-sh ma y-akal-sh	lam akal-at lam akal-Ø	

To illustrate the target grammatical structures, four verbs were selected for the current study. Given that these verbs could be inflected for two different pronouns, there were in total eight different inflected verb forms. Note that, as with the vocabulary items above, these sets of forms could be realized in one or two registers depending on the training condition (see Table 56 in appencix A).

The four verbs selected for Mini-Arabii were: “akal” (to eat), “uhib” (to love), “adrab” (to hit), and “ushil” (to carry). These four verbs were chosen because they are imageable and can

[8] Speakers of Arabic will note that this counterbalancing creates a particle-tense combinations that do not exist in Arabic. I chose to rotating the three negation markers across the registers in both conditions in order to ensure that any peculiarities of particle or tense were not unduly affecting learning outcomes (the goal of counterbalancing stimuli).

be used with a wide variety of nouns and contexts, thus making them suitable for a miniature language. Mini-Arabii verbs were simplified from their original Arabic forms such that there are no underlying stem changes which occur between tenses, thereby avoiding the introduction of morphosyntactic changes unrelated to SLV. The forms were also selected to ensure that the same internal vowel patterns occurred in both lists. Furthermore, in order to avoid confounding the effects of morphosyntactic and lexical variation, the verbs selected for this study only reflected SLV in morphosyntactic inflections; they did not contain any of the variant phonemes described in the vocabulary section above. Participants learned to combine nouns and inflected verbs to create simple subject-verb-object sentences in each tense. They were trained on two of the verbs in the -SLV blocks, and the remaining two verbs in the +SLV blocks. The following are examples of these sentence types in Mini-Arabii:

Present Tense: "Cow doesn't love corn"

- "ba**q**ara la t-uhib **ð**ura" (+SLV – MSA register)
- "ba**ʔ**ara ma-t-uhib-sh **d**ura" (+SLV – ECA register)
- "ba**q**ara lam t-uhib **ð**ura (-SLV)

Past Tense:"Gazelle didn't love fire"

- "**ðʿ**abja la uhib-at ħari:**q**" (+SLV – MSA register)
- "**zʿ**abja ma uhib-at-sh ħari:**ʔ**" (+SLV – ECA register)
- "**ðʿ**abja lam uhib-at ħari:**q**" (-SLV)

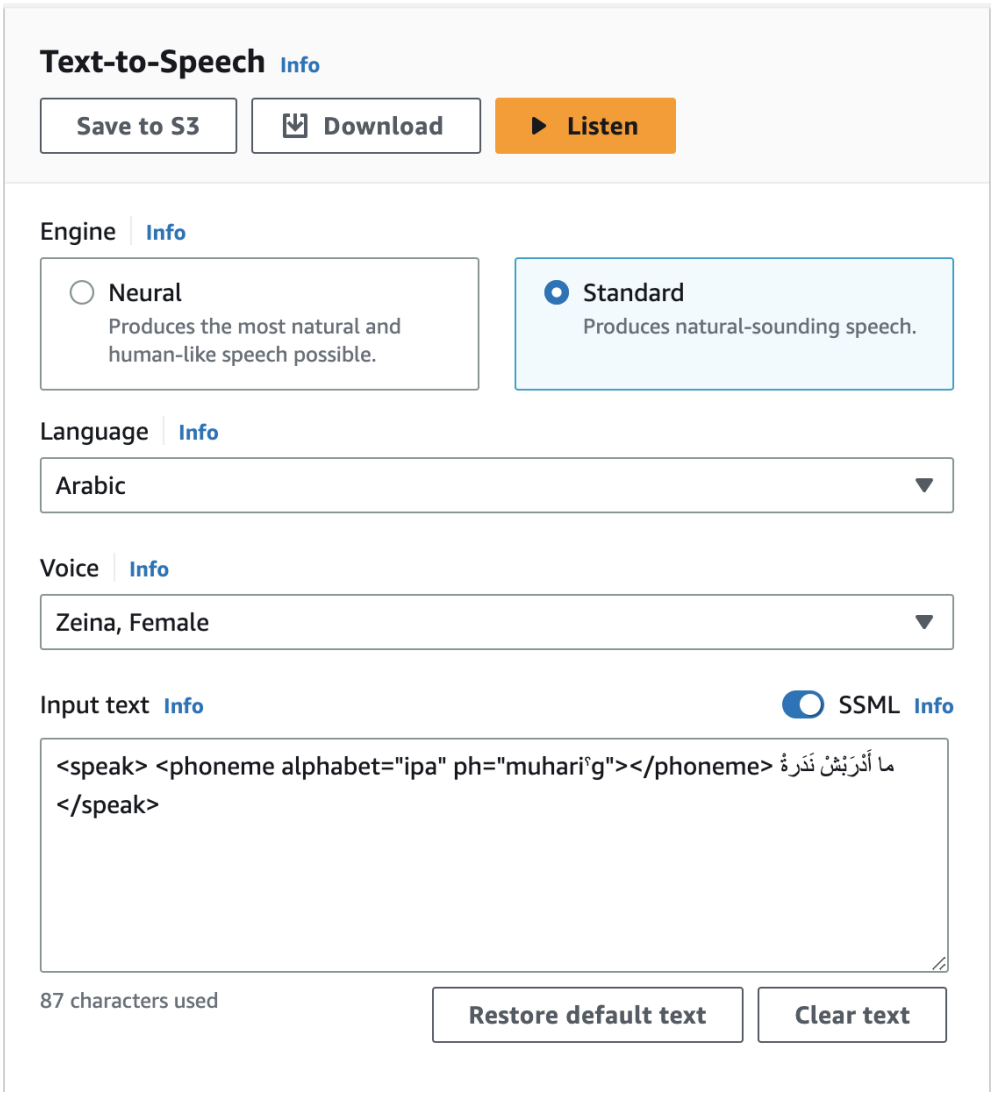
Recall that half of the twelve vocabulary items in each list were semantically plausible subjects, and the other half were semantically plausible objects. To ensure an equal amount of exposure to the L2 vocabulary items in each condition, each grammar experimental block contained six SVO sentences (three for each verb).

Stimuli Creation

All Mini-Arabii stimuli were designed as systematically as possible. First, Mini-Arabii sentences were created using the expand function in tidyr (Wickham & Girlich, 2022). This function joined all viable subject-verb-object combinations, and ensured that no stimuli were erroneously repeated or excluded. To accompany the written stimuli, audio stimuli were created. Previous lines of research on Arabic L2-SLV relied on human talent to create audio stimuli (Huntley, forthcoming; Trentman, 2011; Trentman & S'hiri, 2020), which can invite experimental noise unrelated to the target experimental manipulations. To avoid introducing any undesired variance, the current study relied on audio stimuli created using the Zeina voice in Amazon Polly (Amazon Web Services, 2022). Amazon Polly is a highly customizable text-to-speech service which creates lifelike audio output. Few text-to-speech programs support Arabic speech patterns, and even fewer allow for dialectal phonemes. Using a simple markup language native to the platform (see Figure 2), stimuli which were comparable at all levels of speech (e.g., prosody, tone, volume, etc.) apart from the previously discussed target changes in Mini-Arabii SLV were created.

Figure 2

Sample stimuli creation using within the Amazon Polly platform, customized with simple speech markup language



Experiment Structure

The current experiment consisted of two days of training and one day of testing. Participants were trained on the two sets of vocabulary as well as the two sets of grammar discussed above. Training occurred in two 1-hour lab sessions over the course of two days (one

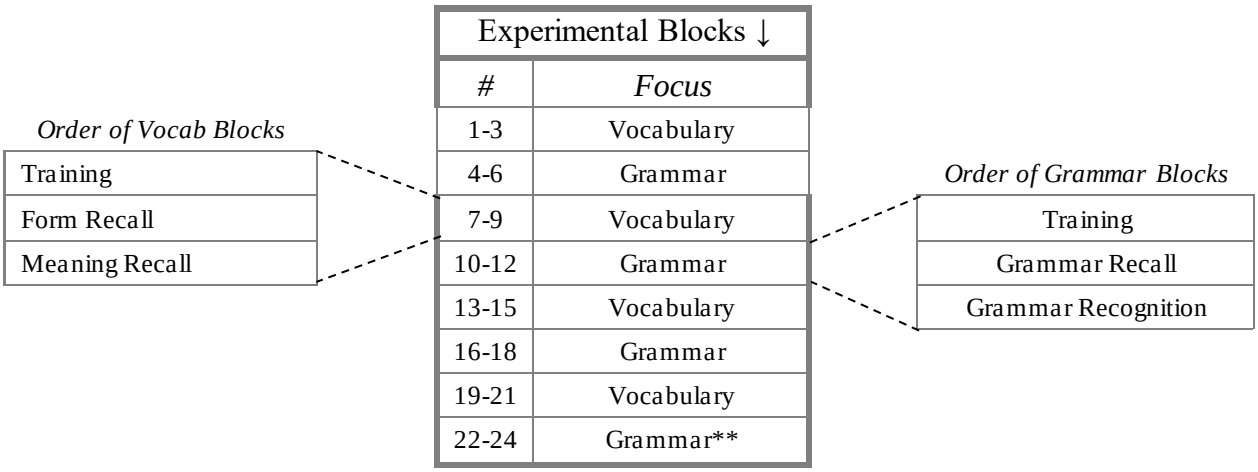
set per session, one session per day). Note, however, that testing was embedded in the training sessions, in order to measure participants' acquisition trajectories (see Figure 3). The third and final lab session on day 3 was the post-testing session.

Each training session consisted of 24 experimental blocks divided between grammar and vocabulary, which are described in detail below. For all sessions (including the post-testing session), participants were told that blocks targeting MSA would be displayed with a blue background and blocks targeting ECA will be displayed with a yellow background. The consistent use of background color was designed to help contextualize the appropriate use of register akin to what is done in integrated Arabic textbooks. For both grammar and vocabulary training, participants only encountered one register per block (that is, registers are not mixed within blocks).

The experiment followed a mixed research design. Each participant learned one set of vocabulary and grammar in a variation condition (integrated or sequential), and one set in the no-variation condition. Hence, curriculum (+SLV or -SLV) was a counterbalanced, within-subjects manipulation in the experiment in which each participant serves as his or her own control. However, +SLV approach (integrated vs. sequential) was a between-subjects experimental manipulation. Half of the participants studied variation in the integrated condition, and half in the sequential condition.

Figure 3

Order of experimental blocks in the first two lab sessions, with testing blocks embedded in the overall training

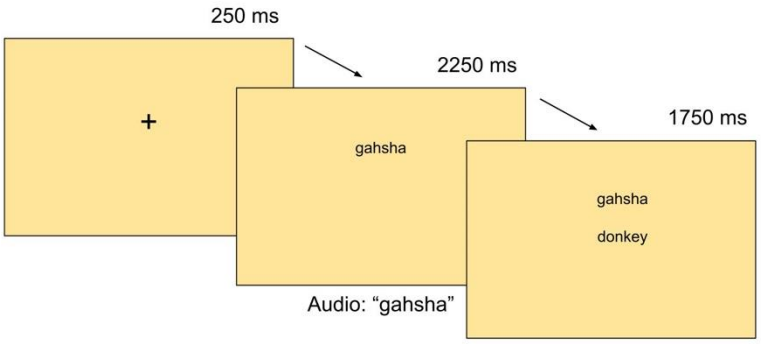


Vocabulary Training

Participants were exposed to target vocabulary in experimental blocks consisting of 12 trials each (one trial per item in each vocabulary set). Trial order was randomized within each block. During training (lab sessions 1 and 2), the vocabulary blocks occurred in groups of three (blocks 1-3, 7-9, 13-15, and 19-21; see Figure 3). The first block in each group of three was the training block, whereas the second and third blocks in each set were the embedded testing blocks (described in the Vocabulary Testing section below). During this block, participants first saw a fixation cross for 250 ms. Next, they saw the written out Arabic word for 2250 ms and heard an audio recording of it (see Figure 4). Finally, they saw the English translation appear below the Arabic word (1750 ms).

Figure 4

Sequence of events in vocabulary training block



Grammar Training

Similar to the vocabulary blocks in the training sessions, grammar blocks also occurred in groups of three (blocks 4-6, 10-12, 16-18, and 22-24; see Figure 3). The first block in each group of three was the training block, whereas the second and third blocks in each set were the embedded testing blocks (described in the Grammar Testing section below). Each grammar block consisted of six trials, three for each of the two verbs per condition. Half of the verbs were grammatically masculine and half grammatically feminine. Vocabulary items acting as subjects and objects were rotated across experimental blocks to ensure equal exposure during the grammar training.

The grammar training blocks introduced the grammar rules in a piecemeal fashion, starting with the conjugated verb, then adding negation, and then adding nouns as subjects and objects. Participants began each training block by reading the explicit rules for the grammar set as laid out in Table A3. In each trial, participants first saw a fixation cross (250 ms), followed by the conjugated verb (2500 ms). The English translation then appeared below the conjugated verb (3000 ms). Next, participants saw the negated conjugated verb (2500 ms), followed by the

English translation appearing below (3500 ms). Finally participants saw the full subject-verb-object sentence in Arabic (3500 ms), followed by its translation appearing below (3500 ms). Each Arabic stimulus was accompanied by an audio recording.

Testing (All Sessions)

Testing occurred both during the training sessions (days 1 and 2) as well as in the post-testing session (day 3). The same types of tests, described in detail below, were used in all three sessions. The only key difference between the testing in the training sessions and in the post-testing session was that, during the training sessions, participants received feedback on their answers (described in detail below). As such, training and testing were integrated during the training sessions. During the post-tests, participants did not receive feedback on their answers. All testing blocks began with explicit instructions on what to expect (see Figure 48 in appendix B).

Vocabulary Testing.

Vocabulary acquisition was measured through two sub-components of lexical knowledge (Nation, 2013): form recall and meaning recall. Form recall and meaning recall were tested in separate experimental blocks, consisting of 12 trials each (one trial per item in each vocabulary set). In the form recall test, participants first saw a fixation cross (250 ms), then saw the question "What is this in Arabic?" (1500 ms). Following the prompt, participants were asked to type out the correct Arabic translation of the provided English word as quickly as possible. Participants had up to 7500 ms to type their translation before the trial timed out and automatically advanced. At the end of each form recall trial the correct answer appeared on the screen for 2000 ms before

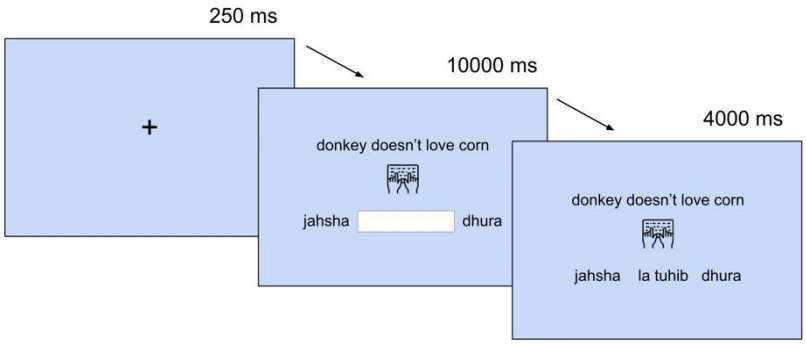
the next trial began. Participants received this style of feedback only during the training sessions on days 1 and 2; on day 3 the correct answer was not displayed. The meaning recall testing mirrored the form recall testing, except that participants were only give 5000 ms to respond before the experiment automatically advanced to the next trial.

Grammar Testing.

Grammar acquisition was operationalized as recall and recognition of negated conjugated verbs embedded in SVO sentences. Recall (translation to Arabic) and recognition (translation to English) were tested in separate experimental blocks, consisting of six trial each. In the production testing blocks, participants first saw a fixation for 250 ms, then saw a screen containing the sentence to be translated, a typing icon, and the translated subject and object with a blank textbox to write the correct verb form (see Figure 5). Participants had up to 10 seconds to type their translation before the trial timed out and automatically advanced. At the end of each form recall trial, the correct answer appeared on the screen for 4000 ms before the next trial began. The meaning recall test mirrored the form recall test, except that participants were only given 5000 ms to respond before the experiment automatically advanced to the next trial.

Figure 5

Sample trial sequence for grammar recall in MSA



The grammar recognition test mirrored the grammar recall testing except that participants were only given 7500 ms to respond before the experiment timed out and advanced to the next trial.

Outcome Measures and Scoring.

The current study explored the development of Mini-Arabii in terms of lexis and grammar. Both types of knowledge were tested productively (English to Arabic) and receptively (Arabic to English). Productive and receptive knowledge were measured both in terms of processing speed (reaction time), as well as accuracy (as a binary measure, 0 or 1). Reaction time is a latent variable measuring how quickly learners can access lexical and grammatical knowledge in real-time (Godfroid, 2020). In practical terms, it may be a useful indicator of fluency development, since learners will ultimately need to efficiently pair form and meaning in order to participate in real-time conversations. Furthermore, to allow for a more nuanced understanding of accuracy, productive measures were also scored using the Levenshtein Distance (Levenshtein, 1966). The Levenshtein Distance quantifies the distance between two sequences based on the minimum number of single-character edits needed. For example, if the desired

string is “CAT” and a participant types “CA”, they would receive a Levenshtein Distance score of 1 (addition of T). If the provided response is “CHAD”, the Levenshtein Distance score would be 2 (deletion of H, substitution of D for T). In both cases, a binary accuracy score would be 0 and would not reflect any of the partial form knowledge that participants had gained. The outcome different outcome measures for both vocabulary and grammar knowledge are visually summarized in Figure 6.

Figure 6

Visual summary of the outcome variables by knowledge type and metric

	Receptive Measures	Receptive Measures

Vocabulary Knowledge	• Accuracy:
	  ○ Binary (0/1)
	  ○ Levenshtein Distance
	• Processing: Reaction Time
		• Accuracy:
		  ○ Binary (0/1)
		• Processing: Reaction Time

Grammar Knowledge	• Accuracy:
	  ○ Binary (0/1)
	  ○ Levenshtein Distance
	• Processing: Reaction Time
		• Accuracy:
		  ○ Binary (0/1)
		• Processing: Reaction Time

Software and Data Quality

Due to the COVID-19 global pandemic, all training and testing was conducted remotely using Gorilla (Anwyl-Irvine et al., 2020; *Gorilla Experiment Builder*, 2021). While online platforms are not necessarily considered the gold standard in psychometric research (e.g., Al-Salom & Miller, 2019), they can capture comparable metrics (Hilbig, 2016; Ruiz et al., 2019)

if paired with the appropriate guardrails to ensure a high level of data quality (Curran, 2016; Sauter et al., 2020; Vaughn et al., 2018). In the case of the current study, the experiment was piloted in two separate rounds to ensure clarity of instructions and visuals, as well as to finalize the timing of all trial sequences and testing timeouts. Furthermore, attention checks (Hauser & Schwarz, 2016) were included during each training block. Please see the "Data Cleaning" section below for more information on data inspection post data-collection.

Procedure

All study procedures and materials were approved by the Institutional Review Board (ID #00006627). Prior to beginning the study, all participants indicated their consent to participate in the three-day study in exchange for compensation of 45 USD.

Participants began the experiment on day 1 by logging in to the Gorilla online platform (*Gorilla Experiment Builder,* 2021) and reading the instructions as provided in Figure 49 in appendix B. These instructions were designed to familiarize participants with the experimental procedure, including the use of color to contextualize register as well as the presence of semi-novel graphemes to represent the Mini-Arabii sound system. They then moved through the experiment as illustrated in Figures 50 (the two training days with embedded testing) and 51 (the post-testing day) in appendix B.

Recall that the experiment followed a mixed-design approach to cover two different levels of comparison: +SLV vs. -SLV, and +SLV-Integrated vs. +SLV-Sequential. All participants encountered -SLV as well as one type of +SLV. Half of the participants will encounter +SLV in the integrated condition, and half in the sequential condition. The use of color in Figures B3-4 illustrate the block sequences of these different experimental conditions. Blocks operationalizing a type of +SLV are colored in blue (representing MSA) and yellow

(representing ECA). Blocks operationalizing a traditional curriculum (-SLV), are colored in green. Finally, vocabulary blocks are depicted by the lighter shading and grammar blocks are depicted by the darker shading.

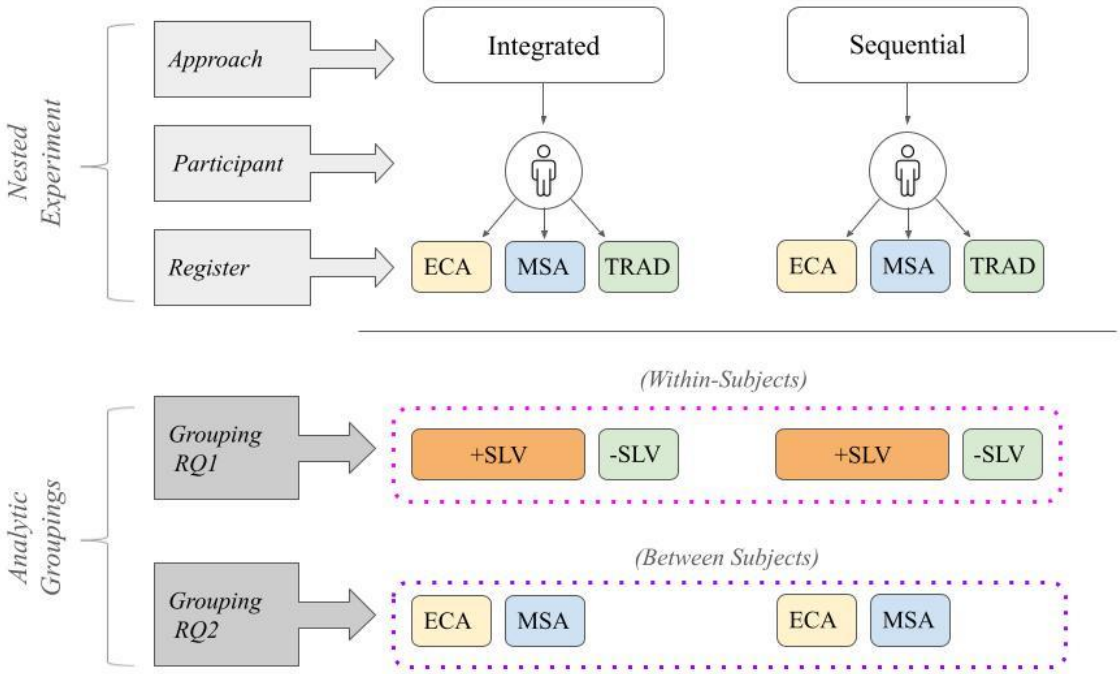
Analyses

Data Structure

The current study utilizes a mixed design experimental structure. Recall that all participants are assigned to one +SLV condition. Thus, training condition (Integrated vs. Sequential) is a between-subjects variable. The variable "register" is nested within "training condition." Each training condition contains three registers: MSA, ECA, and TRAD. MSA and ECA represent registers in +SLV curricula (acquired in varying orders depending on condition), and TRAD represents a single register learned in -SLV curricula. Thus, each participant serves as their own control to compare +SLV outcomes again -SLV outcomes. This mixed experimental design is illustrated in Figure 7:

Figure 7

Illustration of nested data structure for mixed experimental design conditions in the current study



Data Analysis

To account for the hierarchical structure of the data as well as to avoid violating the assumption of independence of errors, mixed effects models were used for data analysis (Baayen et al., 2008; Singmann & Kellen, 2019). Models were constructed using the lmer (Bates, 2010) package in R. Model selection is not a straightforward process, balancing between what is theoretically relevant and what is computationally permissible (Barr et al., 2013; Matuschek et al., 2017). For ease of comparison between measures and relevance to the research questions, the current study adopted the same overall process of selection for all analyses. First, null models were fitted with random effects. The random effects were nested as follows:

1. random intercept for Participant (hypothesizing that participants did not start at the same baseline level);

2. random intercept for Curriculum (RQ1) or Approach (RQ2), nested within Participant (to account for clustering);

Nested null models were compared using a likelihood ratio test, and the simplest model with the best fit was chosen. Then, full mixed effects were fitted. All final models contained the same fixed effects structure because of their theoretical relevance to answering the research questions (Winter, 2019). For RQ1, Curriculum (-SLV vs. +SLV) and Approach were included as independent variables, although only Curriculum is of interest; because Curriculum is nested within Approach (the initial group assignment variable for the between-subjects portion of the experiment), Approach was included to parse out variability unrelated to Curriculum. RQ2 explores only the +SLV subset of the data, so Condition (Integrated vs. Sequential) and Register were included as interacting independent variables. Table 1 spells out the levels of the hierarchical models used in the current study:

Table 1

Nested Levels and Variables in the Current Study's Mixed Effects Models

Sub-index	RQ1		RQ2	
	Level	Variables	Level	Variables
i	Approach	Sequential* Integrated	Approach	Sequential* Integrated
J	Participant		Participant	
K	Curriculum	-SLV* +SLV	Register	MSA* ECA

* = reference variable

Post-hoc assumptions for mixed effects linear regression were checked using the easystats and arsenal packages in R (Heinzen et al., 2021; Lüdecke et al., 2022; Staniak & Biecek, 2019). Models were inspected for normality of distribution of dependent variables

(skewness and kurtosis) and residuals (QQ and residual density plots), multicollinearity (variance inflation index), and influential outliers. All post-hoc assumptions visuals can be found in Appendix C. Any deviations from these assumptions will be discussed in context alongside individual model results.

Finally, planned post hoc comparisons of interest and estimated marginal means were computed using the emmeans package (Lenth, 2022) in R. Estimated marginal means (EMMs) are an especially useful estimate when experimental groups are unequal, because the calculated subgroup means are given equal weight according to their likely distribution in the population at large (Lenth, n.d.). Recall that Huntley (forthcoming) failed to detect significant differences between the +SLV and -SLV conditions for form recognition of vocabulary, leaving open the possibility that the two learning conditions may be equivalent. Hence, all planned post hoc comparisons will be furthermore tested for equivalence using the two one-sided test (TOST) method. This practice follows the advice of Lakens et al., who recommend that "researchers by default perform both a null-hypothesis significance test and an equivalence test on their data, as long as they can justify a SESOI [smallest effect size of interest], in order to improve the falsifiability of predictions in psychological science" (2019, p. 267).

Within the TOST method, the concept of "equivalence" is two-pronged: group outcomes can neither be inferior *nor* superior to one another. In simple terms, group A is considered equivalent to group B if it can be shown that not only are group A's scores not *worse* than group B's, but that they're also not *better*. This concept of equivalence is measured by implementing two one-sided *t*-tests (TOST), one for (non) superiority and the other for (non) inferiority. If either one-sided *t*-test is insignificant (i.e. if either superiority or inferiority is established), then the entire test is insignificant. Standard practice in reporting TOST results is to report the least

significant of the two one-sided *t*-tests (Lakens, 2022). Finally, because the TOST method performs two *t*-tests, to achieve a level of $\alpha = 0.05$ for both superiority *and* inferiority, 90% confidence interval estimates are used (Wellek, 2010, pp. 33–34).

Although tests of equivalence are relatively uncommon in Second Language Acquisition (Godfroid & Spino, 2015), they are frequently used in fields such as medicine where the goal is to show comparable outcomes in treatments (E. Walker & Nowacki, 2011). If a new medication is functionally equivalent to older therapies, then providers can consider other factors (such as cost effectiveness, ease of administration, or associated side effects) when deciding what to prescribe. Unlike traditional significance tests, which use standardized test statistics to determine difference, tests of equivalence rely on situation-specific metrics. Researchers must *a priori* specify the acceptable range of dependent variable outcomes, δ, which would be considered functionally comparable (Lakens, 2017). Given that there is not, to date, any prior research establishing equivalence for the metrics utilized in the current study, the SESOI (smallest effect size of interest) was established according to performance-based estimates of a meaningful difference in outcomes. The threshold for meaningful difference in outcomes according to training is as follows for each of the three general metrics making up the dependent variables of interest:

Mean Log Reaction Time (ms): delta = log100 (within 100 ms)

Mean Levenshtein Distance: delta = 1 (within 1 letter change to arrive at the correct answer)

Mean Accuracy: delta = 0.1 (within 10%)

Data Cleaning

All data were first manually inspected for overall quality. Any participants who left entire experimental blocks unanswered, had excessive timeouts, or else rushed through the experiment by providing illogical quick responses (i.e., repeated letter sequences), were removed. This resulted in the removal of four participants. Next, reaction time data for correct responses was trimmed using the R package trimr (Grange, 2022). Following Jiang (2013, p. 70), all reaction times below 300 ms were first trimmed, followed by all outliers beyond three standard deviations of each individual participant's mean reaction time. Table 2 illustrates the number of observations trimmed at each step.

Table 2

Reaction Time Trimming for Each Dependent Variable

Dependent Variable	NOBS	# Trimmed			NOBS (final)
		Incorrect	< 300 ms	>|3 SD|	
Vocabulary					
Form	2808	2044	1	11	752
Meaning	2808	1063	1	21	1723
Grammar				0	
Recall	1406	1109	2	6	289
Recognition	1404	802	0	8	594

Chapter 3: Results (RQ1: -SLV vs. +SLV)

The first research question compared the learning outcomes of training in a curriculum that incorporates sociolinguistic variation against a curriculum that does not. These curricula were operationalized in the current study as +SLV (studying two registers, either learned simultaneously or back-to-back) and -SLV (studying only one register), respectively. Recall that Mini-Arabii was designed to reflect sociolinguistic variation in both lexical items (nouns) and grammar rules (verb conjugation and negation). Thus, to comprehensively capture the effects of training in a +SLV vs. -SLV curriculum on the development of L2 lexical and grammatical knowledge, a broad array of measures capturing both accuracy and processing speed were employed. Results are organized according to knowledge type and outcome measure.

Word Form Knowledge

Word form knowledge refers to participants' ability to produce the target lexical item. In the context of the study, participants provided the Mini-Arabii translation for the English prompt.

Reaction Time

Descriptive Statistics.

The first metric of word form knowledge was processing speed: how quickly participants were able to produce the correct target form. As shown in Table 3, the descriptives statistics for log-transformed mean reaction time are strikingly similar between the +SLV and -SLV conditions. Both have a mean of 7.57 log ms, with the 95% confidence intervals for -SLV completing encompassing those of +SLV (recall that the vocabulary form knowledge test automatically advanced after 7500 ms, which is 8.92 on the log scale).

Table 3

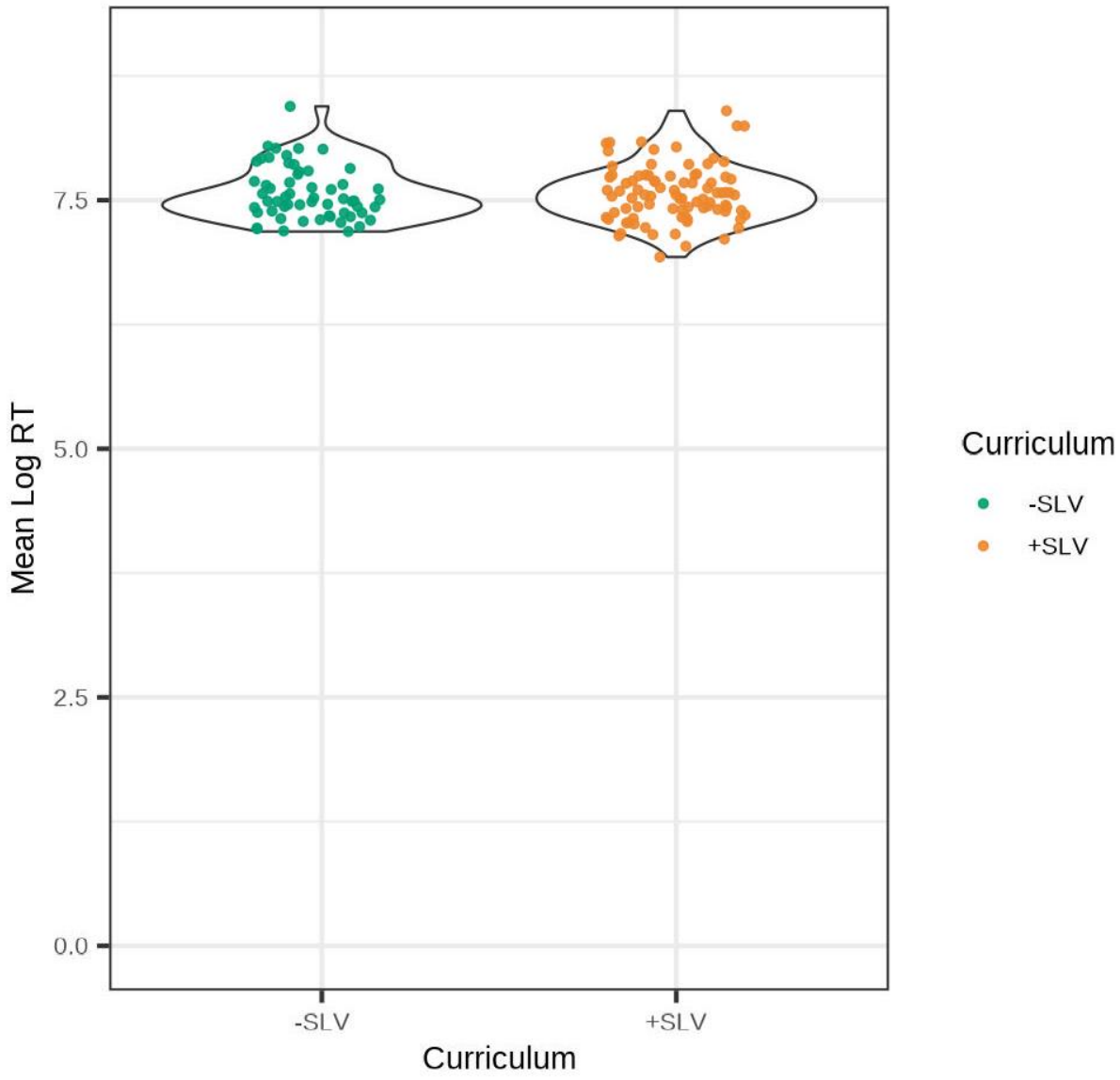
Descriptive Statistics for Vocabulary Form Knowledge: Mean Log Reaction Time

	-SLV	+SLV
N	58	88
Missing	0	0
Mean (CI)	7.57 (7.50, 7.64)	7.57 (7.51, 7.63)
SD	0.27	0.28
Skewness	0.87	0.5
Kurtosis	0.67	0.34

The +SLV condition has a slightly larger standard deviation than that of the -SLV condition (0.28 and 0.27, respectively), which is visually reinforced by the longer spread of +SLV datapoints in Figure 8. Finally, the data appear to be relatively normally distributed according to both a visual check of the datapoints in Figure 8 and the skewness and kurtosis values in Table 3, which are all less than |2| (Lomax & Hahs-Vaughn, 2012).

Figure 8

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Log Reaction Time for Vocabulary Form Knowledge



Analyses.

The results of the linear mixed effects model for vocabulary form knowledge can be seen in Table 4. While there was significant effect of Approach (β_{int} = -0.13, t = -2.53, p = 0.013), it is not of interest for the current research question (recall that Approach was included in the model

to account for the nested nature of the data). However, the significance of Approach as a factor could indicate that, despite random assignments, groups were not equal at baseline. According to the model, participants took about 1% longer[9] to recall items from the +SLV curricular condition than the -SLV curriculum. The inferential effect of Curriculum, however, was not statistically significant (β_{+SLV} = 0.01, t = 0.19, p = 0.848), suggesting that there may not be a practical difference between +SLV and -SLV training at the population level.

Table 4

Results of Linear Mixed Effects Model for Vocabulary Form Knowledge: Mean Log RT

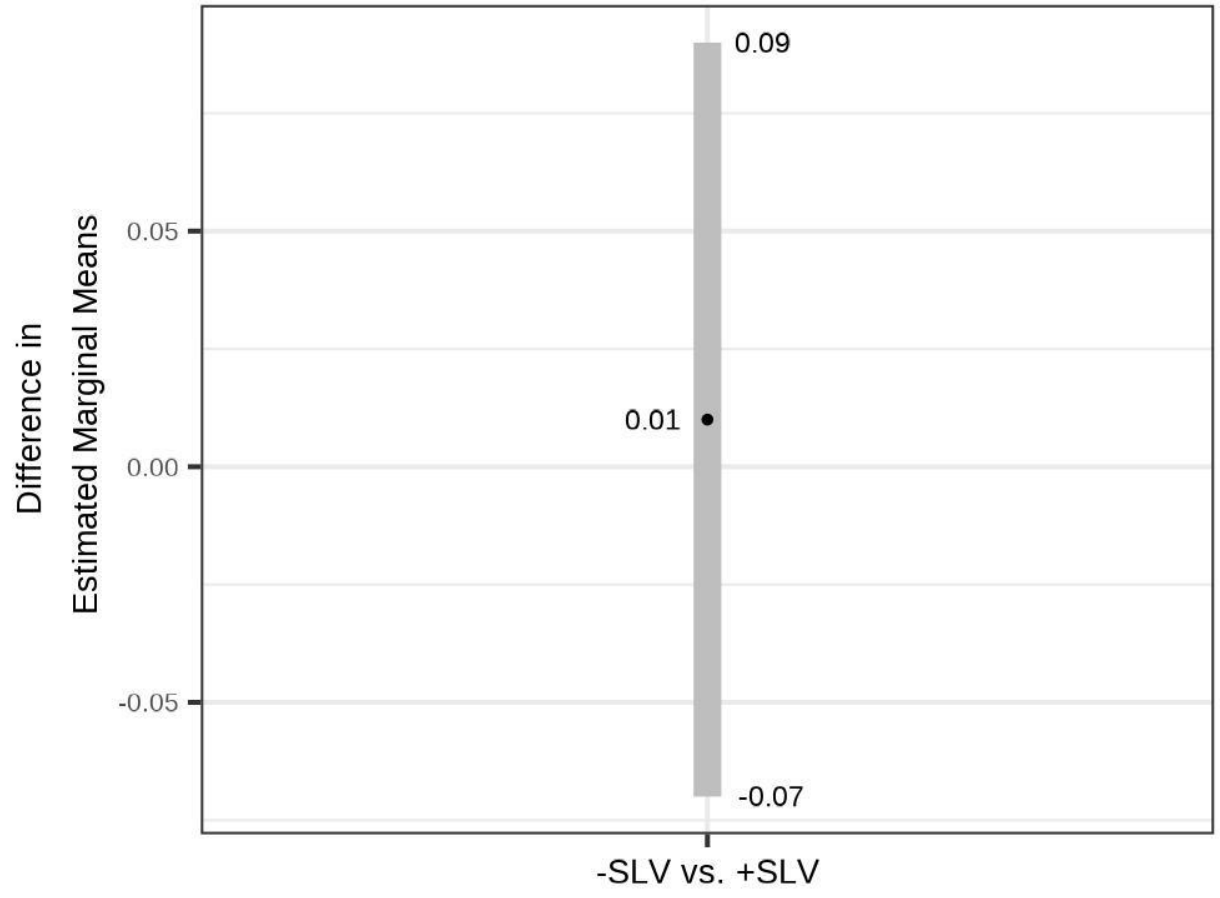
	Log RT				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	7.50	0.05	7.41 - 7.59	165.75	<0.001
Approach [Int]	-0.13	0.05	-0.22 - -0.03	-2.53	0.013
Curriculum [+SLV]	0.01	0.05	-0.09 - 0.11	0.19	0.848
Random Effects					
σ^2	0.03				
$\tau_{00\ participant:curriculum}$	0.05				
ICC	-0.62				
$N_{\ participant}$	65				
$N_{\ curriculum}$	2				
Observations	146				
Marginal R^2 / Conditional R^2	0.050 / 0.642				

[9] The beta estimate is exponentiated to back-transform the difference between the reference level (-SLV) and the level of interest (+SLV) into meaningful units.

In order to test for equivalence, estimated marginal means were calculated for mean log reaction time of word form recall. Figure 9 shows the difference in estimated marginal means between groups with 90% confidence intervals. It largely confirms what was concluded from the descriptive statistics: the -SLV and +SLV outcomes almost completely overlap one another; that is, the difference between them is practically zero.

Figure 9

Difference in Estimated Marginal Means with 90% Confidence Intervals for Vocabulary Form Knowledge: Mean Log RT (equivalence bounds not depicted for visual clarity)



Finally, the estimated marginal means for Curriculum were tested for functional equivalence. Delta was set at 100 ms (4.61 on the log-transformed scale). The presence of effects (that is, group differences) greater than the equivalence range of $\delta = \log(100)$ was rejected (t = -92.64, p <.001, d = 0.06). It can be concluded that participants at the population level would likely be able to recall the correct target word form with functionally equivalent speed regardless of whether they were learned in a +SLV (two registers) or a -SLV (one register) curriculum).

Levenshtein Distance

Descriptive Statistics.

The second metric of word form knowledge was an approximation of accuracy: the Levenshtein Distance. The Levenshtein Distance counts the number of changes needed to arrive at the correct form. As can been seen in Table 5, the descriptive statistics of Levenshtein Distance scores for vocabulary word knowledge are fairly similar between the +SLV and -SLV conditions. While -SLV has an overall lower score (indicating fewer changes needed, on average, to arrive at the correct word form) than +SLV, the 95% confidence intervals around the means overlap to a large extent (-SLV: [1.60, 2.27], +SLV: [1.92, 2.35]).

Table 5

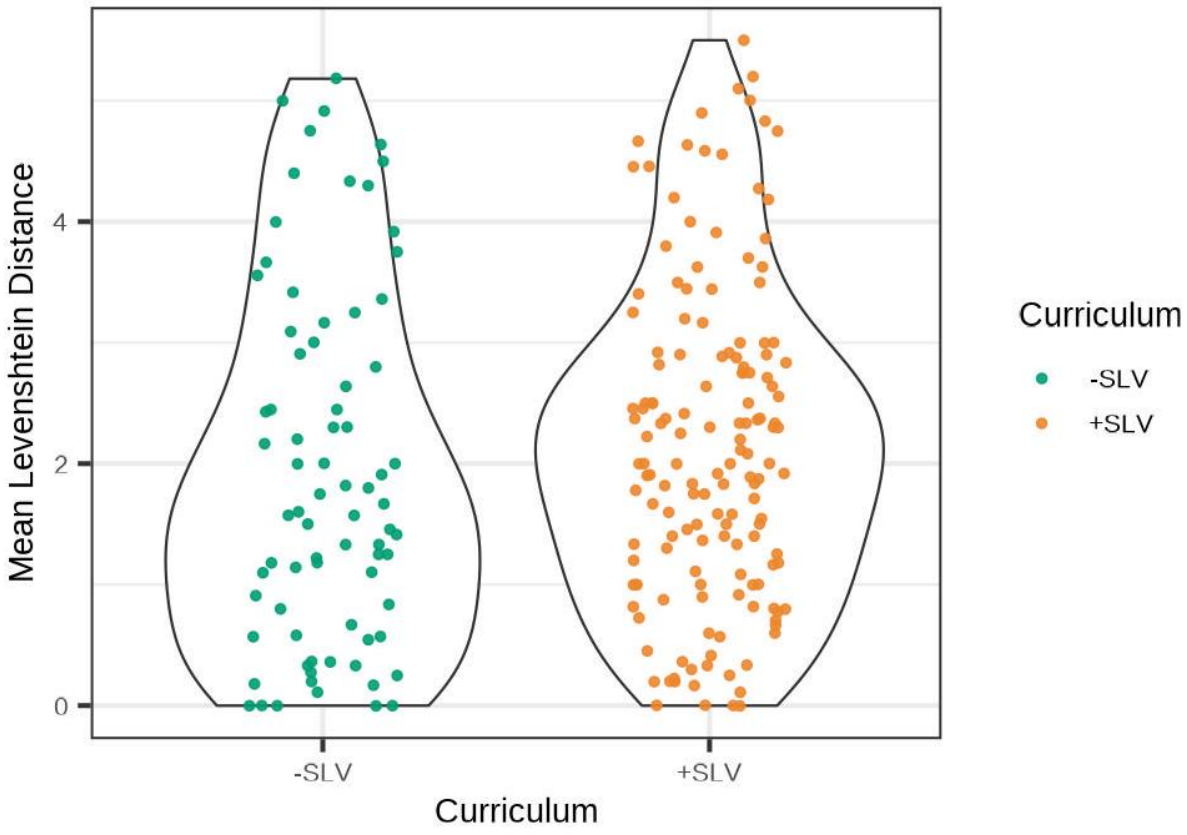
Descriptive Statistics for Vocabulary Form Knowledge: Mean Levenshtein Distance

	-SLV	+SLV
N	77	151
Missing	1	5
Mean (CI)	1.94 (1.60, 2.27)	2.14 (1.92, 2.35)
SD	1.47	1.32
Skewness	0.57	0.47
Kurtosis	-0.71	-0.39

The -SLV condition has a slightly larger standard deviation than that of the +SLV condition (1.47 and 1.32 respectively). On the whole, however, the spread of scores is comparable between conditions as can be seen by the distribution of datapoints in Figure 10. Finally, the data appear to be relatively normally distributed according to both a visual check of the datapoints in Figure 10 and the skewness and kurtosis values in Table 5, which are all less than |2| (Lomax & Hahs-Vaughn, 2012).

Figure 10

Violin Plot Illustrating the Distribution of Mean Levenshtein Distance for Vocabulary Form Knowledge



Analyses.

The results of the linear mixed effects model for vocabulary form knowledge can be seen in Table 6. There was a significant effect of Curriculum on Levenshtein Distance, where training in the +SLV group was associated with an additional 0.22 changes needed to arrive at the correct word form answer (β_{+SLV} = 0.22, t = 2.1, p = 0.037).

Table 6

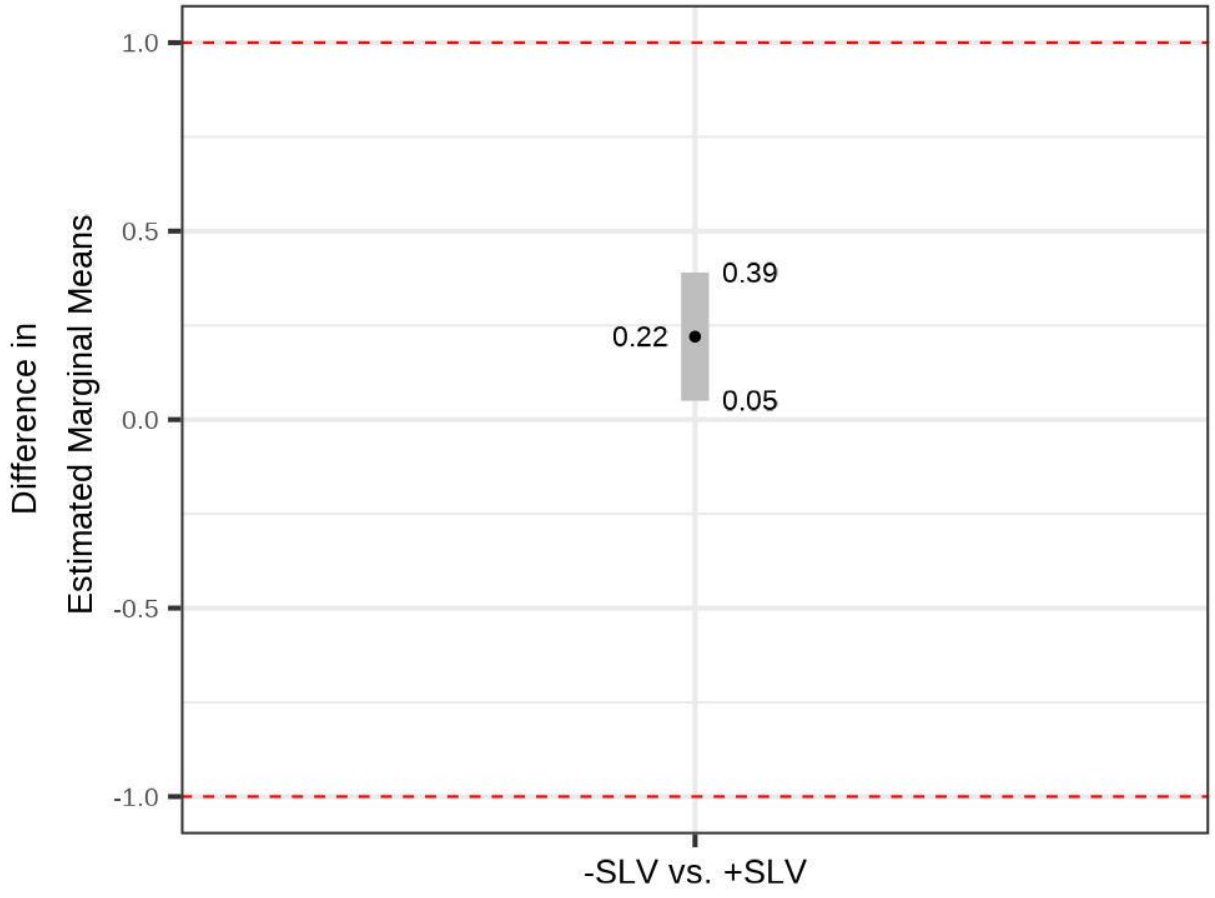
Results of Linear Mixed Effects Model for Vocabulary Form Knowledge: Mean Levenshtein Distance

	Levenshtein Distance				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	2.08	0.22	1.65 - 2.51	9.48	<0.001
Approach [Int]	0.27	0.29	-0.29 - 0.84	0.94	0.347
Curriculum [+SLV]	0.22	0.1	0.01 - 0.42	2.1	0.037
Random Effects					
σ^2	0.47				
$\tau_{00\ participant:curriculum}$	0.06				
$\tau_{00\ participant}$	1.38				
ICC	0.75				
$N_{\ participant}$	77				
$N_{\ curriculum}$	2				
Observations	228				
Marginal R^2 / Conditional R^2	0.015 / 0.758				

Although the results of the linear mixed effect model showed evidence for a significant difference between Curriculum type, the model was still tested for equivalence using the estimated marginal means. Delta was set at 1 (one letter change). With this limit, the null hypothesis that there is indeed a meaningful difference between the two Curriculum groups was rejected (t = -7.5, p < .001, d = 0.32). The outcomes between groups were statistically equivalent.

