Studies in Second Language Acquisition (2024), 1–22
doi:10.1017/S0272263124000214

METHODS FORUM

A comparison of Lab- and Web-based elicited imitation: Insights from explicit-implicit L2 grammar knowledge and L2 proficiency

Kathy Minhye Kim[1], Xiaoyi Liu[2], Daniel R. Isbell[3] and Xiaobin Chen[4]

[1]Boston University; [2]Western New England University; [3]University of Hawai'i at Mānoa; [4]University of Tübingen
Corresponding author: Xiaobin Chen; Email: xiaobin.chen@uni-tuebingen.de

(Received 07 July 2023; Revised 01 February 2024; Accepted 14 February 2024)

Abstract

Elicited imitation (EI) tasks are a practical tool for measuring second language (L2) knowledge and skills. In this study, we implemented a web-based EI task that measures English morphosyntactic knowledge and compared its measurement properties to a traditional laboratory-based EI. A cohort of 149 L2 English learners engaged in the web-based EI task, and 151 participants completed a traditional lab-based EI counterpart. Correlation analyses revealed a significant, comparable relationship between English proficiency and the two EI versions, with the ungrammatical items showing less consistency that neither improved nor harmed the overall EI effectiveness. Factor analyses corroborated the validity of web-programmed EI, with both EI versions relating similarly to time-pressured, implicit knowledge and untimed, explicit knowledge measures. Our results suggest the potential for utilizing web-based EI to substitute lab-based tasks, enabling larger-scale, more diverse sampling. We end with implications for future web-based EI task users and include a coding guideline for customized web-based EI use.

Introduction

Elicited imitation (EI) has a long, productive history in second language acquisition (SLA) as a means of assessing morphosyntactic knowledge (Erlam, 2006; Godfroid & Kim, 2021; Suzuki et al., 2023) and second language (L2) global proficiency (Kostromitina & Plonsky, 2022; Yan et al., 2016). The primary objective of EI is to measure one's ability to reconstruct spoken language. The reconstruction process involves the use of working memory, comprehension, and knowledge of syntax, collectively reflecting individuals' developing L2, known as their interlanguage. This interlanguage system is what researchers aim to measure through EI, serving as a proxy for implicit (or automatized explicit) linguistic knowledge in psycholinguistics and global L2 proficiency in the field of language assessment.

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

The goal of this study is to establish the validity of a low-fee, web-programmed EI that measures English morphosyntactic knowledge. To this end, we examined the comparability, in terms of construct validity, of an EI task delivered in a traditional in-person laboratory setting against an equivalent EI task administered on the web. Drawing on theoretical models and empirical work in L2 psycholinguistics and assessment, we formed two hypotheses. From a view of psycholinguistics, L2 grammar knowledge is composed of at least two types of knowledge: explicit (conscious) and implicit (unconscious). While the exact nature of the EI construct has long been debated (see Godfroid & Kim, 2021; Suzuki et al., 2023), the meaning-focused and time-pressured element of EI requires largely automatic processing of sentences, relying on implicit knowledge (R. Ellis, 2005; Godfroid & Kim, 2021) or more automatized version of explicit knowledge (Suzuki & DeKeyser, 2015; Suzuki et al., 2023). If the web-based (WB) EI were to be as valid as the lab-based (LB) version, we expect to obtain a similar factor structure and loadings for EI scores when situated among several other measures of implicit and explicit morphosyntactic knowledge.

Second, L2 morphosyntactic knowledge, which is measured through the EI task in the current study, is a central component of L2 proficiency (Bachman & Palmer, 1996; Canale & Swain, 1980; Hulstijn, 2015); as such, we expected a significant and comparable relationship between both EIs and L2 proficiency. In reference to Hulstijn's language proficiency model (2011, 2015), EI may pertain mostly to basic language cognition (BLC) while standardized proficiency tests, which were used as a measure of L2 proficiency in this paper, may extend beyond BLC to higher language cognition (HLC). The EI test, for instance, typically uses words and structures that frequently appear in communicative situations and is largely free of differences among L2 speakers in cognition factors such as phonological short-term memory capacity (Kim et al., 2016; Park et al., 2020). It is therefore common for native speakers of the target language to achieve near-perfect scores in EI, reflective of their proficiency in these linguistic elements. Standardized proficiency tests, on the other hand, require the use of HLC that contains more complex words and structures that need not be spoken, pertain to topics other than everyday matters (e.g., academic topics), and is subject to individual differences in education, literacy, and other cognitive abilities (Hulstijn, 2011, pp. 230–231). Importantly, because BLC is situated as a core component *within* HLC proficiency and is entirely encompassed by it, we expect a good amount of shared variance between standardized L2 proficiency test scores and EI.

Heeding calls to probe the validity of measures more rigorously in SLA research (Chapelle, 2021; Révész & Brunfaut, 2021; Sudina, 2021), we address this issue in the context of a web-programmed spoken production task. While WB spoken data collection brings many benefits, it does raise questions about its comparability to LB studies. Among these are construct-*irrelevant* factors, such as environmental distractions or suboptimal and/or varying audio playback conditions, which may threaten construct validity. Also, even when LB and WB versions of an instrument tap into the same construct, how similarly the construct is measured might still vary, influencing the interpretation of scores and study outcomes.