Figure 11

Difference in Estimated Marginal Means with 90% Confidence Intervals for Vocabulary Form Knowledge: Mean Levenshtein Distance (equivalence bounds depicted in red)



These inferential findings are confirmed in Figure 11, which illustrates the difference between the two groups' estimated marginal means with 90% confidence intervals. The red dotted lines show the equivalence bounds. The estimated difference between -SLV and +SLV is fully within the equivalence bounds, indicating that all plausible scores at the population level are indeed within in the pre-determined range of functional equivalence. Thus, the results of the difference and equivalence testing draw different inferential conclusions. Such conflicting results are indeed possible (see E. Walker & Nowacki, 2011 for a discussion). In this case, the difference of 0.22 is considered statistically different (with the test statistic falling beyond the critical value), yet also statistically equivalent (within the functional equivalency bounds of one letter change). As such, no clear inferential conclusions can be drawn. All that can be stated is that, for the sample participants, learning vocabulary in two registers (+SLV) rather than one (-SLV) is associated with an additional 0.2 letter changes needed to arrive at the correct form.

Accuracy

Descriptive Statistics.

The third metric of word form knowledge was accuracy as a binary variable (0/1). Unlike vocabulary form knowledge reaction time and Levenshtein distance, the descriptive statistics shown in Table 7 indicate that participants likely did achieve different levels of mean accuracy depending on Curriculum condition. The 95% confidence intervals are close but do not overlap, with the upper bound of +SLV trailing behind the lower bound of -SLV by 0.02.

Table 7

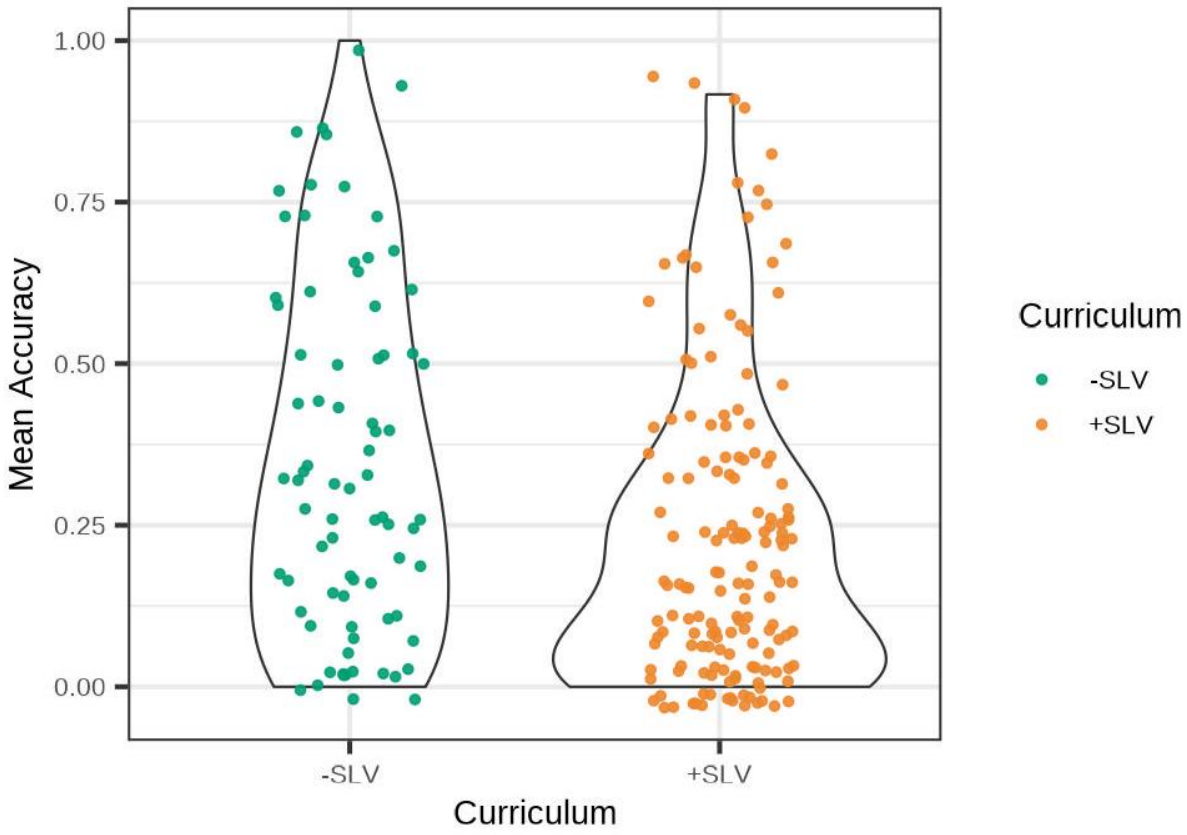
Descriptive Statistics for Vocabulary Form Knowledge: Mean Accuracy

	-SLV	+SLV
N	78	156
Missing	0	0
Mean (CI)	0.35 (0.29, 0.41)	0.23 (0.20, 0.27)
SD	0.27	0.24
Skewness	0.47	1.12
Kurtosis	-0.8	0.59

Although the skewness and kurtosis are within the range of |2| (Lomax & Hahs-Vaughn, 2012), the violin plots displayed in Figure 12 indicate a larger cluster of lower mean accuracy scores in the +SLV condition as compared to the -SLV condition.

Figure 12

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Accuracy by Participant for Vocabulary Form Knowledge



This clustering is confirmed in the model diagnostics (Figure 54 in appendix C), which shows mild deviation between the observed and model-predicted values around 0 and some evidence that the assumptions of homoscedasticity and homogeneity of variance have been violated (wider spread of residuals around the fitted value of 0.4). As such, caution should be used when inferentially interpreting the model results.

Analyses.

The results of the linear mixed effects model for vocabulary form knowledge can be seen in Table 8. There was a significant effect of Curriculum, where training in the +SLV group was associated with an 11% decrease in mean accuracy (β_{+SLV} = 0.11, t = -2.93, p = 0.004).

Table 8

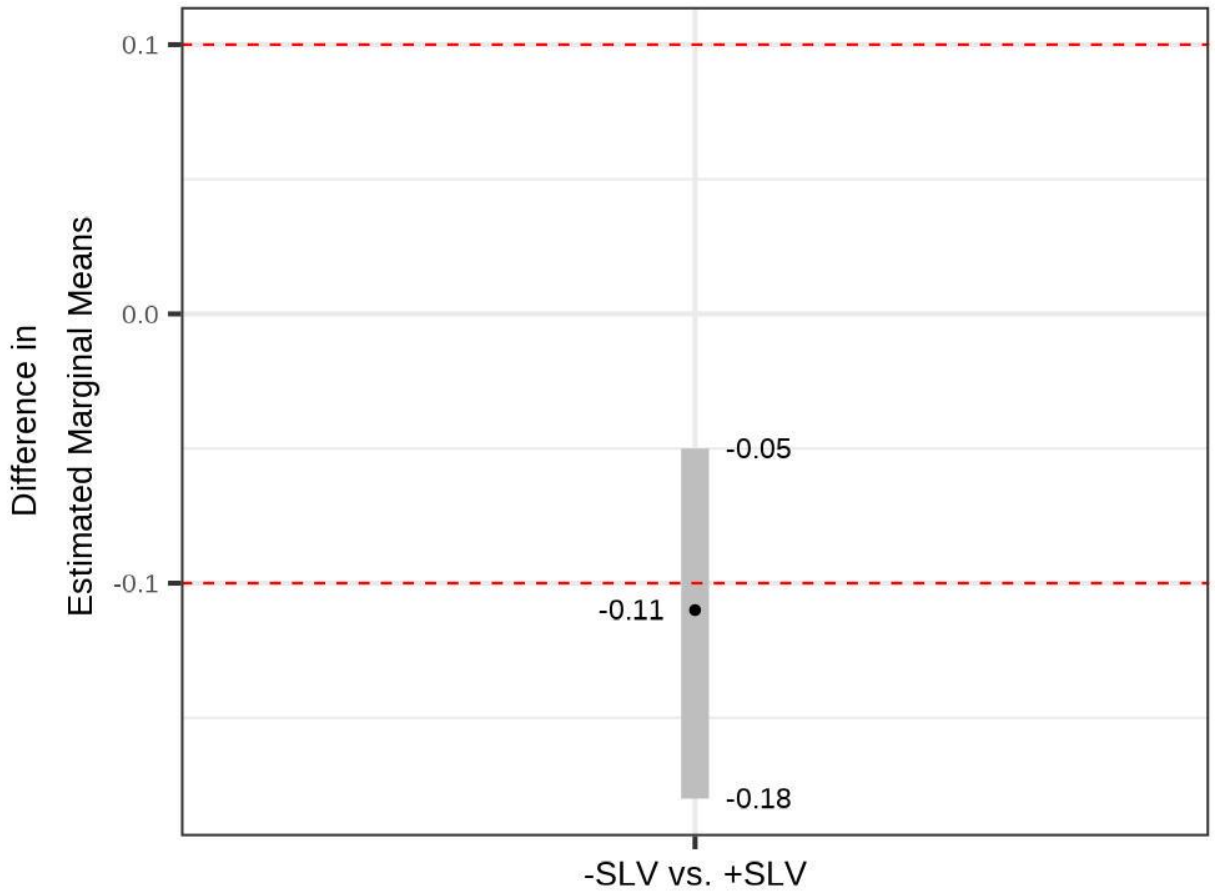
Results of Linear Mixed Effects Model for Vocabulary Form Knowledge: Mean Accuracy

	Accuracy				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	0.37	0.03	0.30 - 0.44	10.89	<0.001
Approach [Int]	-0.04	0.04	-0.12 - 0.03	-1.17	0.242
Curriculum [+SLV]	-0.11	0.04	-0.19 - -0.04	-2.93	0.004
Random Effects					
σ^2	0.02				
$\tau_{00\ participant:register}$	0.04				
ICC	0.63				
$N_{\ participant}$	78				
$N_{\ register}$	2				
Observations	234				
Marginal R^2 / Conditional R^2	0.050 / 0.646				

In keeping with the analytic plan, post-hoc testing for equivalence between estimated marginal means was conducted. Delta was set at a range of within 10% (δ = 0.1). With this limit, the null hypothesis that the groups are not equivalent was retained (t = 0.33, p = .629, d = 0.73).

Figure 13

Difference in Estimated Marginal Means with 90% Confidence Intervals for Vocabulary Form Knowledge: Mean Accuracy (equivalence bounds depicted in red)



This finding is supported by Figure 13, which shows that a portion of estimated difference range between -SLV and +SLV falls beyond the lower equivalence bound. Thus, it can be concluded that there may be a practical difference between +SLV and -SLV training at the population level in terms of mean word form accuracy. Learning vocabulary in two registers (+SLV) rather than one (-SLV) is associated with an 11% decrease in form accuracy.

Word Meaning Knowledge

Word meaning knowledge refers to participants' ability to produce the meaning of the target item. In the context of the study, participants were prompted with the Mini-Arabii noun, and provided the English translation.

Reaction Time

Descriptive Statistics.

The first metric of word meaning knowledge was processing speed: how quickly participants were able to translate the word into English. As shown in Table 9, the descriptives statistics for log-transformed mean reaction time are fairly similar between the +SLV and -SLV conditions. Mean log RT of -SLV is practically equal to that of +SLV (7.41 log ms and 7.43 log ms respectively). Furthermore, the 95% confidence intervals are largely overlapping, which suggests parity between these two curricula.

Table 9

Descriptive Statistics for Vocabulary Meaning Knowledge: Mean Log Reaction Time

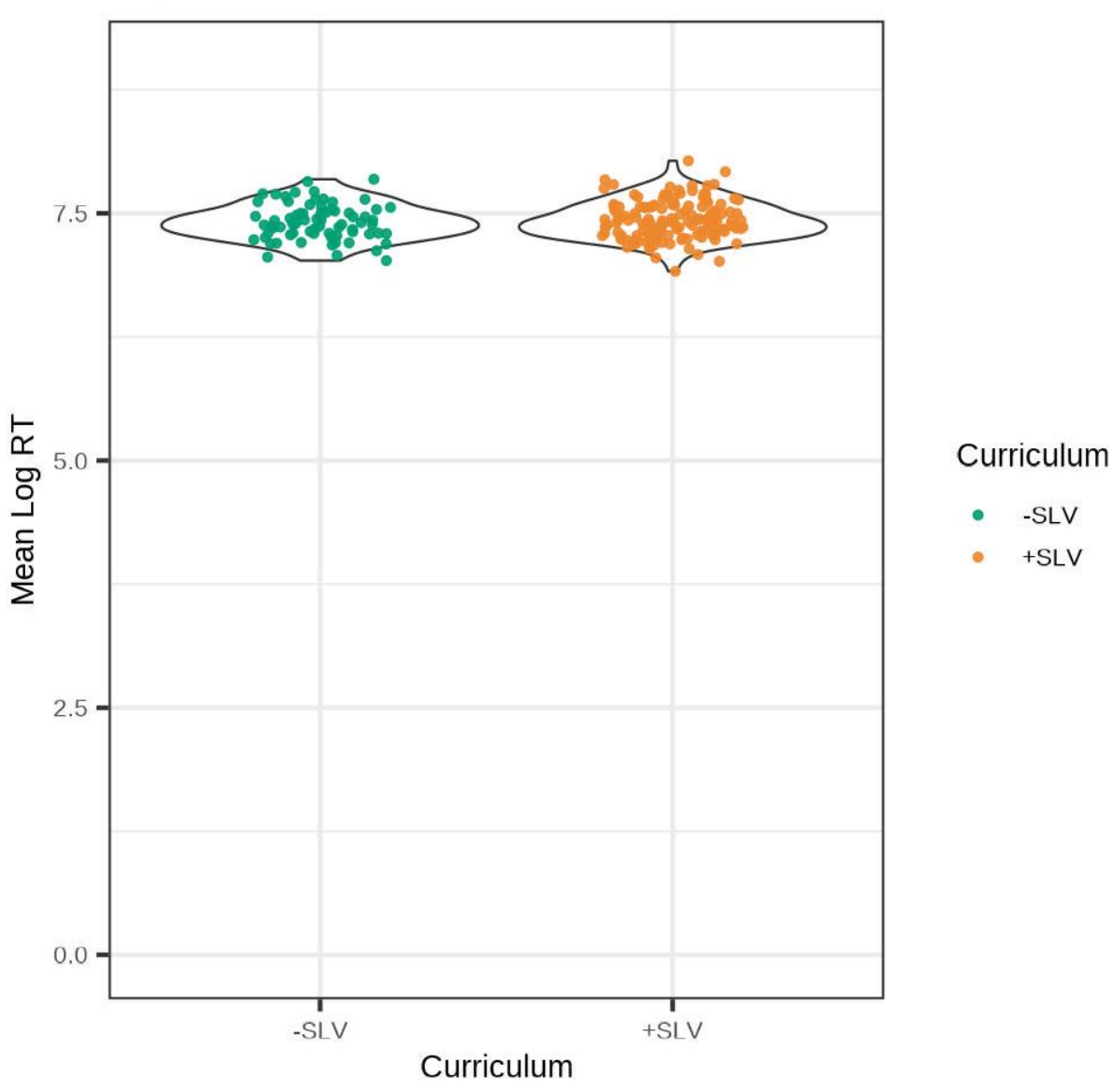
	-SLV	+SLV
N	73	148
Missing	0	0
Mean (CI)	7.41 (7.37, 7.46)	7.43 (7.40, 7.46)
SD	0.18	0.19
Skewness	0.18	0.32
Kurtosis	-0.27	0.15

The +SLV condition has an ever-so-slightly larger standard deviation than that of the -SLV condition (0.19 and 0.18, respectively), which is visually reinforced by the longer spread of +SLV datapoints in Figure 14. Lastly, both the skewness and kurtosis values in Table 9, which

are all less than |2| (Lomax & Hahs-Vaughn, 2012), and a visual check of the datapoints in Figure 14 confirm that the data appear to be normally distributed.

Figure 14

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Log Reaction Time for Vocabulary Meaning Knowledge



Analyses.

The results of the linear mixed effects model for vocabulary meaning knowledge can be seen in Table 10. Similar to the results for form knowledge log RT, there was significant effect of Approach (β_{int} = -0.05, t = -1.86, p = 0.065), although it is not of interest for the current research question (recall that Approach was included in the model for account for the nested nature of the data). Again, the significance of Approach as a factor could indicate that, despite random assignments, groups were not equal at baseline. According to the model, participants took about 2% longer[10] to recall items from the +SLV curricular condition than the -SLV condition. The inferential effect of Curriculum, however, was not statistically significant (β_{+SLV} = 0.02, t = 0.55, p = 0.585), suggesting that there may not be a practical difference between +SLV and -SLV training at the population level.

[10] The beta estimate is exponentiated to back-transform the difference between the reference level (-SLV) and the level of interest (+SLV) into meaningful units.

Table 10

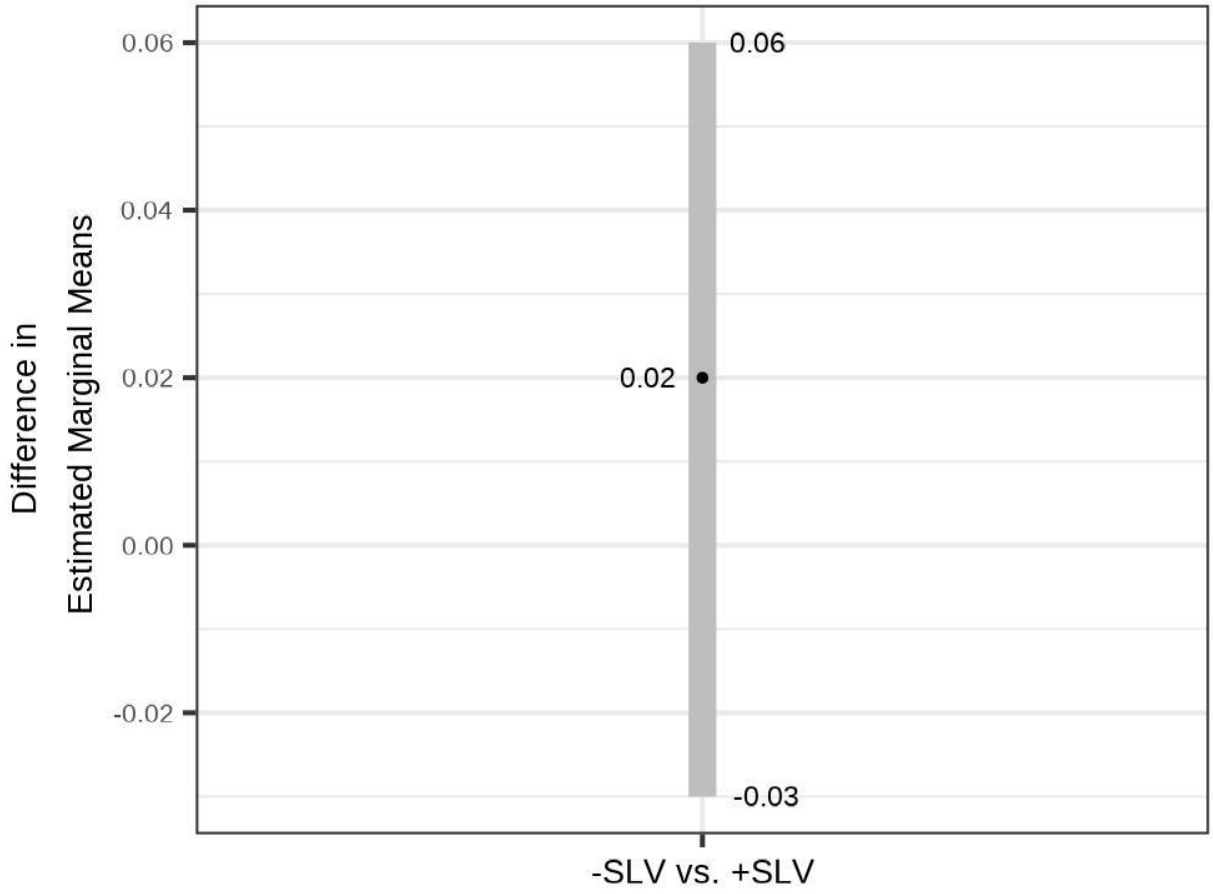
Results of Linear Mixed Effects Model for Vocabulary Meaning Knowledge: Mean Log RT

	Log RT				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	7.44	0.03	7.39 - 7.49	293.71	<0.001
Approach [Int]	-0.05	0.03	-0.11 - 0.00	-1.86	**0.065**
Curriculum [+SLV]	0.02	0.03	-0.04 - 0.07	0.55	0.585
Random Effects					
σ^2	0.02				
$\tau_{00\ participant:curriculum}$	0.02				
ICC	0.53				
$N_{\ participant}$	76				
$N_{\ curriculum}$	2				
Observations	221				
Marginal R^2 / Conditional R^2	0.021 / 0.540				

Following up on the null results of differene testing, mean log reaction time of word form recall was tested for equivalence using estimated marginal means. Figure 15 shows the difference in estimated marginal means between groups with 90% confidence intervals. It largely confirms what was concluded from the descriptive statistics: the -SLV and +SLV outcomes almost completely overlap one another; that is, the difference between them is practically zero.

Figure 15

Difference in Estimated Marginal Means with 90% Confidence Intervals for Vocabulary Meaning Knowledge: Mean Log RT (equivalence bounds not depicted for visual clarity)



For the test of functional equivalence, delta was set at 100 ms (4.61 on the log-transformed scale). The presence of effects greater than the equivalence range was rejected (t = -160.68 , p <.001, d = -0.12). It can be concluded that participants in the population at large would likely be able to recall the correct target word meaning with functionally equivalent speed, regardless of whether they were learned in a +SLV (two registers) or a -SLV (one register) curriculum.

Accuracy

Descriptive Statistics.

The second metric of word meaning knowledge was accuracy as a binary variable (0/1). The descriptive statistics for mean accuracy of vocabulary meaning knowledge can be seen in Table 11. Although the -SLV mean for accuracy is 0.06 higher, the 95% confidence intervals in Table 11 show some overlap in accuracy between curricular conditions. This finding is distinct from the form knowledge accuracy results, in which there was no overlap between conditions (although mean -SLV accuracy was also higher than +SLV accuracy).

Table 11

Descriptive Statistics for Vocabulary Meaning Knowledge: Mean Accuracy

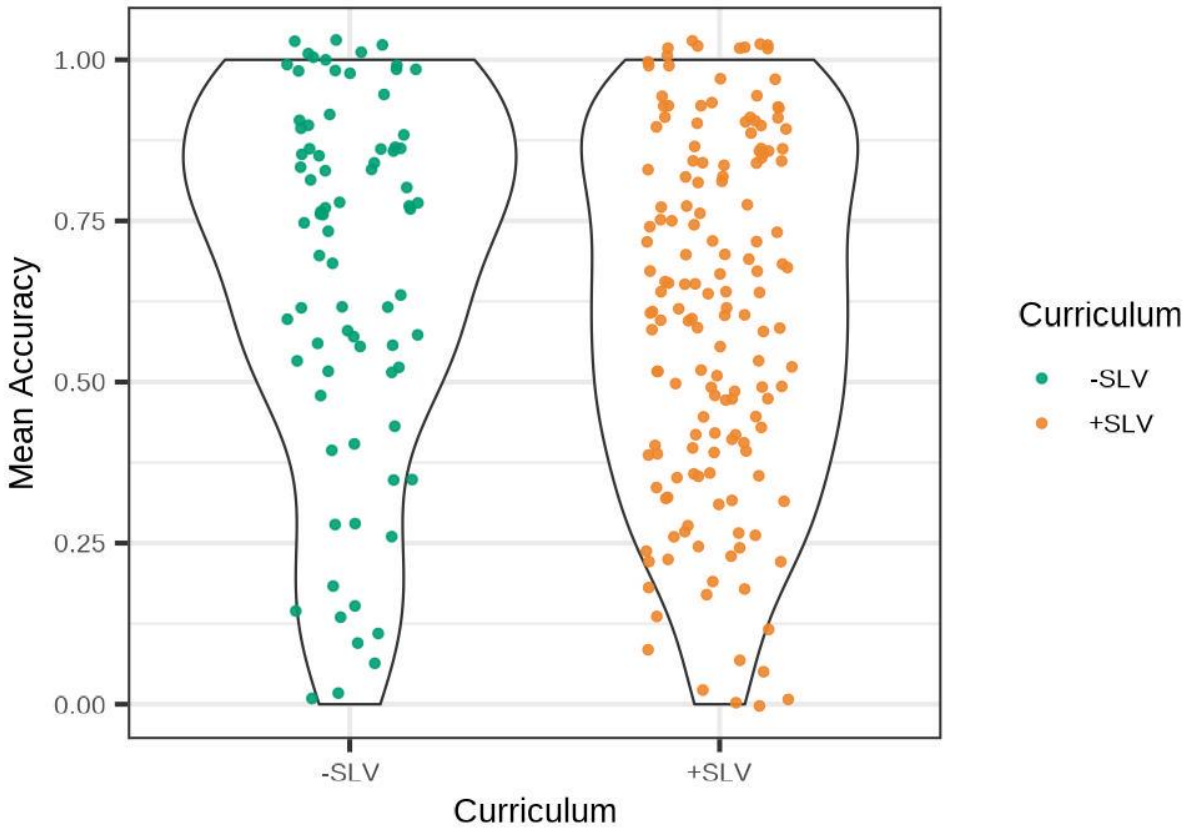
	-SLV	+SLV
N	78	156
Missing	0	0
Mean (CI)	0.66 (0.60, 0.73)	0.60 (0.56, 0.64)
SD	0.29	0.27
Skewness	-0.73	-0.3
Kurtosis	-0.46	-0.88

As with form knowledge accuracy, the -SLV condition has a slightly larger standard deviation than that of the +SLV condition (0.29 and 0.27 respectively). The wider standard deviation higher average scores can be confirmed in the distribution of the datapoints in Figure 16, where -SLV scores are primarily clustered between 0.5 - 1. Furthermore, there seems to have been a ceiling effect in both the -SLV and +SLV curriculum conditions, as the tails are quite short at the upper end of the distribution around 1. This would indicate deviation from normality

of distribution, although the skewness and kurtosis values in Table 11 are all less than |2| (Lomax & Hahs-Vaughn, 2012).

Figure 16

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Accuracy by Participant for Vocabulary Meaning Knowledge



This clustering is confirmed in the model diagnostics (Figure 56 in appendix C), which shows mild deviation between the observed and model-predicted values around 1. As such, caution should be used when inferentially interpreting the model results.

Analyses.

The results of the linear mixed effects model for vocabulary meaning knowledge can be seen in Table 12. Although training in the +SLV group was associated with an 6% decrease in mean accuracy, the effect of +SLV curriculum was not found to be significantly different from -SLV (β_{+SLV} = 0.06, t = -1.45, p = 0.148).

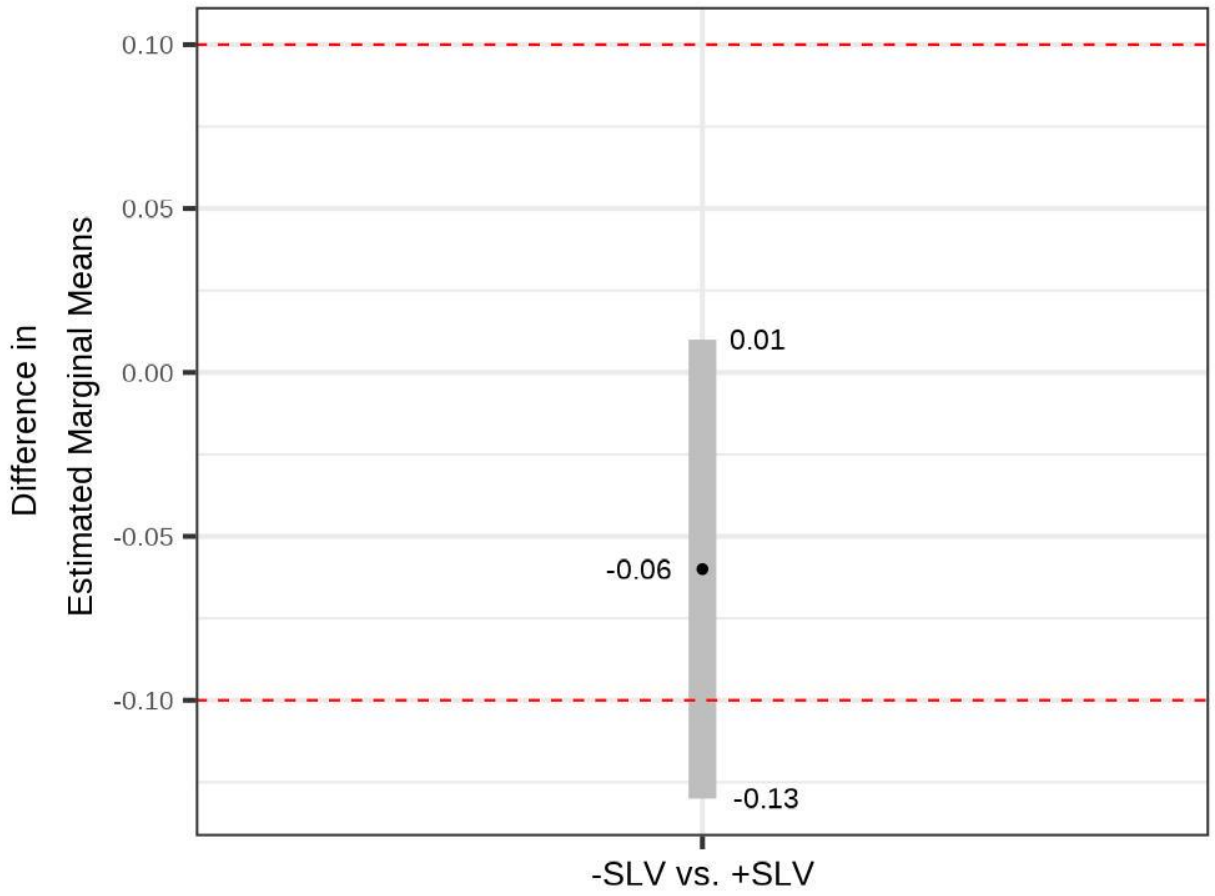
Table 12

Results of Linear Mixed Effects Model for Vocabulary Meaning Knowledge: Mean Accuracy

	Accuracy				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	0.68	0.04	0.61 - 0.76	18.08	<0.001
Approach [Int]	-0.04	0.04	-0.13 - 0.04	-1.01	0.312
Curriculum [+SLV]	-0.06	0.04	-0.15 - 0.02	-1.45	0.148
Random Effects					
σ^2	0.02				
$\tau_{00\ participant:curriculum}$	0.06				
ICC	0.73				
$N_{participant}$	78				
$N_{curriculum}$	2				
Observations	234				
Marginal R^2 / Conditional R^2	0.017 / 0.731				

Figure 17

Difference in Estimated Marginal Means with 90% Confidence Intervals for Vocabulary Meaning Knowledge: Mean Accuracy



Following the advice of Lakens et al. to test for both significance of both difference and equivalence "in order to improve the falsifiability of predictions in psychological science" (2019, p. 267), the linear mixed effect model tested for equivalence using the estimated marginal means. The acceptable level of functional equivalence, delta, was set at a range of within 10% ($\delta = 0.1$). With this limit, the null hypothesis that the groups are not equivalent was retained ($t = -0.85$, $p = .198$, $d = -0.43$). This finding is supported by Figure 17, which shows that a portion of the estimated difference between the -SLV and +SLV conditions falls beyond the lower equivalence

bound. Thus, in the case of word meaning accuracy, neither the test of difference nor equivalence had significant inferential results. All that can be concluded is that, while at the sample level, training in the -SLV curriculum led to 6% more accuracy on average in word meaning knowledge than in the +SLV curriculum, there was nonetheless a high degree of overlap between the mean accoracy scores of the two conditions.

Grammar Recall Knowledge (Negated Sentences)

Grammar recall knowledge refers to participants' ability to produce the correctly conjugated and negated target verb form. In the context of the study, participants were given a full subject-verb-object sentence in English, as well as the Mini-Arabii subject and object translations with a blank in between them. Participants provided the target verb form in the blank. This measure, from a theoretical standpoint, was likely the most difficult for participants, as it required both the verb form retrieval as well as the application of grammar rules.

Reaction Time

Descriptive Statistics.

The first metric of grammar recall knowledge was processing speed: how quickly participants were able to produce the correctly conjugated and negated target verb form. Recall that reaction time data is only based on correctly produced answers. The descriptives statistics for log-transformed mean reaction time of grammar recall knowledge are displayed in Table 13. While the average scores in the -SLV condition are overall faster by 0.09 log ms, the 95% confidence intervals between curricular conditions largely overlap.

Table 13

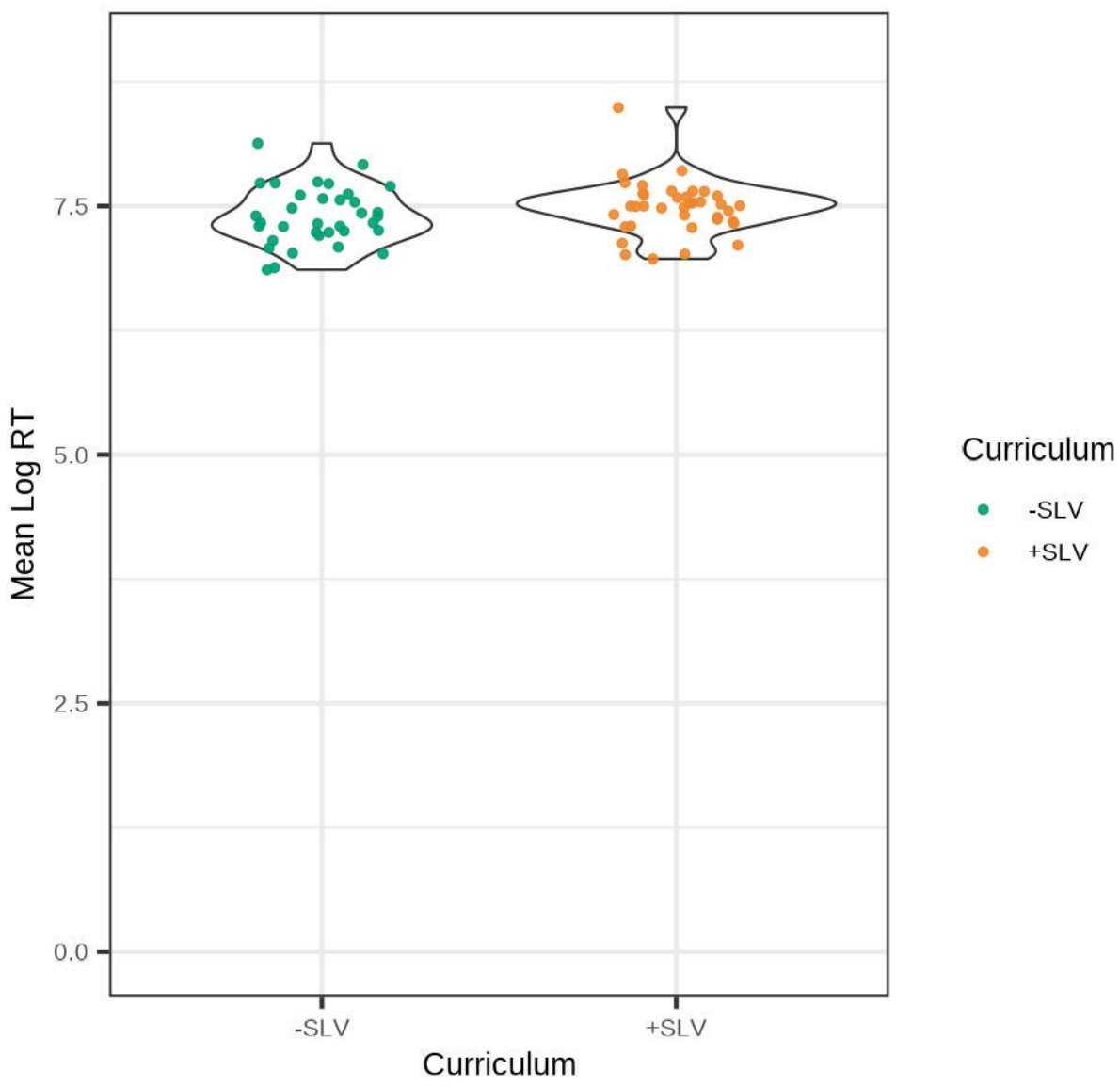
Descriptive Statistics for Grammar Recall Knowledge: Mean Log Reaction Time

	-SLV	+SLV
N	35	38
Missing	0	0
Mean (CI)	7.40 (7.30, 7.50)	7.49 (7.40, 7.58)
SD	0.29	0.27
Skewness	0.68	1.67
Kurtosis	-1.1	1.45

This overlap in the 95% confidence intervals between conditions may be unduly influenced by a left-tail outlier in the +SLV condition (see Figure 18). However, closer inspection of this participant's data did not present any justification for removing them from the sample. Furthermore, skewness and kurtosis values are still all less than |2| (Lomax & Hahs-Vaughn, 2012), suggesting normality of distribution.

Figure 18

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Log Reaction Time for Grammar Recall Knowledge



Analyses.

The results of the linear mixed effects model for grammar recall knowledge can be seen in Table 14. According to the model, participants took about 12% longer[11] to recall items from the +SLV curricular condition than the -SLV curriculum. The inferential effect of Curriculum, however, was not statistically significant (β_{+SLV} = 0.11, t = 1.65, p = 0.104), suggesting that there may not be a practical difference between +SLV and -SLV training at the population level.

Table 14

Results of Linear Mixed Effects Model for Grammar Recall Knowledge: Mean Log Reaction Time

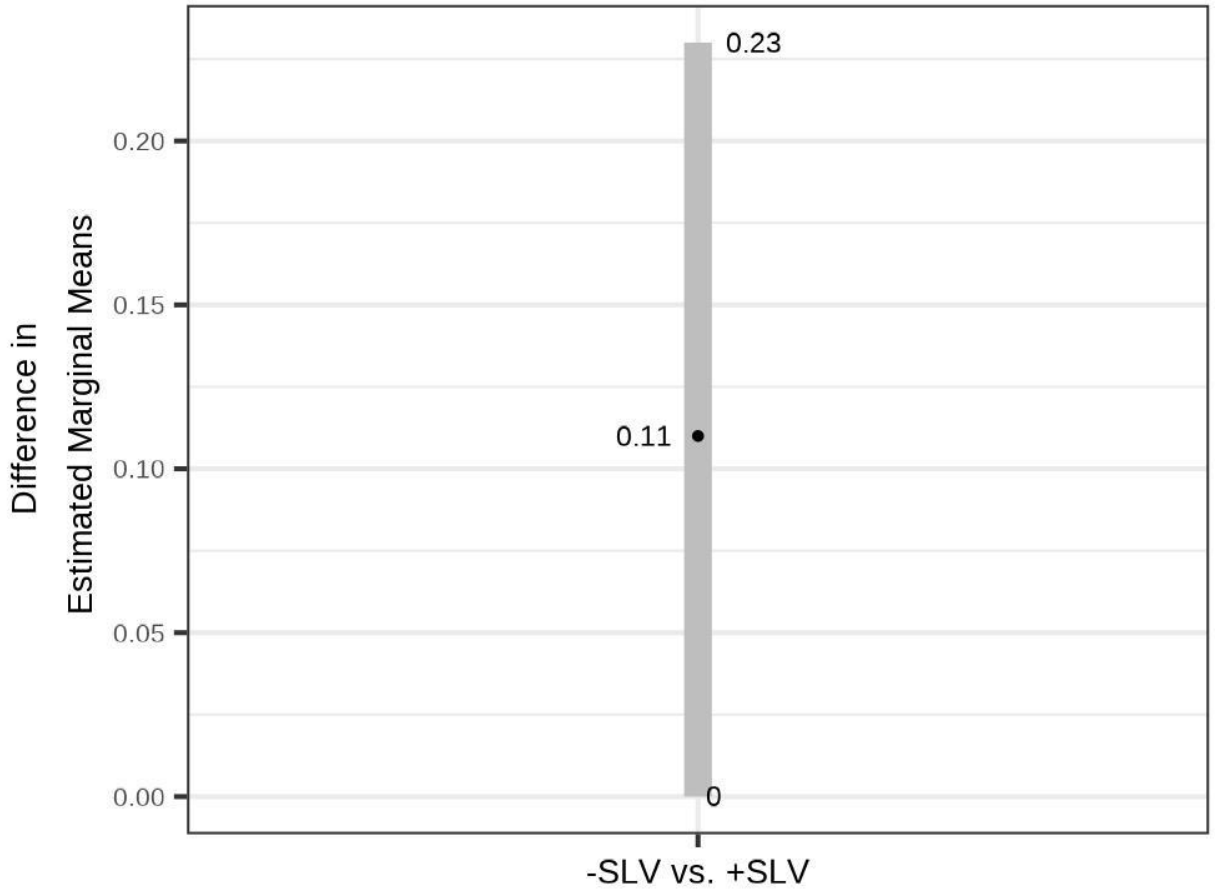
	Log RT				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	7.42	0.06	7.30 - 7.53	129.59	<0.001
Approach [Int]	-0.04	0.07	-0.18 - 0.10	-0.59	0.557
Curriculum [+SLV]	0.11	0.07	-0.02 - 0.25	1.65	0.104
Random Effects					
σ^2	0.03				
$\tau_{00\ participant:curriculum}$	0.05				
ICC	0.59				
$N_{participant}$	50				
$N_{curriculum}$	2				
Observations	73				
Marginal R^2 / Conditional R^2	0.044 / 0.607				

[11] The beta estimate is exponentiated to back-transform the difference between the reference level (-SLV) and the level of interest (+SLV) into meaningful units.

Following up on the null results of difference testing, mean log reaction time of grammar recall was tested for equivalence using estimated marginal means. The difference in the two groups' estimated marginal means with 90% confidence intervals are illustrated in Figure 19.

Figure 19

Difference in Estimated Marginal Means with 90% Confidence Intervals for Grammar Recall Knowledge: Mean Log Reaction Time (equivalence bounds not depicted for visual clarity)



For the test of functional equivalence, delta was set at 100 ms (4.61 on the log-transformed scale). The estimated difference is well within the bounds of equivalence bounds of

δ = +/-log(100). As such, the null hypothesis that training in +SLV and -SLV curricula lead to significantly different reaction times for grammar recall was rejected (t = -64.66 , p <.001, d = 0.06). It can be concluded that, at at the population level, participants would likely be able to recall the correctly conjugated and negated verb with functionally equivalent speed (within 1/10th of a second) regardless curriculum. Learning grammar in two registers (+SLV) rather than one (-SLV) does not appear to come at a cost in terms of grammar production.

Levenshtein Distance

Descriptive Statistics.

The second metric of grammar recall knowledge was an approximation of accuracy: the Levenshtein Distance. The Levenshtein Distance counts the number of changes needed to arrive at the correct form. The descriptive statistics of Levenshtein Distance scores for grammar recall knowledge can been seen in Table 15. The -SLV curricular condition has an overall lower score by 0.64 (indicating fewer changes needed, on average, to arrive at the correct word form) than +SLV. Furthermore, the 95% confidence intervals largely do not overlap (-SLV: [1.81, 2.74], +SLV: [2.58, 3.24]).

Table 15

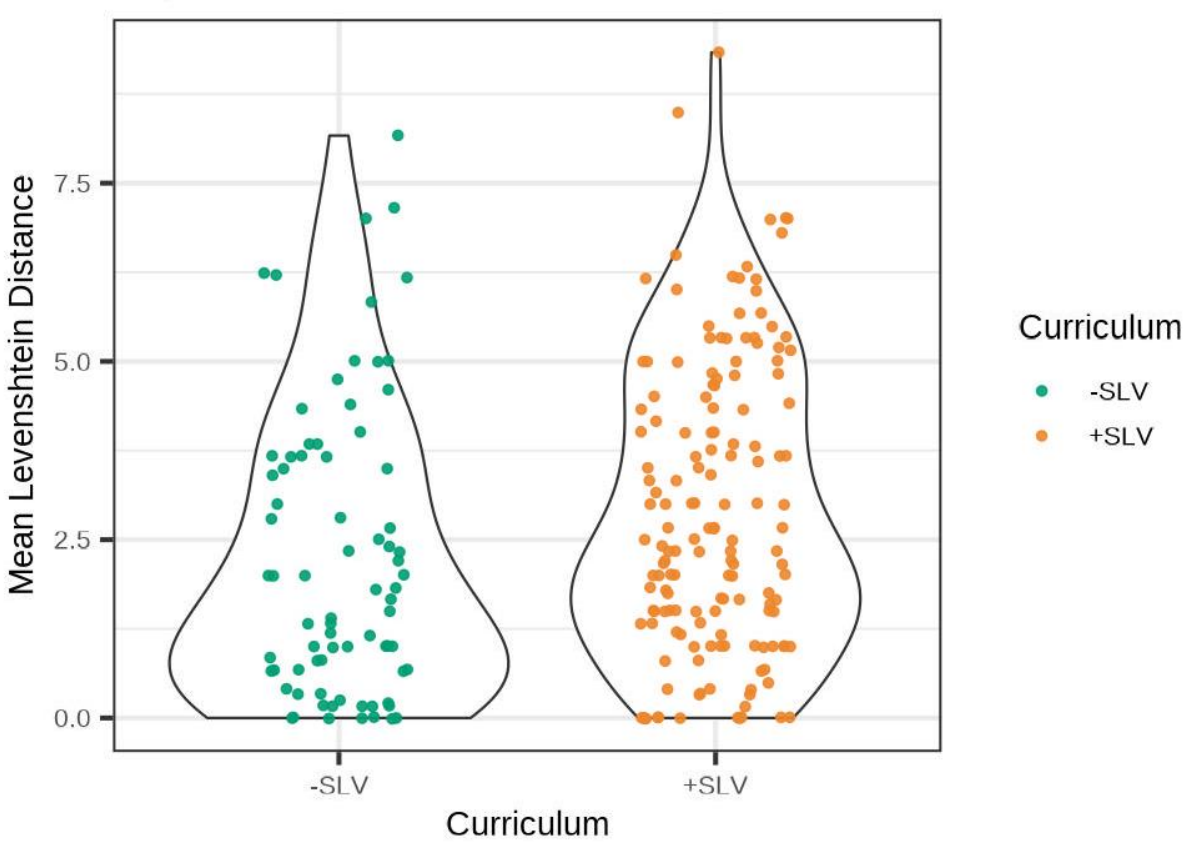
Descriptive Statistics for Grammar Recall Knowledge: Mean Levenshtein Distance

	-SLV	+SLV
N	77	154
Missing	1	2
Mean (CI)	2.27 (1.81, 2.74)	2.91 (2.58, 3.24)
SD	2.04	2.05
Skewness	0.91	0.53
Kurtosis	0.06	-0.37

The overall spread of scores between the two conditions is fairly similar, as indicated by the almost their identical standard deviations (SD_{-SLV} = 2.04; SD_{+SLV} = 2.05). A visual check of the violin plots in Figure 20 shows evidence of deviations from the normal distribution, with a clustering of mean Levenshtein distance scores around 0 in the -SLV condition. On the other hand, the skewness and kurtosis values in Table 15 are all less than |2| (Lomax & Hahs-Vaughn, 2012), which would suggest that deviation from normality is minimal.

Figure 20

Violin Plot Illustrating the Distribution of Mean Levenshtein Distance for Grammar Recall Knowledge



Analyses.

The results of the linear mixed effects model for grammar recall knowledge can be seen in Table 16. There was a significant effect of Curriculum on Levenshtein Distance, where training in the +SLV group was associated with an additional 0.64 changes needed to arrive at the correctly conjugated and negated verb form answer (β_{+SLV} = 0.64, t = 3.14, p = 0.002).

Table 16

Results of Linear Mixed Effects Model for Grammar Recall Knowledge: Mean Levenshtein Distance

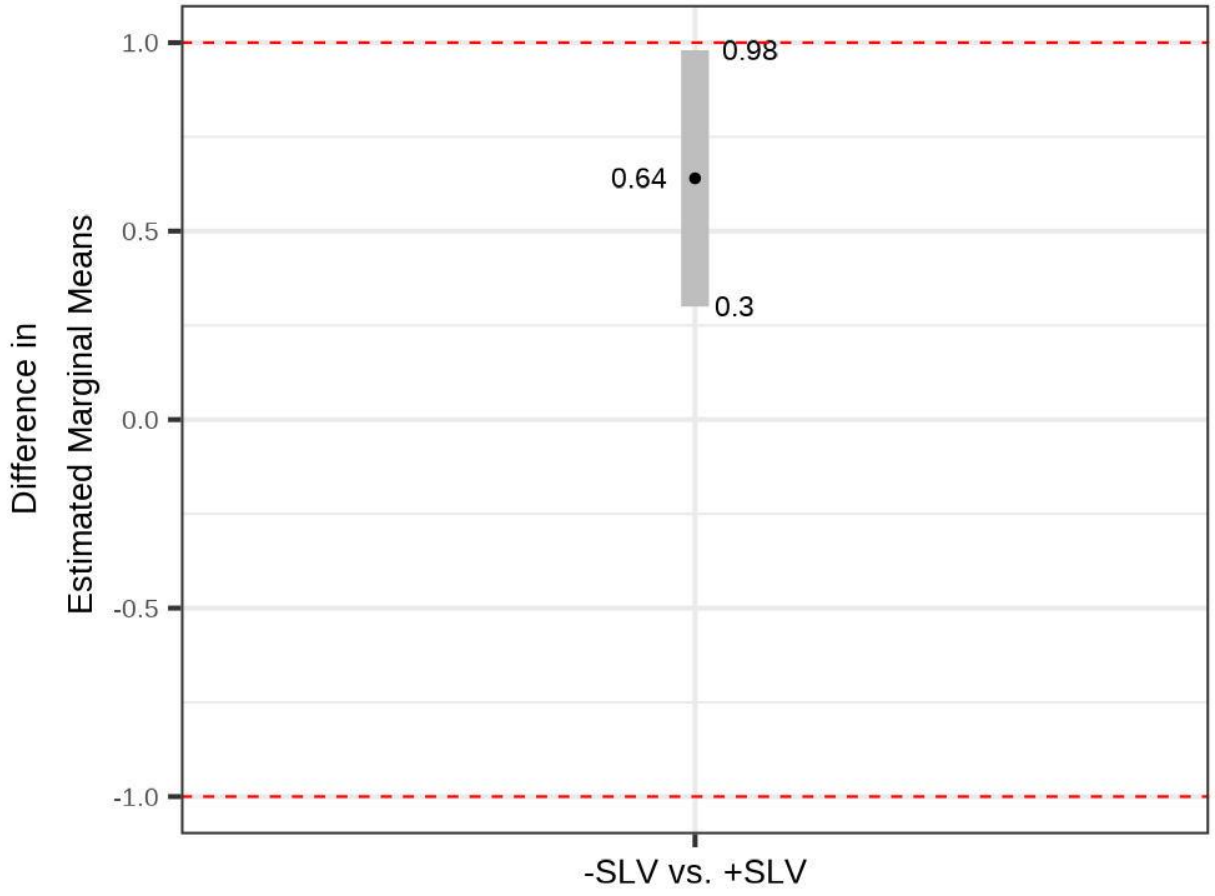
	Levenshtein Distance				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	2.13	0.30	1.55 - 2.72	7.20	<0.001
Approach [Int]	0.30	0.39	-0.47 - 1.07	0.76	0.447
Curriculum [+SLV]	0.64	0.20	0.24 - 1.04	3.14	0.002
Random Effects					
σ^2	1.75				
$\tau_{00\ participant:curriculum}$	0.28				
$\tau_{00\ participant}$	2.18				
ICC	0.58				
$N_{\ participant}$	77				
$N_{\ curriculum}$	2				
Observations	231				
Marginal R^2 / Conditional R^2	0.026 / 0.595				

Finally, the linear mixed effect model was tested for equivalence using the estimated marginal means. Delta was set at 1 (representing 1 letter change needed to arrive at the correct

answer). With this limit, the null hypothesis that there is indeed a difference between the two Curriculum groups was rejected (t = -1.78, p = .039, d = 0.48).