Acknowledging that establishing strict psychometric equivalence (or *concordance*; see Zumbo, 2021) across different forms or administration conditions of a test is a difficult endeavor generally requiring carefully planned studies with large samples, our goal in the current study is to preliminarily compare the construct measurement of WB and LB EI scores by using existing data to examine (1) relationships with conceptually related constructs and (2) a relationship with a criterion variable (American Educational Research Association et al., 2014). For trait-type constructs, such as implicit

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

morphosyntactic knowledge, a critical juncture in evaluating the validity of a test is what Chapelle (2021) describes as an *explanation inference*. This inference has to do with whether test scores are reflective of a well-defined target construct. Evidence to support such an inference can come from examining a nomological network of related concepts, among other sources. In our case, there is theoretical support to suggest that EI scores reflect implicit morphosyntactic knowledge that is distinguishable from explicit knowledge. Theory also suggests that (implicit) morphosyntactic knowledge ought to have a positive correlation of at least moderate strength with language proficiency (Hulstijn, 2011). This relationship should hold for subparts of an EI test composed of grammatical and ungrammatical items (see section, Elicited Imitation in SLA Research), although the relationship may be relatively stronger for grammatical items (Kostromitina & Plonsky, 2022; Yan et al., 2016). Importantly, we wish to know whether these hypothesized relationships bear out in a similar fashion across WB and LB EI tests. Positive evidence of this would suggest that the scores from each version similarly indicate standing on the theoretical construct being measured, namely, implicit morphosyntactic knowledge.

As such, to investigate the comparability of the LB and WB EI tasks, we examined evidence (1) regarding relationships with conceptually related constructs and (2) of a relationship with a relevant criterion variable. In particular, we empirically tested the extent to which an EI administered in one's home yielded relationships to measures of implicit (i.e., oral production [OP], timed grammaticality judgement test [TGJT]) and explicit (i.e., untimed grammaticality judgement test [UGJT] and metalinguistic knowledge test [MKT])] grammar knowledge and L2 proficiency (i.e., standardized language test) similar to those of an EI administered in a lab setting.

Elicited imitation in SLA research

Among several means of measuring L2 implicit knowledge, EI has a long, productive history in SLA (R. Ellis, 2005; Erlam, 2006; Godfroid & Kim, 2021; Isbell & Rogers, 2020). As described by Vinther (2002), R. Ellis (2005), Erlam (2006), and others, EI tasks are characterized by the following elements:

- Stimuli in the form of spoken sentences, designed to vary in overall length or complexity, or to target specific (morpho)syntactic features
- A pause between stimuli playback and subject response
- Scoring based on the accuracy of the repeated stimuli

The EI task format has been adapted to measure global L2 oral proficiency (e.g., Ortega et al., 2002) and, in line with the focus of the current study, L2 (morpho)syntactic knowledge. Because of its fleeting audio stimuli and time constraints on production that require largely automatic processing of language with minimal opportunity to apply metalinguistic knowledge, EI tasks meet R. Ellis' (2005) criteria for operationalizing implicit knowledge.

A common design feature of EI tasks intended to measure morphosyntactic knowledge more narrowly is the introduction of ungrammatical stimuli, which are expected to be corrected in a test taker's repetition as (further) evidence of automatic—and even subconscious—application of target L2 grammatical representations (Erlam, 2006; Erlam & Akakura, 2015). Markman et al. (1975) provided early insights by emphasizing the difference between simple memory-based replication and the application of an "internalized grammar factor" (p. 37). They claim that learners often adjust linguistic

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

structures in ways that align with their inherent understanding of a language's grammar. Corrected imitation of ungrammatical sentences, then, becomes evidence of its assimilation into their internal grammar or interlanguage system, suggesting that learners are actively processing and understanding the content, rather than merely replicating it. Therefore, when a learner, armed with an accurate internal grammar system of the target language, confronts these items, their capacity to recognize and rectify them becomes a potent indicator of their deep-seated understanding of the language.

However, delving further into the literature reveals skepticism of the inconsistency in error correction, especially in the absence of explicit correction instructions, potentially undermining the effectiveness of ungrammatical sentences in EI tasks. In response to these concerns, Yan et al. (2016) conducted a meta-analysis of studies that included ungrammatical items ($k = 9$), and found that the grammaticality of EI stimuli had little impact on measurement: Both measures, with and without ungrammatical stimuli, were similarly sensitive measures of morphosyntactic knowledge thought to be implicit in nature. However, there was a trend of ungrammatical items being less discriminating across proficiency levels than grammatical items ($g = 1.16$ vs. $g = 1.34$), but these differences were comparable (i.e., $Q[1] = 2.87$, $p = 0.09$), suggesting that the inclusion of ungrammatical items neither improved nor harmed the sensitivity of EI. Kostromitina and Plonsky (2022) echoed the concerns about the validity of ungrammatical items and expanded Yan et al.'s meta-analysis by including 4 additional studies (k = 15). They found ungrammatical items to be less discriminating than grammatical ones: mean effect sizes of studies that included and excluded ungrammatical items fall outside their respective confidence intervals (CIs) ($r = .56$, CI [0.46, 0.65] vs. $r = .71$, CI [0.66, 0.76]). The authors concluded that the repetition of ungrammatical sentences did not aid in EI's capacity to assess learner proficiency, particularly when clear correction instructions are not provided. Their research adds another layer to the debate, highlighting the practical challenges in the administration of EI tasks with ungrammatical items.