Figure 21

Difference in Estimated Marginal Means with 90% Confidence Intervals for Grammar Recall Knowledge: Mean Levenshtein Distance



These inferential findings confirmed in Figure 21, which illustrates the estimated difference between the two groups' marginal means with 90% confidence intervals. The estimated difference just barely falls within the upper equivalence boundary. As such, it can be

concluded that training in the +SLV and the -SLV condition are functionally equivalent in terms of Levenshtein distance scores for grammar recall.

Thus, the results of the difference and equivalence testing draw different inferential conclusions. Such conflicting results are indeed possible (see E. Walker & Nowacki, 2011 for a discussion). In this case, the difference of 0.64 is considered statistically different (with the test statistic falling beyond the critical value), yet also statistically equivalent (within the functional equivalency bounds of no more than one letter change to arrive at the correct answer). As such, no clear inferential conclusions can be drawn. All that can be stated is that learning grammar in two registers (+SLV) rather than one (-SLV) is associated with an additional 0.64 letter changes needed to arrive at the correct form.

Accuracy

Descriptive Statistics.

The third metric of grammar recall knowledge was accuracy as a binary variable (0/1). The descriptive statistics for mean accuracy of grammar recall knowledge can be seen in Table 17. Unlike grammar recall reaction time and Levenshtein distance, the descriptive statistics indicate that participants likely did achieve different levels of mean accuracy depending on Curriculum condition. The 95% confidence intervals are close but do not overlap, with the upper bound of +SLV trailing behind the lower bound of -SLV by 0.02.

Table 17

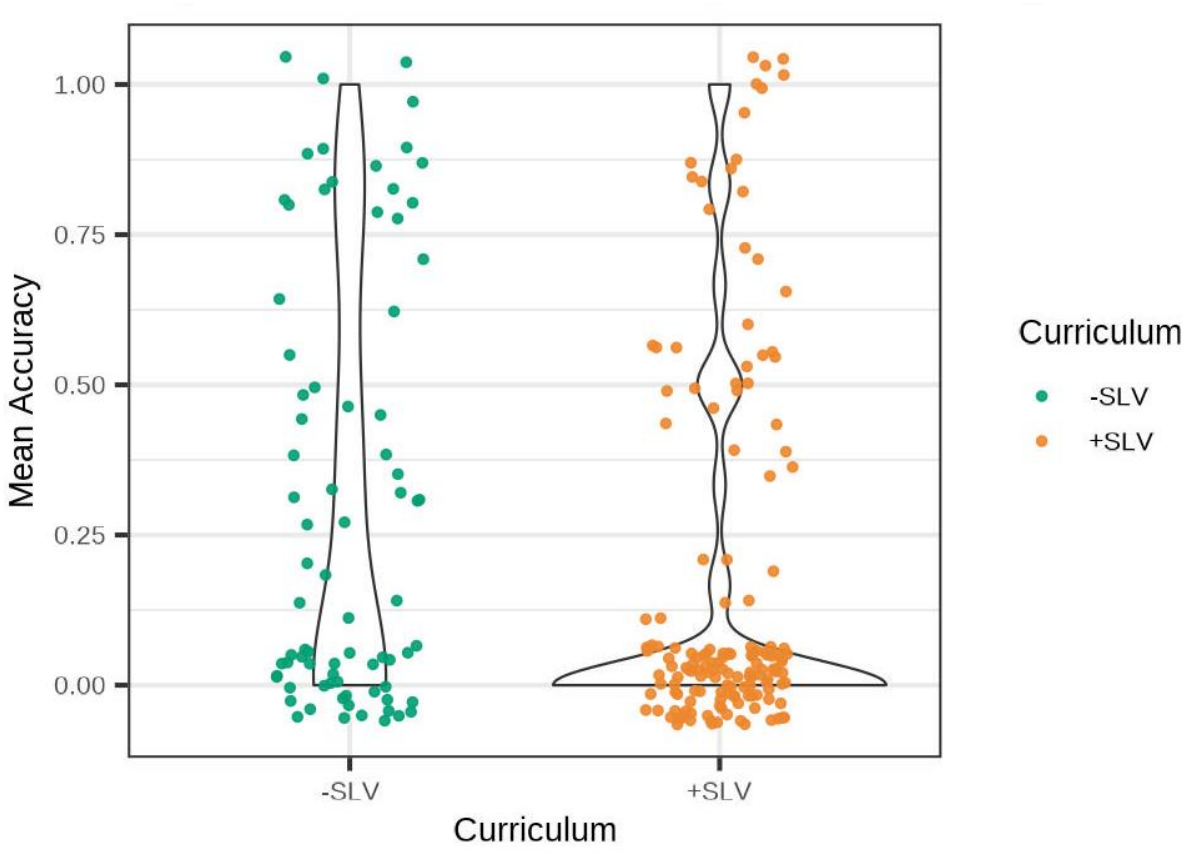
Descriptive Statistics for Grammar Recall Knowledge: Mean Accuracy

	-SLV	+SLV
N	78	156
Missing	0	0
Mean (CI)	0.31 (0.23, 0.39)	0.16 (0.12, 0.21)
SD	0.36	0.3
Skewness	0.68	1.67
Kurtosis	-1.1	1.45

Although the skewness and kurtosis values in Table 17 are all less than |2| (Lomax & Hahs-Vaughn, 2012), a visual inspection of the violin plots in Figure 22 indicates that the data are left skewed to toward 0 in the +SLV condition.

Figure 22

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Accuracy by Participant for Grammar Recall Accuracy



This skew is confirmed in the model diagnostics (Figure 59 in appendix C), which shows a strong deviation between the observed and model-predicted values around 0. Furthermore, the QQ plot of residuals points to evidence of heteroskedasticity. As such, caution should be used when inferentially interpreting the model results.

Analyses.

The results of the linear mixed effects model for grammar recall knowledge can be seen in Table 18. There was a significant effect for Curriculum, where training in the +SLV group was associated with a 14% decrease in mean accuracy (β = -0.14, t = -3.06, p = 0.002).

Table 18

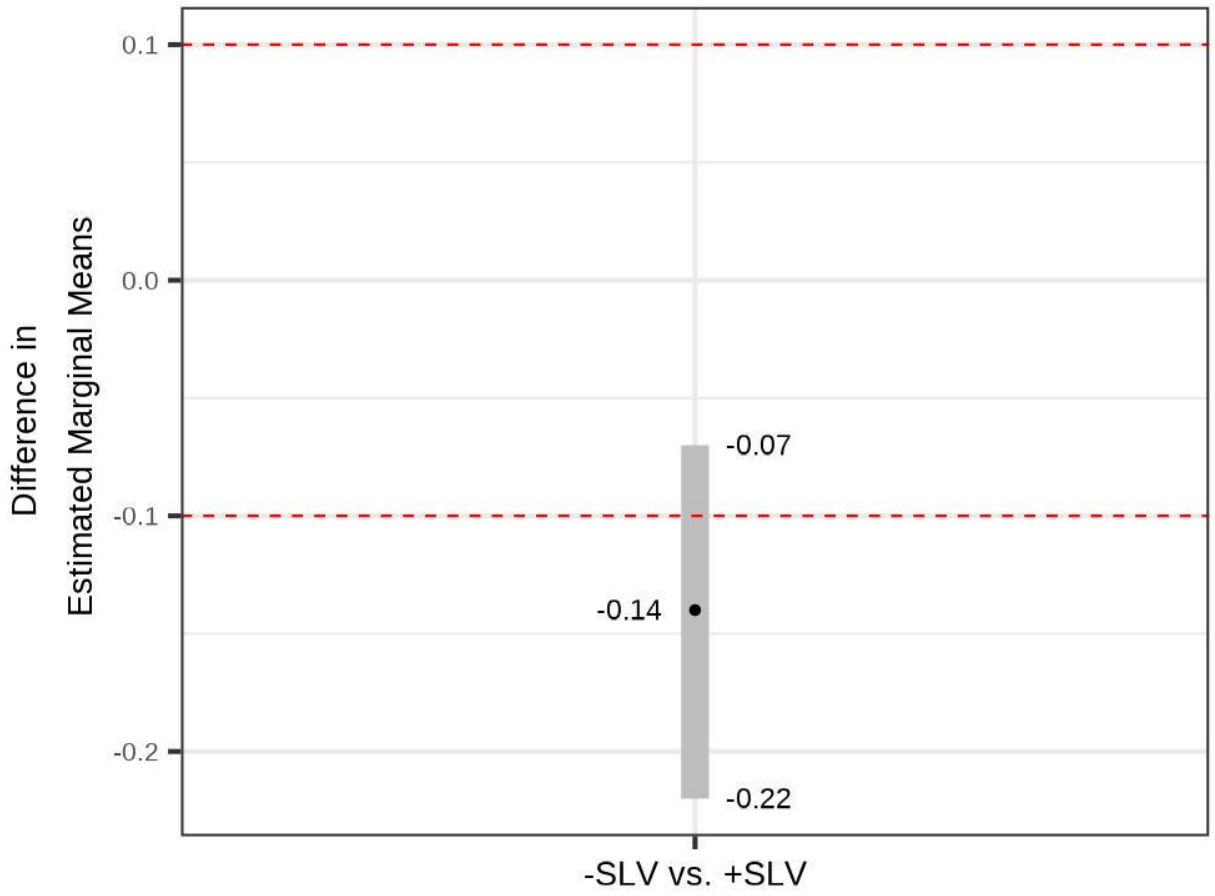
Results of Linear Mixed Effects Model for Grammar Recall Knowledge: Mean Accuracy

	Accuracy				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	0.30	0.04	0.22 - 0.38	7.07	<0.001
Approach [Int]	0.02	0.05	-0.08 - 0.11	0.34	0.738
Curriculum [+SLV]	-0.14	0.05	-0.24 - -0.05	-3.06	0.002
Random Effects					
σ^2	0.07				
$\tau_{00\ participant:curriculum}$	0.04				
ICC	0.35				
$N_{\ participant}$	78				
$N_{\ curriculum}$	2				
Observations	234				
Marginal R^2 / Conditional R^2	0.044 / 0.375				

To further explore the null findings of difference, post-hoc testing for equivalence between estimated marginal means was conducted. Delta was set at a range of within 10% (δ = 0.1). With this limit, the null hypothesis that the groups are not equivalent was retained (t = 3.06, p = .826, d = -0.56).

Figure 23

Difference in Estimated Marginal Means with 90% Confidence Intervals for Grammar Recall Knowledge: Mean Accuracy (equivalence bounds depicted in red)



This finding is supported by Figure 23, which shows that the bulk of the estimated difference between the two groups falls beyond the lower equivalence bound. Thus, it cannot be concluded at a population level that training in a +SLV and -SLV curriculum led to functionally equivalent accuracy in terms of grammar recall. Within the sample data, accuracy in the +SLV curricular condition was over 10% less accurate than in the -SLV curricular condition.

Grammar Recognition Knowledge (Negated Sentences)

Grammar recognition knowledge refers to participants' ability to produce the correct English translation of the conjugated and negated target verb form. In the context of the study, participants were given a full subject-verb-object sentence in Mini-Arabii, as well as the English subject and object translations with a blank in between them. Participants provided the English translation of the verb form in the blank.

Reaction Time

Descriptive Statistics.

The first metric of grammar recognition knowledge was processing speed: how quickly participants were able to translate the negated verb form into English. As can be seen in Table 19, the descriptives statistics for log-transformed mean reaction time are fairly similar between the +SLV and -SLV conditions. Mean log RT of -SLV is somewhat lower than that of +SLV (7.17 log ms and 7.24 log ms respectively). This indicates that reaction time for correctly producing grammar recognition knowledge was, on average, slightly faster in the no-variation condition. However, the 95% confidence intervals overlap to a large extent, suggesting a degree of parity between these two curricula.

Table 19

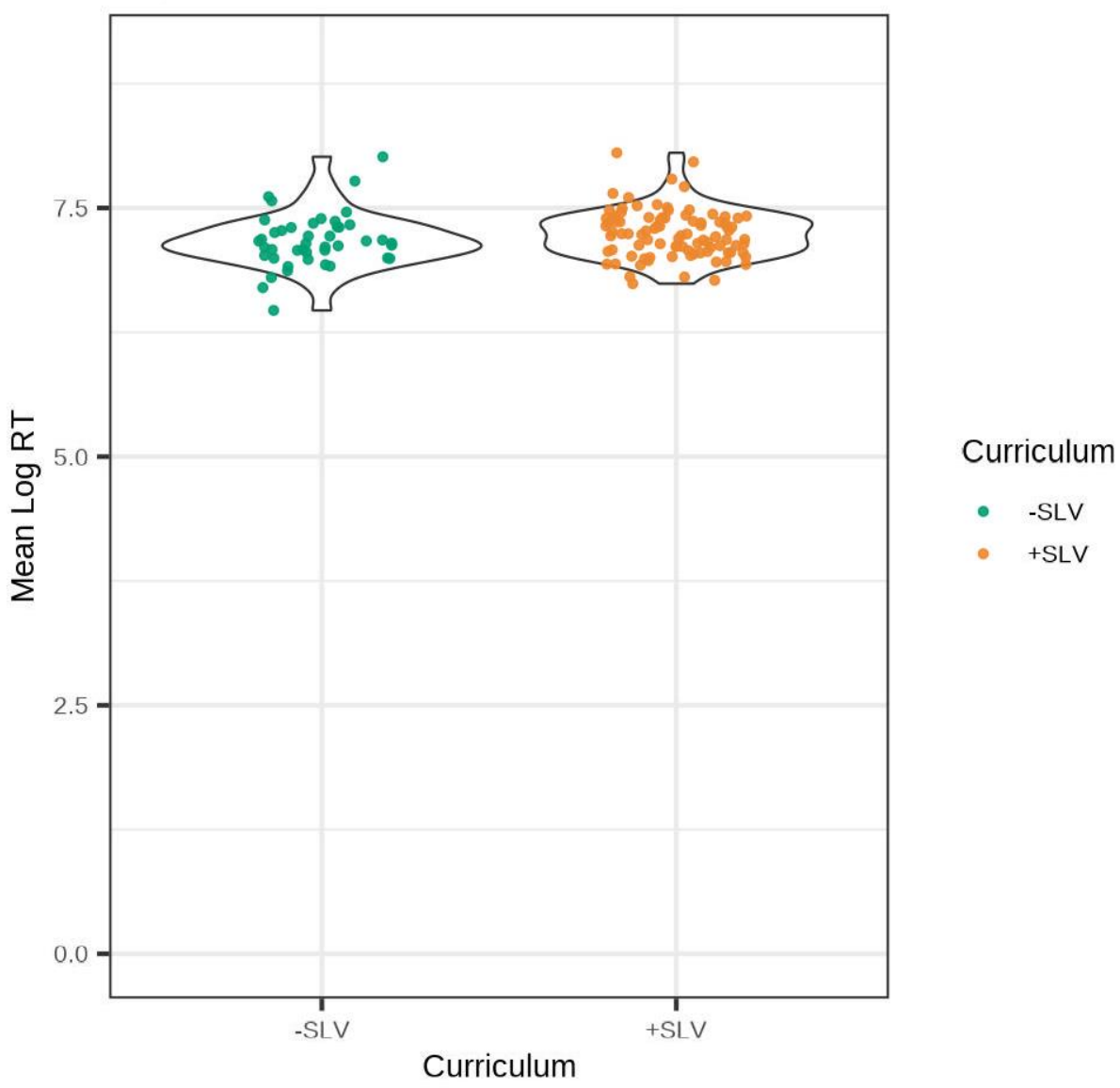
Descriptive Statistics for Grammar Recognition Knowledge: Mean Log Reaction Time

	-SLV	+SLV
N	44	92
Missing	0	0
Mean (CI)	7.17 (7.09, 7.25)	7.24 (7.19, 7.29)
SD	0.27	0.24
Skewness	-0.74	-0.9
Kurtosis	4.66	7.06

Although the spread of datapoints in the violin plots in Figure 24 appear to be somewhat normally distributed, the data are highly leptokurtic, which indicates deviations from normality. Despite this finding, the distribution of residuals along the QQ plot suggest normality.

Figure 24

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Log Reaction Time for Grammar Recognition Knowledge



Analyses.

The results of the linear mixed effects model for grammar recognition knowledge can be seen in Table 20. There was a significant effect of Curriculum, where training in the +SLV

curricular condition was associated with reaction times 7% longer[12] than those in the -SLV condition ($\beta = 0.07$, $t = 2.7$, $p = 0.008$).

Table 20

Results of Linear Mixed Effects Model for Grammar Recognition Knowledge: Mean Log Reaction Time

	Log RT				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	7.21	0.05	7.11 - 7.31	145.22	<0.001
Approach [Int]	-0.07	0.07	-0.20 - 0.06	-1.00	0.322
Curriculum [+SLV]	0.07	0.03	0.02 - 0.13	2.70	0.008
Random Effects					
σ^2	0.01				
$\tau_{00\ participant:curriculum}$	0.01				
$\tau_{00\ participant}$	0.04				
ICC	0.80				
$N_{\ participant}$	49				
$N_{\ curriculum}$	2				
Observations	136				
Marginal R^2 / Conditional R^2	0.033 / 0.804				

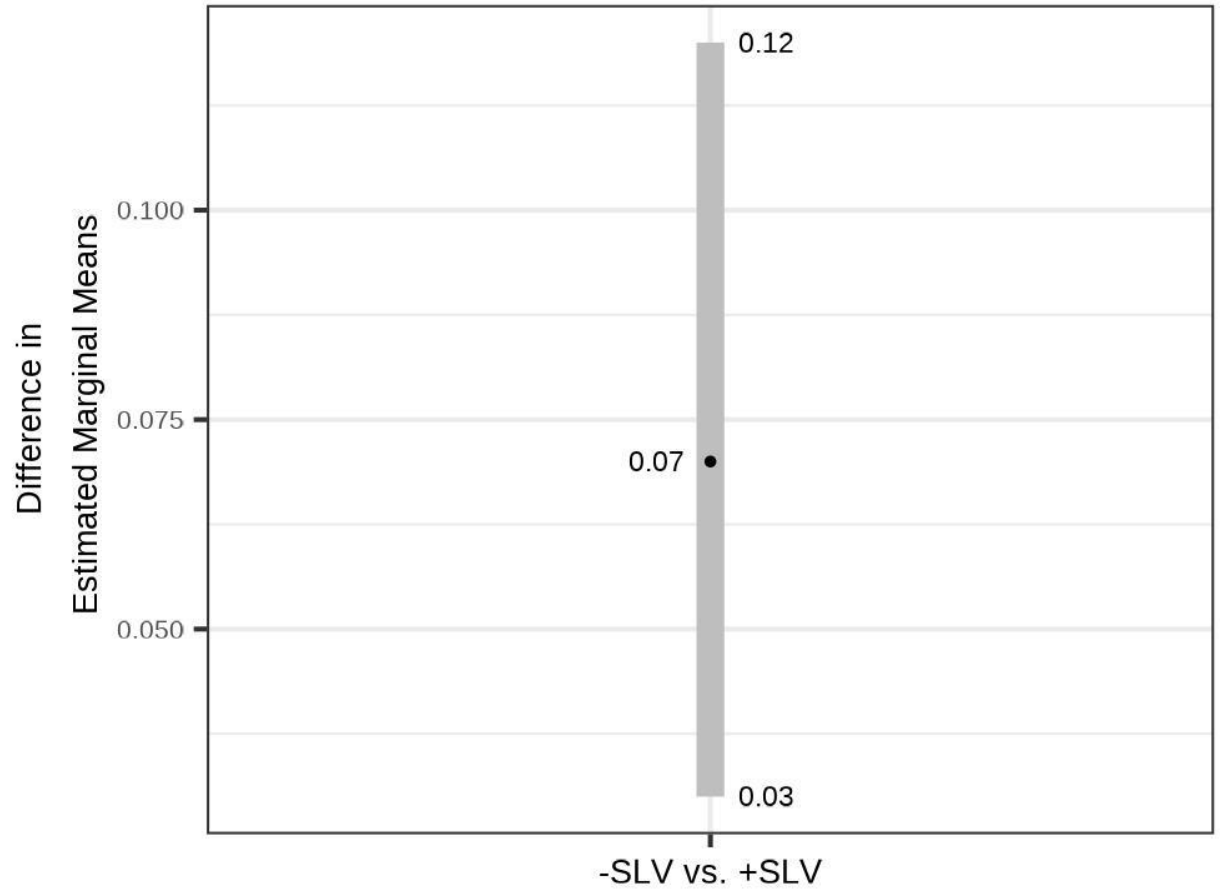
In order to test for equivalence, estimated marginal means were calculated for mean log reaction time of grammar recognition. Figure 25 shows the difference in estimated marginal

[12] The beta estimate is exponentiated to back-transform the difference between the reference level (-SLV) and the level of interest (+SLV) into meaningful units.

means between groups with 90% confidence intervals. Despite the inferential findings of significant difference, the difference between -SLV and +SLV outcomes clearly falls within the equivalence bounds of +/- log(300) (4.61 on the log-transformed scale).

Figure 25

Difference in Estimated Marginal Means with 90% Confidence Intervals for Grammar Recognition Knowledge: Mean Log Reaction Time (equivalence bounds not depicted for visual clarity)



Following the advice of Lakens et al. to test for both significance of both difference and equivalence "in order to improve the falsifiability of predictions in psychological science" (2019, p. 267), the linear mixed effect model tested for equivalence using the estimated marginal means. Delta was set at 100 ms (4.61 on the log-transformed scale). The presence of effects greater than the equivalence range was rejected (t = -167.07, p <.001, d = 0.64). Thus, the test of equivalence produces results that appear, on the surface, contrary to the test of difference: grammar recognition reaction time is both inferentially different *and* equivalent. While this is indeed statistically possible (e.g., Godfroid & Spino, 2015), it effectively means that results cannot be reliably generalized beyond the sample population. What can be concluded is participants in the +SLV curricular condition learned two registers together at a cost of reaction times that were, on average, 7% slower than those in the -SLV condition who only learned one register.

Accuracy

Descriptive Statistics.

The final metric of grammar recognition knowledge was accuracy as a binary variable (0/1). As can been seen in Table 21, the descriptive statistics of mean accuracy for grammar recognition knowledge are fairly similar between the +SLV and -SLV conditions. While -SLV has a slightly lower aggregate mean than +SLV, the 95% confidence intervals largely overlap (-SLV: [0.35, 0.55], +SLV: [0.35, 0.48]).

Table 21

Descriptive Statistics for Grammar Recognition Knowledge: Mean Accuracy

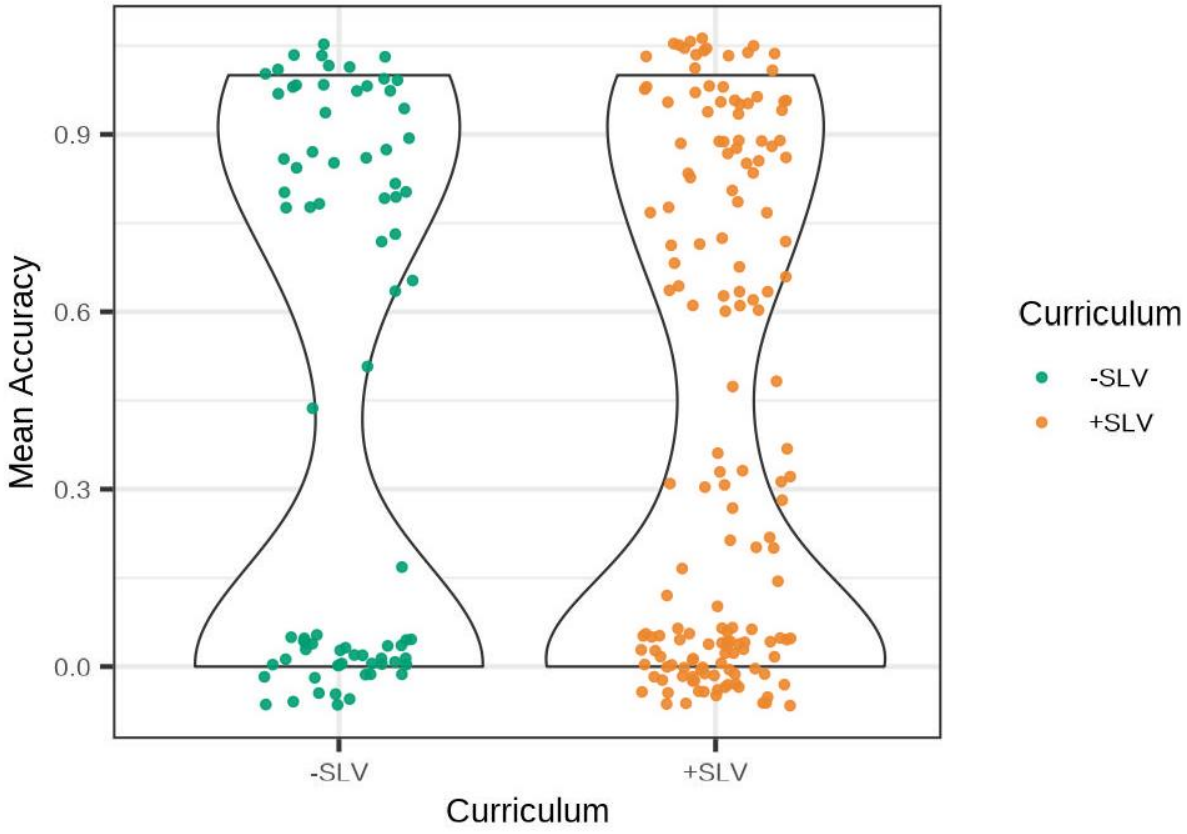
	-SLV	+SLV
N	78	156
Missing	0	0
Mean (CI)	0.45 (0.35, 0.55)	0.42 (0.35, 0.48)
SD	0.45	0.42
Skewness	0.07	0.26
Kurtosis	-1.91	-1.7

Although the skewness and kurtosis values Table 21 are all less than |2| (Lomax & Hahs-Vaughn, 2012), a visual check of the spread of datapoints in Figure 26 indicates that the data are largely polarized – participants either did fairly well or rather poorly on grammar recognition accuracy regardless of curricular condition. This finding is likely due to the fact that many participants forgot to include the apostrophe when typing out the English negated forms "doesn't" and "didn't." This polarization of results is confirmed by the density plot in

0, which has two peaks around 0 and 1, and in the heavy-tailed residuals of the QQ plot. Thus, the inferential model below is likely less accurate in capturing variability among the more average-performing participants.

Figure 26

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Accuracy by Participant for Grammar Recognition



Analyses.

The results of the linear mixed effects model for grammar recall knowledge can be seen in Table 22. According to the model, participants were about 4% less accurate in the +SLV curricular condition than the -SLV curriculum. This finding, however, was not statistically significant (β = -0.04, t = -1.18, p = 0.239), suggesting that there may not be a practical difference between +SLV and -SLV training at the population level in terms of grammar recall accuracy.

Table 22

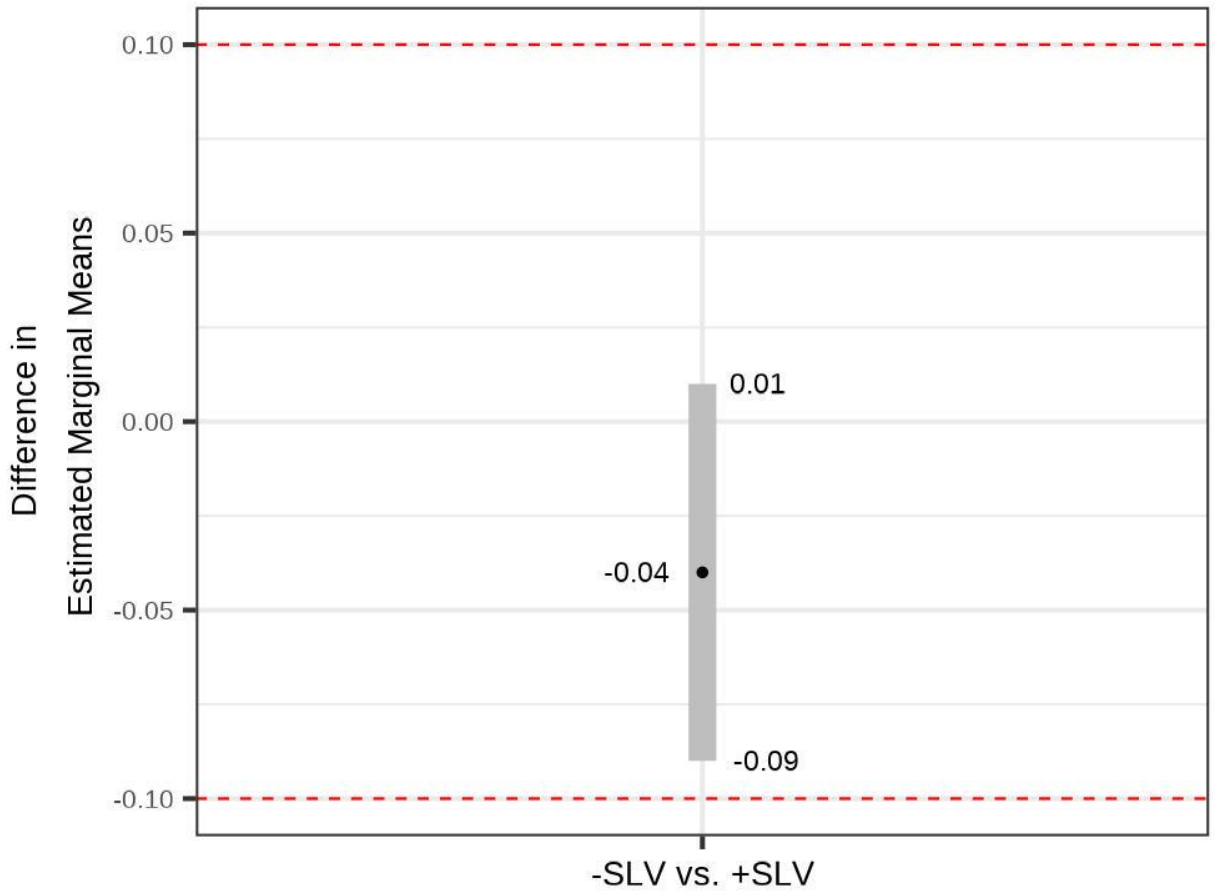
Results of Linear Mixed Effects Model for Grammar Recognition Knowledge: Mean Accuracy

	Accuracy				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	0.42	0.07	0.29 - 0.55	6.39	<0.001
Approach [Int]	0.07	0.09	-0.11 - 0.25	0.80	0.426
Curriculum [+SLV]	-0.04	0.03	-0.10 - 0.02	-1.18	0.239
Random Effects					
σ^2	0.03				
$\tau_{00\ participant:curriculum}$	0.01				
$\tau_{00\ participant}$	0.14				
ICC	0.84				
$N_{\ participant}$	78				
$N_{\ curriculum}$	2				
Observations	234				
Marginal R^2 / Conditional R^2	0.008 / 0.840				

Following up on the null results of difference testing, mean accuracy of grammar recognition was tested for equivalence using estimated marginal means. The estimated difference between the two groups' marginal means with 90% confidence intervals is illustrated in Figure 27. As seen below, the range of plausible differences falls completely within the equivalence bounds.

Figure 27

Difference in Estimated Marginal Means with 90% Confidence Intervals for Grammar Recognition Knowledge: Mean Accuracy



In keeping with the analytic plan, post-hoc testing for equivalence between estimated marginal means was conducted. Delta was set at a range of within 10% ($\delta = 0.1$). The null hypothesis that the groups are not equivalent was rejected ($t = -2.07$, $p = .021$, $d = -0.21$). Thus, it seems that there may not be a practical difference between +SLV and -SLV training at the population level in terms of grammar recognition accuracy. Recall, however, that within the descriptive data participants either performed quite well or fairly poorly, likely due to mistyping of the apostrophe. Given the deviations from normality discussed above, it may be safer to draw

a conclusion from the sample data. Within the sample data, the cost of studying two registers at the same time (+SLV) was associated with, on average, a 3% decrease in accuracy in recalling the correctly conjugated and negated verb.

Summary of Chapter 3 Results: -SLV vs. +SLV

In this chapter, I compared the learning outcomes of training in a curriculum that incorporates sociolinguistic variation against a curriculum that does not. These curricular choices were respectively operationalized as +SLV (studying two registers, either learned simultaneously or back-to-back) and -SLV (studying only one register). The hypothesis was that training in one register (-SLV) would lead to superior outcomes compared to in two registers (+SLV). This was hypothesized because, from a frequency-based approach to learning, participants will have only half as many exposures to each of the two registers in the +SLV curriculum (three for MSA, three for ECA), as they do for the single register in the -SLV curriculum (six exposures).

The results do not support the hypothesis that training in a -SLV curriculum leads to universally superior outcomes for vocabulary and grammar knowledge. Specifically, it seems that incorporating SLV into a curriculum does not come at a significant cost to processing speed. For all four knowledge types measured (productive and receptive vocabulary and grammar), the difference in average log-transformed reaction time between the two curricular conditions was statistically equivalent (less than 100 ms). The only caveat to these processing speed findings is that the difference between grammar recognition outcomes was also statistically significant. Beyond inferential findings, the descriptive differences offer practical insights into the processing speed costs associated with training in a +SLV curriculum. The difference in mean log reaction times ranged from 1% slower (word form knowledge) to 12% slower (grammar

recall knowledge). Finally, the magnitudes of difference between processing speeds in a +SLV and -SLV curriculum were relatively small by SLA-standards (Plonsky & Oswald, 2014), with effect sizes ranging from $d = 0.06$ (vocabulary form knowledge, grammar recall knowledge), to $d = 0.64$. In sum, both inferentially and practically, training in a -SLV curriculum did not seem to lead to superior outcomes in terms of processing speed.

The results for accuracy, both as an absolute binary measure (0/1) and as an approximate measure (the Levenshtein Distance), presented a more nuanced picture. For the productive knowledge measures (vocabulary form and grammar recall knowledge), the difference in absolute accuracy (0/1) between -SLV and +SLV outcomes was statistically significant. Adding an additional register to a curriculum (moving from -SLV to +SLV) was associated with an 11% decrease in accuracy for word form knowledge (a medium effect size of $d = -0.73$), and a 14% decrease in accuracy for grammar recall knowledge (a small effect size of $d = -0.56$). The magnitude of these differences (that is, Cohen's *d*) decreased, however, when adopting an approximate measure of accuracy. For this, I used the Levenshtein Distance, a metric that counts the minimum number of letter changes required to edit an answer so it is correct. Adding an additional register to a curriculum (moving from -SLV to +SLV) was associated with an additional 0.2 letter edits needed to arrive at the correct vocabulary form (a small effect size of $d = 0.32$), and an additional 0.6 letter edits to arrive at the correct verb form (a small effect size of $d = 0.48$). In both cases, however, no inferential conclusions can be drawn from the Levenshtein Distance scores: the difference and equivalence tests were both significant for vocabulary form, and were both insignificant for grammar recall.

For receptive knowledge outcomes, only accuracy as an absolute (0/1) measure was used. The effect of curricular condition (-SLV vs. +SLV) on the development of vocabulary meaning

knowledge was not statistically significant for either difference or equivalence testing. Adding an additional register to a curriculum (moving from -SLV to +SLV) was associated with a 6% decrease in meaning knowledge accuracy (a small effect size of d = -0.43), and a 4% decrease in grammar recognition knowledge accuracy (a small effect size of d = -0.21).

On the whole, the inferential results suggested that studying in a curriculum that incorporates SLV did not come at a cost to processing speed for learners. There did, however, seem to be a cost associated with accuracy as an absolute measure. The association of form and meaning seemed to be less firmly established in a two-register curriculum compared to a one-register curriculum. This difference between training conditions, however, shrank if an approximate measure of accuracy was used. Beyond inferential conclusions, the magnitudes of difference (as calculated by Cohen's *d*) all ranged from small to medium. Thus, the results did not uniformly support the prediction that more is less in L2 acquisition. Incorporating SLV into a curriculum did not lead to universally inferior outcomes compared to a traditional curriculum which ignores SLV.

Chapter 4: Results (RQ2: Sequential vs. Integrated)

The second research question compared the learning outcomes for two competing approaches to incorporating sociolinguistic variation in an L2 curriculum. These approaches are *integrated* (learning two registers side-by-side as a complete language system) and *sequential* (gaining a solid foundation in one register first, and adding another register second). This difference means that there may be global (main) effects of approach, as well as register-specific effects (MSA vs. ECA) within each approach (interactions). Recall that Mini-Arabii was designed to reflect sociolinguistic variation in both lexical items (nouns) and grammar rules (verb conjugation and negation). Thus, to comprehensively capture the effects of training in an Integrated vs. sequential approach on the development of L2 lexical and grammatical knowledge for both MSA and ECA, a broad array of measures capturing both accuracy and processing speed were employed. Results are organized according to knowledge type and outcome measure.

Word Form Knowledge

Word form knowledge refers to participants' ability to produce the target lexical item. In the context of the study, participants provided the Mini-Arabii translation for the English prompt.

Reaction Time

Descriptive Statistics.

The first metric of word form knowledge was processing speed: how quickly participants were able to produce the correct target form. The descriptives statistics for log-transformed mean reaction time are shown in Table 23. All in all, there is a large degree of similarity for form knowledge between the two approaches and their associated registers. The 95% confidence

intervals around the group means for Integrated and Sequential overlap to a large extent, indicating parity between the two +SLV approaches. Within each approach, the MSA register means are the same (μ = 7.6), while the mean of ECA_{int} trails behind ECA_{seq} by 0.05 log ms.

Table 23

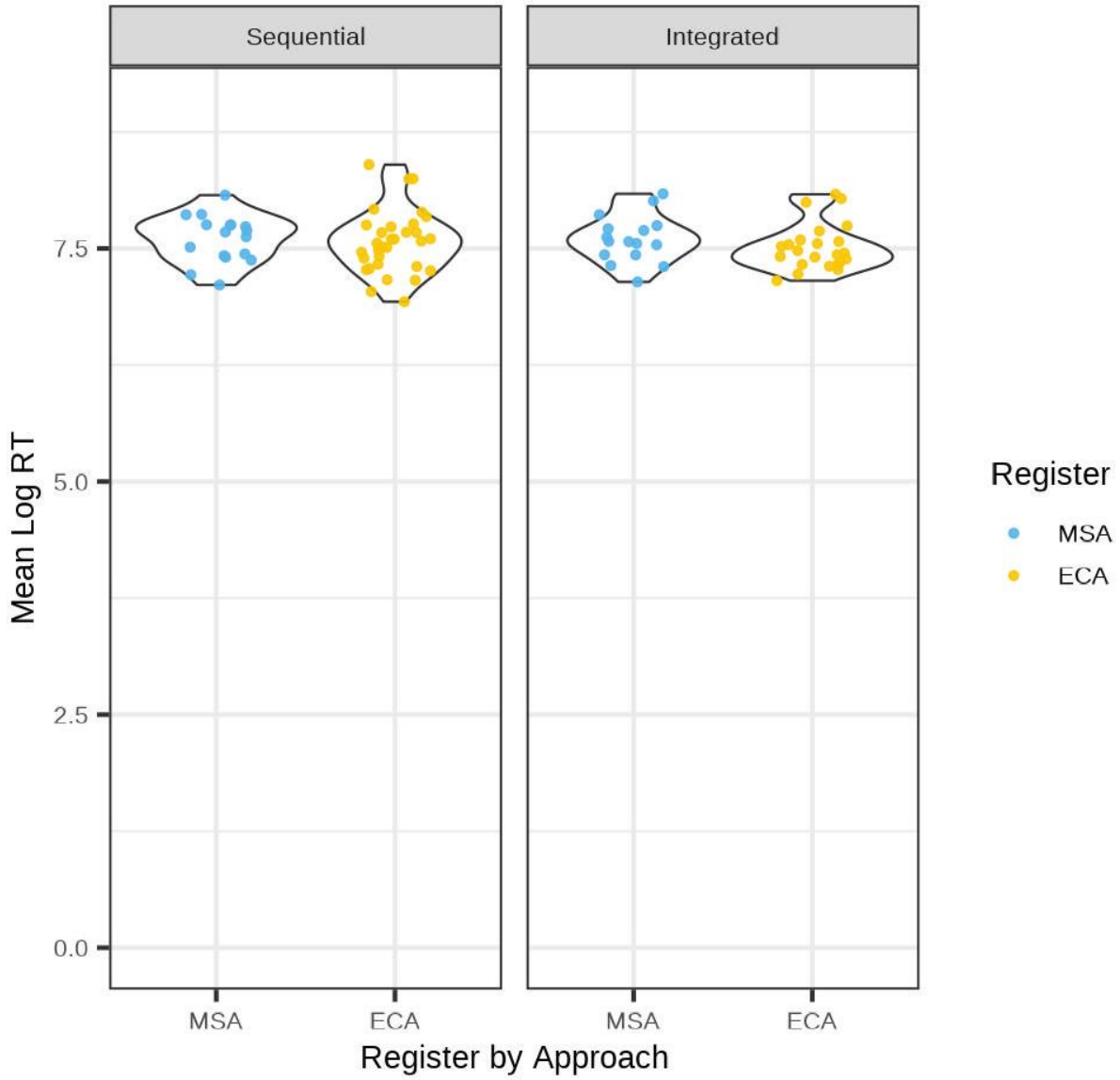
Descriptive Statistics for Vocabulary Form Knowledge: Mean Log Reaction Time

Condition	Combined Registers	MSA only	ECA only
Integrated			
N	39	16	23
Missing	0	0	0
Mean (CI)	7.55 (7.47, 7.63)	7.60 (7.47, 7.73)	7.52 (7.41, 7.62)
SD	0.25	0.25	0.25
Skewness	0.67	0.25	1.05
Kurtosis	-0.14	-0.05	0.58
Sequential			
N	49	17	32
Missing	0	0	0
Mean (CI)	7.58 (7.49, 7.67)	7.60 (7.48, 7.73)	7.57 (7.45, 7.69)
SD	0.31	0.25	0.34
Skewness	0.38	-0.27	0.56
Kurtosis	0.44	-0.24	0.49

All approach-register subgroups have a fairly equivalent spread of data around the mean, apart from ECA_{int} which has both a larger standard deviation as seen in Table 23 and a longer spread of datapoints in Figure 28. Finally, the data appear to be relatively normally distributed according to both a visual check of the datapoints in Figure 28 and the skewness and kurtosis values in Table 23, which are all less than |2| (Lomax & Hahs-Vaughn, 2012).

Figure 28

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Log Reaction Time for Vocabulary Form Knowledge



Analyses.

The results of the linear mixed effects model for vocabulary form knowledge can be seen in Table 24. While there was significant effect of Register (β_{ECA} = -0.1, t = -1.88, p = 0.063), Register as a main effect was not of interest for the current research question. According to the

model, participants were about 5% faster[13] in recalling items learned in the integrated approach compared to Sequential. The inferential effect of Approach, however, was not statistically significant (β_{int} = -0.05, t = -0.55, p = 0.586), suggesting that there may not be a practical difference between the +SLV approaches at the population level.

Table 24

Results of Linear Mixed Effects Model for Vocabulary Form Knowledge: Mean Log RT

	Log RT				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	7.67	0.06	7.54 - 7.79	123.25	<0.001
Approach [Int]	-0.05	0.09	-0.23 - 0.13	-0.55	0.586
Register [ECA]	-0.1	0.05	-0.20 - 0.01	-1.88	**0.063**
Approach x Register	0.01	0.08	-0.15 - 0.16	0.08	0.937
Random Effects					
σ^2	0.02				
$\tau_{00\ participant}$	0.06				
ICC	0.72				
$N_{\ participant}$	57				
Observations	88				
Marginal R^2 / Conditional R^2	0.028 / 0.729				

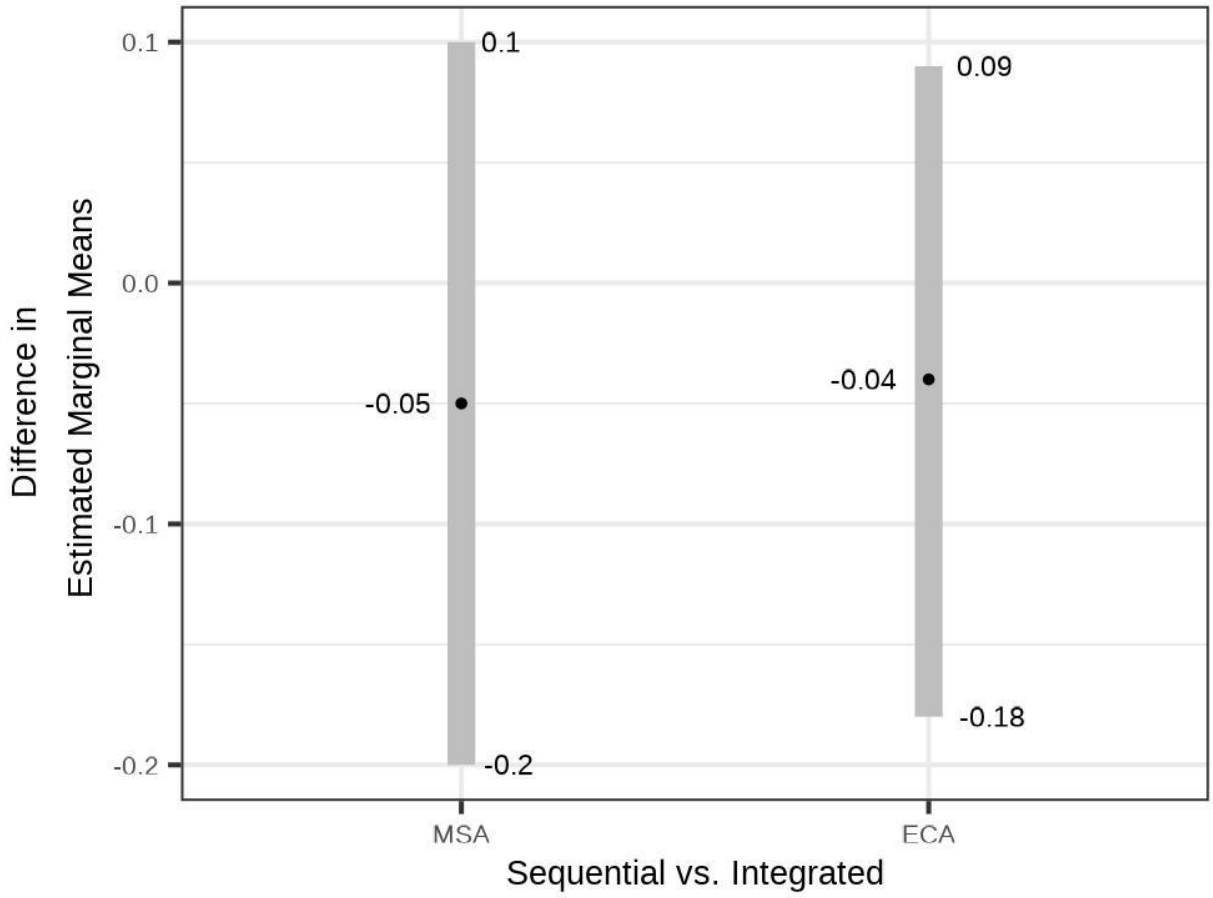
In order to test for equivalence, estimated marginal means were calculated for mean log reaction time of word form recall. Figure 29 illustrates the estimated range of difference between

[13] The beta estimate is exponentiated to back-transform the difference between the reference level (-SLV) and the level of interest (+SLV) into meaningful units.

each Approach x Register sub-group's marginal means with 90% confidence intervals. They largely confirm what was concluded from the descriptive statistics: the difference between the register outcomes for Integrated and Sequential is practically zero.

Figure 29

Difference in Estimated Marginal Means with 90% Confidence Intervals for Vocabulary Form Knowledge: Mean Log RT (equivalence bounds not depicted for visual clarity)



Finally, the estimated marginal means for Approach x Register were tested for functional equivalence. Delta was set at 100 ms (4.61 on the log-transformed scale). Results for comparing each register across approaches are displayed in Table 25.

Table 25

Tests of Difference and Equivalence using Estimated Marginal Means for Vocabulary Form Knowledge: Mean Log RT

Contrast	EMM Diff	*SE*	*df*	Test of Difference			Test of Equivalence	
				t-ratio	*p-value*	*Cohen's d*	*t-ratio*	*p-value*
MSA_{seq} — MSA_{int}	-0.05	0.09	82.13	-0.54	.588	-0.32	-49.77	< **.001**
ECA_{seq} — ECA_{int}	-0.04	0.08	66.15	-0.55	.585	-0.28	-57.24	< **.001**

The presence of effects greater than the equivalence range was rejected for both MSA (t = -49.77 , p <.001) and ECA (t = -57.24, p <.001). It can be concluded that participants at the population level would likely be able to recall the correct target word form with functionally equivalent speed regardless of whether they were learned in a Sequential Approach or an Integrated one.

Levenshtein Distance

Descriptive Statistics.

The second metric of word form knowledge was an approximation of accuracy: the Levenshtein Distance. The Levenshtein Distance counts the number of changes needed to arrive at the correct form. Unlike the descriptive statistics for log RT, the aggregate mean Levenshtein

Distance scores for vocabulary word knowledge differ slightly between the overall Integrated and sequential approaches. This difference is most apparent for ECA knowledge, where ECA_{seq} has slightly lower scores than ECA_{int} (see Table 26). However, the Integrated and Sequential outcomes for the MSA register are fairly similar, with the 95% confidence intervals of MSA_{int} and MSA_{seq} largely overlapping.