In the realm of psycholinguistics, a series of factor analytic studies involving multiple measures of L2 morphosyntactic knowledge (R. Ellis, 2005; R. Ellis & Loewen, 2007; Godfroid & Kim, 2021; Kim & Nam, 2017; Spada et al., 2015), including measures designed to target implicit and those designed to target explicit knowledge, has yielded two key findings. First, EI-based measures of morphosyntax correlate more strongly with other measures of morphosyntax thought to tap into implicit knowledge (e.g., timed grammaticality judgments and oral narratives) compared to measures that clearly access explicit knowledge (e.g., metalinguistic knowledge tests). Second, EI measures load primarily on factors interpreted as reflecting implicit knowledge, though some researchers motivated by skill acquisition theory have argued that EI may be tapping into a third construct referred to automatized explicit knowledge (AEK) rather than truly implicit knowledge (Suzuki, 2017; see also Suzuki & DeKeyser, 2015; Suzuki et al., 2023). Even so, the rapid and efficient access associated with AEK is thought to be largely indistinguishable from implicit knowledge when it comes to successful L2 communication.

Although research has generally supported EI as a measure of implicit L2 knowledge, until this point, most EI tasks have been administered in controlled lab settings, either on paper or by computer (see Erlam & Akakura, 2015), in part due to the LB research tradition in the field of SLA and the complexity of setting up custom environments for running EI tests. Technically, such an environment requires the development of a web application that is, for instance, capable of user authentication,

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

audio playback, and recording file storage/transmission between the user interface and the backend server as well as interactive interfaces for test takers and administrators. Recent advances in online experiment platforms and Internet communications technology more generally have ameliorated these difficulties, making precisely timed and faithfully recorded EI tasks possible online.

Web- and lab-based experimentation

WB experimentation was first introduced in the field of psychology on auditory perception (Welch & Krantz, 1996) and has been used with increasing frequency across disciplines (Birnbaum, 2000). The transition from traditional lab settings to online experimentation has offered researchers practical advantages (Table 1). For one, lifting restrictions on time and location allows data to be drawn from larger and more diverse sample pools, which may circumvent restrictions of traditional experiments, such as low statistical power and convenience sampling. WB experimentation is also more cost-efficient. Without the need for dedicated lab space and supervising and hiring staff, there is a significant reduction in cost.

The flexibility and efficiency of online data collection, however, have raised questions about the validity of web-solicited data. Some claim that web-collected data are inherently different from LB data (Bauer et al., 2012), because a WB format may subvert the purpose of an experiment. It could also limit research approaches due to the differences in processing and response speed (Parsons et al., 2018) and environmental distractors that may add noise to the data (Gagné & Franzen, 2023). Despite some differences worthy of concern, accumulating findings suggest WB tasks are reliable alternatives to LB instruments. Comparable performances have been reported between lab- and web-delivered cognitive measures, such as working memory capacity and

Table 1. Benefits and costs of web experimentation

Benefits
• Fewer logistical concerns (no physical presence needed) • Wider and larger participant access • Fewer motivational confounds (no extra-credit incentives) • Reduction in research, equipment, and employment costs • Reduction of experimenter effects (“clues” that influence participants’ choices) • Ease of replicability (standardized and optimized testing procedures) • Increased ecological validity (experimental setting closer to a natural setting)
Costs and possible solutions

• Higher dropout rates
  ○ Offer rewards, such as immediate feedback or monetary incentives
• Limited control over experimental access, such as multiple submissions
  ○ Restrict IP addresses
  ○ Collect personal identification
• Limited interaction with staff
  ○ Add contact information
  ○ Provide test tutorials in advance
• Higher risk of distractions
  ○ Include progress markers
  ○ Include screening tasks
• Noisy data
  ○ Increase sample size

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

declarative memory (Ruiz et al., 2019), cognitive functioning (Backx et al., 2020; Houben & Wiers, 2008), and psycholinguistics measures, such as lexical decision task (Hilbig, 2016).

In response to technological advances coupled with COVID-19 and the replication crisis (Klein et al., 2014; McManus, 2021), online experimentation has only recently bloomed in L2 research. From recruitment to implementation, various topics of research have been carried out online. In L2 speech perception, for instance, researchers investigated the effectiveness of high variability phonetic training in which both the perception and production data were collected online (Saito et al., 2022). In some studies, researchers crowdsourced listeners through Amazon Mechanical Turk (AMT) and sampled their transcription and speech ratings remotely (Loukina et al., 2015; Nagle, 2019; Nagle & Huensch, 2020). In psycholinguistics, topics such as bilinguals' agreement processing (Lee & Phillips, 2022), pragmatic inferencing by autistic learners (Van Teil & Kissine, 2018), and implicit L2 learning (Kerz et al., 2017), were investigated on the web, enabling the collection of both accuracy and reaction time data. For implementation, many studies have used a range of commercial services, such as PsychoPy, Prolific, Gorilla, LimeSurvey, and Unity. Alternatively, some programmed tasks created customized systems for web delivery by utilizing technologies such as HTML, CSS, JavaScript, SQL, and server-side programming. For crowdsourcing purposes, AMT, Prolific, and Appen (known as CrowdFlower until 2019), served to promote remote recruitment.

Despite the increased transition to WB research, few studies have directly explored the validity of web-solicited data. Hilbig (2016) investigated the reaction time effects of the lexical decision task by software (E-Prime vs. JavaScript) and environment (web vs. lab). To dissect such effects, participants were assigned to one of three conditions: 1) the lab using commercial software, E-Prime; 2) the lab using the JavaScript version; or 3) the web using the same JavaScript version. All conditions yielded the desired word frequency effects, demonstrating the comparability of effects across different conditions and environments. Likewise, results from Ruiz and colleagues (2019) suggested that the web version of a paired associates subtest of the Modern Language Aptitude Test yielded a strong correlation to its LB counterpart (r = .82), affirming its comparability to the web-version. Lastly, in SLA, Nagle and Rehman (2021) aimed to develop a reliable online method for recruiting listeners to rate L2 speech samples for previous research (Nagle, 2019). In their approach, they included both traditional local labs and the digital domain of AMT. Within AMT, participants were either exposed to the specific target dialect or a broader range of dialects. With quality controls included (i.e., prescreening and task timers), the study confirmed that three groups of raters (local lab and two AMT groups) had similar speech ratings, suggesting the high comparability between WB and LB participant recruitment with proper quality validation. Collectively, these results bring prospects for administering language research, including reaction time– and accuracy-based, cognitive, and speech-rating tasks on the web.

To our knowledge, no studies in SLA have explored the validity of web-delivered speech production tasks on the web. With the increased possibility of collecting spoken responses remotely using custom web platforms and services, empirical validation is needed to ensure accurate and reliable use of the web version of EI.

Current study

The current study aims to explore the validity and usefulness of a web-delivered EI task designed to measure English morphosyntactic knowledge. To this end, we examined the comparability, in terms of construct validity, of an EI task delivered in a traditional

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

in-person lab setting against an equivalent EI task administered on the web. The following Research Questions (RQs) guided this study:

RQ1: To what extent do the EI scores administered in the lab and on the web relate to an English proficiency test?

RQ1a: Do grammatical and ungrammatical EI items administered in the lab and on the web exhibit similar relationships to an English proficiency test?

RQ2: To what extent do EI tasks administered in the lab and on the web relate to measures of implicit and explicit English morphosyntactic knowledge?

With regard to RQ1 and RQ1a, we anticipate EI-based measures of morphosyntactic knowledge to exhibit comparable correlations with English proficiency scores across modalities, for both grammatical and ungrammatical items. With regard to RQ2, we expect to obtain a similar factor structure and loadings for EI scores when situated among several other measures of implicit and explicit morphosyntactic knowledge.

Methods

This study drew on data from Kim's dissertation (n = 149; Kim, 2020; Kim & Godfroid, 2023) and Godfroid et al.'s project (n = 151; Godfroid & Kim, 2021; Godfroid et al., 2018). To perform the comparisons of interest, we used datasets of a standardized English proficiency measure and five linguistic measures: EI, OP, UGJT, TGJT, and MKT.

The same instruments of identical items were used in both Kim (2020) and Godfroid et al. (2018), with the exception of the delivery mode of EI and OP. In Kim (2020), the two oral tasks (i.e., EI and OP) were delivered on the web, and other measures (i.e., UGJT, TGJT, and MKT) in the lab, while all data in Godfroid et al. (2018) were collected from measures administered in a lab setting (see section, Web-based Environment, for details).

Participants

In both Kim's WB study and Godfroid et al.'s LB study, the participants were English L2 speakers attending a large university in the American Midwest. Kim's study recruited 149 participants, and Godfroid et al. initially recruited 160. Nine participants in Godfroid et al., who missed all five linguistic measures, were excluded from the analysis, resulting in a final sample of 151. For speaker L1, defined here as self-reported dominant language, the most common was reported as Chinese in both Kim (n = 53) and Godfroid et al. (n = 74), followed by any variety of Spanish (Kim, n = 14; Godfroid et al., n = 9), and Korean (Kim, n = 17; Godfroid et al., n = 10). The remaining 65 (Kim) and 58 (Godfroid et al.), reported 24 (Kim) and 30 (Godfroid et al.) other L1s. Table 2 outlines the demographic information of the participants in each project.

Target structures

Six grammatical features (Table 3) initially developed by Godfroid et al. (2018) were included in both WB and LB research as target structures across five linguistic tasks: (1) third-person singular -s, (2) mass/count nouns, (3) comparatives, (4) embedded questions, (5) passive, and (6) verb complements. These structures were chosen due to their emergence at various stages of SLA, making them suitable for assessing a broad spectrum of English morphosyntactic knowledge.