Table 26

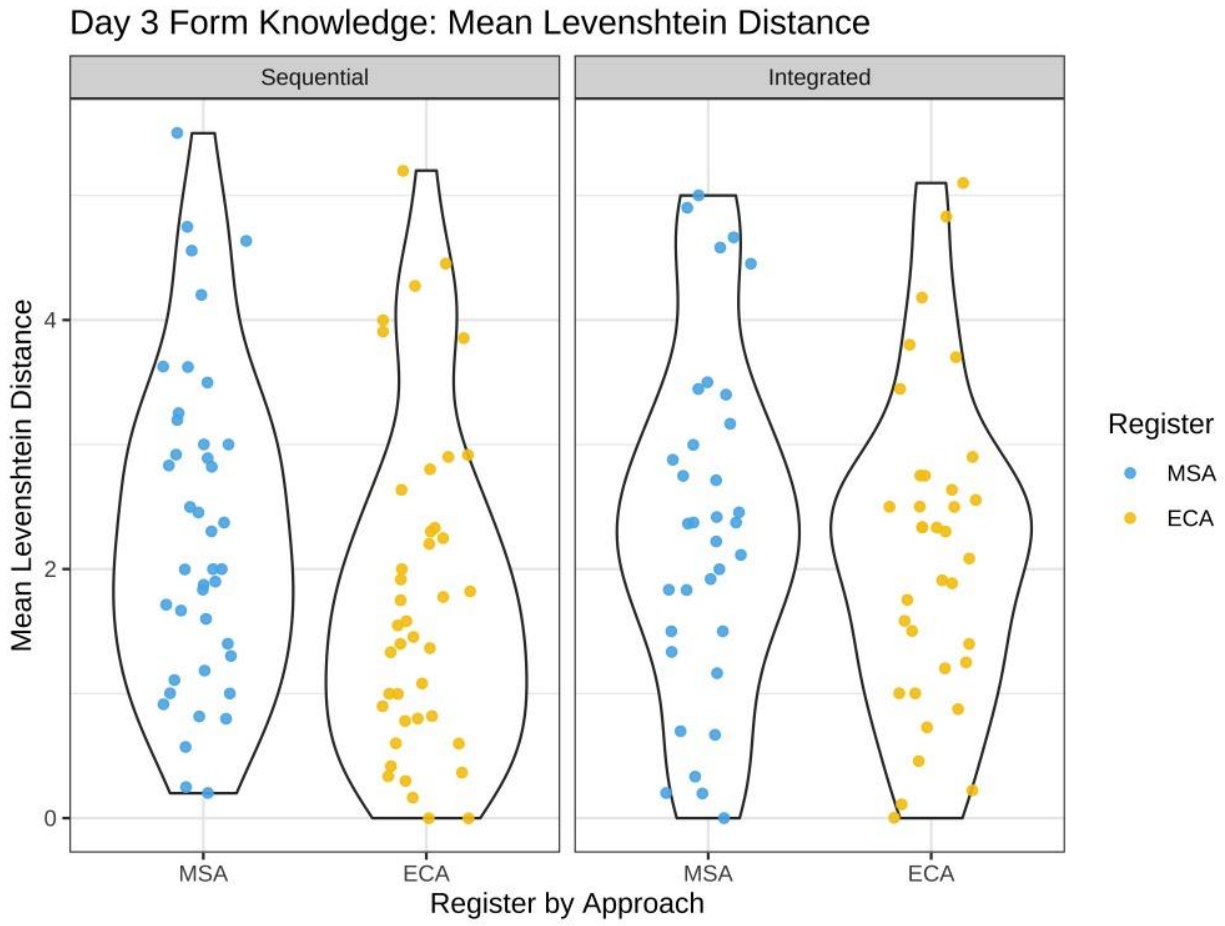
Descriptive Statistics for Vocabulary Form Knowledge: Mean Levenshtein Distance

Condition	Combined Registers	MSA only	ECA only
Integrated			
N	69	34	35
Missing	5	3	2
Mean (CI)	2.24 (1.92, 2.55)	2.35 (1.87, 2.83)	2.13 (1.69, 2.56)
SD	1.32	1.38	1.26
Skewness	0.33	0.22	0.43
Kurtosis	-0.34	-0.49	0.02
Sequential			
N	82	41	41
Missing	0	0	0
Mean (CI)	2.05 (1.76, 2.34)	2.32 (1.91, 2.73)	1.78 (1.37, 2.20)
SD	1.33	1.3	1.32
Skewness	0.6	0.5	0.83
Kurtosis	-0.31	-0.28	0.03

On the whole, the spread of scores is comparable between conditions as can be seen by the distribution of datapoints in Figure 30 as well as their associated standard deviations in Table 26. Finally, the skewness and kurtosis values in Table 26 are all less than |2| (Lomax & Hahs-Vaughn, 2012), indicating normality of distribution.

Figure 30

Violin Plot Illustrating the Distribution of Mean Levenshtein Distance for Vocabulary Form Knowledge



Analyses.

The results of the linear mixed effects model for vocabulary form knowledge can be seen in Table 27. As with log rt, there was a significant effect of Register on Levenshtein Distance, where ECA lexical items required, on average, 0.53 fewer changes needed to arrive at the correct word form answer (β_{ECA} = -0.53, t = -3.95, $p < 0.001$) compared to MSA lexical items. However, the main effect of Register was not of interest to the current research question. Furthermore, the

inferential effect of Approach was not statistically significant (β_{int} = 0.02, t = 0.05, p = 0.957), suggesting that there may not be a practical difference between Integrated and Sequential training at the population level.

Table 27

Results of Linear Mixed Effects Model for Vocabulary Form Knowledge: Mean Levenshtein Distance

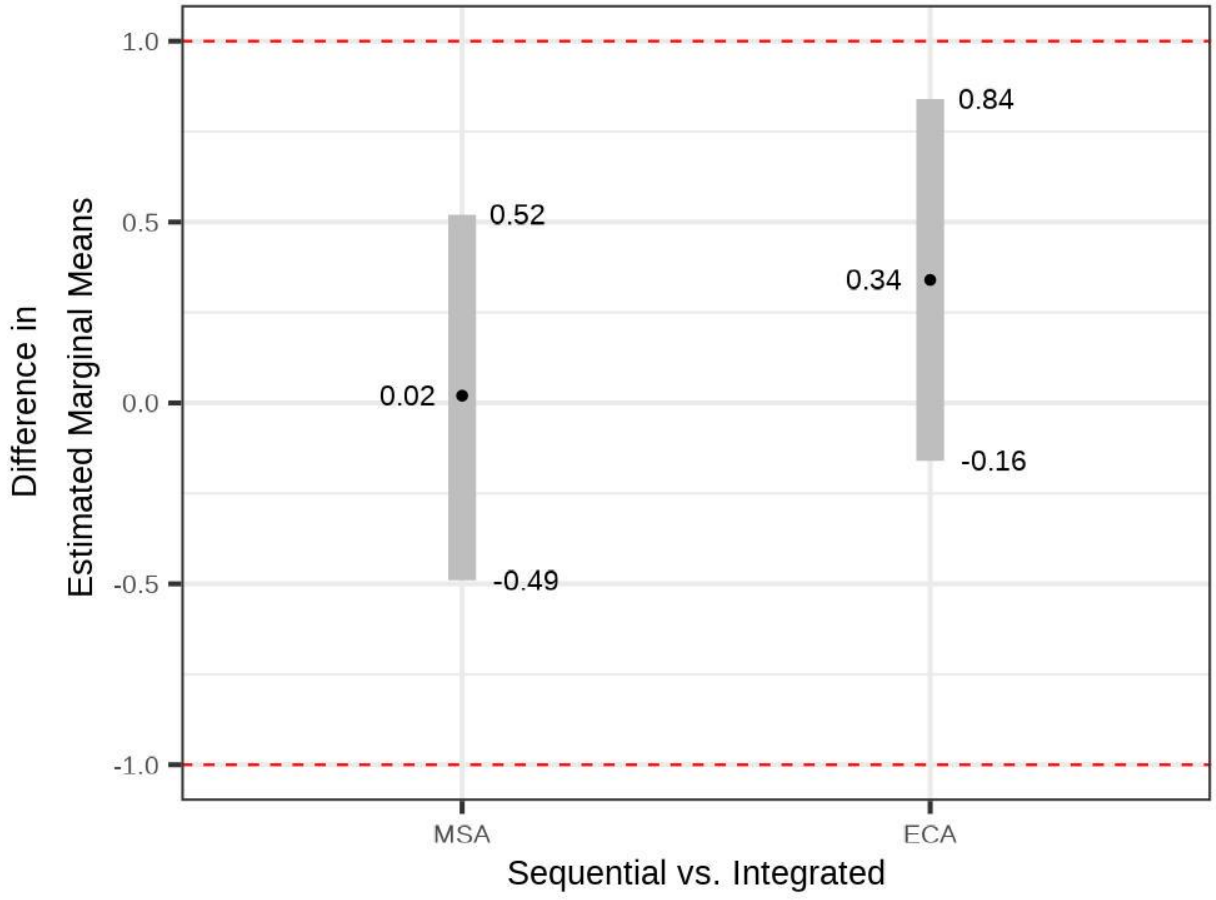
	Levenshtein Distance				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	2.32	0.21	1.91 - 2.72	11.31	<0.001
Approach [Int]	0.02	0.30	-0.58 - 0.62	0.05	0.957
Register [ECA]	-0.53	0.14	-0.80 - -0.27	-3.95	<0.001
Approach x Register	0.33	0.20	-0.07 - 0.72	1.62	0.107
Random Effects					
σ^2	0.38				
$\tau_{00\ participant}$	1.35				
ICC	0.78				
$N_{\ participant}$	76				
Observations	151				
Marginal R^2 / Conditional R^2	0.029 / 0.788				

Following the analytic plan, Register was tested for equivalence across approaches using the estimated marginal means. Figure 31 illustrates the estimated range of difference between each Approach x Register sub-group's marginal means with 90% confidence intervals. The estimated difference is slightly higher for ECA than for MSA, indicating that the differerence

between Sequential and Integrated was even larger in ECA than in MSA. By and large, however, the figure confirms what was seen in the descriptive statistics: that the difference between the two Approaches for each Register is close to zero.

Figure 31

Difference in Estimated Marginal Means with 90% Confidence Intervals for Vocabulary Form Knowledge: Mean Levenshtein Distance



Finally, the estimated marginal means for Approach x Register were tested for functional equivalence. Delta was set at 1 (within a score of 1 change needed). The results are displayed in Table 28.

Table 28

Tests of Difference and Equivalence using Estimated Marginal Means for Vocabulary Form Knowledge: Mean Levenshtein Distance

Contrast	EMM Diff	*SE*	*df*	Test of Difference			Test of Equivalence	
				t-ratio	*p-value*	*Cohen's d*	*t-ratio*	*p-value*
MSA_{seq} — MSA_{int}	0.02	0.30	92.51	0.05	.957	0.03	-3.25	**.001**
ECA_{seq} — ECA_{int}	0.34	0.30	91.65	1.13	.26	0.56	-2.18	**.016**

Since there were no effects beyond the equivalence range, the null hypothesis of difference was rejected for both MSA (t = -3.25 , p <.001) and ECA (t = -2.18 , p = .016). This means that studying in an integrated approach (where multiple registers are taught side-by-side) compared to a sequential approach (where a standard register is taught first, followed by a non-standard one) leads to functionally equivalent in terms of approximating accuracy of form knowledge.

Accuracy

Descriptive Statistics.

The third metric of word form knowledge was accuracy as a binary variable (0/1). The descriptive statistics for mean accuracy of vocabulary form knowledge is shown in Table 29.

Similar to the Levenshtein Distance scores, the mean MSA_{int} and MSA_{seq} scores are nearly identical (0.18 and 0.17, respectively) with almost perfectly overlapping 95% confidence intervals. The ECA scores, on the other hand, have slightly diverging outcomes. Specifically, the mean accuracy for ECA_{int} is lower than ECA_{seq} by 0.08, a difference which is reflected in the spread of their respective 95% confidence intervals as well.

Table 29

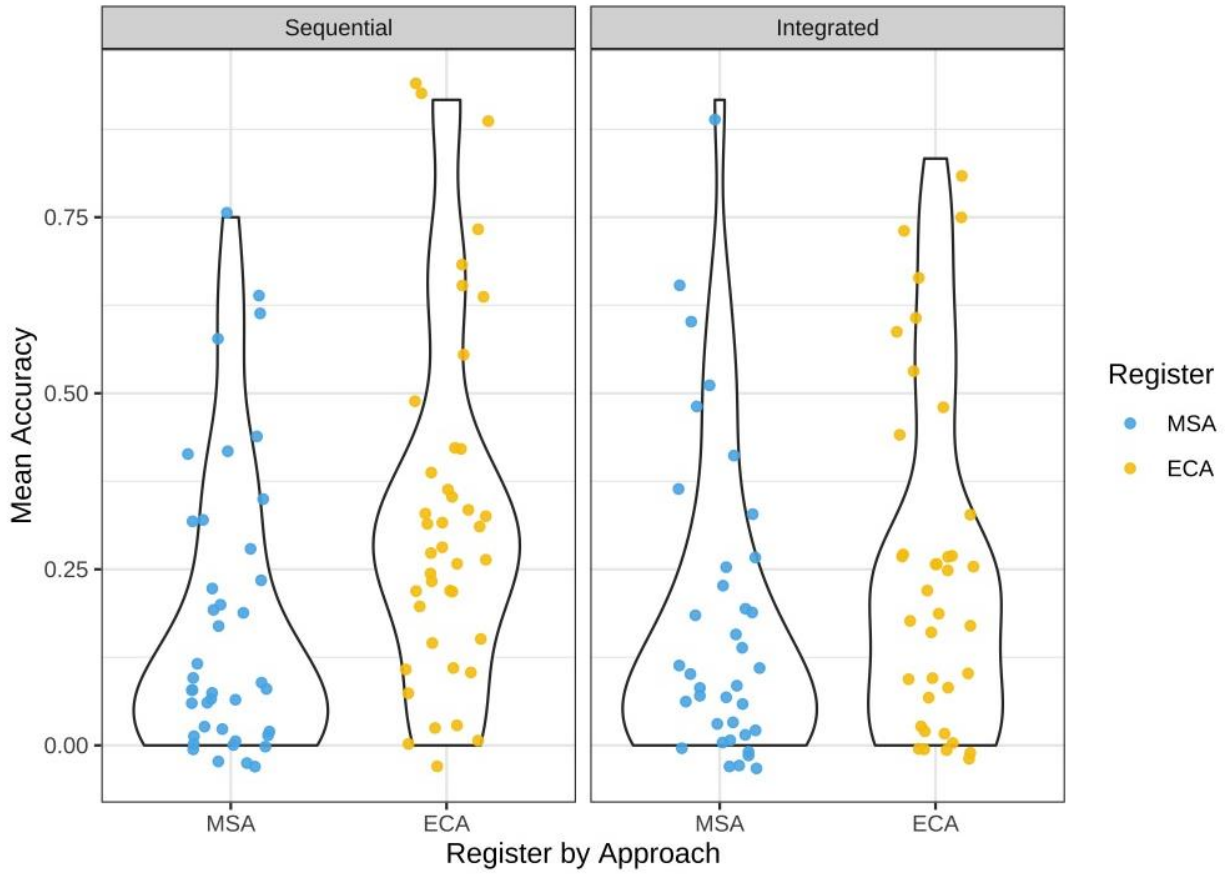
Descriptive Statistics for Vocabulary Form Knowledge: Mean Accuracy

Condition	Combined Registers	MSA only	ECA only
Integrated			
N	74	37	37
Missing	0	0	0
Mean (CI)	0.21 (0.16, 0.27)	0.18 (0.10, 0.25)	0.25 (0.17, 0.33)
SD	0.23	0.22	0.24
Skewness	1.23	1.66	0.95
Kurtosis	0.79	2.65	-0.03
Sequential			
N	82	41	41
Missing	0	0	0
Mean (CI)	0.25 (0.20, 0.31)	0.17 (0.11, 0.24)	0.33 (0.25, 0.41)
SD	0.24	0.2	0.25
Skewness	1.06	1.31	0.88
Kurtosis	0.56	0.94	0.25

Although the skewness and kurtosis are within the range of |2| (Lomax & Hahs-Vaughn, 2012), the violin plots displayed in Figure 32 indicate a large cluster of lower mean accuracy scores. This is particularly true for in the MSA registers. This may indicate the test of word form knowledge was somewhat difficult for participants. The clustering of low scores is confirmed in the density plot in Figure 64 in appendix C, which sharply peaks around 0.

Figure 32

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Accuracy by Participant for Vocabulary Form Knowledge



Analyses.

The results of the linear mixed effects model for accuracy of vocabulary form knowledge can be seen in Table 30. The interaction of Approach x Register was significant ($\beta_{ECA\text{-}int}$ = -0.08, t = -2.1, p = 0.038). Within the context of the current research question, this means that learning ECA in the integrated approach was associated with an 8% decrease in mean accuracy compared to in the sequential approach.

Table 30

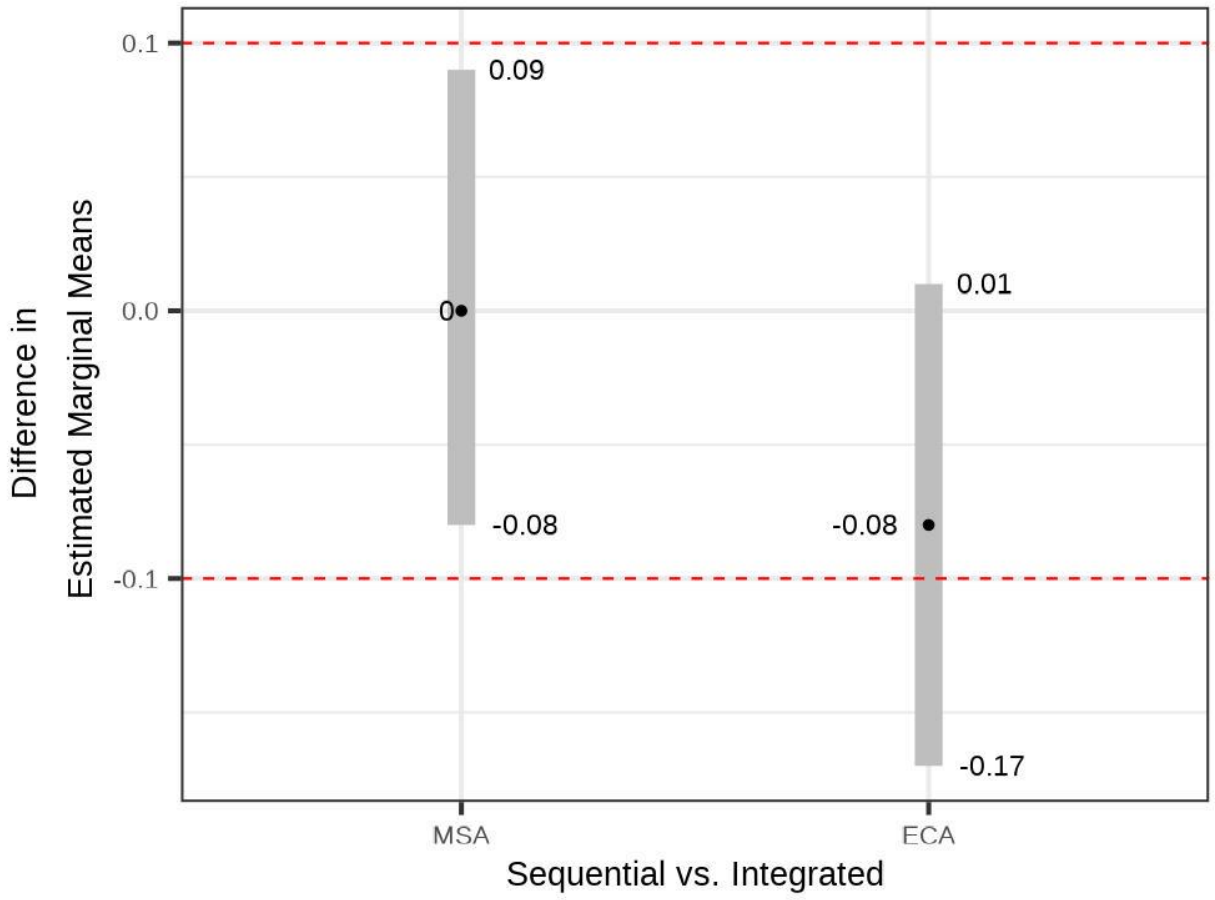
Results of Linear Mixed Effects Model for Vocabulary Form Knowledge: Mean Accuracy

	Accuracy				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	0.17	0.04	0.10 - 0.25	4.84	<0.001
Approach [Int]	0.003	0.05	-0.10 - 0.11	0.06	0.952
Register [ECA]	0.16	0.03	0.10 - 0.21	5.65	<0.001
Approach x Register	-0.08	0.04	-0.16 - -0.00	-2.1	0.038
Random Effects					
σ^2	0.02				
$\tau_{00\ participant}$	0.04				
ICC	0.71				
$N_{\ participant}$	78				
Observations	156				
Marginal R^2 / Conditional R^2	0.074 / 0.727				

In order to test for equivalence, estimated marginal means were calculated for mean accuracy of word form recall. Figure 33 illustrates the estimated range of difference between each Approach x Register sub-group's marginal means with 90% confidence intervals. The figure confirms the findings from the descriptive statistics that ECA_{seq} appears to have a higher accuracy than ECA_{int} (the difference between ECA_{seq} and ECA_{int} is likely negative), and this advantage does not hold for MSA (the difference is squarely estimated around zero).

Figure 33

Difference in Estimated Marginal Means with 90% Confidence Intervals for Vocabulary Form Knowledge: Mean Accuracy



Finally, the estimated marginal means for Approach x Register were tested for functional equivalence. Delta was set at 10% (0.1). Results for comparing each register across approaches are displayed in Table 31.

Table 31

Tests of Difference and Equivalence using Estimated Marginal Means for Vocabulary Form Knowledge: Mean Accuracy

Contrast	EMM Diff	*SE*	*df*	Test of Difference			Test of Equivalence	
				t-ratio	*p-value*	*Cohen's d*	*t-ratio*	*p-value*
MSA_{seq} — MSA_{int}	0.003	0.05	101.53	0.06	.952	0.02	-1.85	**.034**
ECA_{seq} — ECA_{int}	-0.08	0.05	101.53	-1.55	.124	-0.65	-0.36	.361

The test of equivalence between approaches was significant for MSA (t = -1.85 , p = .034). This finding is to be expected, given that the range of estimated MSA_{int} values does not cross the equivalence boundaries in Figure 33. On the other hand, the bottom portion of the estimated difference between ECA_{seq} and ECA_{int} extends below the lower equivalence boundary. Accordingly, the test of equivalence between approaches was rejected for ECA (t = -0.36 , p = .361). Thus, it seems that accuracy of word form knowledge may be functionally equivalent between an Integrated and a sequential approach for MSA (recall that, in a sequential approach, MSA is learned first, followed by ECA). The same cannot necessarily be concluded for ECA knowledge.

Word Meaning Knowledge

Word meaning knowledge refers to participants' ability to produce the meaning of the target item. In the context of the study, participants were prompted with the Mini-Arabii noun, and provided the English translation.

Reaction Time

Descriptive Statistics.

The first metric of word meaning knowledge was processing speed: how quickly participants were able to translate the word into English. As shown in Table 32, the descriptive statistics for log-transformed mean reaction time are fairly similar between the Integrated and sequential approaches. Mean log RT of Integrated is lower than Sequential by 0.04, indicating that reaction time for correctly producing word meaning knowledge was, on average, marginally faster in the Integrated condition. However, the 95% confidence intervals are largely overlapping, which suggests parity between these two approaches. This small advantage in integrated approach processing speed is also reflected in the MSA and ECA register descriptive statistics.

Table 32

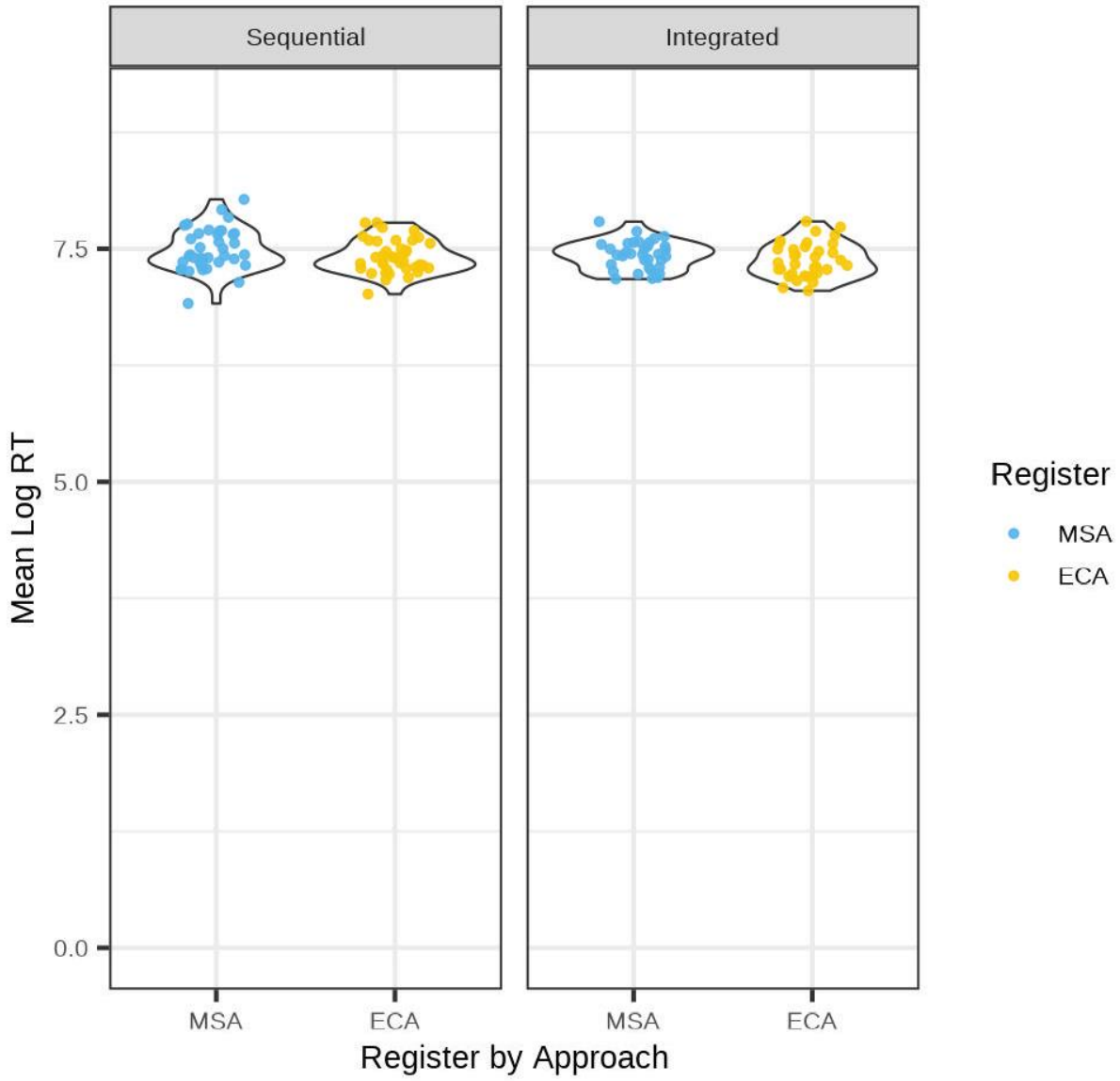
Descriptive Statistics for Vocabulary Meaning Knowledge: Mean Log Reaction Time

Condition	Combined Registers	MSA only	ECA only
Integrated			
N	70	35	35
Missing	0	0	0
Mean (CI)	7.41 (7.37, 7.45)	7.43 (7.38, 7.48)	7.38 (7.32, 7.45)
SD	0.17	0.15	0.19
Skewness	0.17	-0.27	-0.49
Kurtosis	-0.56	-0.79	-0.49
Sequential			
N	78	39	39
Missing	0	0	0
Mean (CI)	7.45 (7.41, 7.50)	7.49 (7.42, 7.56)	7.42 (7.36, 7.48)
SD	0.2	0.22	0.18
Skewness	0.32	0.04	-0.67
Kurtosis	0.3	-0.99	-0.59

All Approach x Register subgroups have a fairly equivalent spread of data around the mean. The only noticeable difference is that MSA_{seq} has both a larger standard deviation as seen in Table 32 and a longer spread of datapoints in Figure 34. Finally, the data appear to be relatively normally distributed according to both a visual check of the datapoints in Figure 34 and the skewness and kurtosis values in Table 32, which are all less than |2| (Lomax & Hahs-Vaughn, 2012).

Figure 34

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Log Reaction Time for Vocabulary Meaning Knowledge



Analyses.

The results of the linear mixed effects model for vocabulary meaning knowledge can be seen in Table 33. While there was significant effect of Register (β_{ECA} = -0.08, t = -2.95, p =

0.004), Register as a main effect was not of interest for the current research question. According to the model, participants were about 4% faster[14] in recalling items learned in the integrated approach compared to the sequential approach. The inferential effect of Approach, however, was not statistically significant (β_{int} = -0.06, t = -1.38, p = 0.169), suggesting that there may not be a practical difference between the two +SLV approaches at the population level.

Table 33

Results of Linear Mixed Effects Model for Vocabulary Meaning Knowledge: Mean Log RT

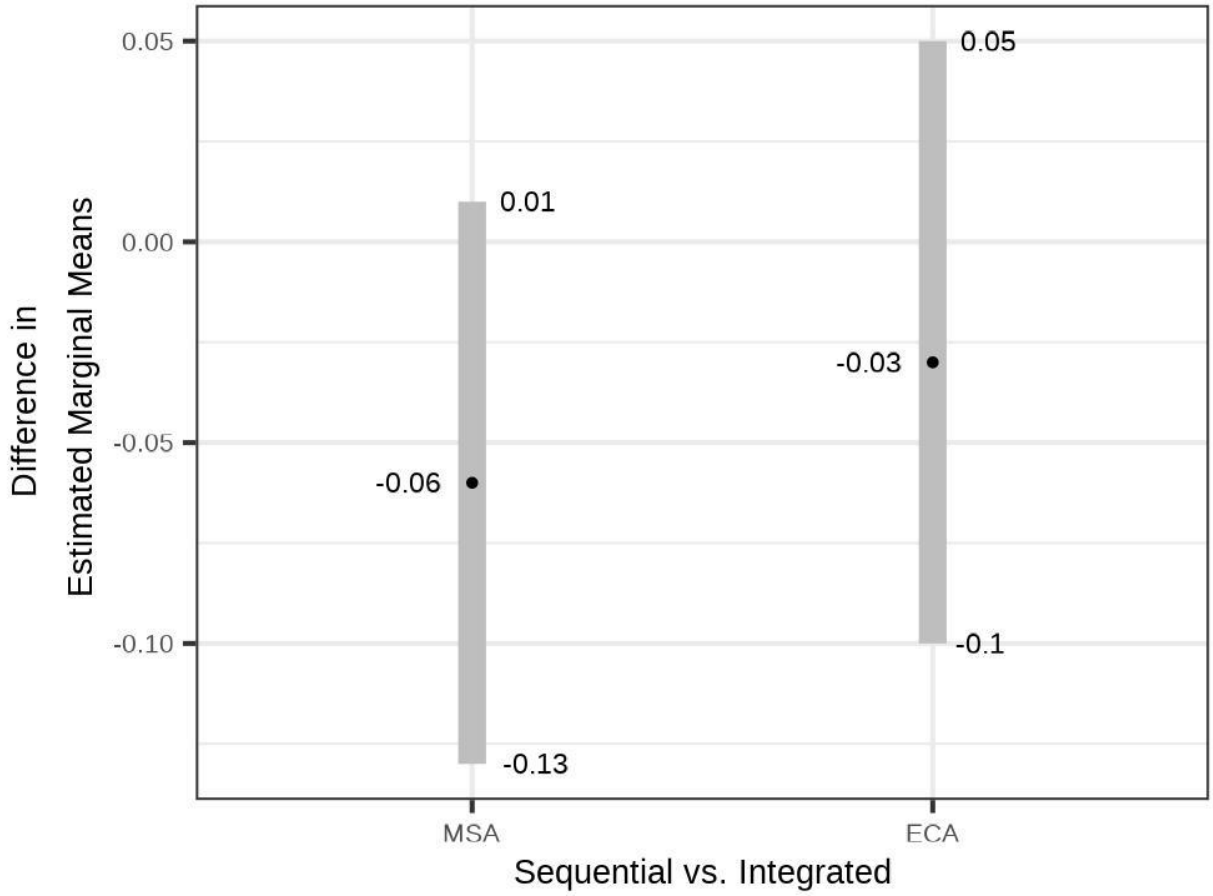
	Log RT				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	7.49	0.03	7.43 - 7.55	248.27	<0.001
Approach [Int]	-0.06	0.04	-0.15 - 0.03	-1.38	0.169
Register [ECA]	-0.08	0.03	-0.13 - -0.03	-2.95	0.004
Approach x Register	0.03	0.04	-0.04 - 0.11	0.87	0.387
Random Effects					
σ^2	0.01				
$\tau_{00\ participant}$	0.02				
ICC	0.60				
$N_{\ participant}$	76				
Observations	148				
Marginal R^2 / Conditional R^2	0.042 / 0.621				

[14] The beta estimate is exponentiated to back-transform the difference between the reference level (-SLV) and the level of interest (+SLV) into meaningful units.

In order to test for equivalence, estimated marginal means were calculated for mean log RT of word meaning recall. Figure 35 illustrates the estimated range of difference between each Approach x Register sub-group's marginal means with 90% confidence intervals. The figure confirms the findings from the descriptive statistics, namely that processing speed for the integrated approach is slightly faster than Sequential, but nonetheless the difference between them is largely zero.

Figure 35

Difference in Estimated Marginal Means with 90% Confidence Intervals for Vocabulary Meaning Knowledge: Mean Log RT (equivalence bounds not depicted for visual clarity)



Finally, the estimated marginal means for Approach x Register were tested for functional equivalence. Delta was set at 100 ms (4.61 on the log-transformed scale). Results for comparing each register across approaches are displayed in Table 34.

Table 34

Tests of Difference and Equivalence using Estimated Marginal Means for Vocabulary Meaning Knowledge: Mean Log RT

Contrast	EMM Diff	*SE*	*df*	Test of Difference			Test of Equivalence	
				t-ratio	*p-value*	*Cohen's d*	*t-ratio*	*p-value*
MSA_{seq} — MSA_{int}	-0.06	0.04	108.53	-1.38	.17	-0.51	-103.61	< **.001**
ECA_{seq} — ECA_{int}	-0.03	0.04	108.53	-0.6	.551	-0.22	-104.39	< **.001**

The presence of effects beyond the equivalence range was rejected for both MSA (t = -103.61, p <.001) and ECA (t = -104.39 , p <.001). It can be concluded that participants at the population level likely be able to recall the correct target word form with functionally equivalent speed regardless of whether they were learned in a Sequential Approach or an Integrated Approach.

Accuracy

Descriptive Statistics.

The second metric of word meaning knowledge was accuracy as a binary variable (0/1). The descriptive statistics for mean accuracy of vocabulary meaning knowledge can be seen in Table 35. In line with findings for form meaning accuracy, the MSA means and 95% confidence intervals are strikingly similar across Approach, while ECA_{int} again trails behind ECA_{seq} by 8%.

Table 35

Descriptive Statistics for Vocabulary Meaning Knowledge: Mean Accuracy

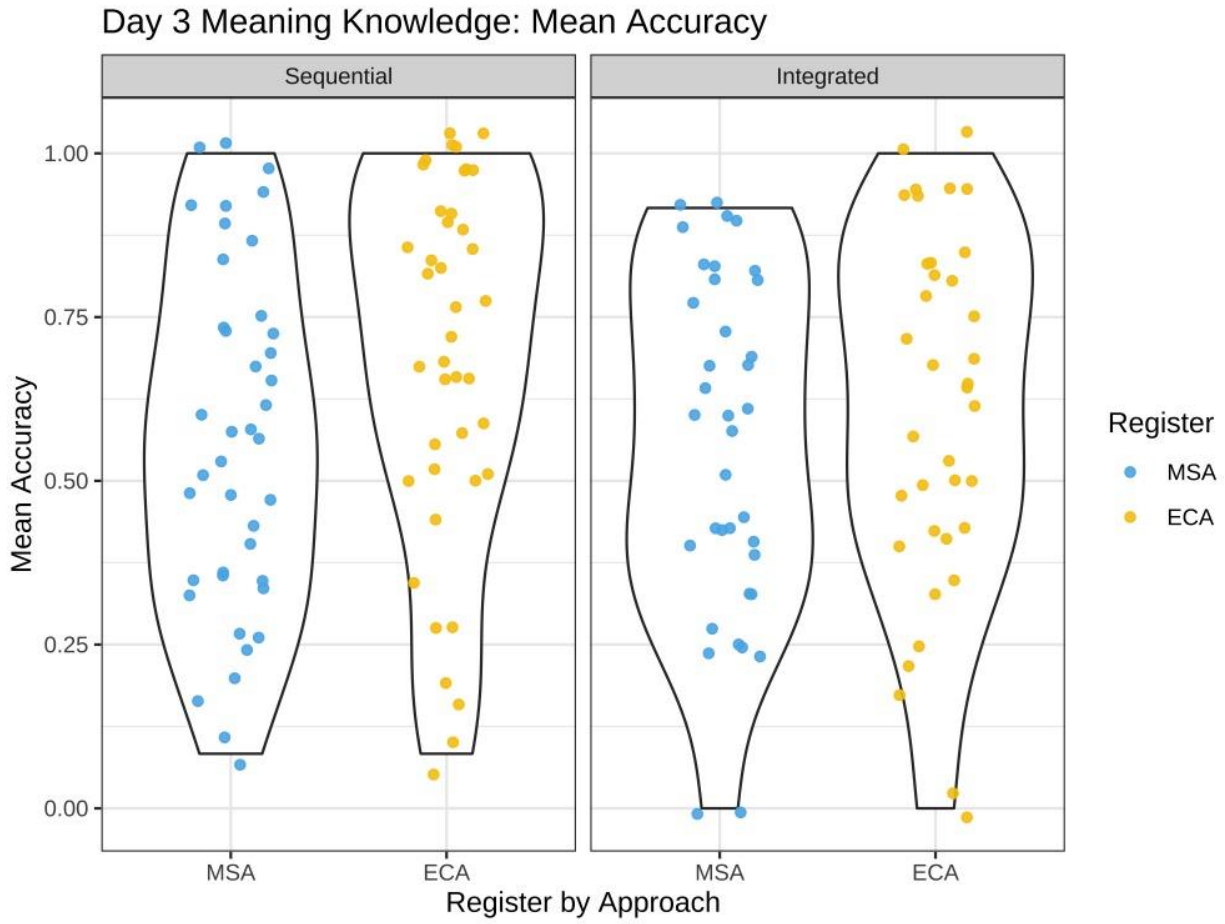
Condition	Combined Registers	MSA only	ECA only
Integrated			
N	74	37	37
Missing	0	0	0
Mean (CI)	0.58 (0.52, 0.64)	0.56 (0.47, 0.64)	0.60 (0.51, 0.69)
SD	0.27	0.26	0.27
Skewness	-0.37	-0.27	-0.49
Kurtosis	-0.7	-0.79	-0.49
Sequential			
N	82	41	41
Missing	0	0	0
Mean (CI)	0.62 (0.56, 0.68)	0.56 (0.47, 0.64)	0.68 (0.59, 0.77)
SD	0.28	0.27	0.28
Skewness	-0.27	0.04	-0.67
Kurtosis	-1.03	-0.99	-0.59

The standard deviation is also fairly similar between approaches and the Approach x Register sub-groups. On the whole, the spread of scores is comparable between conditions as can be seen by the distribution of datapoints in Figure 36. A visual check of the datapoints in Figure 36 reveals a clustering of data at the upper end (bulk of scores closer to 1). However, the data

appear to be relatively normally distributed according to the skewness and kurtosis values in Table 35, which are all less than |2| (Lomax & Hahs-Vaughn, 2012).

Figure 36

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Accuracy by Participant for Vocabulary Meaning Knowledge



Analyses.

The results of the linear mixed effects model for vocabulary meaning knowledge can be seen in Table 36. The interaction of Approach x Register was significant ($\beta_{ECA\text{-}int}$ = -0.08, t = -

1.82, $p = 0.071$). Within the context of the current research question, this means that learning ECA in the integrated approach was associated with an 8% decrease in mean accuracy compared to in the sequential approach.

Table 36

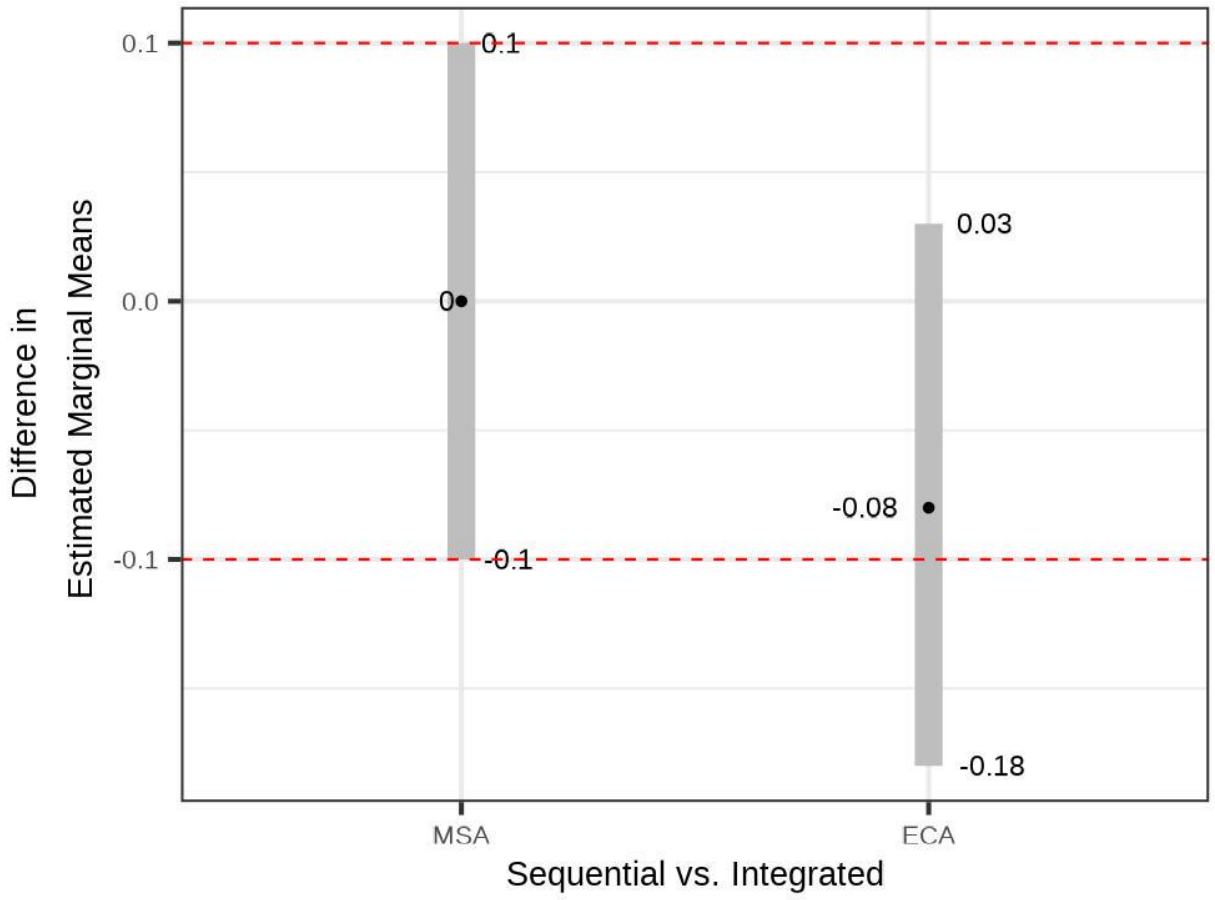
Results of Linear Mixed Effects Model for Vocabulary Meaning Knowledge: Mean Accuracy

	Accuracy				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	0.56	0.04	0.47 - 0.64	13.13	<0.001
Approach [Int]	-0.0006	0.06	-0.12 - 0.12	-0.01	0.992
Register [ECA]	0.12	0.03	0.07 - 0.18	4.28	<0.001
Approach x Register	-0.08	0.04	-0.16 - 0.01	-1.82	**0.071**
Random Effects					
σ^2	0.02				
$\tau_{00\ participant}$	0.06				
ICC	0.77				
$N_{\ participant}$	78				
Observations	156				
Marginal R^2 / Conditional R^2	0.035 / 0.775				

In order to test for equivalence, estimated marginal means were calculated for mean accuracy of word meaning recall. Figure 37 illustrates the estimated range of difference between each Approach x Register sub-group's marginal means with 90% confidence intervals. The estimated difference between MSA scores just barely falls within the equivalence bounds, while difference between ECA scores extends beyond the lower equivalence boundary.

Figure 37

Difference in Estimated Marginal Means with 90% Confidence Intervals for Vocabulary Meaning Knowledge: Mean Accuracy



Finally, the estimated marginal means for Approach x Register were tested for functional equivalence. Delta was set at 10% (0.1). Results for comparing each register across approaches are displayed in Table 37.

Table 37

Tests of Difference and Equivalence using Estimated Marginal Means for Vocabulary Meaning Knowledge: Mean Accuracy

Contrast	EMM Diff	*SE*	*df*	Test of Difference			Test of Equivalence	
				t-ratio	*p-value*	*Cohen's d*	*t-ratio*	*p-value*
MSA_{seq} — MSA_{int}	<-0.01	0.06	95.75	-0.01	.992	<0.00	-1.61	.055
ECA_{seq} — ECA_{int}	-0.08	0.06	95.75	-1.25	.213	-0.59	-0.37	.357

Since the estimated range of difference between groups crossed beyond the equivalence range, the alternative hypothesis of equivalence (neither one inferior nor superior to the other) was rejected for both MSA (t = -1.61 , p = 0.55) and ECA (t = -0.37 , p = .357) across Approach types. Thus, the tests of equivalence produce results that appear, on the surface level, contrary to the tests of difference: vocabulary meaning accuracy is neither inferentially different *nor* equivalent. While this is indeed statistically possible (e.g., Godfroid & Spino, 2015), it effectively means that results cannot be reliably generalized beyond the sample population. What can be concluded is that, in the sample population, the ability to correctly translate Arabic words studied in the integrated approach was, on average less accurate than in the sequential approach, and this effect was greater for ECA than for MSA.

Grammar Recall Knowledge (Negated Sentences)

Grammar recall knowledge refers to participants' ability to produce the correctly conjugated and negated target verb form. In the context of the study, participants were given a full subject-verb-object sentence in English, as well as the Mini-Arabii subject and object

translations with a blank in between them. Participants provided the target verb form in the blank. This measure, from a theoretical standpoint, was likely the most difficult for participants, as it required both the verb form retrieval as well as the application of grammar rules.

Reaction Time

Descriptive Statistics.

The first metric of grammar recall knowledge was processing speed: how quickly participants were able to produce the correctly conjugated and negated target verb form. Recall that reaction time data is only based on correctly produced answers. In the case of grammar recall, this means that a very small subset of the original data was used (38 observations, out of an original 154).

The descriptives statistics for log-transformed mean reaction time of grammar recall knowledge are displayed in Table 38. The overall scores are quite similar between the approaches both globally as well as for each register. One noteable difference is that the 95% confidence interval for MSA_{seq} is much wider than that MSA_{int}, such that the estimated true mean for MSA_{int} is completely contained within the MSA_{seq} confidence intervals.

Table 38

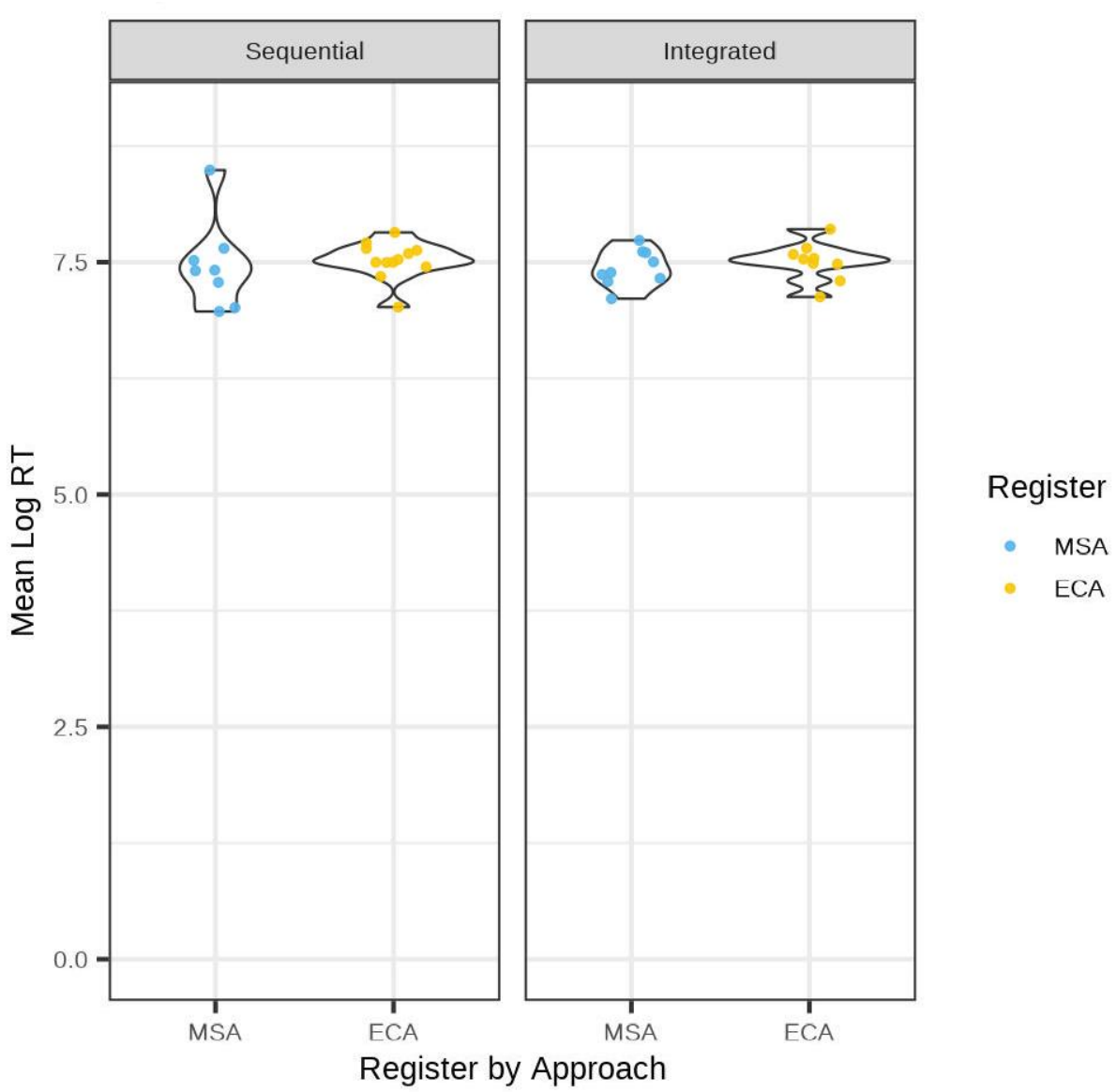
Descriptive Statistics for Grammar Recall Knowledge: Mean Log Reaction Time

Condition	Combined Registers	MSA only	ECA only
Integrated			
N	18	9	9
Missing	0	0	0
Mean (CI)	7.47 (7.37, 7.57)	7.44 (7.29, 7.59)	7.51 (7.35, 7.66)
SD	0.2	0.19	0.21
Skewness	-0.16	1.89	1.44
Kurtosis	-0.14	2.56	0.46
Sequential			
N	20	8	12
Missing	0	0	0
Mean (CI)	7.50 (7.35, 7.65)	7.47 (7.07, 7.87)	7.52 (7.39, 7.65)
SD	0.33	0.48	0.2
Skewness	1.06	1.8	1.43
Kurtosis	3.81	2.27	0.57

The standard deviations for the integrated approach as well as its associated register sub-groups are strikingly similar, ranging from 0.19 - 0.21. Conversely, there is a much wider range of scores for the sequential approach (SD MSA_{seq}: 0.48; SD MSA_{int}: 0.2) as can be seen in Figure 38. This likely reflects the small sample size of the trimmed reaction time data for grammar recall. Finally, the overall sequential approach data and the MSA_{seq} data are somewhat leptokurtic, with kurtosis values greater than 2 (Lomax & Hahs-Vaughn, 2012).