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

Table 2. Background information of the L2 speakers in the lab- and web-based environments

	N	Mean	*SD*	Min	Max	Skew	Kurtosis
	Lab–based (n = 151): Godfroid, Kim, Isbell, & Hui (2018) Dataset						
Residence Length: Months	145	44.33	32.34	2	240	2.72	11.71
Age of Arrival	145	19.96	6.31	0	38	–0.28	2.27
Age of Onset	145	8.09	3.75	0	25	0.92	2.21
	Web–based (n = 149): Kim (2020) Dataset						
Residence Length: Months	137[a]	34.78	33.11	1	216	2.11	6.89
Age of Arrival	137[a]	22.92	6.05	3	40	0.12	0.46
Age of Onset	138[a]	8.16	3.88	2	30	1.87	7.27

[a]Discrepancies in the number of participants are due to missing background data.

Table 3. Six target structures

Structure	Example
Third Person Singular –S	* *The old woman enjoy watching many different famous movies.*
Mass/Count Noun	* *The boy had rices in his lunch box.*
Comparative Adjective	* *It is more harder to learn Korean than to learn English.*
Embedded Question	* *He wanted to know why had he studied for the exam.*
Passive	* *The flowers were pick last winter for the festival.*
Verb Complement	* *Jim is told his parents want buying a new house.*

Instruments

Lab-based elicited imitation

The experimental stimuli of the LB EI developed by Godfroid et al. (2018) consisted of 24 critical items (half grammatical and half ungrammatical) and eight filler items. The 24 critical items were counterbalanced by grammaticality yielding two counterbalanced lists (form A and form B). The sentences ranged from 6 to 13 words. Prior to the test phase, four practice sentences (two grammatical and two ungrammatical) were provided. Participants were instructed to listen to each sentence, judge its plausibility, and repeat the sentence in "correct English." At the beginning of each trial, a fixation cross (+) appeared for 500 milliseconds. The sentence was then played with a speaker icon appearing on the screen. After each sentence, a plausibility question followed (e.g., Do you agree, disagree, or are you unsure about the content of the statement?). Participants had four seconds to decide. A microphone icon then appeared on the screen with a beep sound and the text, "Please repeat now." They had eight seconds to repeat the sentence. The responses were recorded by a research assistant with a recorder. The task was programmed on Superlab 5.0. An overall accuracy score was calculated based on the correct usage of the target forms in obligatory contexts.

Web-based elicited imitation

In the WB EI, two changes were made from the LB EI. First, to enhance test item reliability, only one form of the counterbalanced versions of LB EI (form B) was employed. Second, we added four practice sentences (resulting in eight sentences, four grammatical and four ungrammatical) and model responses to two practice sentences.

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

These changes were made to compensate for the lack of immediate assistance from researchers to address queries. Aside from these adjustments, all other conditions remained identical to the LB EI.

The task and the trial sequences of the WB EI were identical to the LB EI. During the repetition trial, an additional text, "Your voice is now being recorded." was included under the microphone icon. We also designed a progress marker (e.g., 1/32) to inform participants of their progress throughout the experiment. At no point did the task instructions or the model responses include explicit guidance to focus on forms or provide metalinguistic feedback on the linguistic features. The task was programmed as a custom-made Java web application, which consisted of a backend server with Java Servlets doing user authentication and data storage in an SQL database as well as a web front-end accessible from the participants' browsers. As in the LB EI, correct usage of the target forms in obligatory contexts was used as an outcome measure.

Oral production

Both Kim's and Godfroid et al.'s OP tasks instructed participants to read and retell a picture-cued short story containing the target structures (Godfroid et al., 2018; Godfroid & Kim, 2021). The pictures provided illustrations of the main context and assisted in the later story retrieval. Each picture was followed by one to four sentences. Participants were asked to read the story twice without time constraints, with no note-taking allowed. Subsequently, they were asked to retell the story within 2.5 minutes, in as much detail as possible, while the pictures remained on the screen. The overall accuracy score was calculated by dividing the number of times a target structure was needed by the number of its correctly used instances. The WB OP tasks were programmed using the same technical setup as the EI tasks. In the current paper, only the WB EI was validated, given its wider use and greater technical sophistication compared to OP (i.e., prompt presentation and timing of responses).

Timed and untimed grammaticality judgment tests

The LB computerized written grammaticality judgment tests (GJTs) instructed participants to read a sentence and determine its grammaticality under timed or untimed conditions. The TGJT consisted of 24 critical items (half grammatical and half ungrammatical) and 16 filler items (Godfroid et al., 2018; Godfroid & Kim, 2021). In the UGJT, we only included 12 ungrammatical items, following R. Ellis (2005), and 16 fillers. The TGJT asked participants to make a decision under a time limit, which was set relative to the sentence lengths. The time limit was determined by computing the average audio length of sentences of the same length and adding 50% of its median (Godfroid et al., 2018; Godfroid & Kim, 2021). The UGJT was equivalent to the TGJT, but a time constraint was not imposed. In this way, participants were able to employ explicit knowledge to help with their decisions. Correct responses in both tests were awarded one point.