Figure 38

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Log Reaction Time for Grammar Recall Knowledge



Analyses.

The results of the linear mixed effects model for grammar recall knowledge can be seen in Table 39. According to the model, participants were about 2% faster[15] in recalling items learned in the integrated approach compared to the sequential approach. The inferential effect of Approach, however, was not statistically significant (β_{int} = -0.02, t = -0.13, p = 0.897), suggesting that there may not be a practical difference between the +SLV approaches at the population level.

[15] The beta estimate is exponentiated to back-transform the difference between the reference level (-SLV) and the level of interest (+SLV) into meaningful units.

Table 39

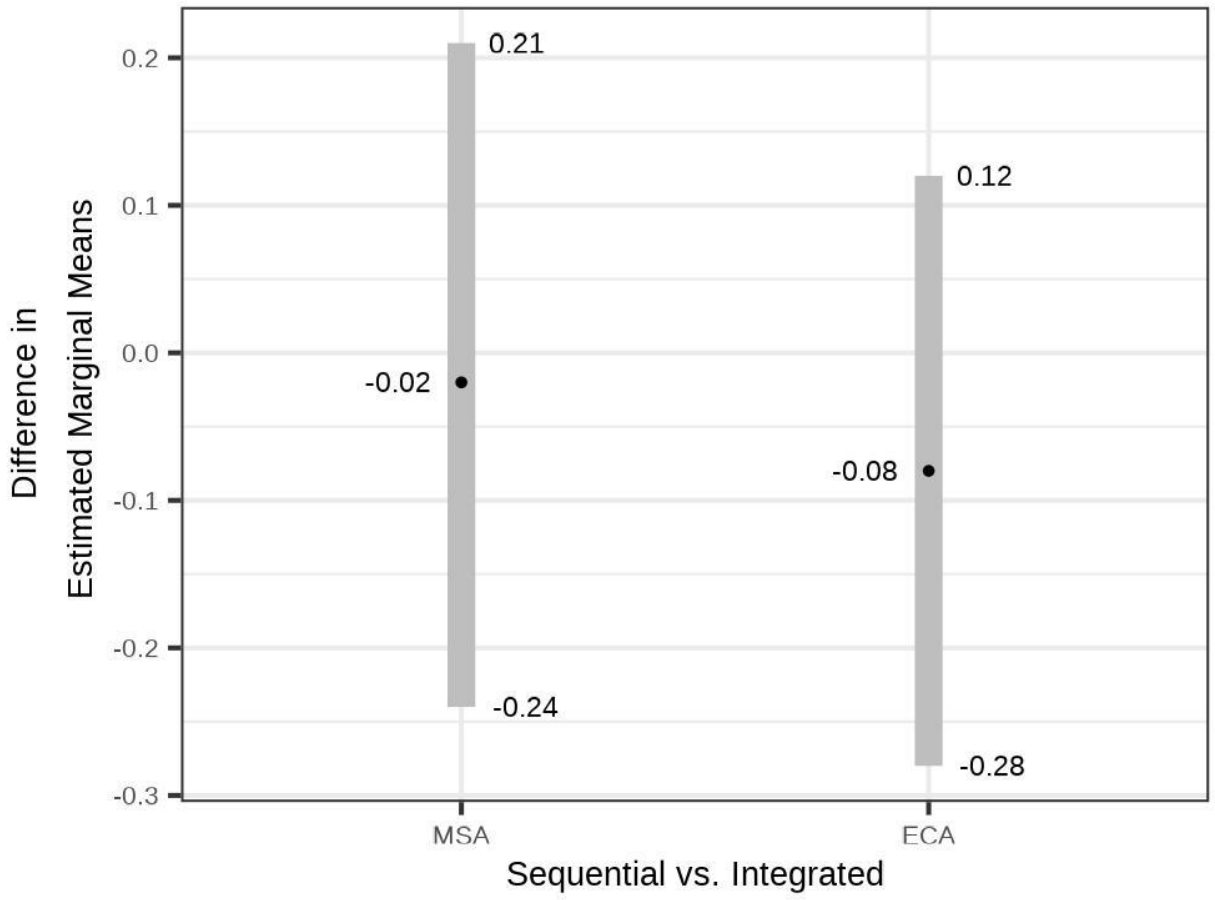
Results of Linear Mixed Effects Model for Grammar Recall Knowledge: Mean Log Reaction Time

	Log RT				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	7.48	0.09	7.29 - 7.67	81.19	<0.001
Approach [Int]	-0.02	0.13	-0.28 - 0.24	-0.13	0.897
Register [ECA]	0.1	0.1	-0.10 - 0.29	0.99	0.332
Approach x Register	-0.06	0.14	-0.35 - 0.22	-0.45	0.657
Random Effects					
σ^2	0.03				
$\tau_{00\ participant}$	0.05				
ICC	0.6				
$N_{\ participant}$	29				
Observations	38				
Marginal R^2 / Conditional R^2	0.026 / 0.611				

In order to test for equivalence, estimated marginal means were calculated for mean log RT of grammar recall. Figure 39 illustrates the estimated range of difference between each Approach x Register sub-group's marginal means with 90% confidence intervals. The figure confirms the findings from the descriptive statistics; specifically, that processing speed for each Approach and their associated register sub-groups are, on the whole, fairly similar (that is, th estimated difference between them is around zero).

Figure 39

Difference in Estimated Marginal Means with 90% Confidence Intervals for Grammar Recall Knowledge: Mean Log Reaction Time (equivalence bounds not depicted for visual clarity)



Finally, the estimated marginal means for Approach x Register were tested for functional equivalence. Delta was set at 100 ms (4.61 on the log-transformed scale). Results for comparing each register across approaches are displayed in Table 40.

Table 40

Tests of Difference and Equivalence using Estimated Marginal Means for Grammar Recall Knowledge: Mean Log Reaction Time

Contrast	EMM Diff	*SE*	*df*	Test of Difference			Test of Equivalence	
				t-ratio	*p-value*	*Cohen's d*	*t-ratio*	*p-value*
MSA_{seq} — MSA_{int}	-0.02	0.13	33.70	-0.13	.9	-0.09	-34.85	< **.001**
ECA_{seq} — ECA_{int}	-0.08	0.12	33.87	-0.66	.514	-0.45	-37.37	< **.001**

The presence of effects greater than the equivalence range was rejected for both MSA (t = -34.85 , p <.001) and ECA (t = -37.37 , p <.001). It can be concluded that participants at the population level would likely be able to produce the correct translation of the verb form with functionally equivalent speed regardless of whether they were learned in a Sequential Approach or an Integrated Approach.

Levenshtein Distance

Descriptive Statistics.

The second metric of grammar recall knowledge was an approximation of accuracy: the Levenshtein Distance. The Levenshtein Distance counts the number of changes needed to arrive at the correct form. The descriptive statistics of Levenshtein Distance scores for grammar recall knowledge can been seen in Table 41. The sequential approach has an overall lower average score than the integrated approach (2.76 and 3.08 respectively), indicating that fewer changes were needed, on average, to arrive at the correct verb form in the sequential approach. This difference was mirrored in the MSA and ECA registers within each approach as well.

Nonetheless, the 95% confidence intervals do somewhat overlap across each level of comparison.

Table 41

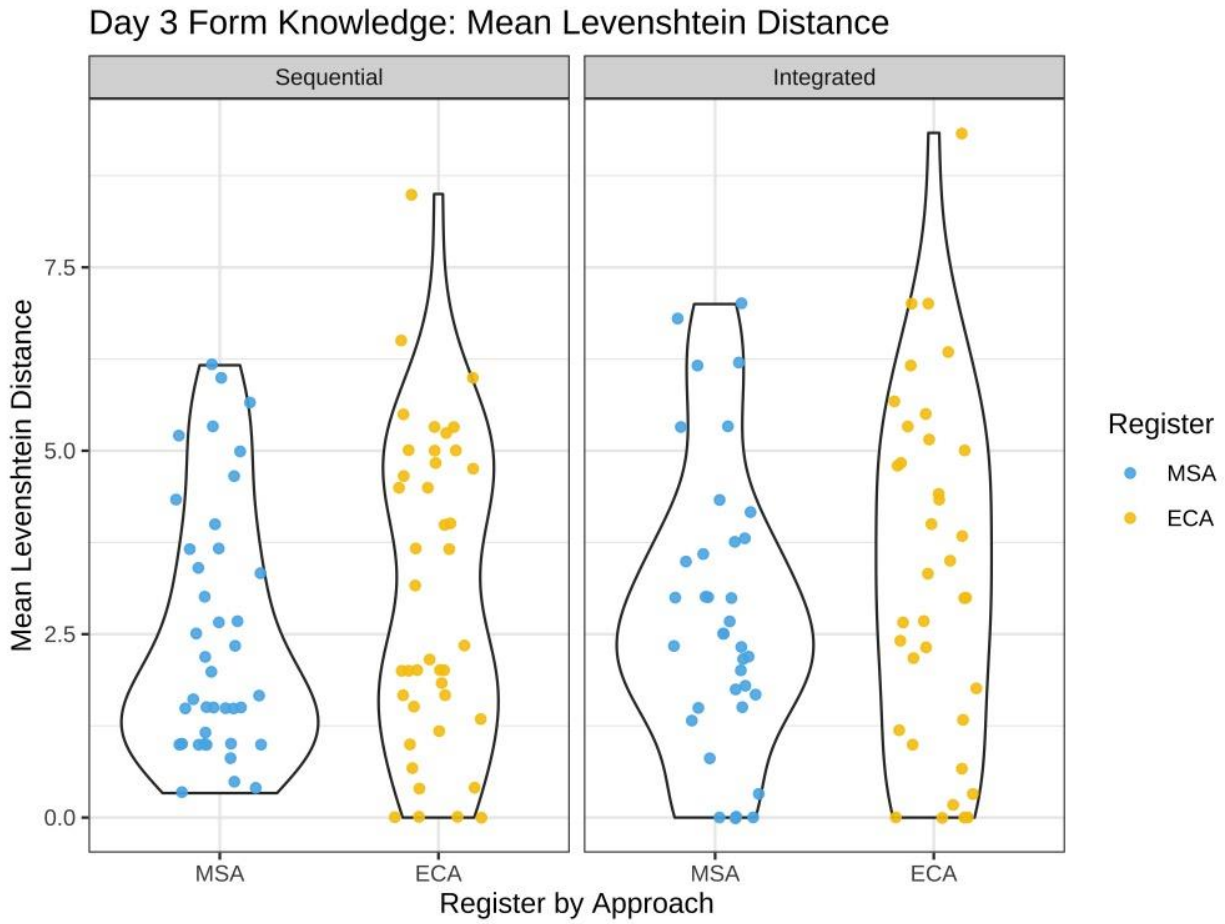
Descriptive Statistics for Grammar Recall Knowledge: Mean Levenshtein Distance

Condition	Combined Registers	MSA only	ECA only
Integrated			
N	72	36	36
Missing	2	1	1
Mean (CI)	3.08 (2.57, 3.59)	2.82 (2.17, 3.47)	3.34 (2.53, 4.15)
SD	2.17	1.92	2.38
Skewness	0.47	0.55	0.33
Kurtosis	-0.28	-0.22	-0.44
Sequential			
N	82	41	41
Missing	0	0	0
Mean (CI)	2.76 (2.34, 3.19)	2.47 (1.93, 3.01)	3.06 (2.39, 3.73)
SD	1.94	1.7	2.12
Skewness	0.56	0.79	0.32
Kurtosis	-0.55	-0.58	-0.63

As can be seen from the standard deviation scores in Table 41, the integrated approach and its associated register sub-groups have a greater degree of variation of mean scores than their Sequential counterparts. This difference is reflected in the longer spread of Integrated Approach datapoints in Figure 40. The skewness and kurtosis values in Table 41, which are all less than |2| (Lomax & Hahs-Vaughn, 2012), and the overall distribution of datapoints in Figure 40 indicate that the data are relatively normally distributed.

Figure 40

Violin Plot Illustrating the Distribution of Mean Levenshtein Distance for Grammar Recall Knowledge



Analyses.

The results of the linear mixed effects model for grammar recall knowledge can be seen in Table 42. There was a significant effect of Register on Levenshtein Distance, where negated ECA verb production required, on average, 0.59 fewer changes to arrive at the correct answer (β_{ECA} = 0.59, t = 2.07, p = 0.04) compared to MSA verbs. However, the main effect of Register was not of interest to the current research question. Within the model, training in the integrated

approach was associated with an additional 0.35 changes needed to arrive at the correct verb form. As was the case with grammar recall processing speed, however, the inferential effect of Approach was not statistically significant (β_{int} = 0.35, t = 0.74, p = 0.459).

Table 42

Results of Linear Mixed Effects Model for Grammar Recall Knowledge: Mean Levenshtein Distance

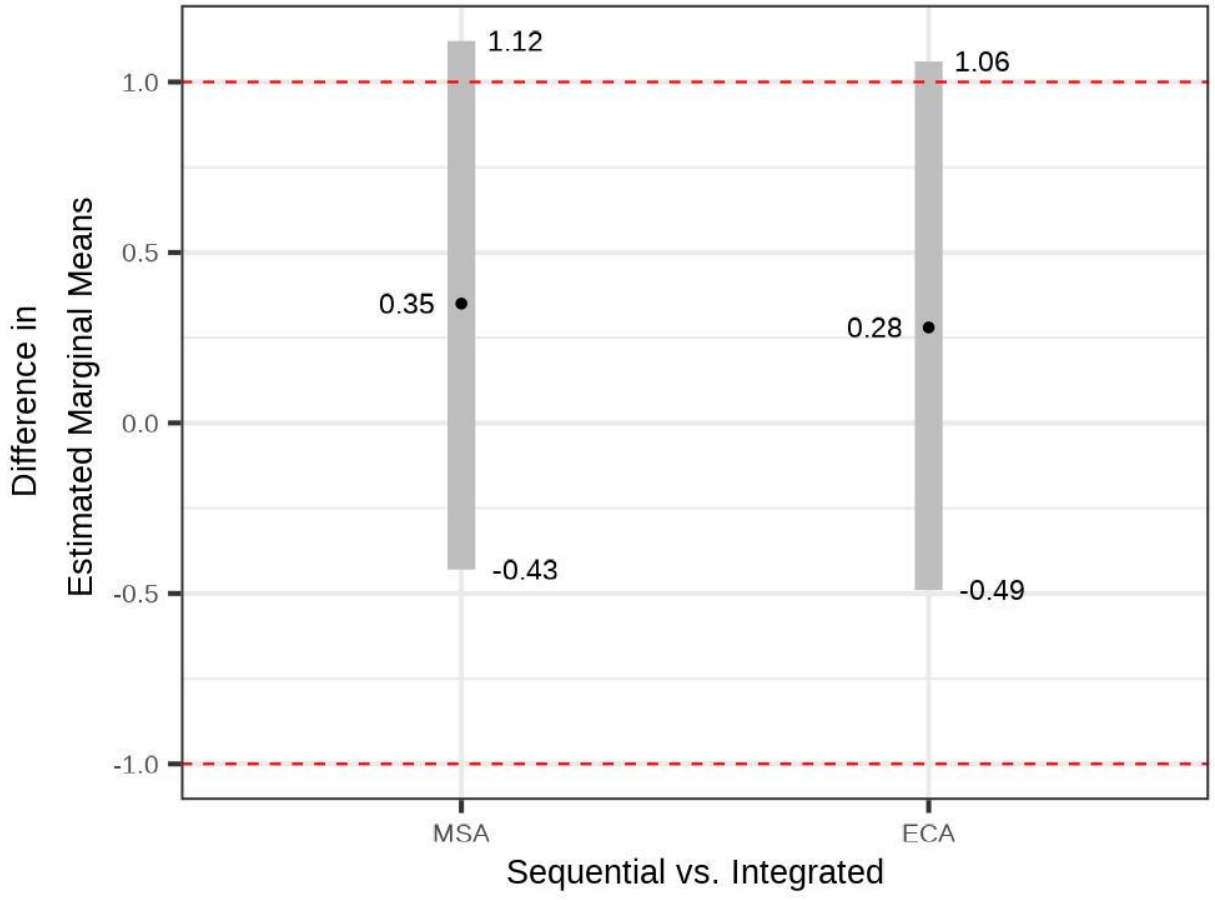
	Levenshtein Distance				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	2.47	0.32	1.84 - 3.1	7.75	<0.001
Approach [Int]	0.35	0.47	-0.58 - 1.27	0.74	0.459
Register [ECA]	0.59	0.28	0.03 - 1.14	2.07	0.04
Approach x Register	-0.06	0.41	-0.88 - 0.75	-0.15	0.88
Random Effects					
σ^2	1.64				
$\tau_{00\ participant}$	2.53				
ICC	0.61				
$N_{\ participant}$	77				
Observations	154				
Marginal R^2 / Conditional R^2	0.024 / 0.616				

Following the analytic plan, Register was tested for equivalence across Approach using the estimated marginal means. Figure 41 illustrates the estimated range of difference between each Approach x Register sub-group's marginal means with 90% confidence intervals. The figure confirms what was seen in the descriptive statistics: although the estimated difference

crosses zero, this difference is skewed slightly towards positive numbers. This indicates a potential advantage for the reference level scores (Sequential) over the comparison level scores (Integrated). Furthermore, the estimated ranges differences for ECA and MSA extend beyond the upper equivalence boundary.

Figure 41

Difference in Estimated Marginal Means with 90% Confidence Intervals for Grammar Recall Knowledge: Mean Levenshtein Distance



Finally, the estimated marginal means for Approach x Register were tested for functional equivalence. Delta was set at 1 (within a score of 1). The results are displayed in Table 43.

Table 43

Tests of Difference and Equivalence using Estimated Marginal Means for Grammar Recall Knowledge: Mean Levenshtein Distance

Contrast	EMM Diff	*SE*	*df*	Test of Difference			Test of Equivalence	
				t-ratio	*p-value*	*Cohen's d*	*t-ratio*	*p-value*
MSA_{seq} — MSA_{int}	0.35	0.47	109.65	0.74	.46	0.27	0.12	.547
ECA_{seq} — ECA_{int}	0.28	0.47	109.65	0.61	.545	0.22	-0.02	.493

Since there were effects greater than the equivalence range, the alternative hypothesis of equivalence (neither one inferior nor superior to the other) was rejected for both MSA (t = 0.12 , p = 0.547) and ECA (t = -0.02 , p = .493) across Approach types. Thus, the tests of equivalence produce results that appear, on the surface level, contrary to the tests of difference: Levenshtein Distance score is neither inferentially different *nor* equivalent for grammar recall knowledge. While this is indeed statistically possible (e.g., Godfroid & Spino, 2015), it effectively means that results cannot be reliably generalized beyond the sample population. What can be concluded is that, in the sample population, verb forms studied in the integrated approach required, on average, an additional 0.23 steps to arrive at the correct answer compared to the sequential approach.

Accuracy

Descriptive Statistics.

The third metric of grammar recall knowledge was accuracy as a binary variable (0/1). The descriptive statistics for mean accuracy of grammar recall knowledge can be seen in Table 44. Similar to grammar recall reaction time and Levenshtein distance, the descriptive statistics indicate that participants achieved somewhat similar levels of mean accuracy regardless of Approach, both globally and at the register sub-level. The 95% confidence intervals of all comparisons across Approach overlap to a large degree, although the mean Integrated scores are slightly higher than their Sequential counterparts.

Table 44

Descriptive Statistics for Grammar Recall Knowledge: Mean Accuracy

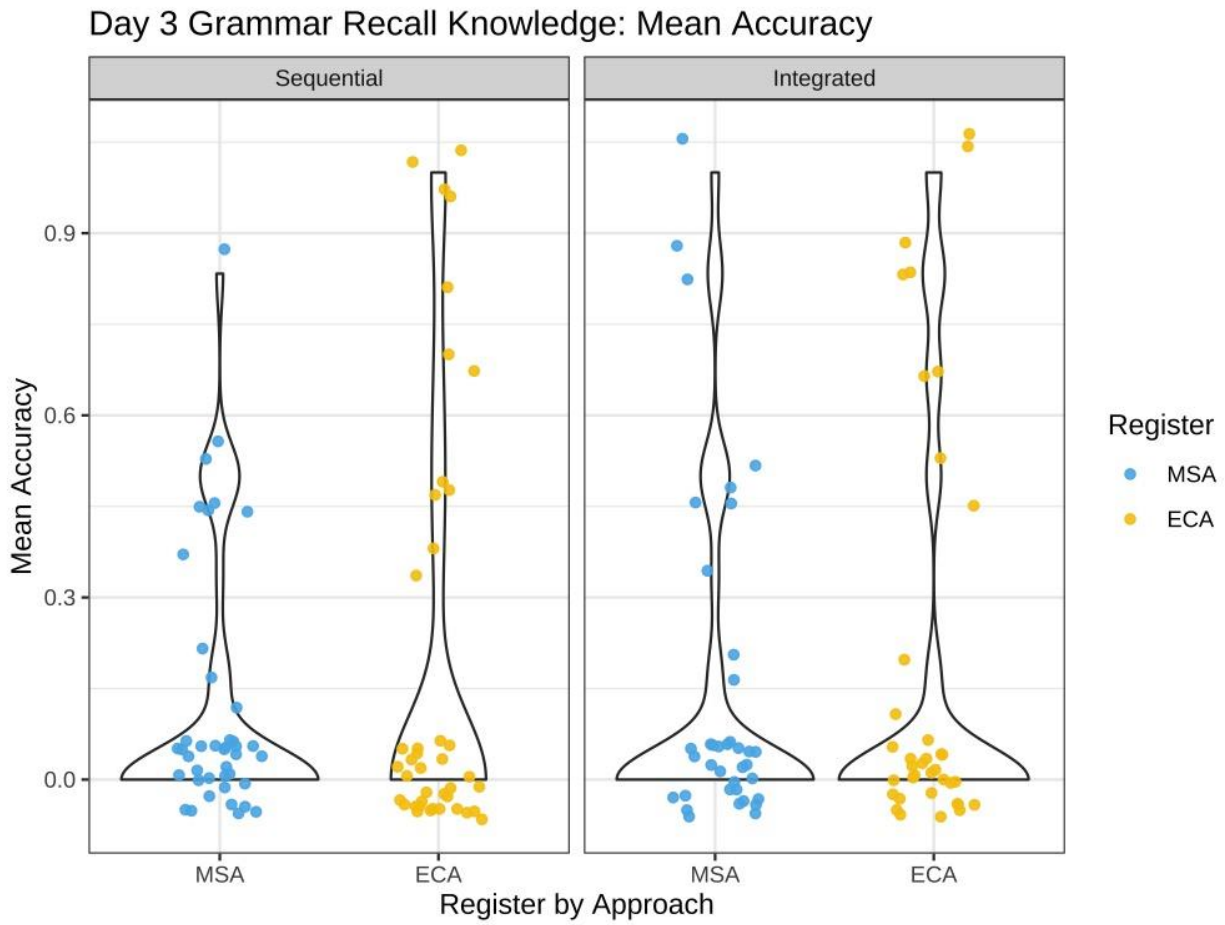
Condition	Combined Registers	MSA only	ECA only
Integrated			
N	74	37	37
Missing	0	0	0
Mean (CI)	0.17 (0.10, 0.24)	0.14 (0.05, 0.24)	0.19 (0.08, 0.31)
SD	0.31	0.28	0.34
Skewness	1.62	1.89	1.44
Kurtosis	1.18	2.56	0.46
Sequential			
N	82	41	41
Missing	0	0	0
Mean (CI)	0.16 (0.09, 0.22)	0.11 (0.05, 0.18)	0.20 (0.09, 0.31)
SD	0.29	0.22	0.35
Skewness	1.74	1.8	1.43
Kurtosis	1.89	2.27	0.57

The results of MSA_{seq} are slightly leptokurtic, with a kurtosis value greater than 2 (Lomax & Hahs-Vaughn, 2012). Furthermore, a visual inspection of the violin plots in Figure 42

indicates that the data are left skewed to toward 0 for all Approach-Register sub-groups. This cluster of scores around zero is likely due to the fact that many participants forgot to include the apostrophe when typing out the English negated forms “doesn’t” and “didn’t.”

Figure 42

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Accuracy by Participant for Grammar Recall Accuracy



The model diagnostics (Figure 69 in appendix C) confirm this evidence of skew, with deviation between the observed and model-predicted values around 0. Likewise, the QQ plot of

residuals points to evidence of heteroskedasticity. As such, caution should be used when inferentially interpreting the model results.

Analyses.

The results of the linear mixed effects model for mean accuracy of grammar recall knowledge can be seen in Table 45. According to the model, participants were about 3% more accurate in recalling items learned in the integrated approach compared to the sequential approach. The inferential effect of Approach, however, was not statistically significant (β_{int} = 0.03, t = 0.45, p = 0.656), suggesting that there may not be a practical difference between the two +SLV approaches at the population level.

Table 45

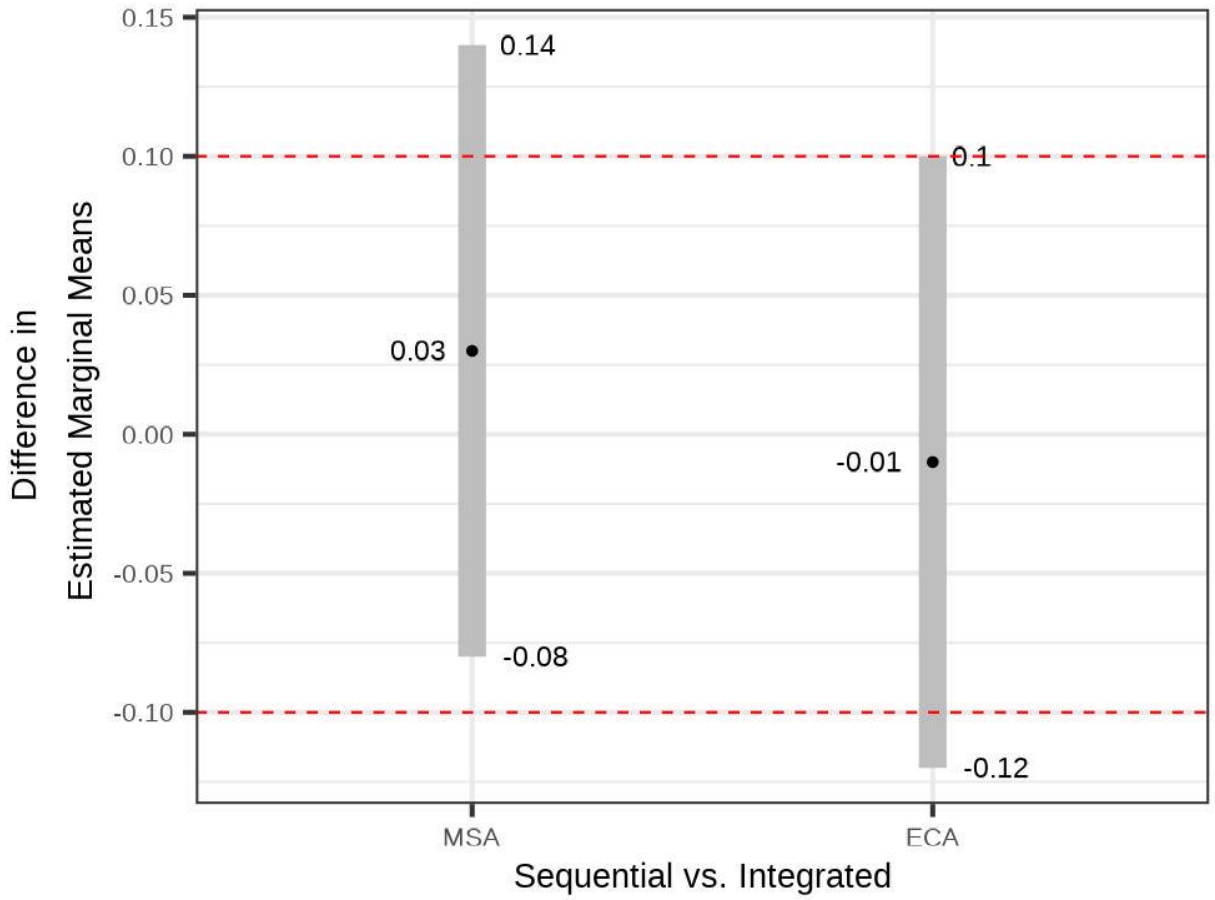
Results of Linear Mixed Effects Model for Grammar Recall Knowledge: Mean Accuracy

	Accuracy				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	0.11	0.05	0.02 - 0.21	2.43	0.016
Approach [Int]	0.03	0.07	-0.10 - 0.16	0.45	0.656
Register [ECA]	0.09	0.05	-0.02 - 0.20	1.63	0.104
Approach x Register	-0.04	0.08	-0.20 - 0.12	-0.50	0.616
Random Effects					
σ^2	0.06				
$\tau_{00\ participant}$	0.03				
ICC	0.32				
$N_{\ participant}$	78				
Observations	156				
Marginal R^2 / Conditional R^2	0.015 / 0.327				

In order to test for equivalence, estimated marginal means were calculated for mean accuracy of word form recall. Figure 43 illustrates the estimated range of difference between each Approach x Register sub-group's marginal means with 90% confidence intervals. The figure confirms the findings from the descriptive statistics; namely, that there is a high degree of overlap between the Integrated and sequential approaches (both difference estimates cross zero). However, the estimated range of difference between MSA_{seq} and MSA_{int} is higher than the upper equivalence boundary (indicating that integrated scores are higher). The opposite is true for ECA_{seq} and ECA_{int}: the estimated difference falls below the lower equivalence boundary (indicating that integrated scores are lower here).

Figure 43

Difference in Estimated Marginal Means with 90% Confidence Intervals for Grammar Recall Knowledge: Mean Accuracy



Finally, the estimated marginal means for Approach x Register were tested for functional equivalence. Delta was set at 10% (0.1). Results for comparing each register across approaches are displayed in Table 46.

Table 46

Tests of Difference and Equivalence using Estimated Marginal Means for Grammar Recall Knowledge: Mean Accuracy

Contrast	EMM Diff	*SE*	*df*	Test of Difference			Test of Equivalence	
				t-ratio	*p-value*	*Cohen's d*	*t-ratio*	*p-value*
MSA_{seq} — MSA_{int}	0.03	0.07	138.18	0.45	.656	0.12	-1.03	.153
ECA_{seq} — ECA_{int}	-0.01	0.07	138.18	-0.14	.888	-0.04	-1.33	.093

Since there were effects greater than the equivalence range, the alternative hypothesis of equivalence (neither one inferior nor superior to the other) was rejected for both MSA (t = -1.03 , p = 0.153) and ECA (t = -1.33 , p = .093) across Approach types. Thus, the tests of equivalence produce results that appear, on the surface level, contrary to the tests of difference: grammar recall accuracy is neither inferentially different *nor* equivalent. While this is indeed statistically possible (e.g., Godfroid & Spino, 2015), it effectively means that results cannot be reliably generalized beyond the sample population. What can be concluded is that, in the sample population, MSA verbs studied in the integrated approach (where MSA and ECA are learned side by side) were, on average, 3% more accurate than in the sequential approach (where MSA is learned first, followed by ECA). The effect, while minimal, is the opposite for ECA: ECA verbs were, on average, recalled 1% less accurately than in the Integrated condition.

Grammar Recognition Knowledge (Negated Sentences)

Grammar recognition knowledge refers to participants' ability to produce the correct translation of the conjugated and negated target verb form. In the context of the study, participants were given a full subject-verb-object sentence in Mini-Arabii, as well as the English subject and object translations with a blank in between them. Participants provided the English translation of the verb form in the blank.

Reaction Time

Descriptive Statistics.

The first metric of grammar recognition knowledge was processing speed: how quickly participants were able to translate the negated verb form into English. The descriptives statistics for log-transformed mean reaction time are displayed in Table 47. The Integrated processing speeds are slightly faster than those of the sequential approach and its associated registers. All in all, however, there is a large degree of similarity for grammar recall knowledge between the approach groups and registers within each group as indicated by the overlapping 95% confidence intervals around the mean.

Table 47

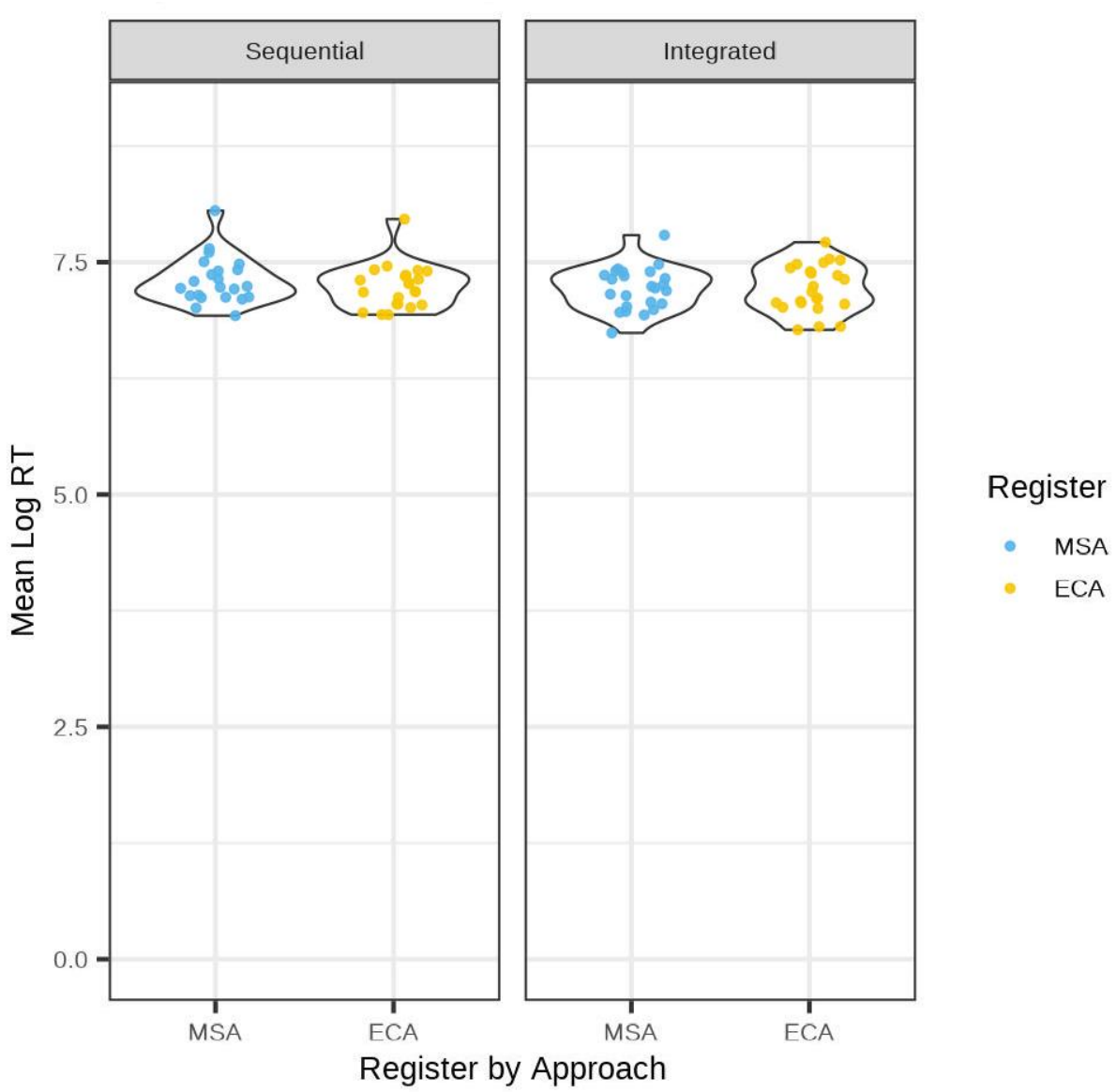
Descriptive Statistics for Grammar Recognition Knowledge: Mean Log Reaction Time

Condition	Combined Registers	MSA only	ECA only
Integrated			
N	47	24	23
Missing	0	0	0
Mean (CI)	7.22 (7.15, 7.29)	7.22 (7.12, 7.31)	7.21 (7.10, 7.33)
SD	0.24	0.23	0.26
Skewness	0.44	0.11	0.05
Kurtosis	0.3	-1.86	-1.78
Sequential			
N	45	22	23
Missing	0	0	0
Mean (CI)	7.27 (7.20, 7.34)	7.30 (7.19, 7.42)	7.24 (7.14, 7.34)
SD	0.24	0.25	0.23
Skewness	-1.59	0.34	0.58
Kurtosis	9.44	-0.1	1.07

The spread of scores is fairly similar between both Approaches and their register sub-groups (standard deviations ranging from 0.23 - 0.25). This parity in the spread of scores is verified with a visual check of the violin plots in Figure 44, where the data appear to extend across a similar range (bearing in mind that, as indicated above, that the Sequential register log RTs are slightly slower than their Integrated counterparts). The overall sequential approach data are quite leptokurtic, with a kurtosis value that extends 7.44 units beyond the threshold of |2| (Lomax & Hahs-Vaughn, 2012). Note that the QQ plot of residuals in Figure 70 in appendix C has also fairly heavy tails. This indicates that caution should be used when interpreting the strength the inferential model findings present below.

Figure 44

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Log Reaction Time for Grammar Recognition Knowledge



Analyses.

The results of the linear mixed effects model for grammar recognition knowledge can be seen in Table 48. According to the model, participants were about 8% faster[16] in recalling items learned in the integrated approach compared to the sequential approach. The inferential effect of Approach, however, was not statistically significant (β_{int} = -0.08, t = -1.19, p = 0.236), suggesting that there may not be a practical difference between the two +SLV approaches at the population level.

[16] The beta estimate is exponentiated to back-transform the difference between the reference level (-SLV) and the level of interest (+SLV) into meaningful units.

Table 48

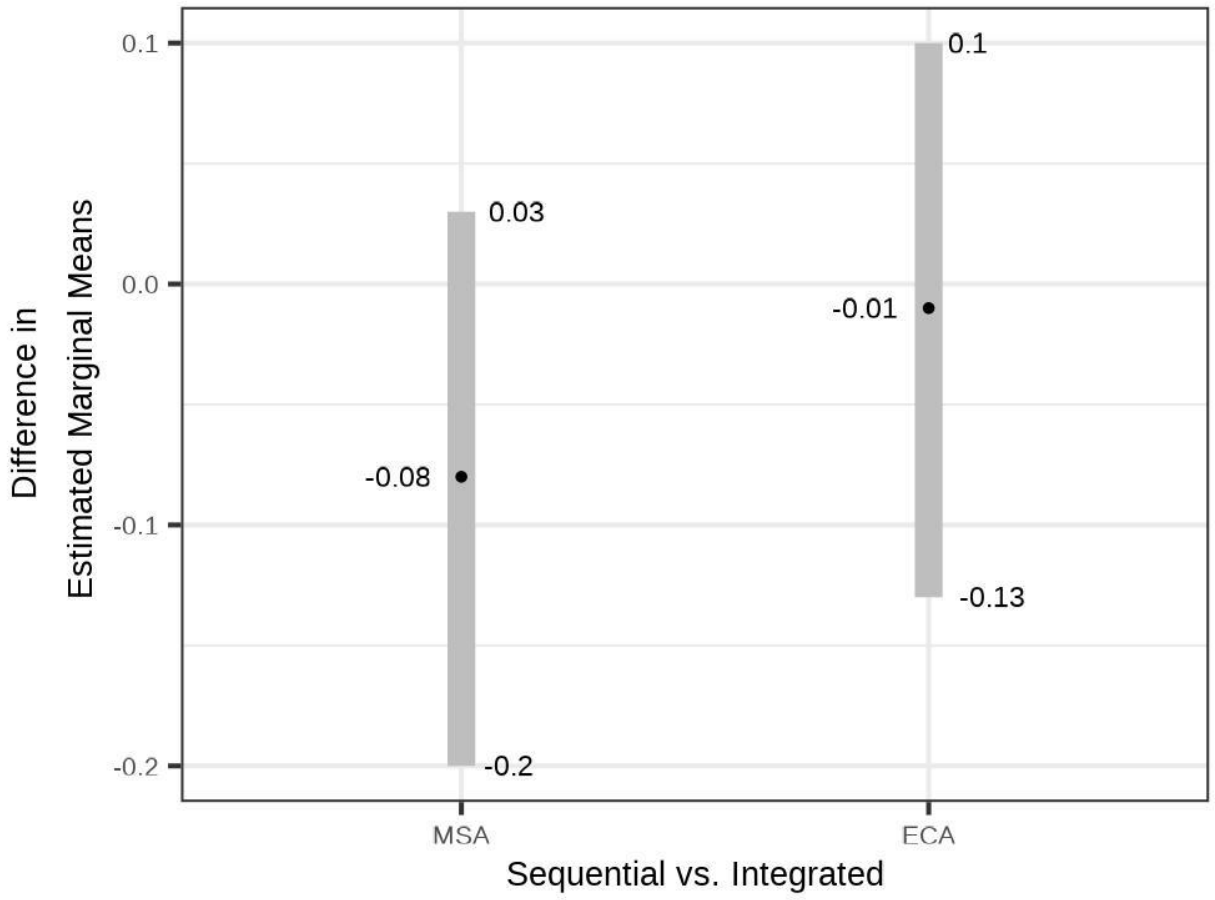
Results of Linear Mixed Effects Model for Grammar Recognition Knowledge: Mean Log Reaction Time

	Log RT				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	7.3	0.05	7.20 - 7.40	146.48	<0.001
Approach [Int]	-0.08	0.07	-0.22 - 0.06	-1.19	0.236
Register [ECA]	-0.07	0.03	-0.13 - 0.00	-1.95	0.055
Approach x Register	0.07	0.05	-0.02 - 0.16	1.47	0.144
Random Effects					
σ^2	0.01				
$\tau_{00\ participant}$	0.05				
ICC	0.79				
$N_{\ participant}$	48				
Observations	92				
Marginal R^2 / Conditional R^2	0.019 / 0.793				

In order to test for equivalence, estimated marginal means were calculated for mean log RT of grammar recognition. Figure 45 illustrates the estimated range of difference between each Approach x Register sub-group's marginal means with 90% confidence intervals. Although the estimated differene between Sequential and integrated approach outcomes is higher for ECA than MSA, both fall within the equivalence bounds of +/- 100 ms (4.61 on the log-transformed scale.

Figure 45

Difference in Estimated Marginal Means with 90% Confidence Intervals for Grammar Recognition Knowledge: Mean Log Reaction Time (equivalence bounds not depicted for visual clarity)



Finally, the estimated marginal means for Approach x Register were tested for functional equivalence. Delta was set at 100 ms (4.61 on the log-transformed scale). Results for comparing each register across approaches are displayed in Table 49.

Table 49

Tests of Difference and Equivalence using Estimated Marginal Means for Grammar Recognition Knowledge: Mean Log Reaction Time

Contrast	EMM Diff	*SE*	*df*	Test of Difference			Test of Equivalence	
				t-ratio	*p-value*	*Cohen's d*	*t-ratio*	*p-value*
MSA_{seq} — MSA_{int}	-0.08	0.07	57.22	-1.19	.237	-0.76	-64.69	< **.001**
ECA_{seq} — ECA_{int}	-0.01	0.07	57.16	-0.21	.837	-0.13	-65.70	< **.001**

The presence of effects beyond the equivalence range was rejected for both MSA (t = -64.69 , p <.001) and ECA (t = -65.70 , p <.001). This result suggests participants at the population level would likely be able to recall the correct translation of the grammar form with functionally equivalent speed regardless of whether they were learned in a Sequential Approach or an Integrated Approach.

Accuracy

Descriptive Statistics.

The final metric of grammar recognition knowledge was accuracy as a binary variable (0/1). The descriptive statistics for mean accuracy of grammar recall knowledge can be seen in Table 50. Unlike to the log reaction time measures for grammar recall knowledge, the Sequential accuracy scores are slightly superior to the Integrated ones. However, on the whole, all aggregate mean scores and their 95% confidence intervals overlap to a certain degree.

Table 50

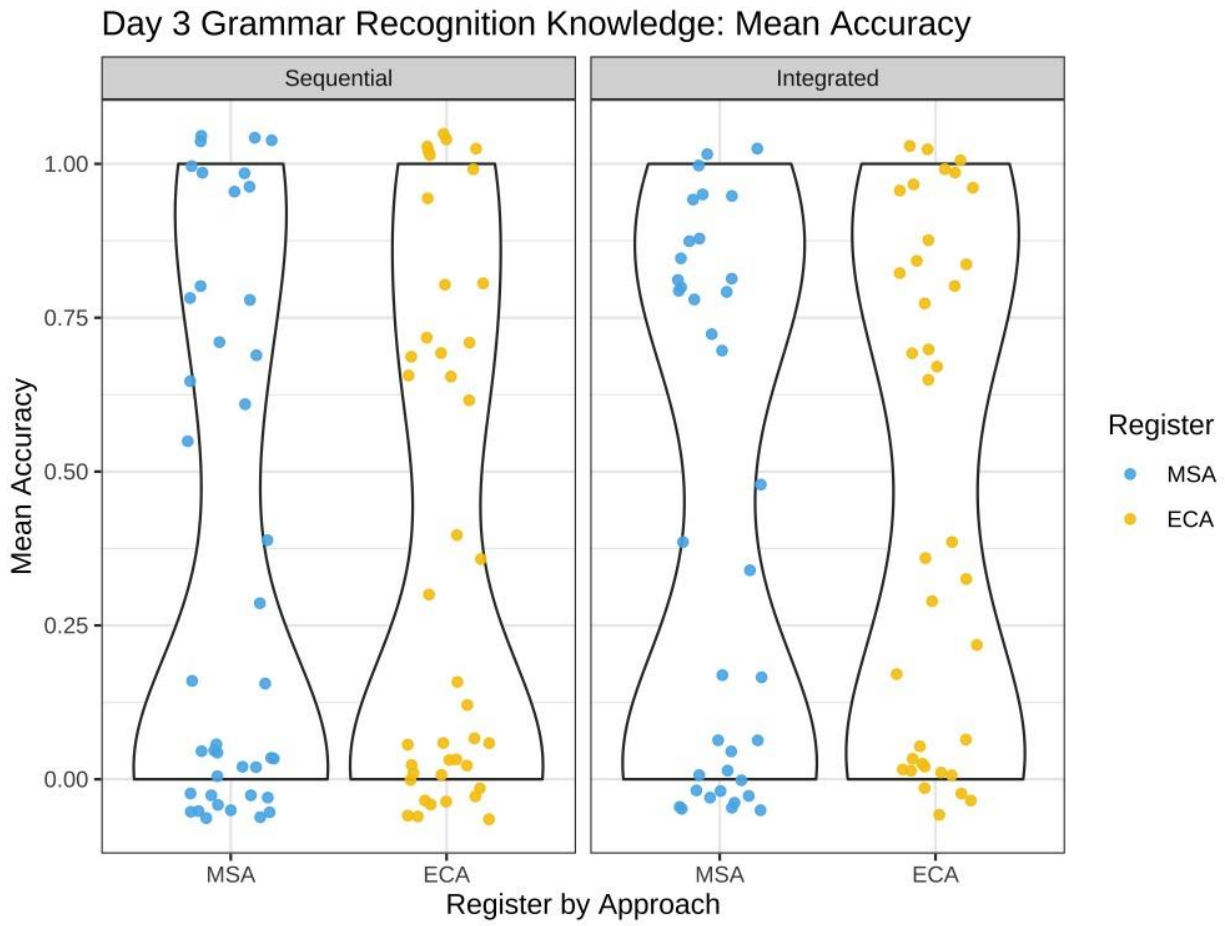
Descriptive Statistics for Grammar Recognition Knowledge: Mean Accuracy

Condition	Combined Registers	MSA only	ECA only
Integrated			
N	74	37	37
Missing	0	0	0
Mean (CI)	0.45 (0.36, 0.55)	0.44 (0.30, 0.58)	0.47 (0.33, 0.61)
SD	0.42	0.42	0.42
Skewness	0.08	0.11	0.05
Kurtosis	-1.78	-1.86	-1.78
Sequential			
N	82	41	41
Missing	0	0	0
Mean (CI)	0.38 (0.29, 0.47)	0.38 (0.25, 0.52)	0.38 (0.25, 0.51)
SD	0.42	0.43	0.42
Skewness	0.43	0.45	0.42
Kurtosis	-1.58	-1.63	-1.58

All curriculum-register subgroups have surprisingly equivalent spreads of data around their associated means, with standard deviations ranging from 0.42 - 0.43. According to the skewness and kurtosis values in Table 51, which are all less than |2| (Lomax & Hahs-Vaughn, 2012), the data appear to be relatively normally distributed. However, a visual check of the spread of datapoints in Table 51 indicates a polarization of accuracy scores (as was the case with grammar recognition mean accuracy scores for RQ1, which compared +SLV and -SLV curricula). This finding is confirmed by the density plot in Figure 71 in appendix C, which has two peaks around 0 and 1, and in the heavy-tailed residuals of the QQ plot. Thus, the inferential model below is likely less accurate in capturing variability among the more average-performing participants whose mean scores would fall between these two extremes.

Figure 46

Violin Plots with Jittered Overlay Illustrating the Distribution of Mean Accuracy by Participant for Grammar Recognition



Analyses.

The results of the linear mixed effects model for grammar recognition knowledge can be seen in Table 52. According to the model, participants were about 6% more accurate in correctly translating the negated, conjugated verb forms studied in the integrated approach compared to in the sequential approach. The inferential effect of Approach, however, was not statistically significant (β_{int} = 0.06, t = 0, p = 1), suggesting that there may not be a practical difference between the two +SLV approaches at the population level.

Table 52

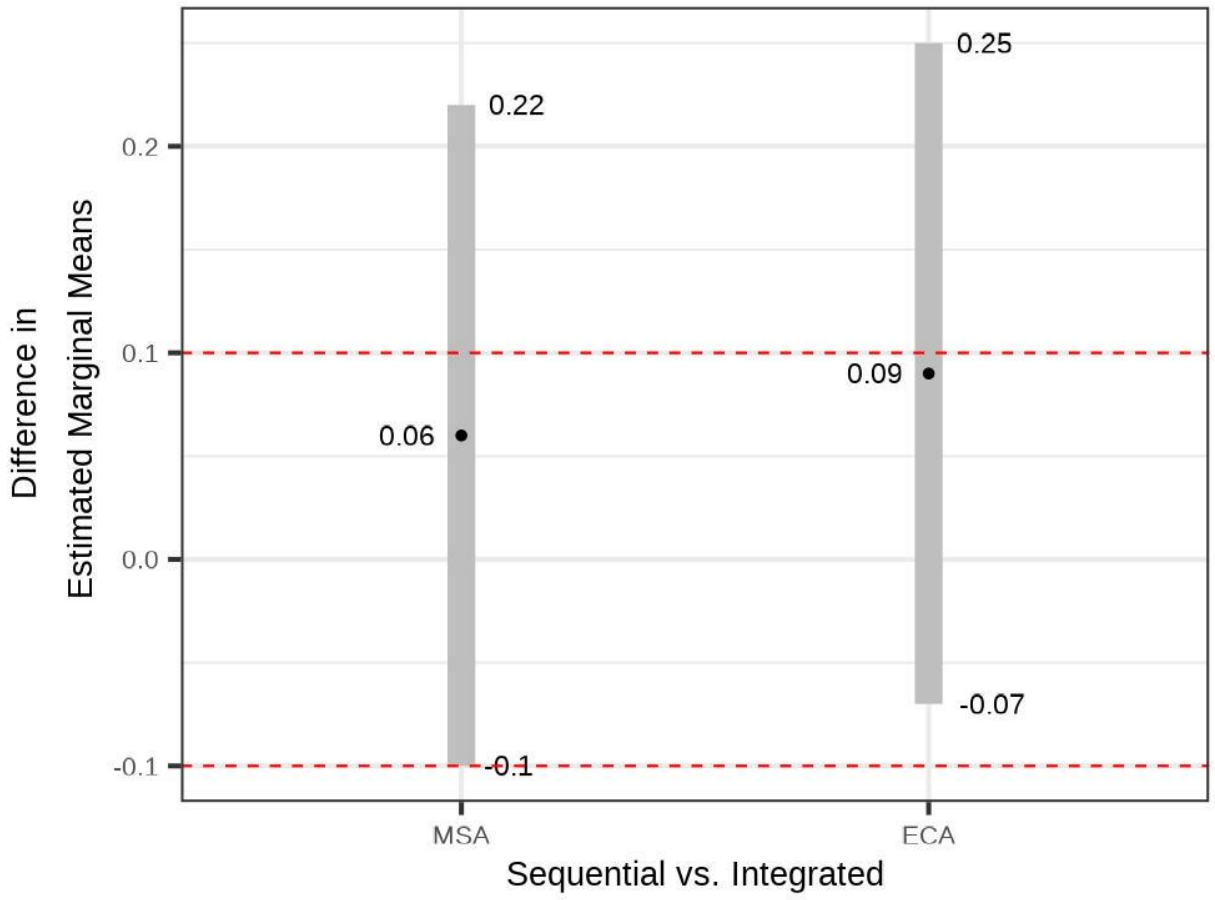
Results of Linear Mixed Effects Model for Grammar Recognition Knowledge: Mean Accuracy

	Accuracy				
Coefficient	*Estimates*	*SE*	*CI (90%)*	*t-value*	*p-value*
Intercept	0.38	0.07	0.25 - 0.51	5.80	<0.001
Approach [Int]	0.06	0.10	-0.13 - 0.25	0.62	0.536
Register [ECA]	< -0.01	0.04	-0.08 - 0.08	-0.00	1.000
Approach x Register	0.03	0.06	-0.08 - 0.14	0.48	0.630
Random Effects					
σ^2	0.03				
$\tau_{00\ participant}$	0.15				
ICC	0.83				
$N_{\ participant}$	78				
Observations	156				
Marginal R^2 / Conditional R^2	0.008 / 0.830				

In order to test for equivalence, estimated marginal means were calculated for mean accuracy of grammar recall. Figure 47 illustrates the estimated range of difference between each Approach x Register sub-group's marginal means with 90% confidence intervals. The figure confirms the findings from the descriptive statistics; namely, that the integrated approach appears to have higher accuracy than the sequential approach (the potential difference is largely positive). As such, the estimated range of score differences extends beyond the upper equivalence boundary for MSA and ECA.

Figure 47

Difference in Estimated Marginal Means with 90% Confidence Intervals for Grammar Recognition Knowledge: Mean Accuracy



Finally, the estimated marginal means for Approach x Register were tested for functional equivalence. Delta was set at 10% (0.1). Results for comparing each register across approaches are displayed in Table 53.

Table 53

Tests of Difference and Equivalence using Estimated Marginal Means for Grammar Recognition Knowledge: Mean Accuracy

Contrast	EMM Diff	*SE*	*df*	Test of Difference			Test of Equivalence	
				t-ratio	*p-value*	*Cohen's d*	*t-ratio*	*p-value*
MSA_{seq} — MSA_{int}	0.06	0.10	90.11	0.62	.537	0.34	-0.42	.336
ECA_{seq} — ECA_{int}	0.09	0.10	90.11	0.90	.369	0.49	-0.14	.443

Since there were effects greater than the equivalence range, the alternative hypothesis of equivalence (neither one inferior nor superior to the other) was rejected for both MSA (t = -0.42 , p = 0.336) and ECA (t = -0.14, p = .443) across Approach types. Thus, the tests of equivalence produce results that appear, on the surface level, contrary to the tests of difference: vocabulary meaning accuracy is neither inferentially different *nor* equivalent. While this is indeed statistically possible (e.g., Godfroid & Spino, 2015), it effectively means that results cannot be reliably generalized beyond the sample population. What can be concluded is that, in the sample population, verbs studied in the integrated approach were, on average, translated into English approximately 3% less accuratey than those learned in the sequential approach.

Summary of Results

The aim of this chapter was to examine statistically how multiple sociolinguistic registers can best be included in a curriculum. Is it better to incorporate sociolinguistic variation through a sequential approach, allowing students to gain a solid foundation in one register first and adding another register second? Or does an integrated approach, learning two registers side-by-side as a

complete language system, lead to superior outcomes? Furthermore, what effect do these choices have on individual registers? Based on Usage-based approaches to language acquisition, I hypothesized that training in an integrated approach (side by side) would lead to superior outcomes than training in a sequential approach (one register after the other). This is because studying registers side by side will help build up the associative strength of how sociolinguistic variants are contextually connected. Conversely, focusing solely on one register first may lead to an unintentional type of "blocking effect", where the existence prior of knowledge makes it more difficult to associate new variant forms with established form-meaning-function units.

The results did not support the hypothesis that training in an integrated approach leads to universally superior outcomes compared to in a sequential approach. In terms of processing speed, participants who trained in the integrated approach provided correct answers more quickly compared to those who trained in the sequential approach. This was true for both MSA and ECA register outcomes across all four knowledge types (productive and receptive knowledge of vocabulary and grammar). However, the magnitudes of these differences (Cohen's *d*) ranged from small (*d* = -0.09 for knowledge of MSA vocabulary meaning) to medium-small (*d* = -0.76 for knowledge of MSA grammar recognition) by SLA-standards (Plonsky & Oswald, 2014). Furthermore, the difference in average log-transformed reaction times between the two approaches was statistically equivalent (less than 100 ms) for both MSA and ECA register outcomeså across all four knowledge types. Thus, although training in an integrated approach was associated with faster reaction times, the inferential findings suggested that the two approaches are functionally equivalent in terms of processing speed.

The results for accuracy, both as an absolute binary measure (0/1) and as an approximate measure (the Levenshtein Distance), present a more nuanced picture. The productive knowledge

measures (vocabulary form and grammar recall knowledge) were the only ones where descriptive results differed between MSA and ECA knowledge. Specifically, absolute accuracy (0/1) was marginally higher for MSA items studied in the integrated approach, yet marginally lower for ECA items studied in the integrated approach. However, none of these comparisons were inferentially significant for difference or equivalence apart from knowledge of MSA vocabulary forms (statistical equivalence between integrated and sequential approaches, p = 0.034).

Switching to the Levenshtein Distance (the number of edits needed to correct an answer) as an approximate measure of accuracy erased any register-specific differences across the two approaches. For vocabulary form knowledge, the Levenshtein Distance outcomes were marginally worse for participants who trained in the integrated approach. This was true for both MSA (0.02 additional edits required, d = 0.03) and ECA (0.34 additional edits required, d = 0.56) stimuli. However, these differences were small enough to be considered statistically equivalent (falling within the pre-determined range of 1 letter edit). For grammar recall knowledge, the inferential comparisons of Levenshtein Distance scores between approaches were not significant for either difference or equivalence. The descriptive differences between these two approaches, however, offer practical insights into the approximate accuracy costs associated with training in an integrated approach versus a sequential one. Similar to the findings for vocabulary form knowledge, the Levenshtein Distance outcomes for grammar production were also marginally worse for participants who trained in the integrated approach compared to those who trained in the sequential approach. Training in an integrated approach was associated with an 0.4 additional edits required, d = 0.03) for MSA stimuli, and an additional 0.34 edits

required (d = 0.56) for ECA stimuli. Both of these differences are considered small by SLA field-specific standards (Plonsky & Oswald, 2014)

Finally, for receptive knowledge measures, only accuracy as an absolute measure (0/1) was used. Since all inferential results were nonsignificant for both difference and equivalence, descriptive results will be discussed. The descriptive results for vocabulary meaning and grammar recognition knowledge point to opposite conclusions. For vocabulary meaning knowledge, participants who trained in an integrated approach were slightly less accurate for both MSA (by 0.06%, $d \sim 0$) and ECA (by 8%, d = 0.59) items compared those who studied in the sequential approach. For grammar recognition knowledge, the opposite was true: participants who trained in an integrated approach were slightly *more* accurate for both MSA (by 6%, d = 0.34) and ECA (by 9%, d = 0.49) verbs compared to their sequential approach peers. The magnitudes of difference for all receptive knowledge comparisons between approaches are considered small.

In sum, the results of the inferential analyses suggest that studying in either approach, integrated or sequential, will lead to functionally equivalent processing speed for learners. The differences in mean log reaction time all fall within the range of functional equivalence for both MSA and ECA registers. For accuracy measures, the only inferential conclusion that can be drawn is that choosing one approach over the other does not seem to affect the development of approximate (Levenshtein Distance) word form knowledge, nor absolute (0/1) word form knowledge for MSA register vocabulary. For all other remaining measures of accuracy, no inferential conclusions could be drawn. Beyond inferential conclusions, the magnitudes of difference (as calculated by Cohen's d) were all considered small. On the whole, the results do not support the hypothesis that learners who study in an integrated approach will universally

outperform learners who study in a sequential approach. From a psycholinguistic perspective, neither approach seems to more thoroughly establish of form-meaning-function connections in the mind of the learner. For those form-meaning-function connections which have been established however, it seems that they can be recall with equal speed regardless of whether they were learned in a sequential approach or an integrated one.

Chapter 5: Discussion and Conclusion

Results for RQ1: Studying in Two Registers Rather than One Leads to Slightly Less Accuracy But Comparable Processing Speed

The first research question investigated the impact of studying in one register (-SLV) versus in two (+SLV). Curriculum type (-SLV vs. +SLV) was a within-subjects variable in the experiment. Outcomes for acquiring the lexis and grammar of Mini-Arabii were measured in terms of accuracy as well as processing speed. On the whole, the results suggest that adding a second register to the curriculum comes at a cost of accuracy. For productive knowledge of word form as well as grammar, mean absolute accuracy (that is, accuracy as a binary 0/1 variable) was higher for items learned in a -SLV curricular condition than in a +SLV one. Specifically, participants' ability to recall +SLV Arabic vocabulary forms was, on average, 11% less accurate than -SLV forms. For productive grammar knowledge (providing the correct Arabic verb form), the gap in mean absolute accuracy was 14%. Not only were these differences statistically significant in the linear mixed effects models, they were furthermore beyond the pre-determined functional equivalence bounds of +/- 10%. However, caution should be taken when extrapolating these results to a broader population, as the linear mixed effects model diagnostics for absolute accuracy showed some violations of the assumptions of regression.

In order to gain a more nuanced understanding of productive knowledge accuracy, the Levenshtein Distance (the number of character edits needed to transform an incorrect answer to a correct one) was adopted as a secondary metric. The results of the linear mixed effects models found a statistically significant difference between +SLV and -SLV outcomes. Specifically, +SLV answers required, on average, 0.2 more changes than -SLV items to arrive at the correct lexical form, and 0.6 more changes to arrive at the correctly conjugated and negated verb form.

However, the pre-determined equivalence range for Levenshtein Distance was +/- 1; that is, if the estimated difference between group scores was 1 or less, then the groups could be considered functionally equivalent. As such, the outcomes of the two curricular conditions were also found to be significantly equivalent as 0.2 and 0.6 are both less than 1.

Finally, across the board, the results for processing speed (log rt) were remarkably similar between -SLV and +SLV lexical and grammatical knowledge. The pre-determined range of functional equivalence was 100 ms (4.61 on the log-transformed scale); that is, if the estimated group outcomes were within 100 ms of each other, the groups could be considered to have achieved equivalent processing speeds. Within the sample data, +SLV processing time was marginally slower than -SLV by 1% (0.01 log ms) for word form knowledge and 2% (0.02 log ms) for word meaning knowledge. The gap was slightly larger for grammar knowledge, where +SLV log rt was, on average, 11% (0.11 log ms) slower for recall of negated, conjugated verbs and 7% (0.07 log ms) slower for translation of those verbs. However, all of these differences fell well within the equivalence bounds. As such, the inferential tests of equivalence were significant for processing speed across all knowledge sub-types of Mini-Arabii. Furthermore, the inferential tests of difference were all non-significant apart from grammar recognition. The complementary results of difference and equivalence testing provide support that these processing speed outcomes are truly comparable rather than due to Type I error alone (Lakens et al., 2019). Thus, on the whole, it seems that studying two registers of Mini-Arabii rather than just one leads to slightly less accuracy but comparable processing speed for both vocabulary and grammar knowledge.

Results for RQ2: Integrated and Sequential Approaches to SLV Lead to Comparable Processing Speeds; But the Jury is Out on Accuracy

The second research question explored competing approaches to incorporating SLV into a curriculum: a sequential approach, where MSA is learned first and ECA is added later, and an integrated approach, where the two registers are studied side by side. Approach (Sequential vs. Integrated) was a between-subjects variable in the experiment. Outcomes for acquiring the lexis and grammar of ECA and MSA within Mini-Arabii were measured in terms of accuracy (both as a binary outcome and through the Levenshtein Distance) as well as processing speed. As discussed in the literature review, applying the key constructs of Usage-Based Approaches (UBA) to L2-SLV would suggest that studying two registers side by side would lead to superior results than learning them sequentially. Theoretically, an integrated approach would allow sociolinguistic variation to be incorporated into the form-meaning-function units of language from the outset. Furthermore, the cognitive task of mapping multiple forms to a single concept is potentially less difficult than having to retool prior knowledge.

The results of the second research question, however, do not neatly fit into these hypotheses adapted from UBA constructs. Because the outcomes of all the inferential tests for difference were not significant, only the equivalence testing will be discussed here. The clearest findings come from processing speed results. Across the board for all lexical and grammatical measures of Mini-Arabii knowledge, reaction time in the sample data was slightly faster in the Integrated than in the sequential approach. This small advantage held for both MSA and ECA items, ranging from 1% faster (ECA grammar recognition) to 8% faster (MSA grammar recognition as well as ECA grammar recall). Furthermore, all of these findings fall within the

pre-determined equivalence bounds (estimated group difference in response time of no more than 100 ms). As such, equivalence was statistically significant.

The results for accuracy, both as an approximate (Levenshtein Distance) and absolute (0/1) measure, are less clear-cut. The results from the Levenshtein Distance sample data show a slight advantage for the Sequential condition for both MSA and ECA. This advantage ranges from 0.02 fewer letter changes needed up to 0.34 letters changes needed (both from word form knowledge, for MSA and ECA comparisons across Approach respectively). These findings are all within the pre-established equivalence bounds (estimated group difference no more than 1) and, as such, are all statistically significant for equivalence.

On the other hand, when accuracy is measured as a binary outcome (0/1), the small global advantage for sequential approach training disappears. For productive knowledge (word form and grammar recall), accuracy is slightly higher for MSA items learned in the integrated approach, and slightly lower for ECA items. The extent of this advantage was limited, however, ranging from 0.3% (MSA_{int} for word form knowledge) to 8% (ECA_{seq} for word form knowledge) more accurate. Of these productive knowledge sub-types, however, the only estimated difference that fell within the equivalence bounds (and was thus statistically significant) was word form accuracy for MSA. Meanwhile, the trends are ran in opposite directions for receptive vocabulary knowledge and receptive grammar knowledge. For word meaning knowledge, accuracy was marginally higher for words learned in the sequential approach (0.06% higher for MSA items, 8% higher for ECA items). For grammar recognition knowledge, accuracy was marginally higher for verb forms studied in the integrated approach (6% higher for MSA items, 9% higher for ECA items). Across the board, however, the tests of equivalence for receptive knowledge accuracy were not statistically significant. In sum, studying sociolinguistic variation in an integrated

approach as opposed to a Sequential one lead to comparable processing speed. However, there was no clear advantage to either approach in terms of accuracy.

Pedagogical Implications

Falling above or below a somewhat arbitrary threshold of statistical significance may not seem immediately meaningful in real-world settings or classroom teaching. For that reason, the results of this dissertation were also presented in terms of quantifiable differences in outcomes. Up until now, language teachers who were interested in incorporating SLV into their classrooms were mainly limited to anecdotes about how much more "difficult" it would be for students. With this dissertation, teachers can now get a sense of the actual potential magnitude of differences in outcomes between these curricular choices. For example, when choosing to add an additional register to a curriculum (moving from -SLV to +SLV), teachers can expect that students will be between 1-12% slower depending on the specific skill. All of these delays, however, are less than 1/10th of a second. Likewise, students will likely be 4-14% less accurate (again, depending on the domain being measured). In terms of error correction, if the average -SLV student produces words that are incorrect by one letter, the average +SLV will produce words than are incorrect by 1.2-1.6 letters. In other words, the differences in spelling accuracy between the two conditions amounts to less than one letter.

What about teachers are interested in incorporating SLV into their lessons, but are not sure whether to pursue an integrated or sequential approach? The results suggest that students who learn in an integrated approach will be 1-8% faster, although again these differences in outcomes are less than 1/10th of a second. Neither approach seems superior in terms of developing students' accuracy—depending on the language skill, students who studied in the

integrated condition may be 9% more accurate or 8 % less accurate than their sequential approach peers. From an error correction standpoint, the differences in spelling accuracy between the two conditions is less than one letter.

In real-world terms, these differences in outcomes between curricula and approach seem quite small compared to the benefits of gaining SLV awareness and building communicative and pragmatic competence from the outset (Canale & Swain, 1980; Nassif & Al Masaeed, 2020). The results do not support the notion that incorporating SLV is overwhelmingly difficult for students; rather, it seems that learners are indeed capable of handling multiple registers without a dramatic loss in performance. However, the time constraints of trying to fit dialects into an already packed curriculum is all too real. Teachers can and should interpret the results of this dissertation for themselves to weigh the costs and benefits of incorporating SLV into their curricula.

Methodological and Theoretical Implications

As discussed in the literature review, the majority of L2-SLV research comes from observational studies that use variationist approaches to document the patterns of language use for intermediate and advanced learners of Western European languages. The goal of the current student was to broaden the scope of L2-SLV inquiry in three distinct ways. First, by expanding the population of interest to novice learners. Not only is awareness of SLV a key skill at the beginning stages of language learning (ACTFL, 2012a), but furthermore novice-level students tend to make up the bulk of classroom-based language learners (Heidrich-Uebel et al., forthcoming). As such, this is an understudied population within the domain of L2-SLV.

The second area of expansion was in applying rigorous psycholinguistic research methods to explore cognitive issues lying at the heart of L2-SLV. Studying L2-SLV within a tightly-controlled experimental design allowed for the effects of quantity (one versus two registers) and ordering (integrated versus sequential) to be isolated. This expansion is key for understanding both L2-SLV and for theories of adult L2 learning in general. As noted by Geeslin (2011), "researchers interested in cognitive models of language must explore how it is possible for the human mind to handle the storage and production of variable structures…" (p. 462). The results suggest that, for novice learners, mapping an additional variant form (learning two registers rather than one) results in marginally less firmly established form-meaning-function associations in terms of the ability to produce the correct lexical or grammatical form (absolute accuracy). Once these form-meaning-function relationships have been established, however, they can be recalled and recognized with functionally equivalent speed. The conclusion that adding an additional register does not come at a cost of processing speed mirrors the vocabulary reaction time data for Huntley (forthcoming) (although note that Huntley used a different analytic framework, difference rather than equivalence testing). Extending this line of inquiry into issues of ordering, whether to study multiple SLV registers sequentially or as an integrated whole, resulted in similar conclusions. Unlike what would be hypothesized from Usage-Based approaches, establishing one form-meaning-function connection first does not seem to "block" (Ellis & Sagarra, 2010) the association of additional variant forms in terms of slower processing speeds. The results were less straightforward for accuracy in recalling and recognizing these form-meaning-function units. Neither approach appeared to have a clear advantage in terms of the ability to produce the correct form or meaning of these variants.

Finally, the current study broadened the scope of L2-SLV research beyond Western European languages by targeting features of Arabic. As such, the results of this study expand the generalizability of L2-SLV findings to typologically diverse languages. The diglossic nature of Arabic, featuring variation at every linguistic level, makes it an ideal language to use for studying L2-SLV. The current study contributed methodologically to L2-SLV research through the creation of Mini-Arabii, a miniature language (Cross et al., 2020; Mueller, 2006) in which the grammatical and lexical variant features of Arabic are carefully counterbalanced. Mini-Arabii will allow researchers to continue probing cognitive variables of interest that affect the acquisition of L2-SLV in a tightly controlled experimental setting.

Limitations and Future Directions

The results of the current study contribute to our understanding of *if* and *how* sociolinguistic variation can be acquired by *ab initio* L2 learners. Although the study fills gaps in our collective knowledge of L2-SLV, it is not without its limitations. While target language exposure was considerably longer in this study than in other artificial- or miniature-language studies, the fairly low learning outcomes for grammar knowledge indicate that three days may not be enough time to adequately capture the acquisition process. Future studies could extend the amount of exposure to more evenly assess how morphological variants are acquired by learners.

Another potential limitation was the timing of the post-tests. Post-testing in the current study occurred 24 hours after learning. This timing is, by artificial and miniature-language research standards, "delayed". However, from studies on spaced vs. massed practice we know that differences in training types may only become evident after longer periods of time (e.g.,

Nakata & Suzuki, 2019). Future studies should consider adding further post-tests (i.e. one-week delay) to capture the effects of training types on long-term retention of L2-SLV.

Conclusion

The current study utilizes rigorous psycholinguistic research methods to answer questions that are both cognitive and pedagogical in nature. For example, a teacher may wonder if dialects can be introduced into the curriculum without totally confusing students. A psycholinguist would, in turn, ask how the human mind processes, stores, and accesses variable structures.

To begin answering these questions in a way that can speak to both teachers and researchers, I developed Mini-Arabii, a miniature language that mimics the lexical and grammatical sociolinguistic variation of Arabic. Participants studied Mini-Arabii over the course of two days in training conditions that operationalize the curricular choices faced by teachers every day. Results from testing on the third day indicate that learners can indeed acquire variant lexical and grammatical structures. While they may not perform as accurately as they would in a traditional curriculum where sociolinguistic variation is ignored, they are able to produce the correct forms and meanings with comparable speed. This comparability in speed holds true whether sociolinguistic variation is incorporated in a sequential manner or integrated side by side (although the relative accuracy of these two approaches remains to be determined).

Ultimately, it is up to individual researchers and teachers to decide whether or not it is worth incorporating sociolinguistic variation into their respective practices. It is my hope that Mini-Arabii and equivalence testing will be useful tools to expand the L2-SLV research agenda into new methodological and analytic directions. Finally, any documented "costs" associated with adding sociolinguistic variation to an L2 curriculum are potentially outweighed by what

was not measured here: the fact that learners can access the tools they need to achieve communicative and pragmatic competence from day one.

REFERENCES

ACTFL. (2012a). *ACTFL proficiency guidelines*. https://www.actfl.org/uploads/files/general/ACTFLProficiencyGuidelines2012.pdf

ACTFL. (2012b). *ACTFL proficiency guidelines for Arabic*. https://www.actfl.org/resources/actfl-proficiency-guidelines-2012/arabic

Al Masaeed, K. (2022). Bidialectal practices and L2 Arabic pragmatic development in a short-term study abroad. *Applied Linguistics, 43*(1), 88–114. https://doi.org/10.1093/applin/amab013

Al Masaeed, K., Abourehab, Y., Azaz, M., Nassif, L., Al Ani, S., Trentman, E., & S'hiri, S. (2022). *Translingual approaches in world language education: Perspectives from Arabic Learning contexts*. American Association of Applied Linguistics, Pittsburgh, PA.

Al-Batal, M. (Ed.). (2018). *Arabic as one language: Integrating dialect in the Arabic language curriculum*. Georgetown University Press.

Alhawary, M. T. (2013). Arabic second language acquisition research and second language teaching: What the teacher, textbook writer, and tester need to know. *Al-'Arabiyya*, *46*, 23–35.

Al-Salom, P., & Miller, C. J. (2019). The problem with online data collection: Predicting invalid responding in undergraduate samples. *Current Psychology, 38*(5), 1258–1264. https://doi.org/10.1007/s12144-017-9674-9

Al-Wer, E., & Horesh, U. (Eds.). (2019). *The Routledge handbook of Arabic sociolinguistics*. Routledge.

Al-Wer, E., Jong, R. E. de, & Holes, C. (Eds.). (2009). *Arabic dialectology: In honour of Clive Holes on the occasion of his sixtieth birthday*. Brill.

Amazon Web Services. (2022). *Amazon Polly* [Computer software]. Amazon Web Services. https://docs.aws.amazon.com/polly/index.html

Anttila, A. (2006). Variation and phonological theory. In J. K. Chambers & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (Second Edition). Wiley-Blackwell, is an imprint of John Wiley.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Arndt, J. (2012). Paired-associate learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 2551–2552). Springer.

Arts, T. (Ed.). (2014). *Oxford Arabic dictionary: Arabic-English · English-Arabic* (First edition). Oxford University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Badawi, E.-S. (1973). *Mustawayāt al-'arabīya al-mu'āṣira fī Miṣr (Levels of contemporary Arabic in Egypt)*. Dār al-Ma'ārif.

Badawi, E.-S. (1985). Educated spoken Arabic: A problem in teaching arabic as a foreign language. In K. R. Jankowsky (Ed.), *Scientific and Humanistic Dimensions of Language* (p. 15). John Benjamins Publishing Company. https://doi.org/10.1075/z.22.09bad

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bassiouney, R. (2009). *Arabic sociolinguistics*. Edinburgh Univ. Press.

Bates, D. (2010). *lme4: Mixed-effects modeling with R*. Springer. http://lme4.r-_forge.r-_project.org/book/

Bayley, R., & Tarone, E. (2012). Variationist perspectives. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 60–75). Routledge.

Bedinghaus, R. (2015). *The effect of exposure to phonological variation on perceptual categorization and lexical access in second language Spanish: The case of /s/-aspiration in western Andalusian Spanish* [Indiana University]. UMI Number: 3712091.

Blanc, H. (1964). *Communal dialects in Baghdad*. Harvard University Press.

Boberg, C. (2000). Geolinguistic diffusion and the U.S.–Canada border. *Language Variation and Change*, *12*(1), 1–24. https://doi.org/10.1017/S0954394500121015

Brustad, K. (2000). *The syntax of spoken Arabic: A comparative study of Moroccan, Egyptian, Syrian, and Kuwaiti dialects*. Georgetown University Press.

Brustad, K., Al-Batal, M., & Tūnisī, 'Abbās. (2011). *Al-Kitaab fii ta'allum al-'Arabiyya =: A textbook for beginning Arabic* (3rd ed). Georgetown University Press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *I*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Cook, S. V., Pandža, N. B., Lancaster, A. K., & Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01345

Cross, Z. R., Zou-Williams, L., Wilkinson, E. M., Schlesewsky, M., & Bornkessel-Schlesewsky, I. (2020). Mini Pinyin: A modified miniature language for studying language learning and incremental sentence processing. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01473-6

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

DeKeyser, R. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition, 19*(2), 195–221. https://doi.org/10.1017/S0272263197002040

DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge Handbook of Instructed Second Language Acquisition* (1st ed., pp. 15–32). Routledge. https://doi.org/10.4324/9781315676968

Ellis, N. C. (2009). Optimizing the input: Frequency and sampling in Usage-based and Form-Focused learning. In M. H. Long & C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 139–158). Wiley-Blackwell. https://doi.org/10.1002/9781444315783.ch9

Ellis, N. C., & Collins, L. (2009). Input and second language acquisition: The roles of frequency, form, and function. *The Modern Language Journal, 93*(3), 329–335. https://doi.org/10.1111/j.1540-4781.2009.00893.x

Ellis, N. C., & Sagarra, N. (2010). The bounds of adult language acquisition: Blocking and learned attention. *Studies in Second Language Acquisition, 32*(4), 553–580. https://doi.org/10.1017/S0272263110000264

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition*, *33*(4), 589–624. https://doi.org/10.1017/S0272263111000325

Ferguson, C. A. (1959). Diglossia. *Word, 15*, 325–340.

Ferguson, C. A. (1991). Diglossia revisited. *Southwest Journal of Linguistics, 10*(1), 214–234.

Fishman, J. A. (1967). Bilingualism with and without diglossia; diglossia with and without bilingualism. *Journal of Social Issues*, *23*(2), 29–38. https://doi.org/10.1111/j.1540-4560.1967.tb00573.x

Friederici, A. D., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *Proceedings of the National Academy of Sciences, 99*(1), 529–534. https://doi.org/10.1073/pnas.012611199

Gagolewski, M. (2022). **stringi**: Fast and portable character string processing in *r*. *Journal of Statistical Software*, *103*(2). https://doi.org/10.18637/jss.v103.i02

Geeslin, K. L. (2011). Variation in L2 Spanish: The state of the discipline. *Studies in Hispanic and Lusophone Linguistics*, *4*(2). https://doi.org/10.1515/shll-2011-1110

Geeslin, K. L. (2018). Variable structures and sociolinguistic variation. In P. A. Malovrh & A. G. Benati (Eds.), *The Handbook of Advanced Proficiency in Second Language Acquisition* (1st ed., pp. 547–565). Wiley. https://doi.org/10.1002/9781119261650.ch28

Geeslin, K. L. (2022, September 30). *Personal communication* [ZOOM].

Geeslin, K. L., & Garrett, J. (2018). Variationist research methods and the analysis of second language data in the study abroad context. In C. Sanz & A. Morales-Front (Eds.), *The Routledge Handbook of Study Abroad Research and Practice* (1st ed., pp. 17–35). Routledge. https://doi.org/10.4324/9781315639970-1

Geeslin, K. L., & Long, A. Y. (2014). *Sociolinguistics and second language acquisition: Learning to use language in context*. Routledge.

Godfroid, A. (2020). Sensitive measures of vocabulary knowledge and processing: Expanding Nation's framework. In S. A. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 433–453). New York, NY : Routledge.

Godfroid, A., & Spino, L. A. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence. *Language Learning*, *65*(4), 896–928. https://doi.org/10.1111/lang.12136

Gorilla Experiment Builder. (2021). [Computer software]. https://gorilla.sc/

Grange, J. (2022). *trimr: An implementation of common response time trimming methods* (1.1.1) [Computer software]. https://CRAN.R-project.org/package=trimr

Hashem-Aramouni, E. (2011). *The impact of diglossia on Arabic language instruction in higher education: Attitudes and experiences of students and instructors in the U.S.* [Unpublished doctoral dissertation]. California State University.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hayes-Harb, R., & Masuda, K. (2008). Development of the ability to lexically encode novel second language phonemic contrasts. *Second Language Research; London*, *24*(1), 5–33. http://dx.doi.org.proxy2.cl.msu.edu/10.1177/0267658307082980

Heidrich-Uebel, E., Kronenberg, F. A., & Sterling, S. (Eds.). (forthcoming). *Language program vitality in the United States: From surviving to thriving*. Springer.

Heinzen, E., Sinnwell, J., Atkinson, E., Gunderson, T., & Dougherty, G. (2021). *arsenal: An arsenal of "R" functions for large-scape statistical summaries* (3.6.3) [Computer software]. https://CRAN.R-project.org/package=arsenal

Hilbig, B. E. (2016). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*, *48*(4), 1718–1724. https://doi.org/10.3758/s13428-015-0678-9

Høigilt, J., & Mejdell, G. (2017). *The politics of written language in the Arab world: Writing change*. Brill.

Holes, C. (1984). Bahraini dialects: Sectarian differences exemplified through texts. *Zeitschrift Für Arabische Linguistik*, *13*, 27–67.

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties* (Rev. ed). Georgetown University Press.

Housen, A., & Simoens, H. (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition*, *38*(2), 163–175. https://doi.org/10.1017/S0272263116000176

Huntley, E. (forthcoming). Does studying multiple sociolinguistic varieties of a second language impact learning outcomes? Investigating the simultaneous acquisition of vocabulary in both standard and Egyptian Arabic. *Critical Multilingualism Studies*.

Husseinali, G. (2006). Who is studying Arabic and why? A survey of Arabic students' orientations at a major university. *Foreign Language Annals*, *39*(3), 395–412. https://doi.org/10.1111/j.1944-9720.2006.tb02896.x

Ibrahim, R., & Aharon-Peretz, J. (2005). Is literary Arabic a second language for native Arab speakers? Evidence from semantic priming study. *Journal of Psycholinguistic Research*, *34*(1), 51–70. https://doi.org/10.1007/s10936-005-3631-8

Jiang, N. (2013). *Conducting reaction time research in second language studies* (1st ed.). Routledge. https://doi.org/10.4324/9780203146255

Kennedy Terry, K. M. (2017). Contact, context, and collocation: The emergence of sociostylistic variation in L2 French learners during study abroad. *Studies in Second Language Acquisition*, *39*(3), 553–578. https://doi.org/10.1017/S0272263116000061

Khalil, S. (2020). A delineation of variation in Arabic between fuṣḥá and Egyptian 'āmmīyah. *The Language Scholar*, *6*, 28. https://languagescholar.leeds.ac.uk/wp-content/uploads/sites/3/2020/05/Language-Scholar-Special-Issue-6-Arabic-1.pdf

King, S. (2020). From African American vernacular English to African American language: Rethinking the study of race and language in African Americans' speech. *Annual Review of Linguistics*, *6*(1), 285–300. https://doi.org/10.1146/annurev-linguistics-011619-030556

Labov, W. (1972). The social stratification of (r) in New York City department stores. In *Sociolinguistic patterns* (11. print, pp. 43–54). Univ. of Pennsylvania Press.

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D. (2022). *Improving your statistical inferences* (v1.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.6409077

Lakens, D., Scheel, A., & Isager, P. (2019). *Equivalence testing for psychological research: A tutorial*. https://doi.org/10.17605/OSF.IO/QAMC6

Leap, W. (Ed.). (1995). *Beyond the lavender lexicon: Authenticity, imagination, and appropriation in lesbian and gay languages*. Gordon and Breach.

Lenth, R. V. (2022). *emmeans: Estimated marginal means, aka least-squares means* (1.8.3) [Computer software]. https://CRAN.R-project.org/package=emmeans

Lenth, R. V. (n.d.). *Vignette: Basics of estimated marginal means*. https://cran.r-project.org/web/packages/emmeans/vignettes/basics.html

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Linford, B. (2016). *The second-language development of dialect-specific morpho-syntactic variation in Spanish during study abroad* (No. 10130845) [Indiana University]. ProQuest Dissertations and Theses.

Lomax, R. G., & Hahs-Vaughn, D. L. (2012). *An introduction to statistical concepts* (3. ed). Routledge.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Wiernik, B. M., & Makowski, D. (2022). *easystats: Framework for easy statistical modeling, visualization, and reporting* [CRAN]. https://easystats.github.io/easystats/

Mandler, G., & Huttenlocher, J. (1956). The relationship between associative frequency, associative ability and paired-associate learning. *The American Journal of Psychology*, *69*(3), 424. https://doi.org/10.2307/1419045

Mathieu, L. (2016). The influence of foreign scripts on the acquisition of a second language phonological contrast. *Second Language Research*, *32*(2), 145–170. https://doi.org/10.1177/0267658315601882

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Mejdell, G. (2017). Diglossia. In E. Benmamoun & R. Bassiouney (Eds.), *The Routledge Handbook of Arabic Linguistics* (1st ed., pp. 332–344). Routledge. https://doi.org/10.4324/9781315147062-18

Modern Language Association. (2016). *Language enrollment database, 1958–2016*. Language Enrollment Database, 1958–2016. https://apps.mla.org/flsurvey_search

Morgan-Short, K. (2020). Insights into the neural mechanisms of becoming bilingual: A brief synthesis of second language research with artificial linguistic systems. *Bilingualism: Language and Cognition, 23*(1), 87–91. https://doi.org/10.1017/S1366728919000701

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience, 24*(4), 933–947. https://doi.org/10.1162/jocn_a_00119

Mueller, J. L. (2006). L2 in a nutshell: The investigation of second language processing in the miniature language model. *Language Learning, 56*, 235–270. https://doi.org/10.1111/j.1467-9922.2006.00363.x

Nakata, T., & Suzuki, Y. (2019). Mixing grammar exercises facilitates long-term retention: Effects of blocking, interleaving, and increasing practice. *The Modern Language Journal*, modl.12581. https://doi.org/10.1111/modl.12581

Nassif, L. (2019). Salience in the noticing and production of L2 Arabic forms. *Foreign Language Annals*, *52*(2), 433–457. https://doi.org/10.1111/flan.12387

Nassif, L., & Al Masaeed, K. (2020). Supporting the sociolinguistic repertoire of emergent diglossic speakers: Multidialectal practices of L2 Arabic learners. *Journal of Multilingual and Multicultural Development*, 1–15. https://doi.org/10.1080/01434632.2020.1774595

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nydell, M. K. (1993). *From modern standard Arabic to the Egyptian dialect: Conversion course*. DLS Press. http://hdl.handle.net/2027/

Owens, J. (2003). Arabic dialect history and historical linguistic mythology. *Journal of the American Oriental Society, 123*(4), 715. https://doi.org/10.2307/3589965

Owens, J. (2013). *The Oxford handbook of Arabic linguistics*. Oxford University Press.

Palmer, J. (2007). Arabic diglossia: Teaching only the standard variety is a disservice to students. *Arizona Working Papers in SLA & Teaching*, *14*, 111–122. https://journals.uair.arizona.edu/index.php/AZSLAT

Palva, H. (1984). A general classification for the Arabic dialects spoken in Palestine and Transjordan. *Studia Orientalia Electronica*, *55*(18), 357–376. https://journal.fi/store/article/view/49783

Parkinson, D. L. (1985). Proficiency to do what? Developing oral proficiency in students of modern standard Arabic. *Al-'Arabiyya*, *18*(1/2), 11–43. JSTOR. http://www.jstor.org/43195739

Petersson, K. (2004). Artificial syntactic violations activate Broca's region. *Cognitive Science*, *28*(3), 383–407. https://doi.org/10.1016/j.cogsci.2003.12.003

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Pozzi, R., & Bayley, R. (2020). The development of a regional phonological feature during a semester abroad in Argentina. *Studies in Second Language Acquisition*, 1–24. https://doi.org/10.1017/S0272263120000303

Rebuschat, P., Monaghan, P., & Schoetensack, C. (2021). Learning vocabulary and grammar from cross-situational statistics. *Cognition*, *206*, 104475. https://doi.org/10.1016/j.cognition.2020.104475

Regan, B. (2022). Individual differences in the acquisition of language-specific and dialect-specific allophones of intervocalic /d/ by L2 and heritage Spanish speakers studying abroad in Sevilla. *Studies in Second Language Acquisition*, 1–28. https://doi.org/10.1017/S027226312200002X

Regan, V., Howard, M., & Lemée, I. (2009). *The acquisition of sociolinguistic competence in a study abroad context*. Multilingual Matters. https://doi.org/10.21832/9781847691583

Rehner, K., Mougeon, R., & Nadasdi, T. (2003). The learning of sociolinguistic variation by advanced FSL learners: The case of nous versus on in immersion French. *Studies in Second Language Acquisition*, *25*(1), 127–156. https://doi.org/10.1017/S0272263103000056

Ruiz, S., Chen, X., Rebuschat, P., & Meurers, D. (2019). Measuring individual differences in cognitive abilities in the lab and on the web. *PLOS ONE*, *14*(12), e0226217. https://doi.org/10.1371/journal.pone.0226217

Ryding, K. C. (2013). *Teaching and learning Arabic as a foreign language: A guide for teachers*. Georgetown University Press.

Sagarra, N., & Ellis, N. C. (2013). From seeing adverbs to seeing verbal morphology: Language experience and adult acquisition of L2 tense. *Studies in Second Language Acquisition*, *35*(2), 261–290. https://doi.org/10.1017/S0272263112000885

Saiegh-Haddad, E. (2003). Linguistic distance and initial reading acquisition: The case of Arabic diglossia. *Applied Psycholinguistics; New York*, *24*(3), 431–451. http://search.proquest.com/llba/docview/200950524/abstract/552D8CA679AA435BPQ/9

Saiegh-Haddad, E. (2004). The impact of phonemic and lexical distance on the phonological analysis of words and pseudowords in a diglossic context. *Applied Psycholinguistics; New York*, *25*(4), 495–512. https://search-proquest-com.proxy1.cl.msu.edu/llba/docview/200866184/abstract/9F84A9EED98B4AA6PQ/8

Saiegh-Haddad, E., & Geva, E. (2008). Morphological awareness, phonological awareness, and reading in English-Arabic bilingual children. *Reading and Writing; Dordrecht*, *21*(5), 481–504. http://dx.doi.org.proxy2.cl.msu.edu/10.1007/s11145-007-9074-x

Saiegh-Haddad, E., & Taha, H. (2017). The role of morphological and phonological awareness in the early development of word spelling and reading in typically developing and disabled Arabic readers. *Dyslexia; Bracknell*, *23*(4), 345–371. http://dx.doi.org.proxy2.cl.msu.edu/10.1002/dys.1572

Sauter, M., Draschkow, D., & Mack, W. (2020). *Building, hosting, recruiting: A brief introduction to running behavioral experiments online* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/tr76d

Schiff, R., & Saiegh-Haddad, E. (2018). Development and relationships between phonological awareness, morphological awareness and word reading in spoken and standard Arabic. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00356

Schmidt, L. B. (2020). Role of developing language attitudes in a study abroad context on adoption of dialectal pronunciations. *Foreign Language Annals*, 1–22. https://doi.org/10.1111/flan.12489

Schmidt, R. (1974). *Sociostylistic variation in spoken Egyptian Arabic: A re-examination of the concept of diglossia* (No. 302694284) [Doctoral dissertation, Brown University]. ProQuest Dissertations & Theses Global.

Showalter, C. E., & Hayes-Harb, R. (2013). Unfamiliar orthographic information and second language word learning: A novel lexicon study. *Second Language Research*, *29*(2), 185–200. https://doi.org/10.1177/0267658313480154

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. Spieler & E. Schumacher (Eds.), *New Methods in Cognitive Psychology* (1st ed., pp. 4–31). Routledge. https://doi.org/10.4324/9780429318405-2

Soliman, R. (2014). *Arabic cross-dialectal conversations with implications for the teaching of Arabic as a second language* (uk.bl.ethos.651233) [University of Leeds]. White Rose eTheses Online. http://etheses.whiterose.ac.uk/id/eprint/9119

Staniak, M., & Biecek, P. (2019). The landscape of R packages for automated exploratory data analysis. *The R Journal*, *11*(2), 347. https://doi.org/10.32614/RJ-2019-033

Tagarelli, K. M., Shattuck, K. F., Turkeltaub, P. E., & Ullman, M. T. (2019). Language learning in the adult brain: A neuroanatomical meta-analysis of lexical and grammatical learning. *NeuroImage*, *193*, 178–200. https://doi.org/10.1016/j.neuroimage.2019.02.061

The Douglas Fir Group. (2016). A transdisciplinary framework for sla in a multilingual world. *The Modern Language Journal*, *100*(S1), 19–47. https://doi.org/10.1111/modl.12301

Trentman, E. (2011). L2 Arabic dialect comprehension: Empirical evidence for the transfer of familiar dialect knowledge to unfamiliar dialects. *L2 Journal*, *3*(1), 22–49. https://doi.org/10.5070/l2319068

Trentman, E. (2013). Imagined Communities and Language Learning During Study Abroad: Arabic Learners in Egypt: Imagined Communities and Language Learning During Study Abroad. *Foreign Language Annals*, *46*(4), 545–564. https://doi.org/10.1111/flan.12054

Trentman, E., & S'hiri, S. (2020). The mutual intelligibility of Arabic dialects: Implications for the language classroom. *Critical Multilingualism Studies*, *8*(1), 104–134.

Ullman, M. T. (2001a). The Declarative/Procedural Model of lexicon and grammar. *Journal of Psycholinguistic Research*, *30*(1), 37–69. https://doi.org/10.1023/A:1005204207369

Ullman, M. T. (2001b). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition*, *4*(2), 105–122. https://doi.org/10.1017/S1366728901000220

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*(1), 231–270. https://doi.org/10.1016/j.cognition.2003.10.008

Vanpee, K. (Forthcoming). Multidialectal approaches and social justice pedagogy: Toward linguistically and culturally diversified Arabic curricula. *Critical Multilingualism Studies*.

Vaughn, K. E., Cone, J., & Kornell, N. (2018). A user's guide to collecting data online. In H. Otani & B. L. Schwartz (Eds.), *Handbook of Research Methods in Human Memory* (1st ed., pp. 354–373). Routledge. https://doi.org/10.4324/9780429439957-20

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, *26*(2), 192–196. https://doi.org/10.1007/s11606-010-1513-8

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning*, *70*(S2), 221–254. https://doi.org/10.1111/lang.12395

Watson, J. C. E. (2002). *The phonology and morphology of Arabic*. Oxford University Press.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC. https://doi.org/10.1201/EBK1439808184

Wickham, H. (2022). *stringr: Simple, consistent wrappers for common string operations* (1.5.0) [Computer software]. https://CRAN.R-project.org/package=stringr

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R* (1st ed.). Routledge. https://doi.org/10.4324/9781315165547

Younes, M. (2015). *The integrated approach to Arabic instruction*. Routledge.

Younes, M. (2018). The Cornell Arabic program model. In M. Al-Batal (Ed.), *Arabic as one language: Integrating dialect in the Arabic language curriculum* (pp. 23–35). Georgetown University Press.

Younes, M., & Huntley, E. (2019). From MSA-only to an integrated curriculum. In E. Al-Wer & U. Horesh (Eds.), *The Routledge handbook of Arabic sociolinguistics* (pp. 288–299). Routledge.

Younes, M., Weatherspoon, M., Featherstone, J., & Huntley, E. (2019). *'Arabiyyat al-naas fii masr (part one): An introductory course in Arabic* (1st edition). Routledge.

Zahler, S. (2018). *The relationship between working memory and sociolinguistic variation in first and second languages: The case of Spanish subject pronouns* (No. 10841553) [Indiana University]. ProQuest Dissertations & Theses Global.