Metalinguistic knowledge tests

The MKT asked participants to read 12 sentences, each containing one grammatical error. Participants were then asked to identify, correct, and provide detailed explanations for each error. Correct responses to each sentence earned one point, allowing a maximum score of 12 points on the MKT. Because explanations required participants to use their most explicit linguistic knowledge, we only considered the explanation part for analysis.

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

English proficiency measure

To obtain participants' general English proficiency, individuals self-reported their most recent Test of English as a Foreign Language (TOEFL) scores through a background questionnaire. In the interest of inclusivity, the ordinal studies (Godfroid et al., 2018; Kim, 2020) also accepted scores of other standardized tests. In Godfroid et al., the International English Language Testing System (IELTS) was included, and in Kim, IELTS and both the Common European Framework of Reference for Languages and the Test of English for International Communication were accepted. In all cases, the scores were converted into equivalent TOEFL scores using the conversion reference provided by the Educational Testing Service (https://www.ets.org/toefl/institutions/scores/compare/).

Procedure

Lab-based environment

Participants enrolled in the LB research (Godfroid et al., 2018; Godfroid & Kim, 2021) completed the tasks one-on-one with a research assistant in a language research lab. Participants started with two tasks (i.e., OP and EI) that draw attention to the meaning of sentences and proceeded to measures that direct attention to forms (i.e., GJTs and MKT). This order was meant to minimize recognition of the target forms in the meaning-based, implicit knowledge measures.

Web-based environment

As previously mentioned, participants in the WB experiment (Kim, 2020; Kim & Godfroid, 2023) performed two oral tasks (i.e., OP and EI) on the web and the GJTs and MKT in the lab setting (Table 4). The decision to conduct the OP and EI tasks online was made to enhance research efficiency, especially by reducing the extensive labor involving one-on-one sessions with a research assistant. The GJTs and MKT were purposefully conducted in the lab to prevent learners from accessing external resources, such as metalinguistic explanations of grammar on websites or in grammar textbooks. Ten days prior to the lab visits, participants received an e-mail with step-by-step directions on accessing and completing the web tasks. This instructional tutorial was recorded using the New Screen Recording function in QuickTime Player (version 10.5). Seven days prior to the lab visits, participants received another e-mail containing a unique web link with a personalized code directed to an interface for the web tasks. On the interface, participants were given general instructions (e.g., find a quiet room for 30 minutes and use Chrome or Firefox). Following the completion of the consent form, background questionnaire, and OP, participants proceeded with the EI. The TGJT,

Table 4. Sequence and settings of web- and lab-based tasks

Web–based: Kim (2020) Dataset		Lab–based: Godfroid, Kim, Isbell, & Hui (2018) Dataset	
Measures	Setting	Measures	Setting
Background questionnaire		Background questionnaire	
Oral production	Web	Oral production	
Elicited imitation		Elicited imitation	Lab
TGJT and UGJT	Lab	TGJT and UGJT	
Metalinguistic knowledge test		Metalinguistic knowledge test	

Note: TFJT, timed grammatical judgment test; UGJT, untimed grammatical judgment test.

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

UGJT, and MKT tasks were conducted in the lab after participants had completed the web tasks in the convenience of their homes.

Analysis

To investigate RQ1 and RQ1a (i.e., the relationship between EIs and standardized English proficiency scores), Pearson correlation coefficients were conducted between the TOEFL scores and two versions of the EI—the WB and LB EIs—as well as their item types: grammatical items (EI_G) and ungrammatical items (EI_UG). We then examined the corresponding 95% CIs and performed a Fisher's *r*-to-*z* transformation (Preacher, 2002) to statistically confirm the relationships (Cumming & Finch, 2005).

RQ2 addressed the extent to which both WB EI and LB EI are associated with other measures of implicit and explicit knowledge. To this end, we configured two separate confirmatory factor analysis (CFA) models with Kim's WB dataset (2020) and Godfroid et al.'s LB dataset (2018), respectively. Two latent constructs, implicit and explicit knowledge, were configured with the implicit knowledge factor indicated by EI, OP, and TGJT and the explicit knowledge factor indicated by UGJT and MKT (R. Ellis, 2005; R. Ellis & Loewen, 2007).

We evaluated the CFA models based on the global goodness of fit, the indices of which provide a (global) summary of the acceptability of the model; that is, whether the model has been properly specified. Model fit indices considered were the χ^2 statistic (and corresponding degree of freedom and *p* value); root mean square error of approximation (RMSEA), corrected for model complexity, taking sample size into account, and its 90% CI; standardized root mean square residual (SRMR); and comparative fit index (CFI), which compares the fitted model to a base model with no parameter restrictions. We followed Hu and Bentler's (1999) and Kline's (2016) guidelines for fit interpretation (i.e., RMSEA lower-bound CI ≤ 0.06, which yields a nonsignificant *p* value, SRMR ≤ 0.08, and CFI ≥. 95). Full-information maximum likelihood estimation was used to evaluate CFA models, and robust maximum likelihood estimator method was used to accommodate multivariate normality assumption violations. All CFA analyses were carried out in *R* version 4.2.0, using the lavaan package.

Results

We first present the descriptive statistics of the TOEFL scores and the linguistic measures in both WB and LB conditions (Table 5). The average TOEFL score for the participants was 93.03 (*SD* = 13.11) in the WB condition and 96.77 (*SD* = 8.86) in the LB condition. The reliability scores of the linguistic tasks ranged from. 51 to. 96, and the skewness and kurtosis values were within the acceptable ranges (±2), indicating the normality of most of the tasks.

RQ1 and RQ1a: Relationship with English proficiency

In addressing RQ1, we carried out correlation analyses between the TOEFL scores and two versions of the EI tasks separately for the grammatical (EI_G), ungrammatical (EI_UG), and combined EI items.

Before stratifying by grammaticality, we first observed the overall EI scores that combined both grammatical and ungrammatical items. Moderate and significant

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

Table 5. Descriptive results of the lab- and web-based tasks

	N	Mean	*SD*	Min	Max	Skewness	Kurtosis	k	Reliability[a]
	Lab–based: Godfroid, Kim, Isbell, & Hui (2018)								
TOEFL	117	96.77	8.86	77.00	119.00	0.43	–0.14	—	—
EI	118	0.72	0.16	0.38	1.00	–0.02	–0.83	21[c]	Form A =. 70; Form B =. 82
OP	132	0.89	0.10	0.48	1.00	–1.25	1.65	250–word story	.96[b]
TGJT	125	0.61	0.14	0.13	0.96	–0.19	0.25	24	.51
UGJT	125	0.62	0.24	0.00	1.00	–0.48	–0.54	12	.65
MKT	141	0.51	0.22	0.00	1.00	–0.38	–0.29	12	.70
	Web–based: Kim (2020)								
TOEFL	149	93.03	13.11	60	120	–0.28	–0.55	—	—
EI	139	0.65	0.15	0.21	0.96	–0.33	–0.04	23[c]	Form B =. 67
OP	123	0.89	0.13	0.33	1.00	–1.93	4.74	250–word story	.93[b]
TGJT	141	0.58	0.15	0.21	1.00	0.17	–0.36	21[c]	.68
UGJT	133	0.63	0.21	0.08	1.00	–0.46	–0.56	12	.68
MKT	141	0.36	0.25	0.00	1.00	0.62	–0.33	12	.79

Note: EI = elicited imitation; Max = maximum; Min = minimum; MKT = metalinguistic knowledge test, only the rule was used; OP = oral production; TOEFL = Test of English As a Foreign Language; TGJT = timed grammatical judgment test; UGJT = untimed grammatical judgment test, only ungrammatical items were used.
[a]Alpha;
[b]Pearson *r* inter-rater reliability;
[c]Several items were removed due to poor discrimination.

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

correlations were found between TOEFL and LB EI (r = .60, 95% CI [0.44, 0.72], n = 83) as well as TOEFL and WB EI (r = .52, 95% CI [0.39, 0.63], n = 139). Figure 1 visually represents these relationships. The correlation between LB -EI and TOEFL (r = .60) falls within the 95% CI of the corresponding WB EI and TOEFL pair (95% CI [0.39, 0.63]). Similarly, the WB EI's relation to the TOEFL score (r = .52) falls within the CI of the LB EI and TOEFL pair (95% CI [0.44, 0.72]). Their comparable associations were supported by the nonsignificant Fisher's test results (p = .393), failing to reject a null hypothesis (H_0 = the strength of EI-TOEFL correlations are comparable between the administrative conditions).

When stratified by grammaticality, different patterns were observed. As shown in Figure 1 (left), the grammatical items in both EI versions yielded similar associations to their respective TOEFL scores, returning nonsignificant Fisher's test results (p = .164). However, we observed differences between the lab and web versions of the ungrammatical EI items (although not reaching statistical significance, p = 0.062), with the web version correlating weaker to TOEFL scores (r = .37, 95% CI [0.22, 0.50]) than those of the lab version (r = .57, 95% CI [0.41, 0.70]). This trend of ungrammatical items showing less discrimination than the grammatical items is pronounced only in the web condition and not in the lab condition (Figure 1, right), suggesting that the administrative environment of the web might introduce variability in how ungrammatical items are processed and performed by participants.