APPENDIX A: TARGET ITEMS

Table 54

Simplified phonology and transliteration in Mini-Arabii (MSA word length mean = 5.95, SD = 0.81; ECA word length mean = 5.41, SD 0.98)

	MSA Item	Translit.	Length	ECA Item	Translit.	Length	"Translation"
1.	/**d͡ʒ**ibna/	jibna	5	/**g**ibna/	gibna	5	"cheese"
2.	/**d͡ʒ**aħʃa/	jahsha	6	/**g**aħʃa/	gahsha	6	"donkey"
3.	/ni**d͡ʒ**ma/	nijma	5	/ni**g**ma/	nigma	5	"star"
4.	/ʃa**d͡ʒ**ara/	shajara	7	/ʃa**g**ara/	shagara	7	"tree"
5.	/tazalː**ud͡ʒ**/	tazaluj	7	/tazalːu**g**/	tazalug	7	"skiing"
6.	/muharːi**d͡ʒ**/	muharij	7	/muharːi**g**/	muharig	7	"clown"
7.	/**q**irda/	qirda	5	/**ʔ**irda/	irda	5	"monkey"
8.	/**q**iʃtˤa/	qishta	6	/**ʔ**iʃtˤa/	ishta	6	"cream"
9.	/ba**q**ara/	baqara	6	/ba**ʔ**ara/	ba'ara	6	"cow"
10.	/wara**q**a/	waraqa	6	/wara**ʔ**a/	wara'a	6	"paper"
11.	/ħariː**q**/	hariq	5	/ħariː**ʔ**/	hari'	5	"fire"
12.	/ħalːa**q**/	halaq	5	/ħalːa**ʔ**/	hala'	5	"barber"
13.	/**θ**aʕlab/	thalab	6	/**t**aʕlab/	talab	5	"fox"
14.	/**ð**ura/	dhura	5	/**d**ura/	dura	4	"corn"
15.	/na**ð**ˤːaːra/	nadhara	7	/na**d**ˤːaːra/	nadara	6	"glasses"
16.	/ʕa**ð**ˤm/	yadhm	5	/ʕa**d**ˤm/	yadm	4	"bone"
17.	/ʃaħaː**ð**/	shahadh	7	/ʃaħaːt/	shahat	6	"beggar"
18.	/nabiː**ð**/	nabidh	6	/nabiː**t**/	nabit	5	"wine"
19.	/**ð**ˤabja/	dhabya	6	/**z**ˤabja/	zabya	5	"gazelle"
20.	/**ð**ˤarf/	dharf	5	/**z**ˤarf/	zarf	4	"envelope"
21.	/tim**θ**aːl/	timthal	7	/tim**s**aːl/	timsal	6	"statue"
22.	/bi**ð**ra/	bidhra	6	/bi**z**ra/	bizra	5	"seed"
23.	/waːʕi**ð**ˤ/	wayidh	6	/waːʕi**z**ˤ/	wayiz	5	"preacher"
24.	/tilmiː**ð**/	tilmidh	7	/tilmiː**z**/	tilmiz	6	"student"

Table 55

Target Nouns in Mini-Arabii

Phonetic Change	MSA word	ECA word	"Translation"	Grammatical Gender	Ave. Conc. *	Syllable Length
/d͡ʒ/ → /g/	1. jibna	gibna	"cheese"	fem.	4.7	2
	2. jahsha	gahsha	"donkey"	fem.	5	2
	3. nijma	nigma	"star"	fem.	4.69	2
	4. shajara	shagara	"tree"	fem.	5	2
	5. tazaluj	tazalug	"skiing"	masc.	4.42	3
	6. muharij	muharig	"clown"	masc.	4.9	3
/q/ → /ʔ/	7. qirda	irda	"monkey"	fem.	4.9	2
	8. qishta	ishta	"cream"	fem.	4.83	2
	9. baqara	ba'ara	"cow"	fem.	4.96	3
	10. waraqa	wara'a	"paper"	fem.	4.93	3
	11. hariq	hari'	"fire"	masc.	4.68	2
	12. halaq	hala'	"barber"	masc.	4.59	2
Interdental to Alveolar	13. thalab	talab	"fox"	masc.	4.97	2
	14. dhura	dura	"corn"	fem.	4.96	2
/θ/ → /t/	15. nadhara	nadara	"glasses"	fem.	4.92	3
/ð/ → /d/	16. yadhm	yadm	"bone"	masc.	4.9	1
/ðˤ/ → /dˤ/	17. shahadh	shahat	"beggar"	masc.	4.59	2
	18. nabidh	nabit	"wine"	masc.	4.79	2
Interdental to Sibilant	19. dhabya	zabya	"gazelle"	fem.	4.72	2
	20. dharf	zarf	"envelope"	masc.	4.93	1
/θ/ → /s/	21. timthal	timsal	"statue"	masc.	4.93	2
/ð/ → /z/	22. bidhra	bizra	"seed"	fem.	4.71	2
/ðˤ/ → /zˤ/	23. wayidh	wayiz	"preacher"	masc.	4.7	2
	24. tilmidh	tilmiz	"student"	masc.	4.92	2

* Average Concreteness Rating means obtained from Brysbaert et al. (2014)

Table 56

Mini-Arabii negated verb forms in counterbalanced training groups

A. Training Group A

		Present Tense (-SLV)	Past Tense (+SLV)	
			MSA	ECA
To Eat	*she*	lam t-akal	la akal-at	ma akal-at-sh
	he	lam y-akal	la akal-Ø	ma akal-Ø-sh
To Love	*she*	lam t-uhib	la uhib-at	ma uhib-at-sh
	he	lam y-uhib	la uhib-Ø	ma uhib-Ø-sh
To Hit	*she*	lam t-adrab	la adrab-at	ma adrab-at-sh
	he	lam y-adrab	la adrab-Ø	ma adrab-Ø-sh
To Carry	*she*	lam t-ushil	la ushil-at	ma ushil-at-sh
	he	lam y-ushil	la ushil-Ø	ma ushil-Ø-sh

Table 56 (cont'd)

B. Training Group B

		Present Tense (+SLV)		Past Tense (-SLV)
		MSA	ECA	
To Eat	*she*	la t-akal	ma t-akal-sh	lam akal-at
	he	la y-akal	ma y-akal-sh	lam akal-Ø
To Love	*she*	la t-uhib	ma t-uhib-sh	lam uhib-at
	he	la y-uhib	ma y-uhib-sh	lam uhib-Ø
To Hit	*she*	la t-adrab	ma t-adrab-sh	lam adrab-at
	he	la y-adrab	ma y-adrab-sh	lam adrab-Ø
To Carry	*she*	la t-ushil	ma t-ushil-sh	lam ushil-at
	he	la y-ushil	ma y-ushil-sh	lam ushil-Ø

APPENDIX B: EXPERIMENT PROCEDURE

Figure 48

Instructions that participants received prior to beginning testing block (pictured here: form recall knowledge for-SLV vocabulary)

A.

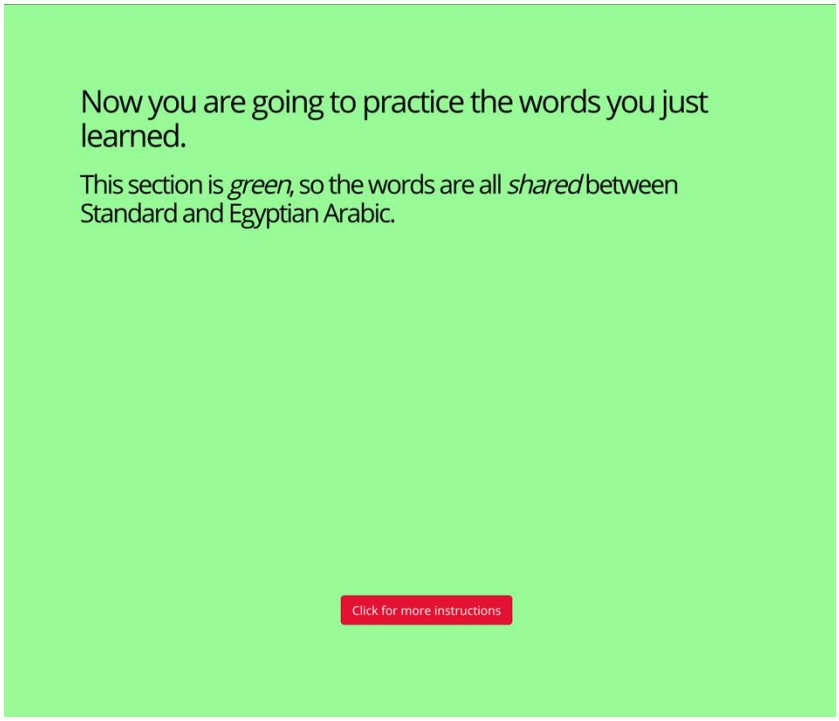
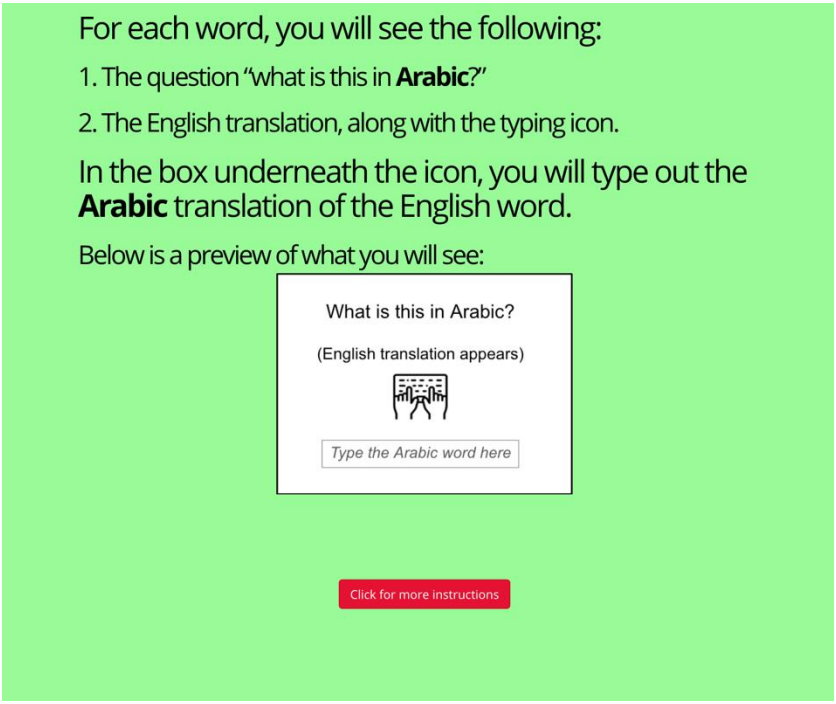


Figure 48 (cont'd)

B.



C.

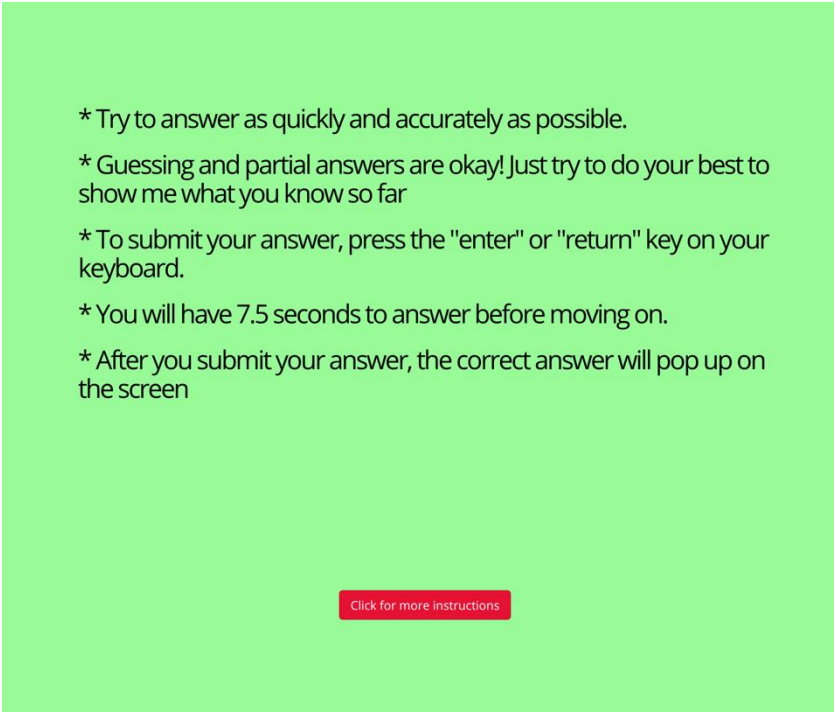


Figure 48 (cont'd)

D.

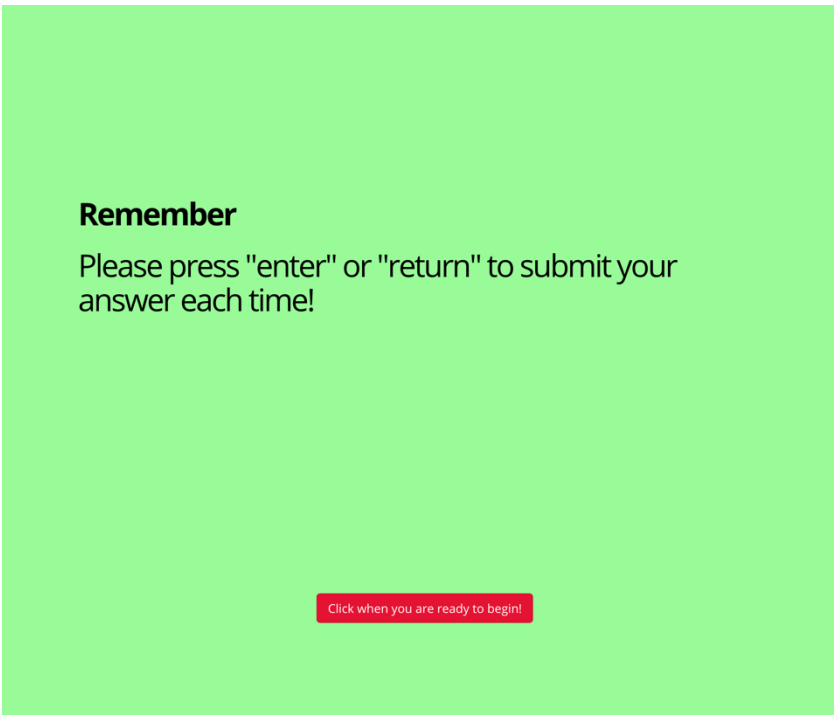


Figure 49

Instructions that participants received prior to beginning the experiment

A.

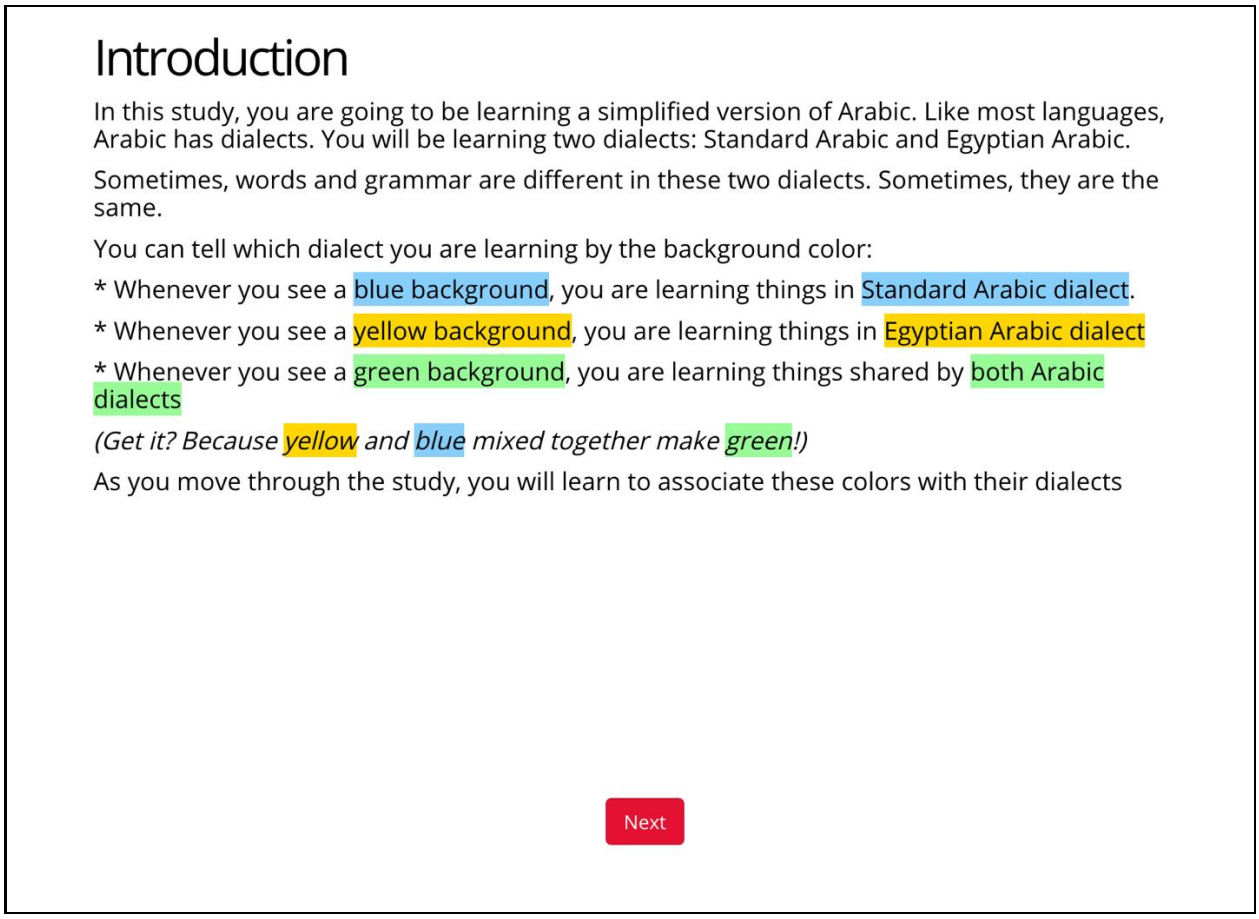


Figure 49 (cont'd)

B.

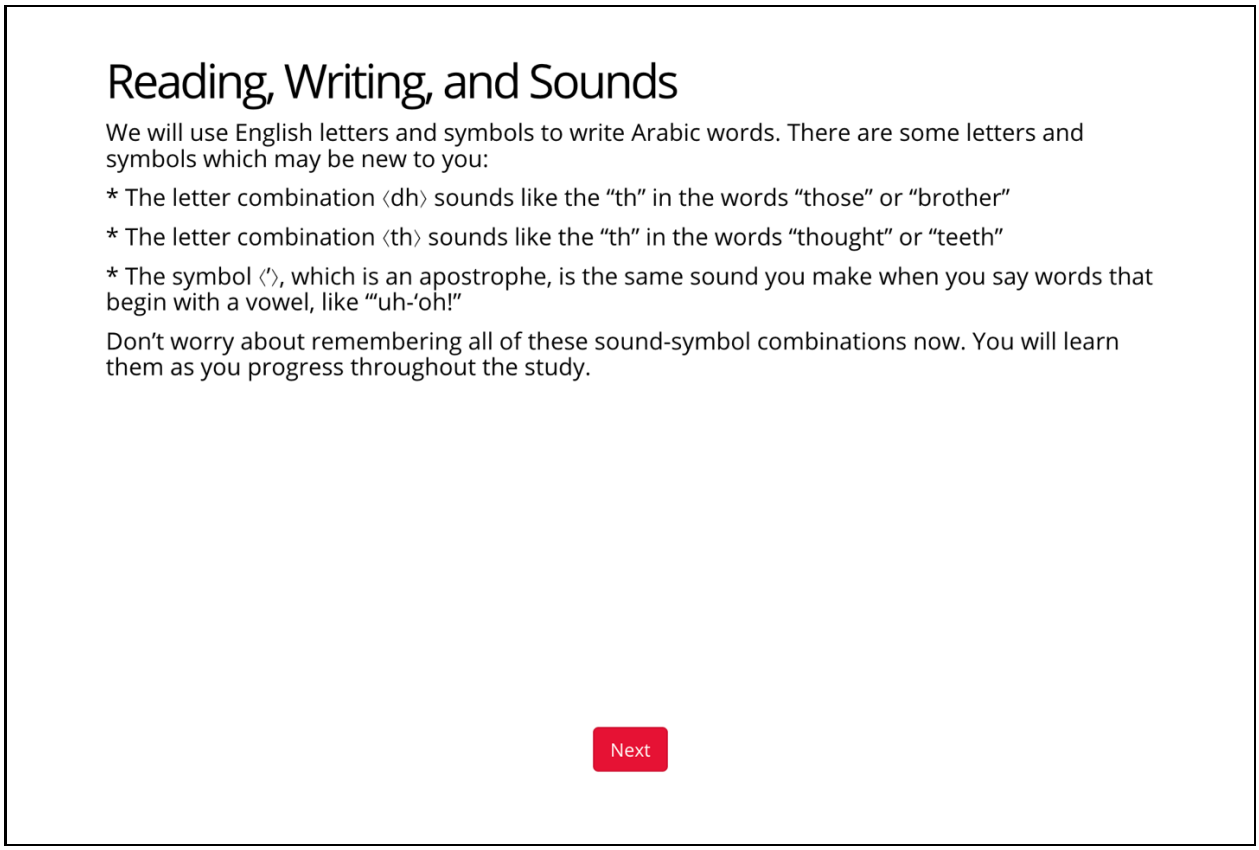


Figure 49 (cont'd)

C.

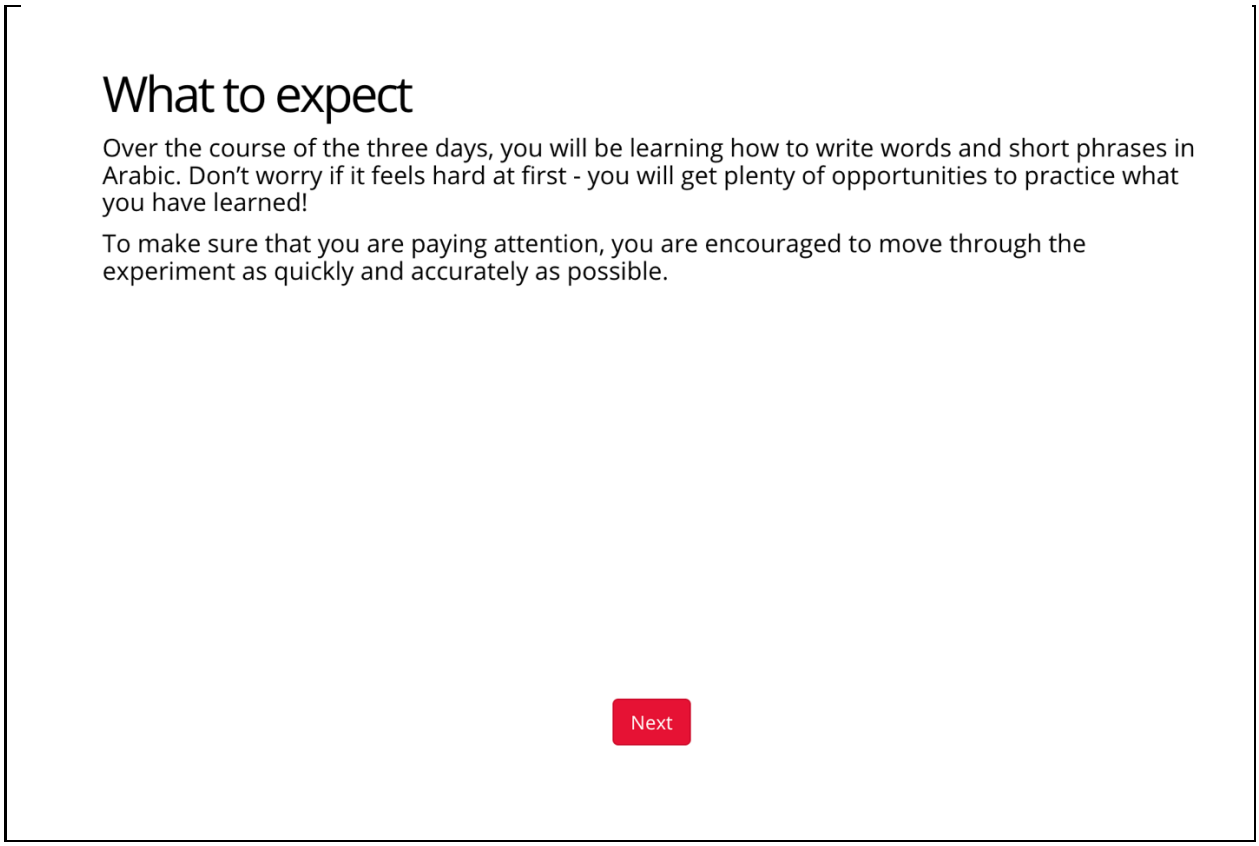


Figure 49 (cont'd)

D.

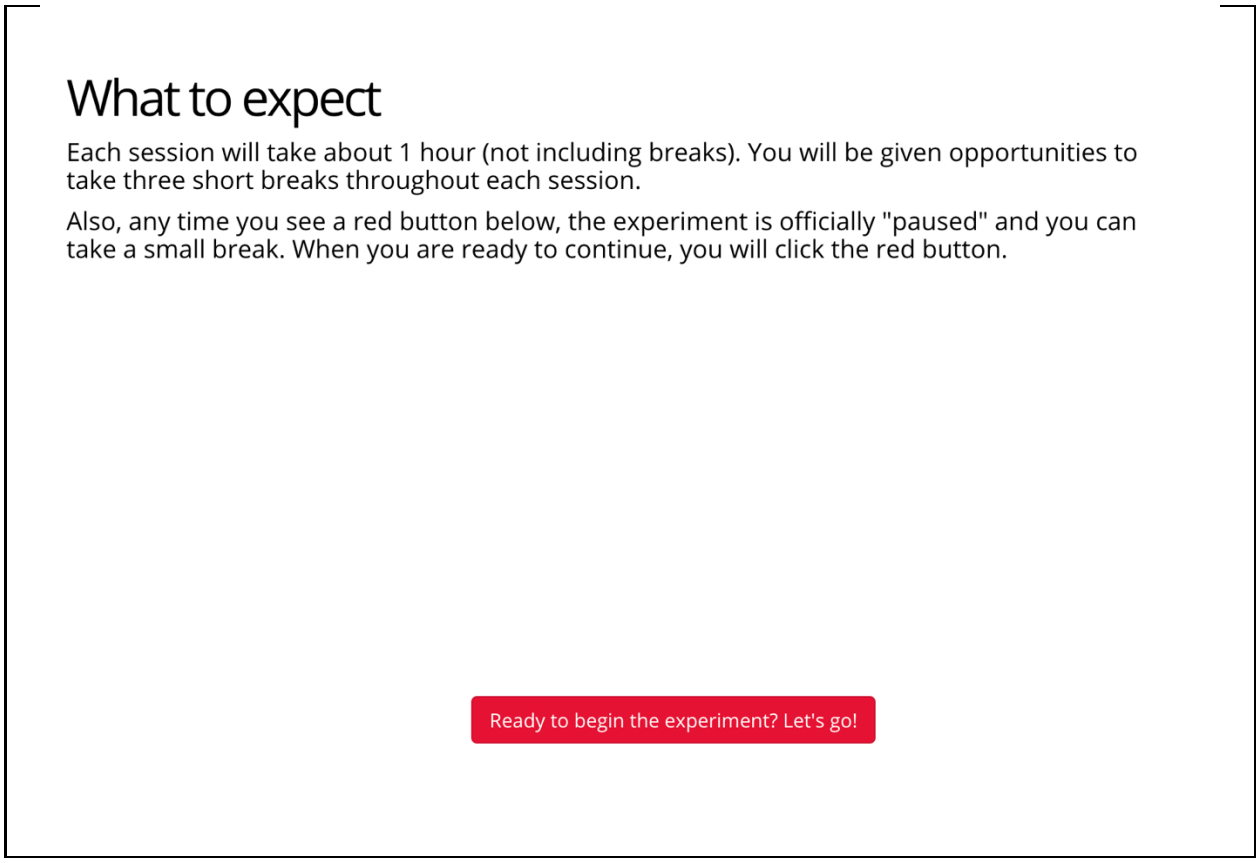


Figure 50

Training Procedure (days 1 and 2). Lighter colored blocks focus on vocabulary, darker colored blocks focus on grammar. For participants in training group A, variation occurs in vocab set 1 and past tense (in the blue MSA and yellow ECA blocks); vocab set 2 and present tense do no vary (in the "no SLV" green blocks). For participants in training group B, variation occurs in vocab set 2 and present tense (in the blue MSA and yellow ECA blocks); vocab set 1 and past tense do not vary (in the "no SLV" green blocks)

SESSION 1 (TRAINING)

**Order of Vocab Blocks*

Vocab Training
Form Test
Meaning Test

Blocks ↓		Condition ↓	
#	*Focus*	*Integrated*	Sequential
1-3	Vocab*	MSA	MSA
4-6	Grammar**	MSA	MSA
7-9	Vocab*	no SLV	no SLV
10-12	Grammar**	no SLV	no SLV
13-15	Vocab*	ECA	MSA
16-18	Grammar**	ECA	MSA
19-21	Vocab*	no SLV	no SLV
22-24	Grammar**	no SLV	no SLV

SESSION 2 (TRAINING)

*** Order of Grammar Blocks*

Grammar Training
Recall Test
Recognition Test

Blocks ↓		Condition ↓	
#	*Focus*	*Integrated*	*Sequential*
1-3	Vocab*	MSA	ECA
4-6	Grammar**	MSA	ECA
7-9	Vocab*	no SLV	no SLV
10-12	Grammar**	no SLV	no SLV
13-15	Vocab*	ECA	ECA
16-18	Grammar**	ECA	ECA
19-21	Vocab*	no SLV	no SLV
22-24	Grammar**	no SLV	no SLV

Figure 51

Posttesting (day 3). Lighter colored blocks focus on vocabulary, darker colored blocks focus on grammar. For participants in training group A, variation occurs in vocab set 1 and past tense (in the blue MSA and yellow ECA blocks); vocab set 2 and present tense do no vary (in the no-SLV green blocks). For participants in training group B, variation occurs in vocab set 2 and present tense (in the blue MSA and yellow ECA blocks); vocab set 1 and past tense do not vary (in the no-SLV green blocks)

SESSION 3 (POST-TESTS)

^*Order of Vocab Blocks*	Blocks ↓		(All conditions)
Form Test	#	Focus	
Meaning Test	1-2	Vocab^	MSA
	3-4	Grammar^^	MSA
^^ *Order of Grammar Blocks*	5-6	Vocab^	no SLV
Recall Test	7-8	Grammar^^	no SLV
Recognition Test	9-10	Vocab^	ECA
	11-12	Grammar^^	ECA

APPENDIX C: INFERENTIAL MODEL DIAGNOSTICS

Figure 52

RQ1: Word Form Log RT

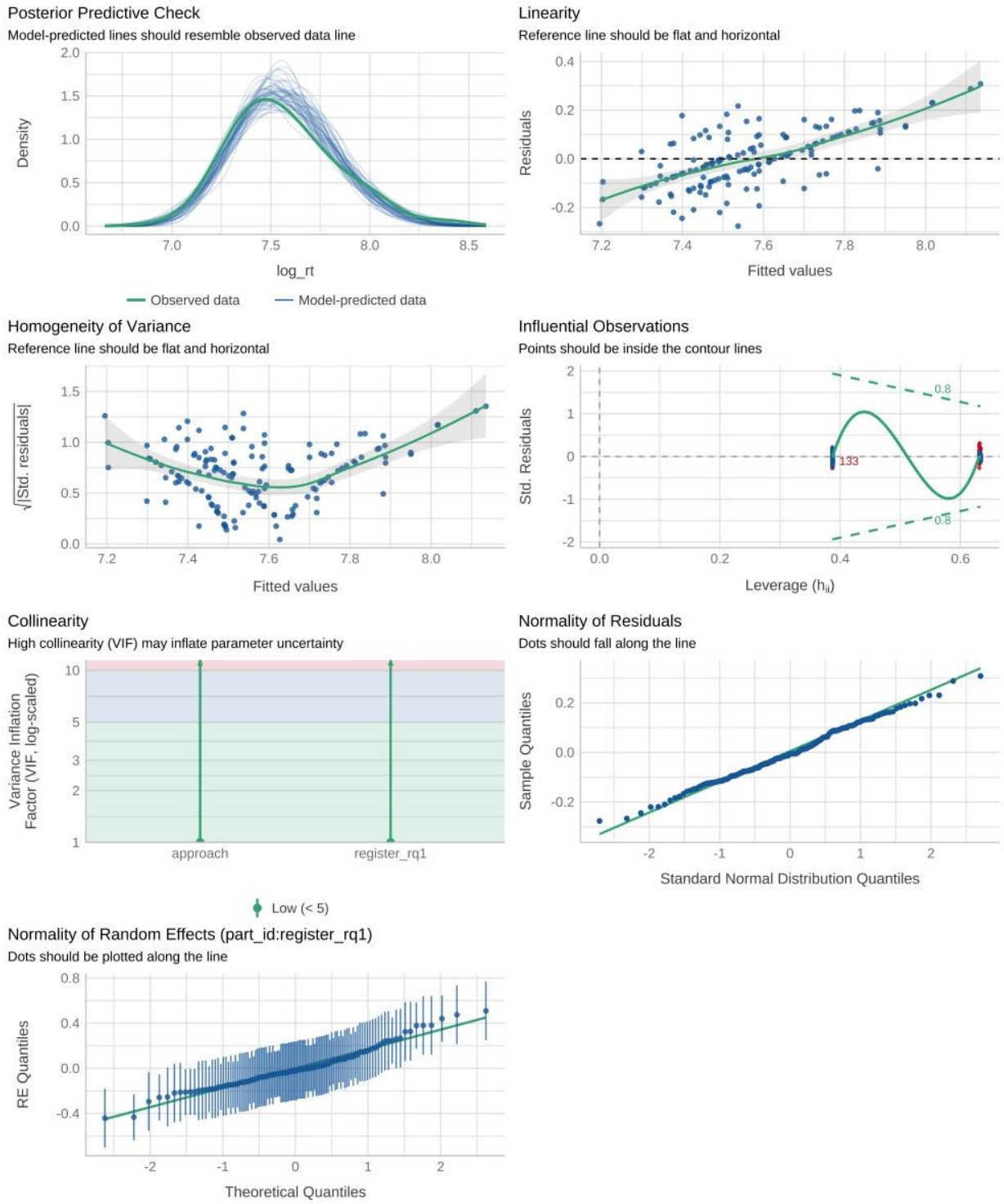


Figure 53

RQ1: Word Form Levenshtein Distance

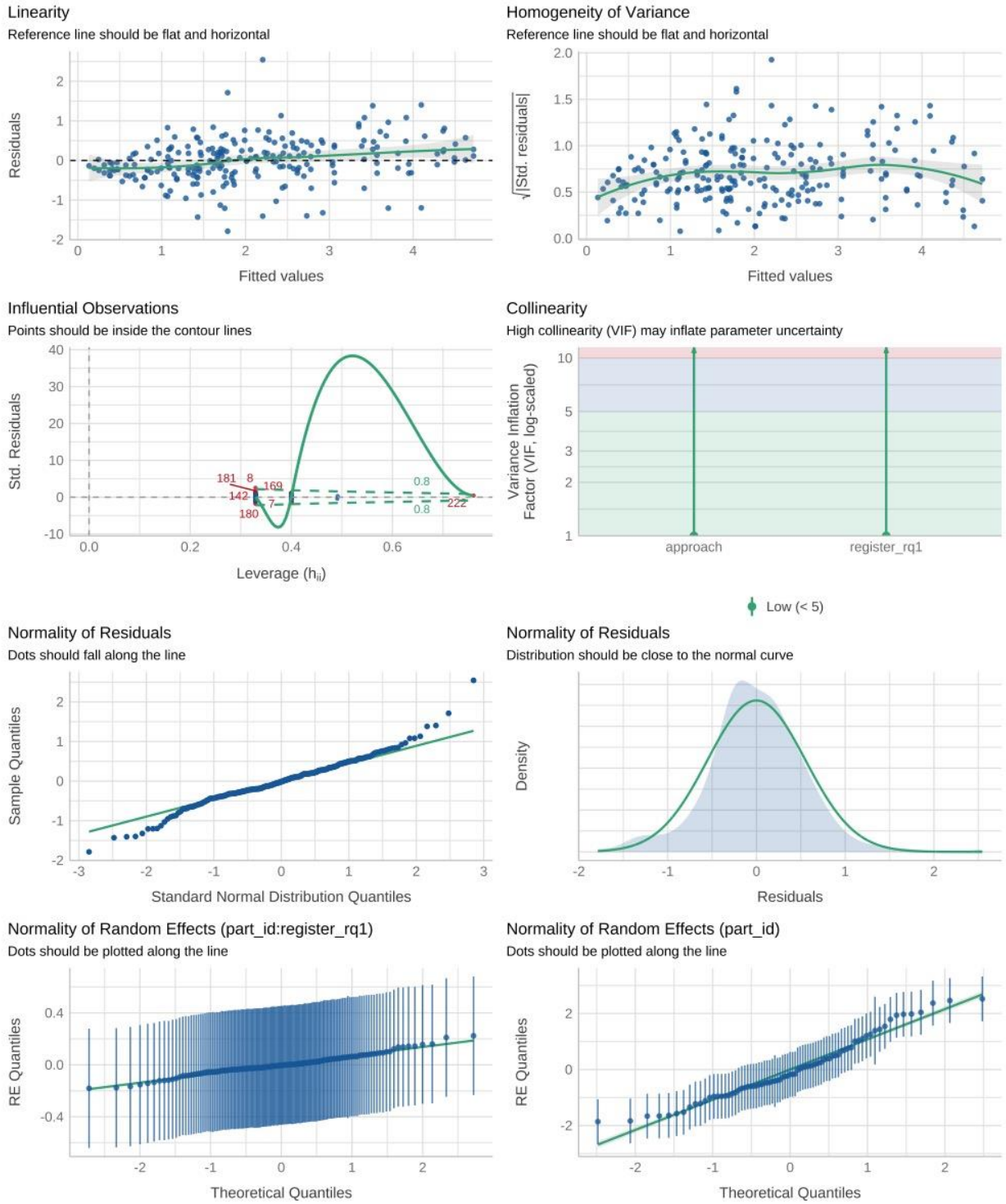


Figure 54

RQ1: Word Form Accuracy

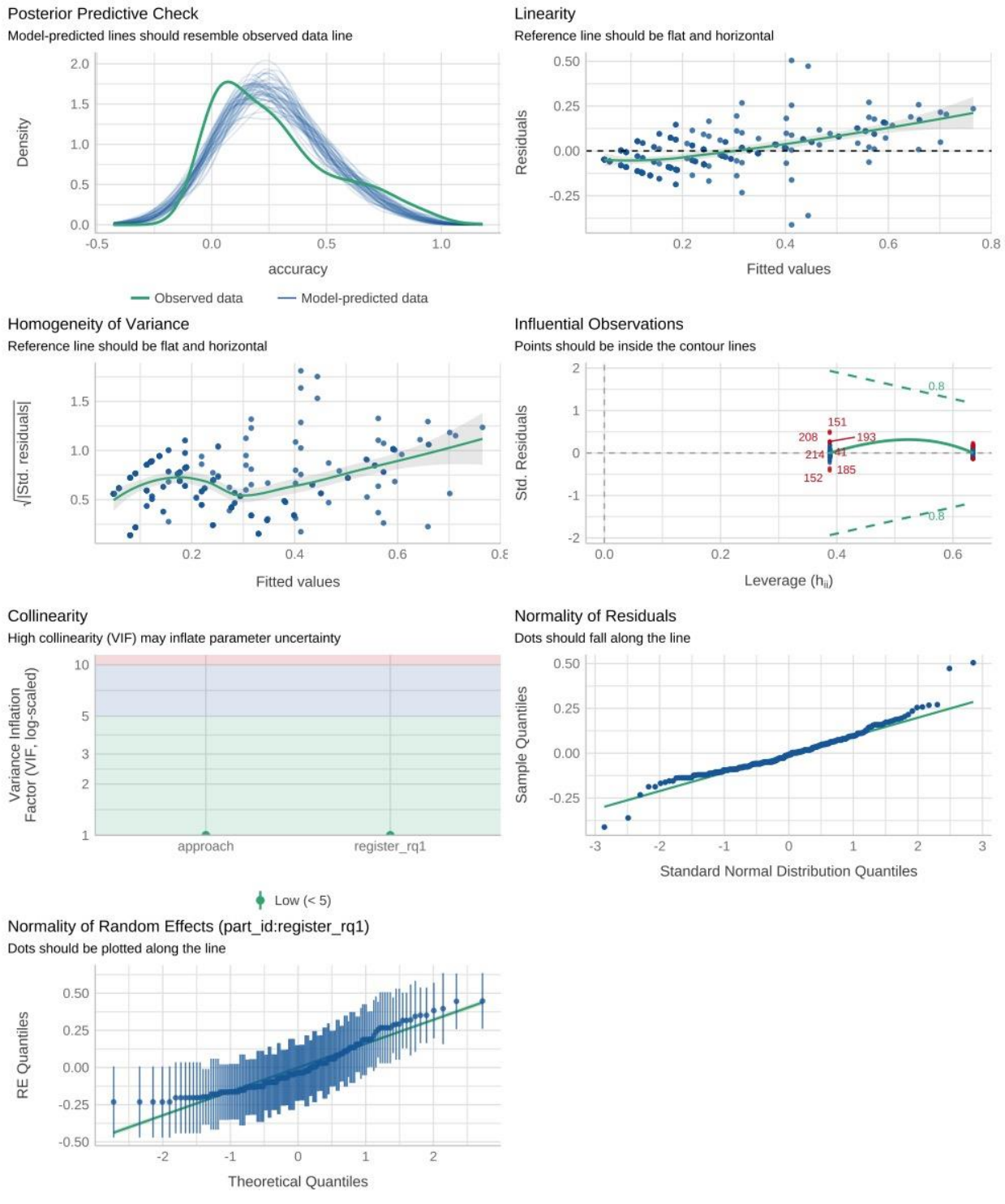


Figure 55

RQ1: Word Meaning Log RT

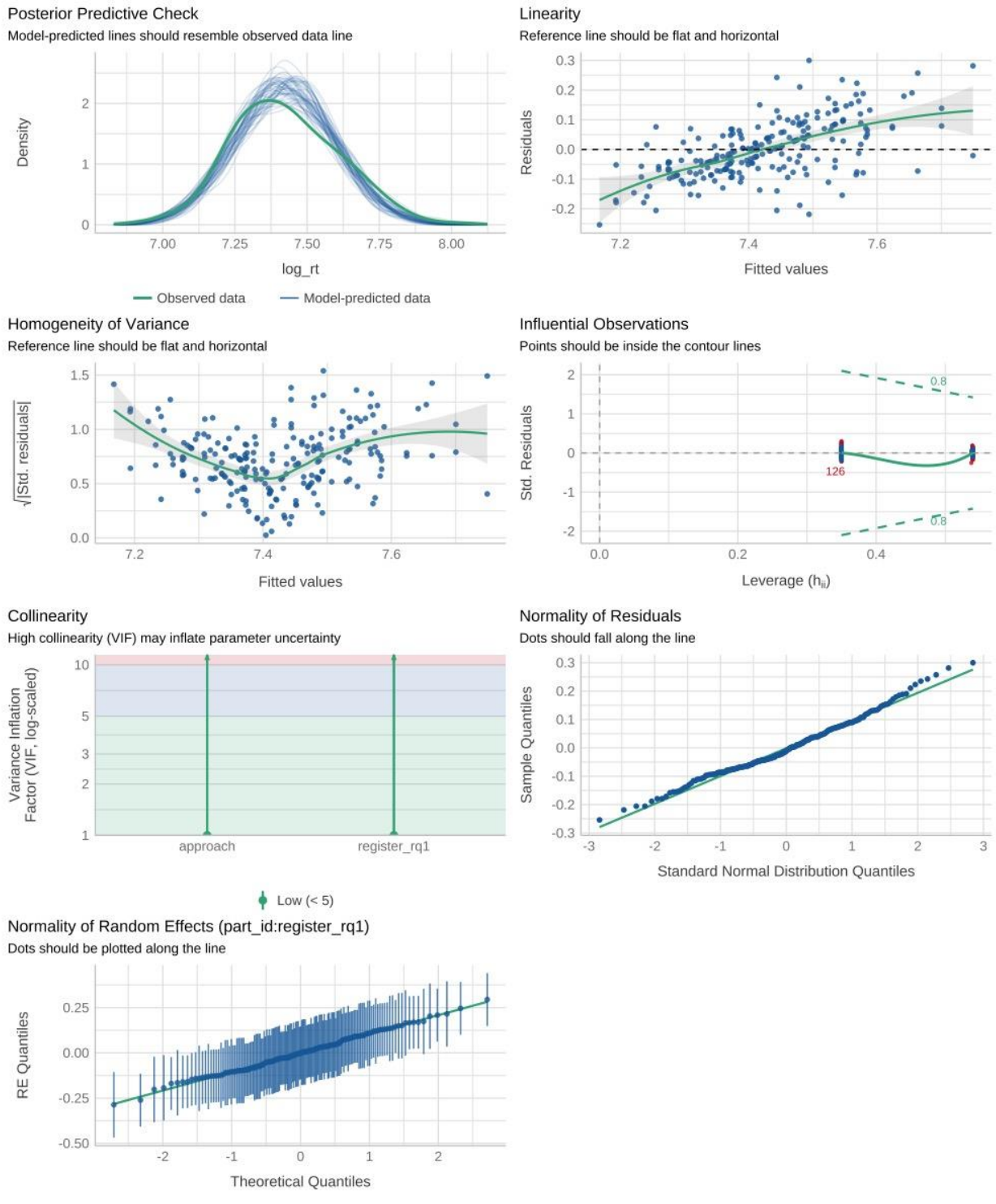


Figure 56

RQ1: Word Meaning Accuracy

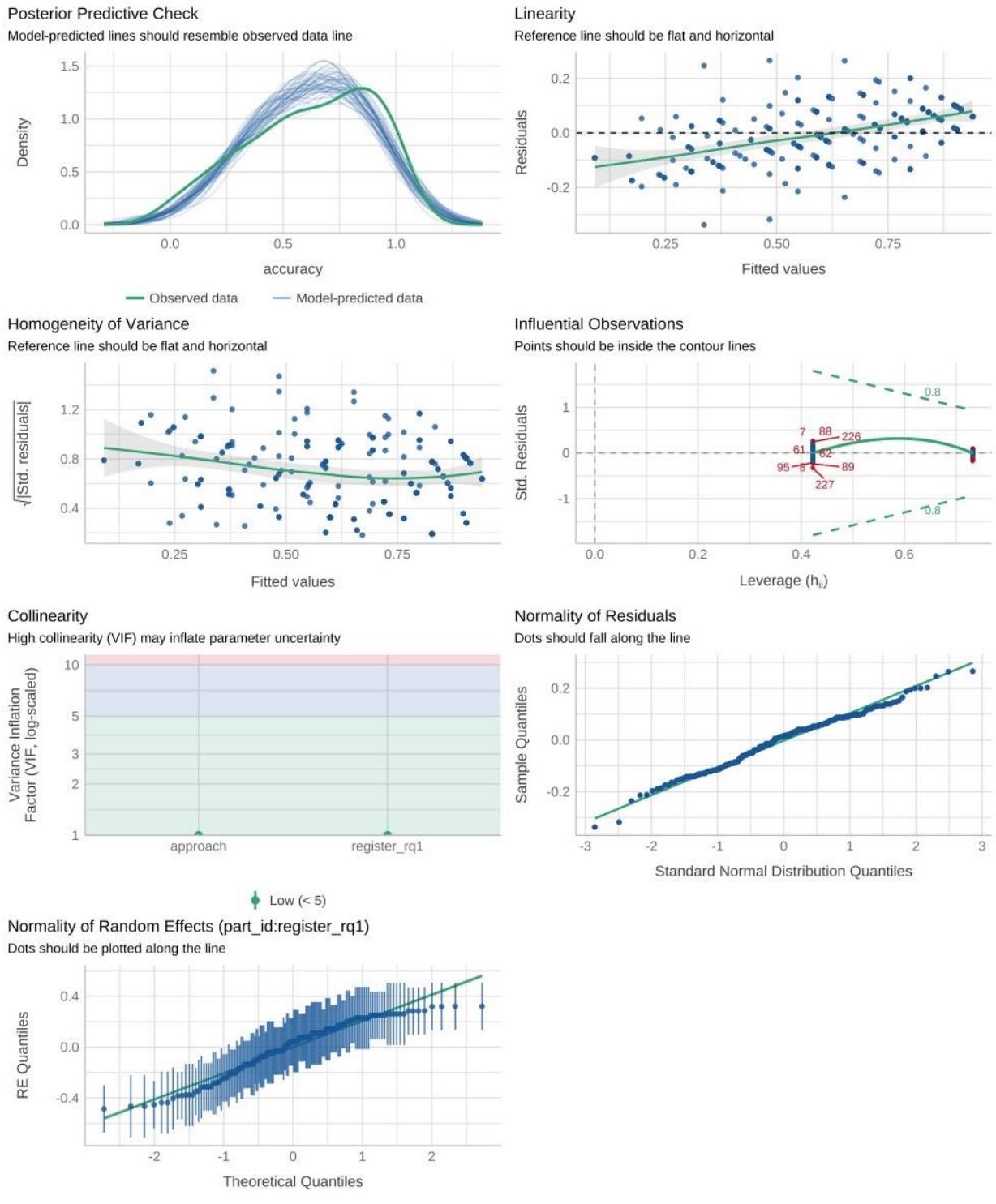


Figure 57

RQ1: Grammar Recall Log RT

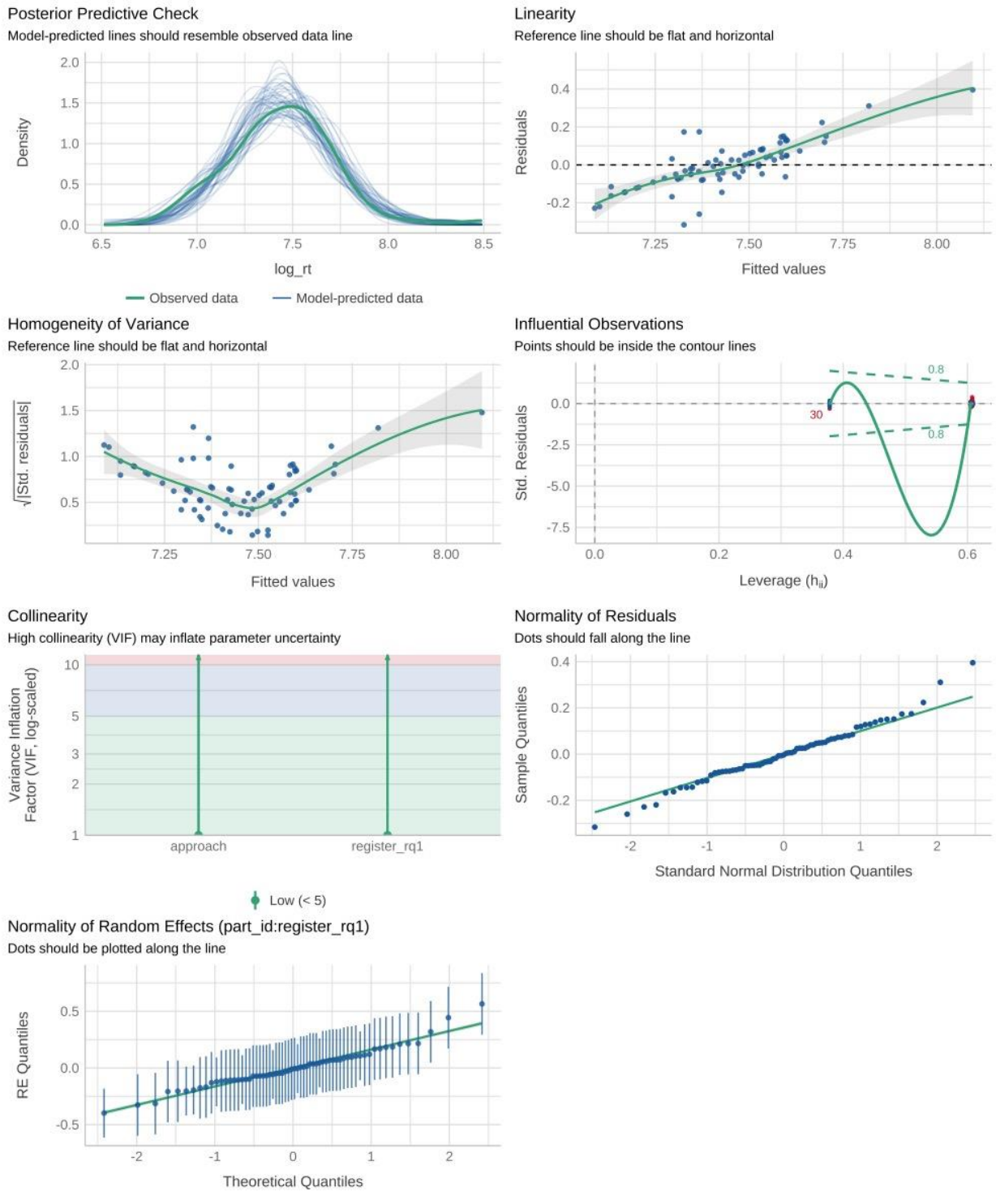


Figure 58

RQ1: Grammar Recall Levenshtein Distance

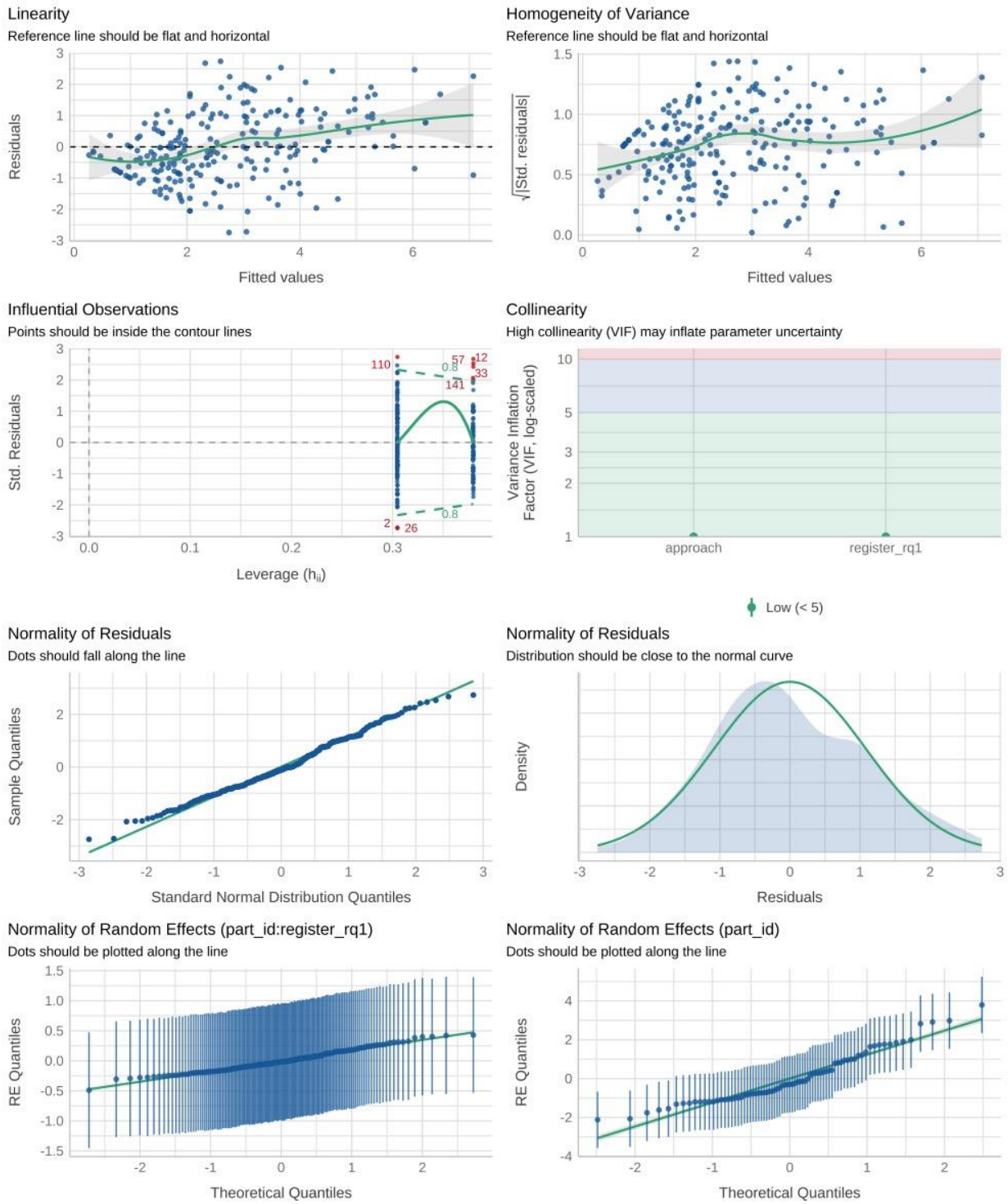


Figure 59

RQ1: Grammar Recall Accuracy

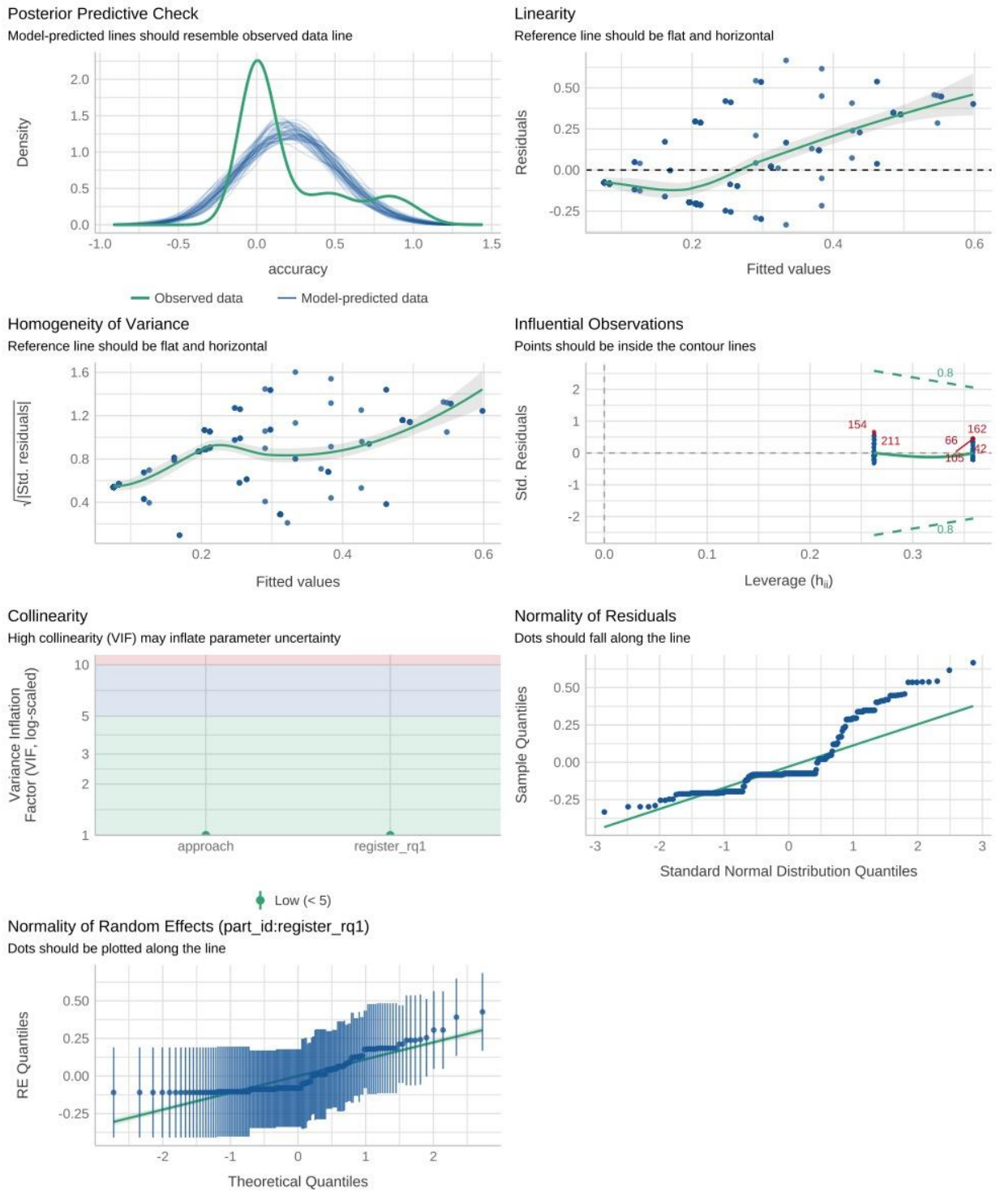


Figure 60

RQ1: Grammar Recognition Log RT

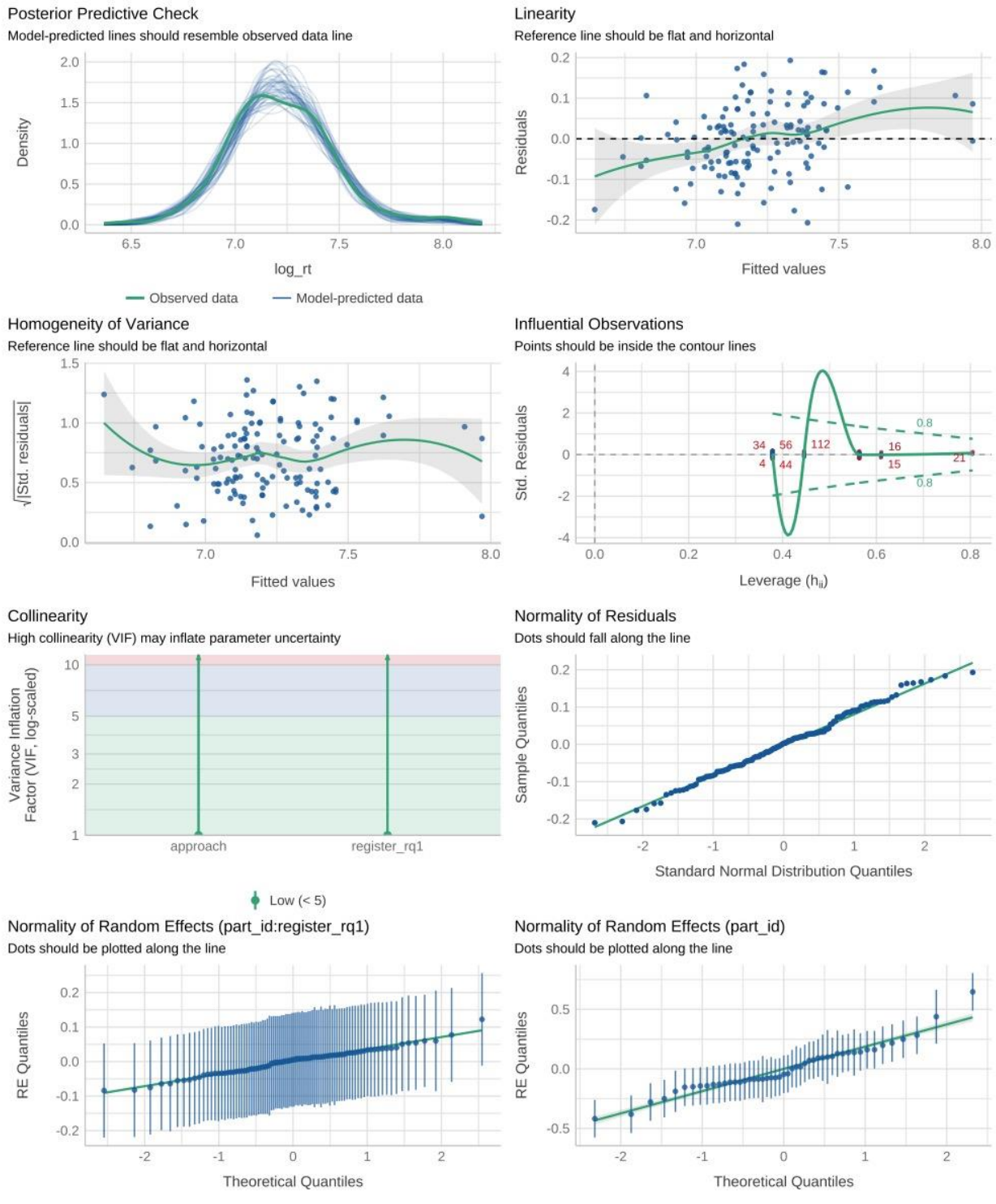


Figure 61

RQ1: Grammar Recognition Accuracy

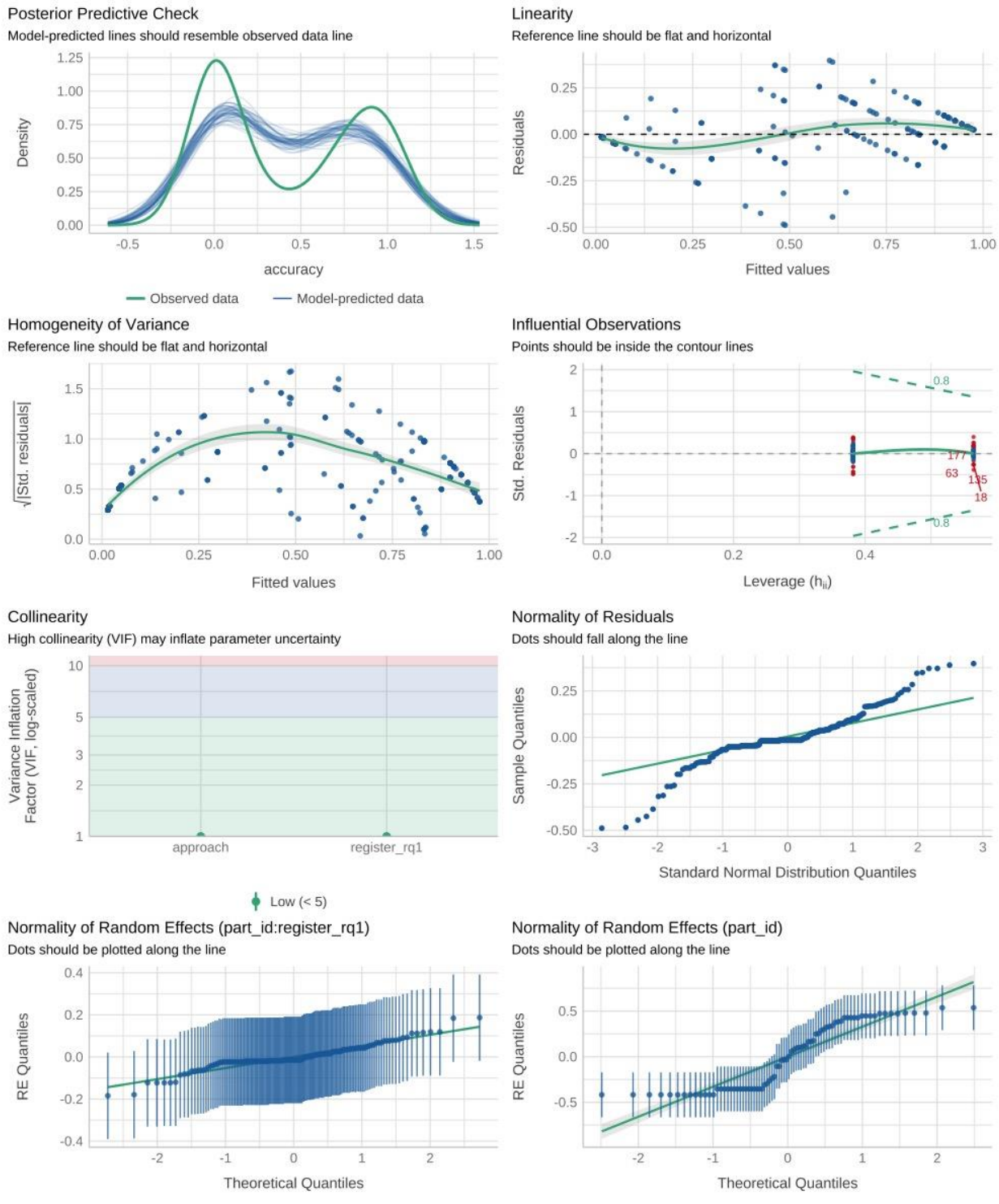


Figure 62

RQ2: Word Form Log RT

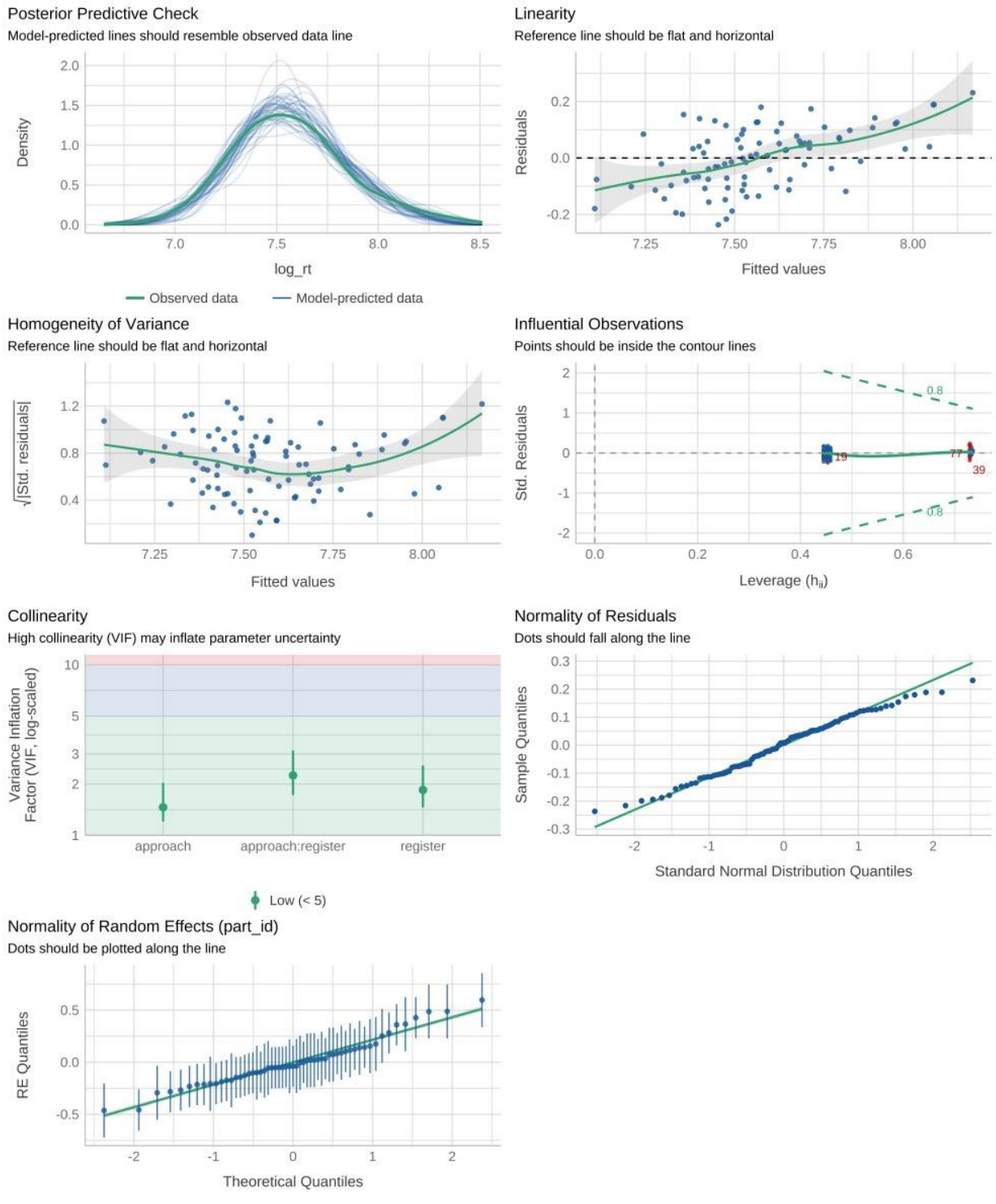


Figure 63

RQ2: Word Form Levenshtein Distance

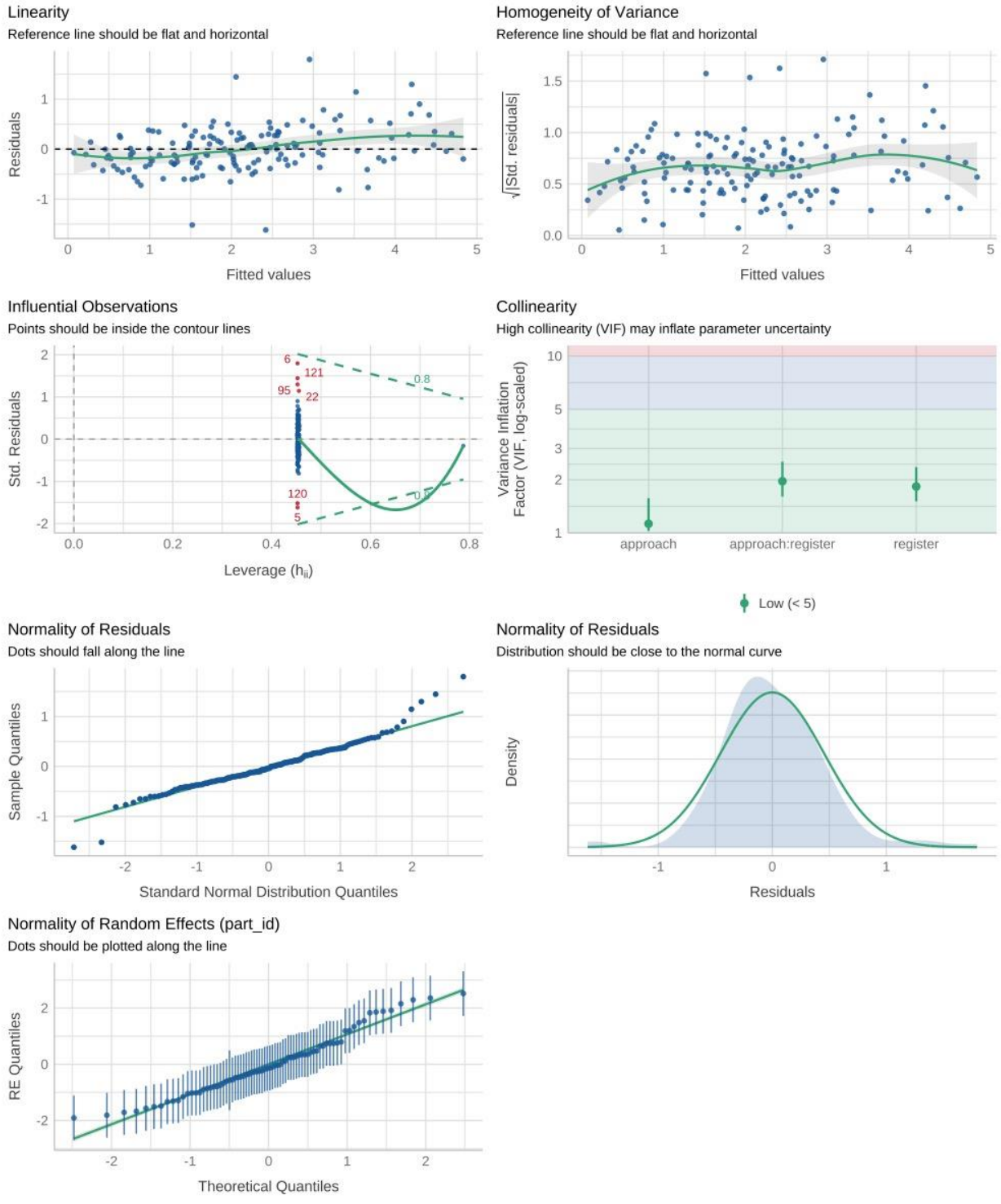


Figure 64

RQ2: Word Form Accuracy

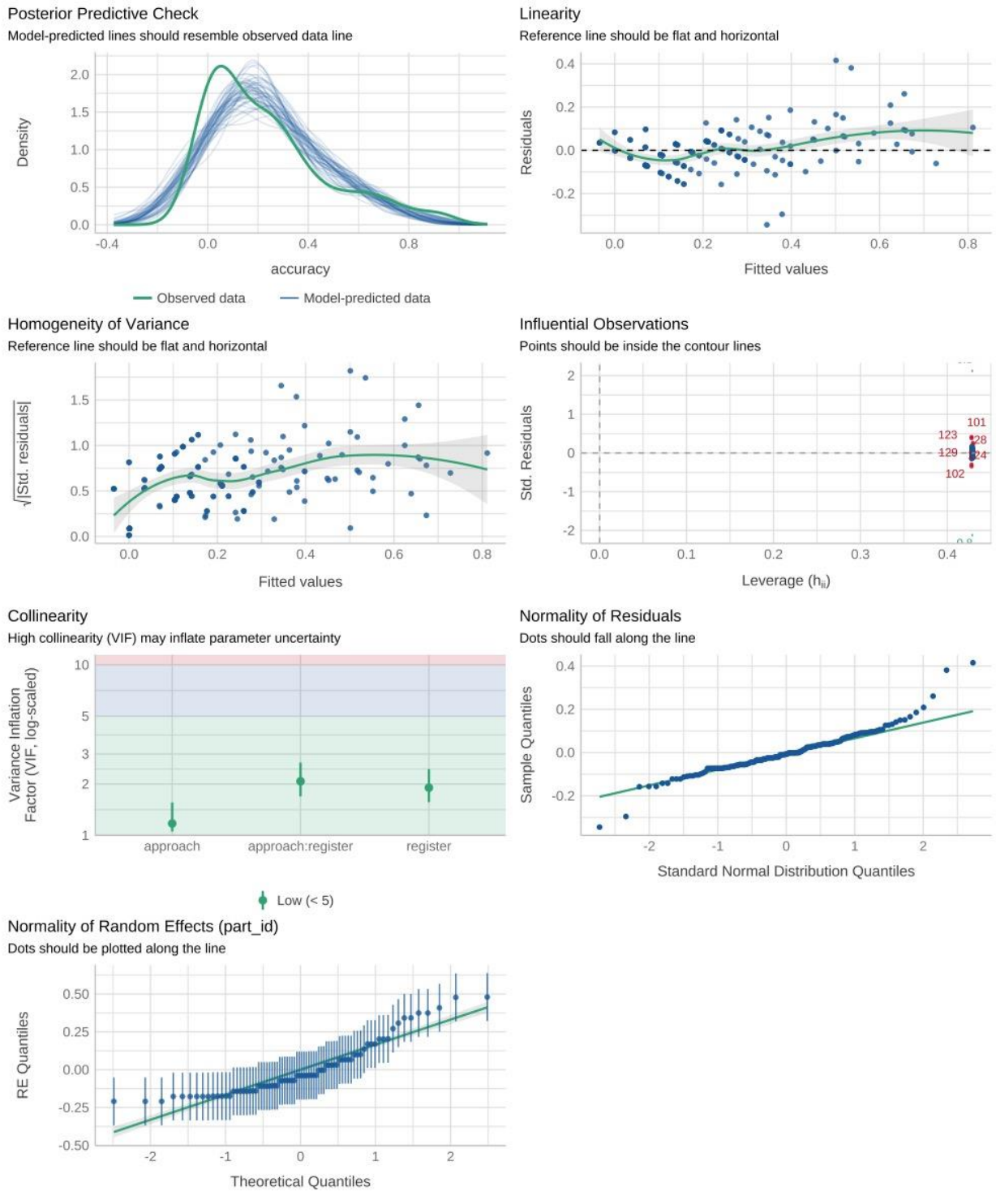


Figure 65

RQ2: Word Meaning Log RT

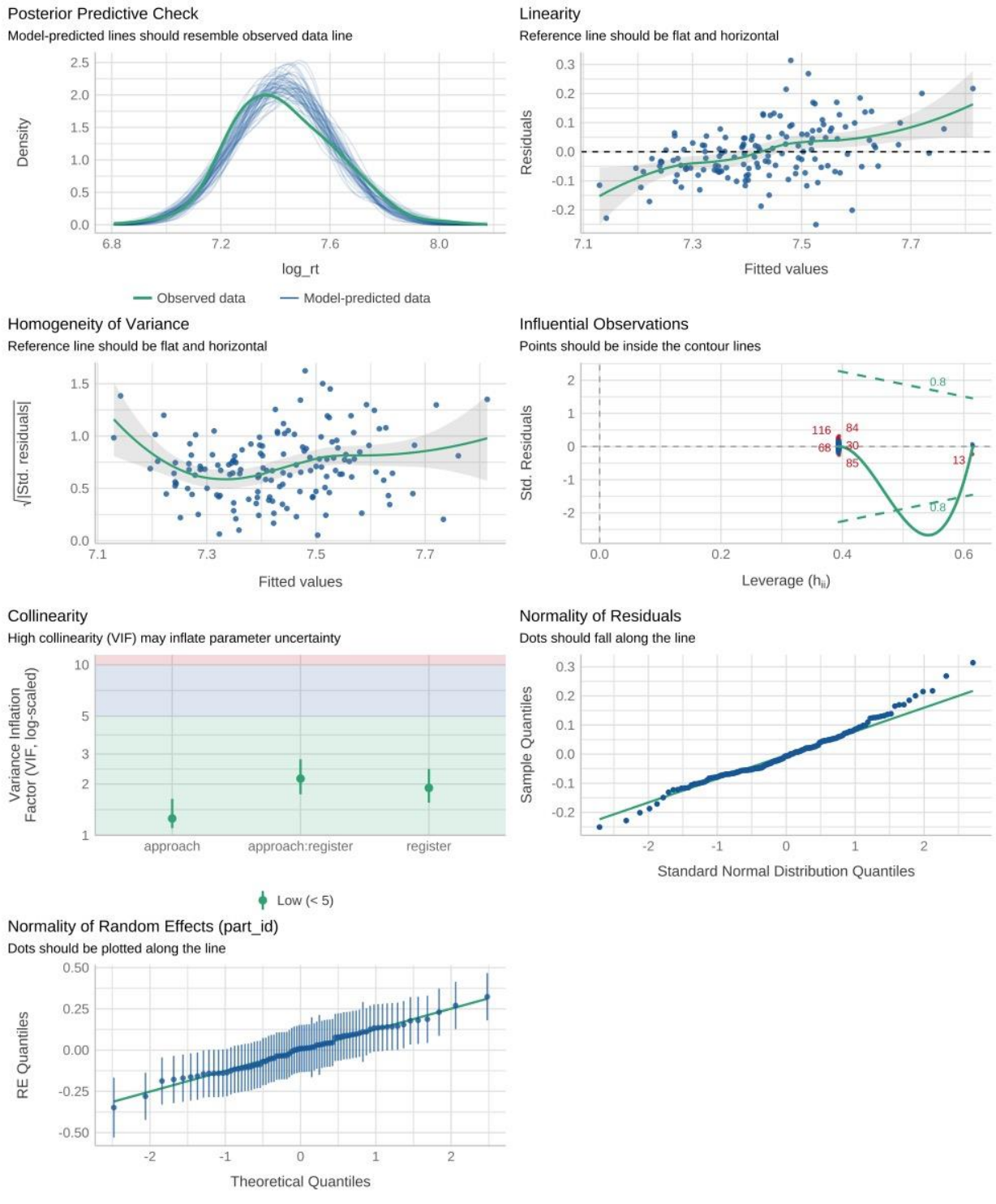


Figure 66

RQ2: Word Meaning Accuracy

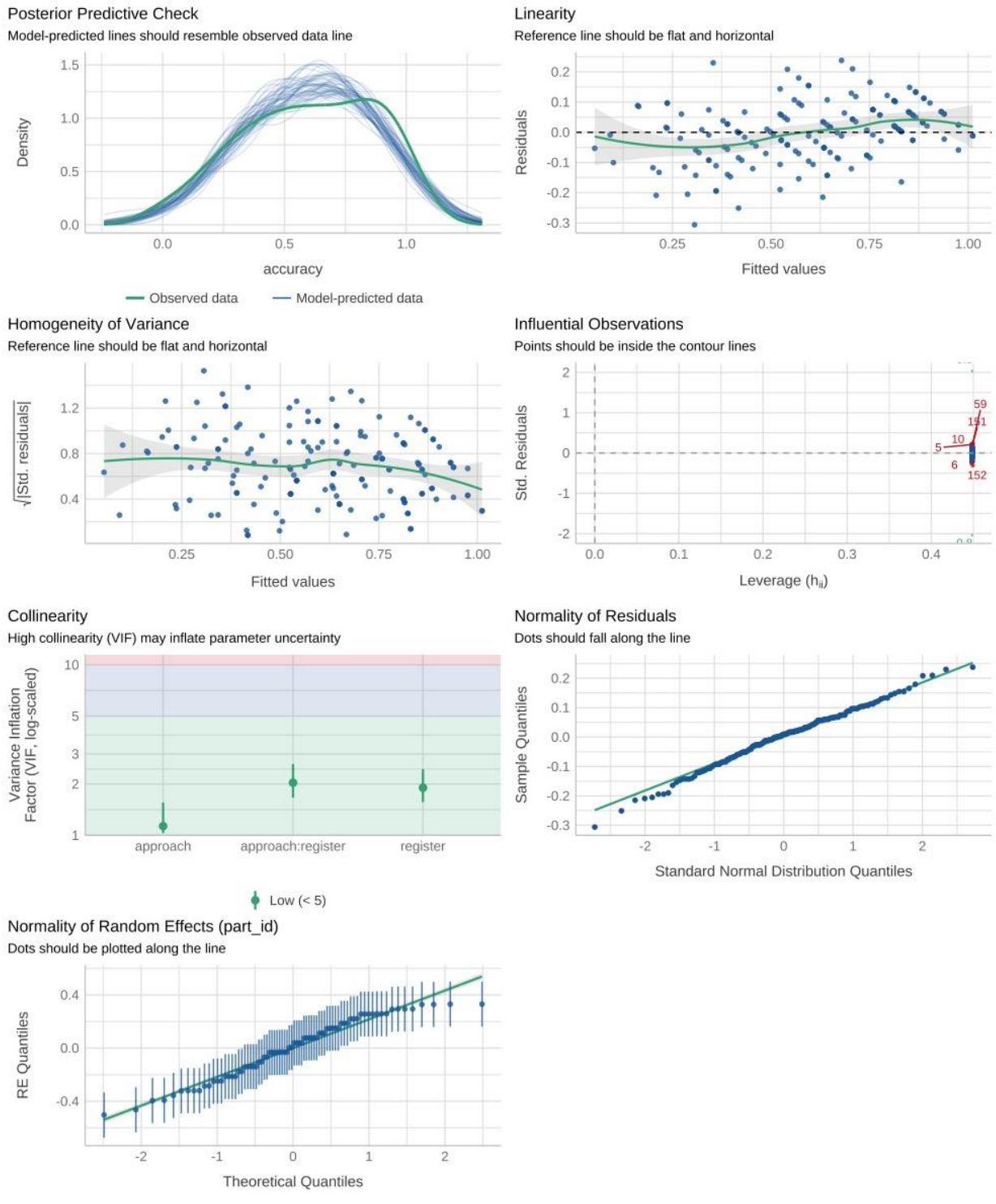


Figure 67

RQ2: Grammar Recall Log RT

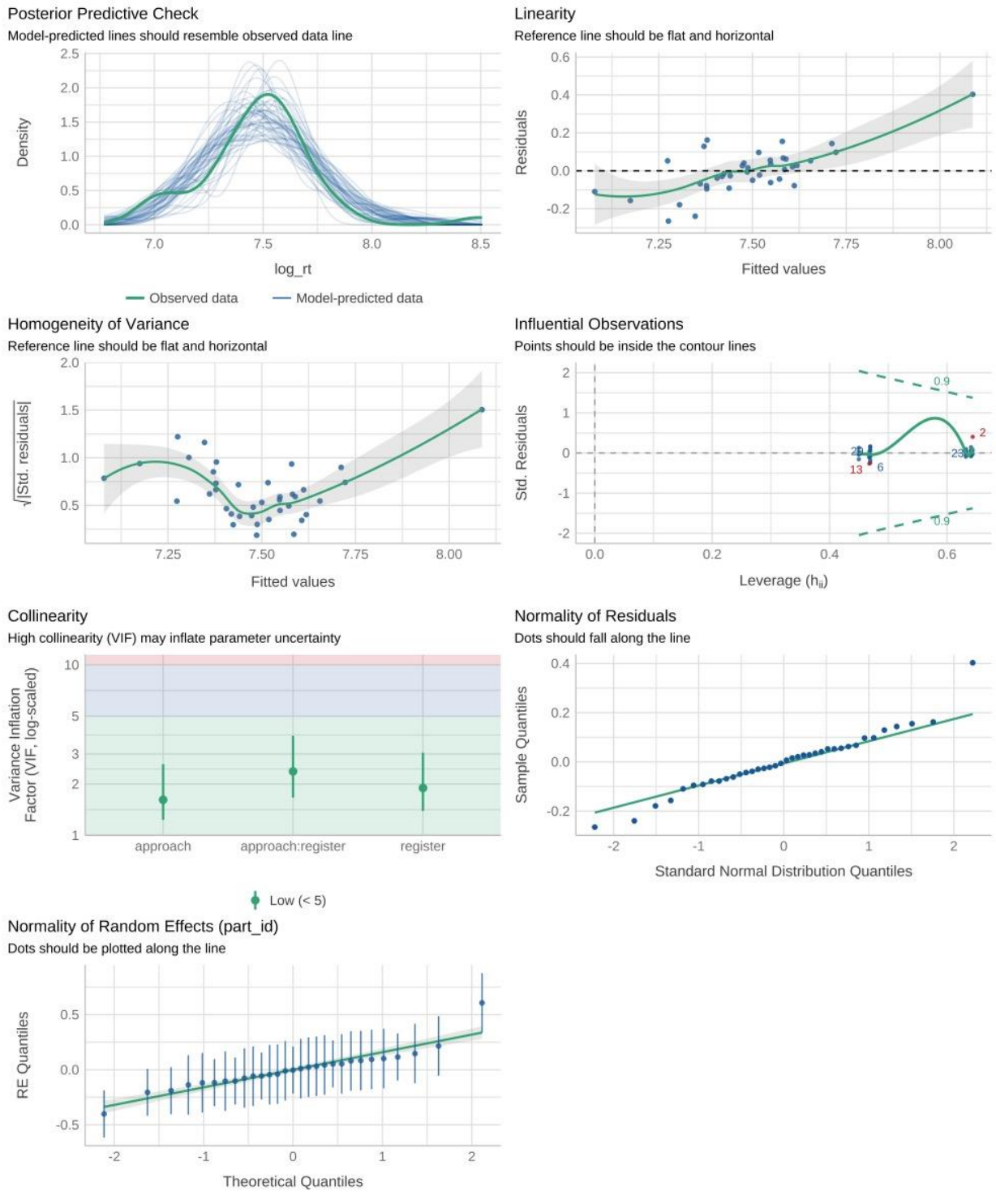


Figure 68

RQ2: Grammar Recall Levenshtein Distance

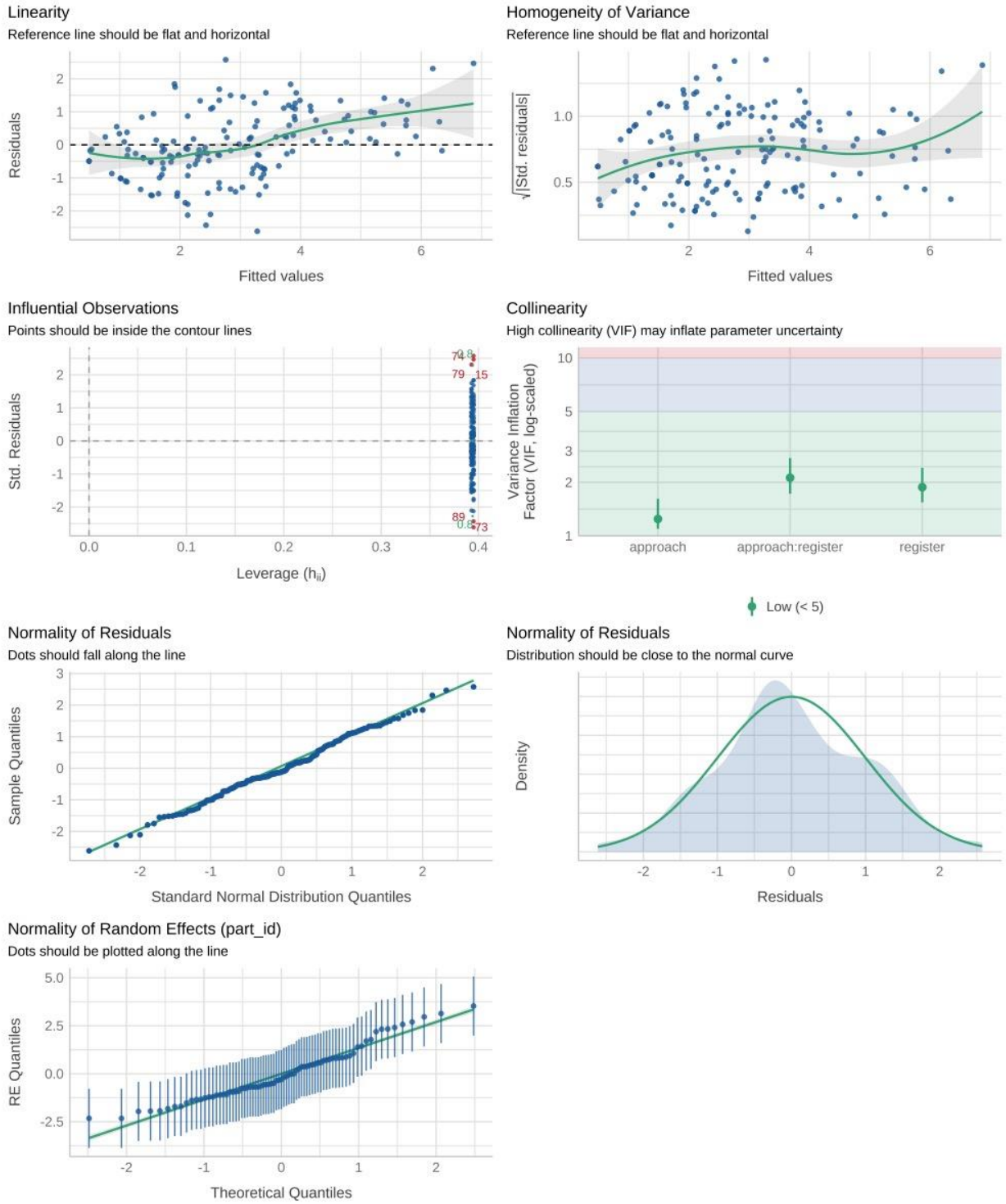


Figure 69

RQ2: Grammar Recall Accuracy

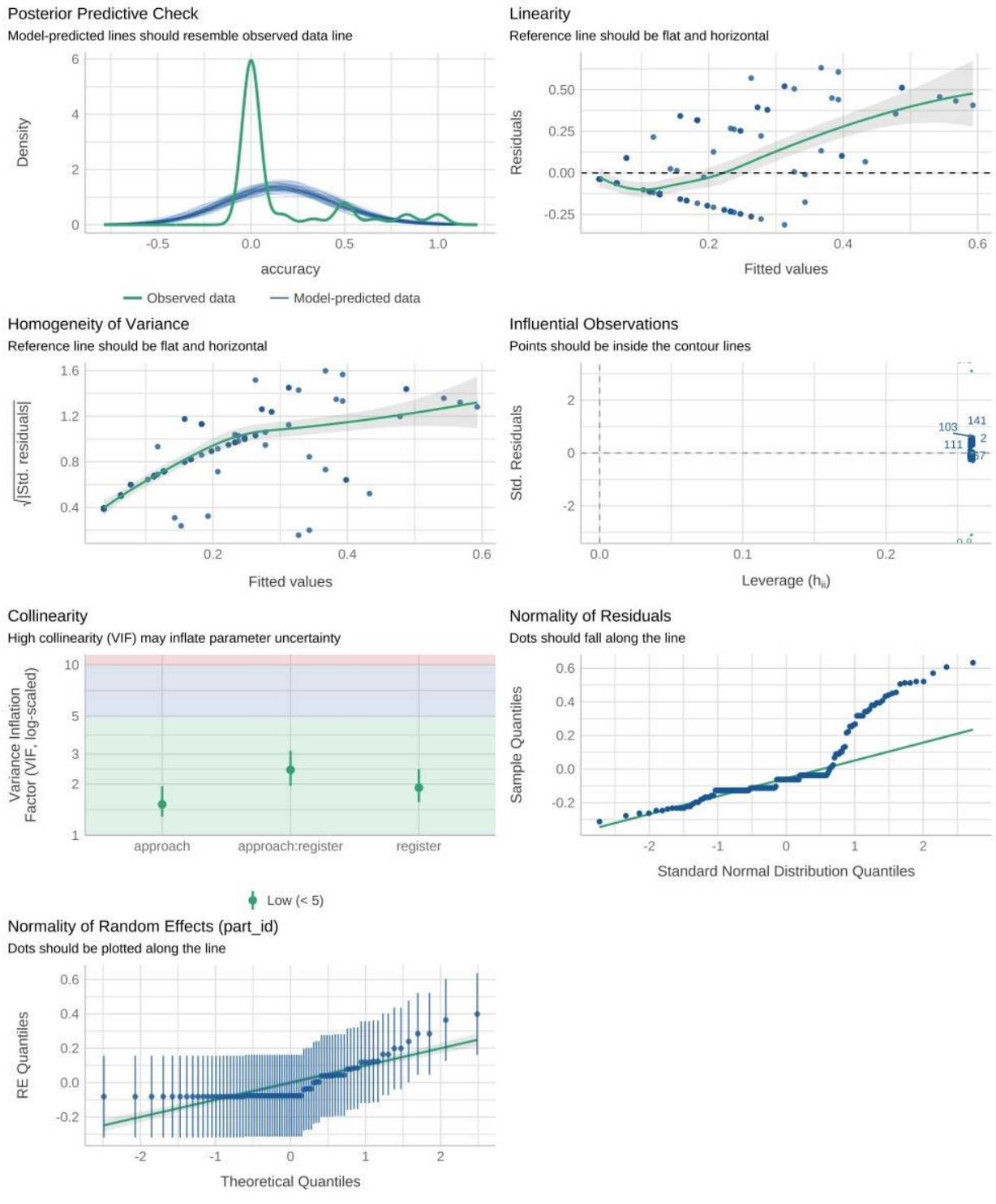


Figure 70

RQ2: Grammar Recognition Log RT

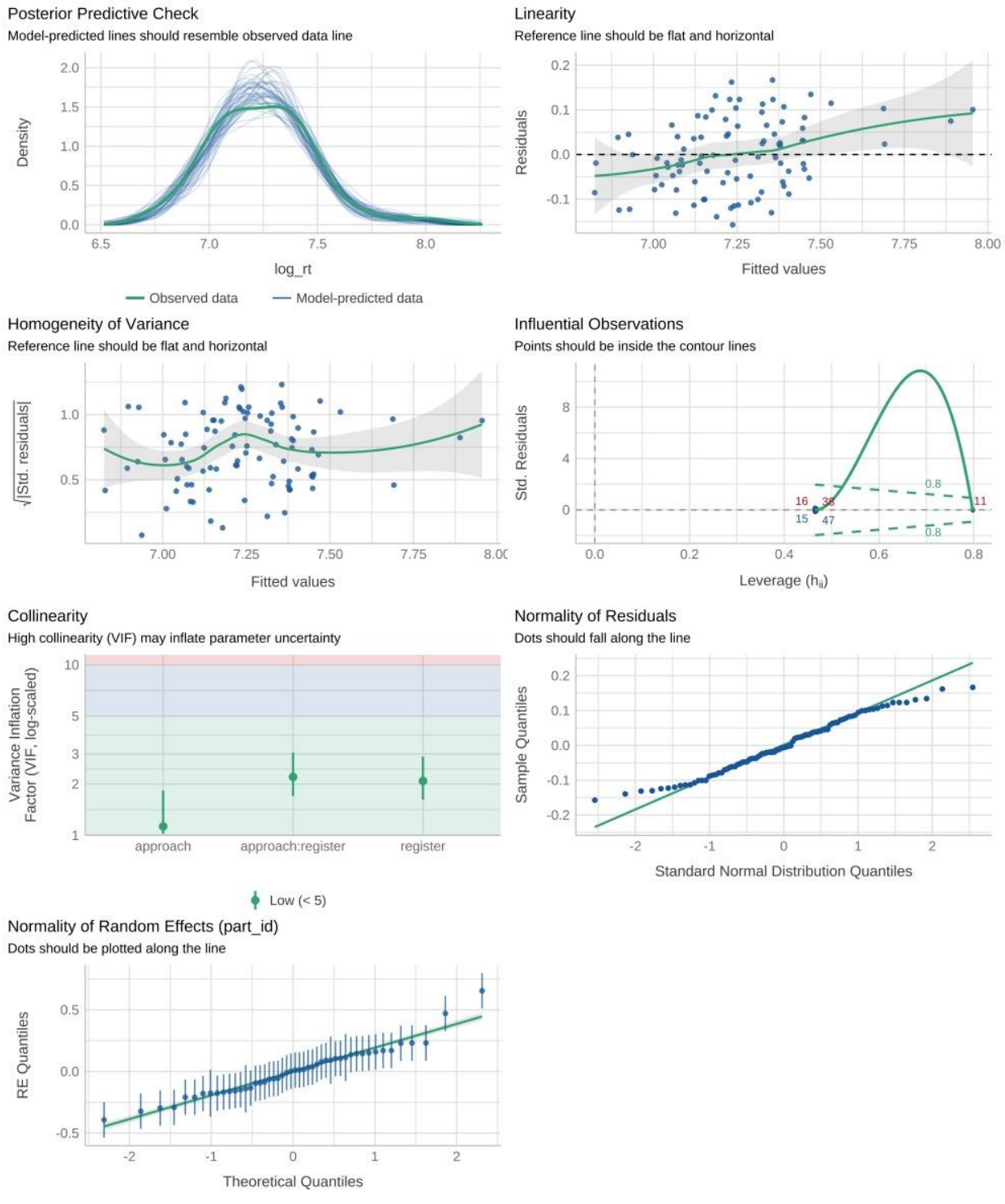


Figure 71

RQ2: Grammar Recognition Accuracy