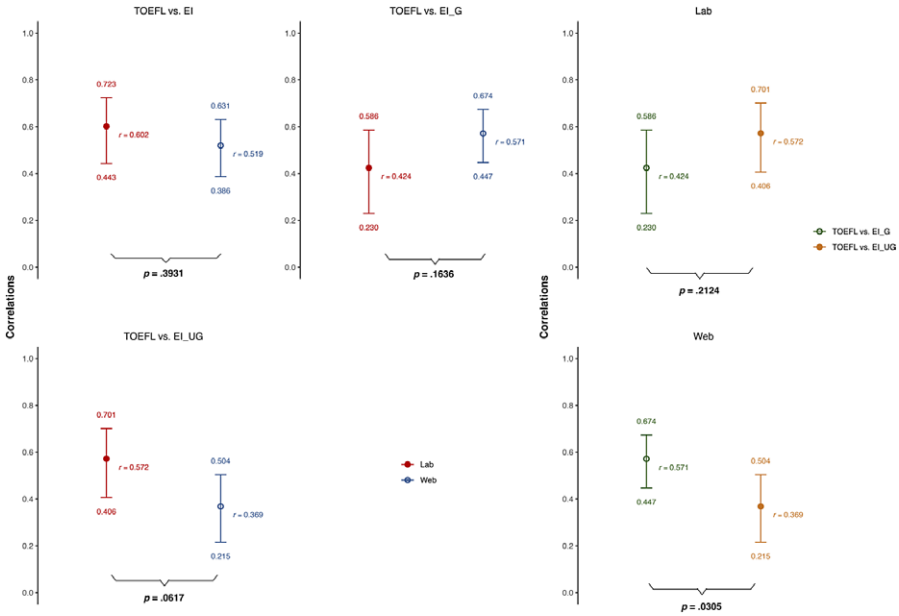

Figure 1. TOEFL and EI correlational values, along with a 95% CI and Fisher's r-to-z transformation, by conditions (left) and by item types (right).
Note: EI = elicited imitation; EI_G = EI with only grammatical items; EI_UG = EI with only ungrammatical items.

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

RQ2: relationship with explicit and implicit measures

To address RQ2 (i.e., the relationship between EI and explicit and implicit L2 knowledge measures), we first explored the associations between EIs and the five linguistic measures. Figure 2 presents the correlation results for all five linguistic measures separately for the WB (top) and LB (bottom) tasks. Small to moderate relationships

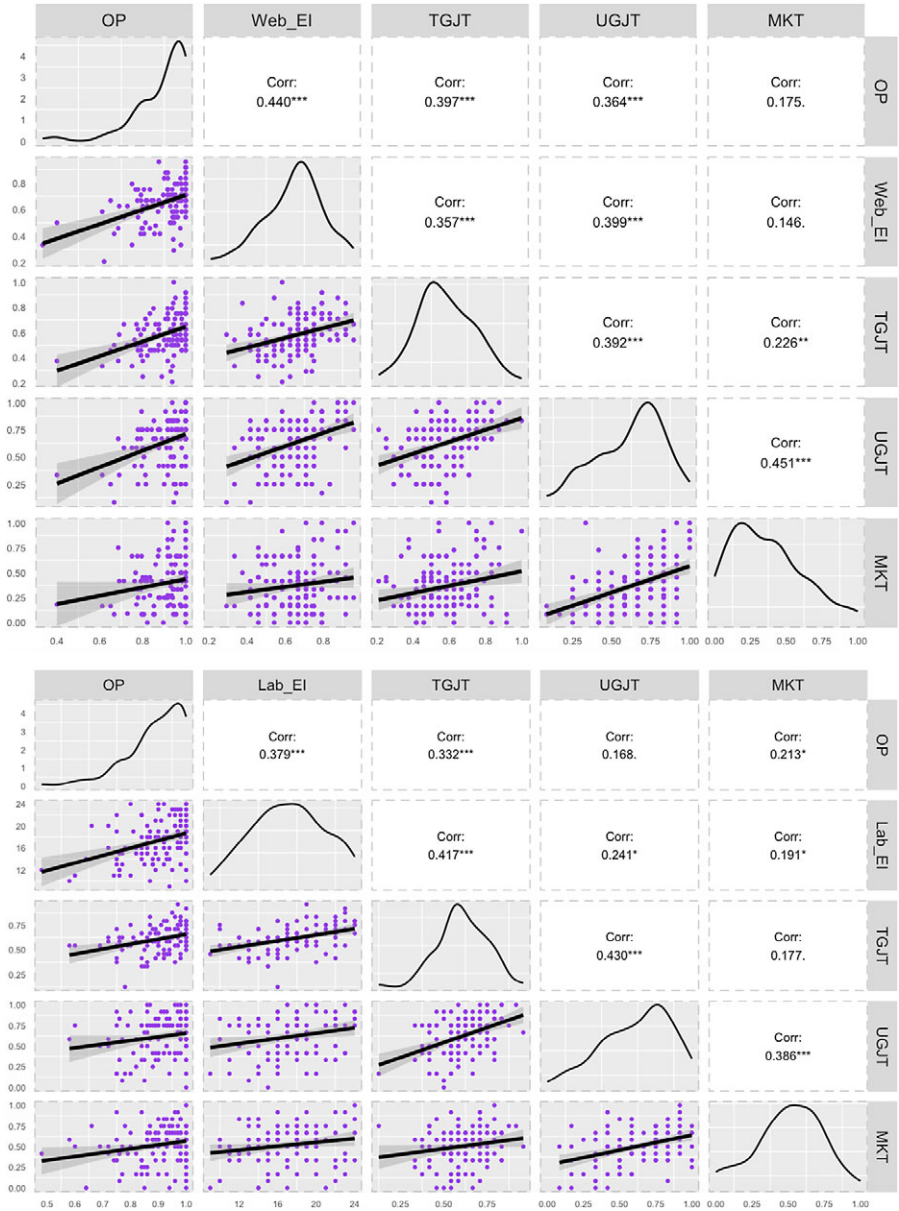

Figure 2. Correlations of linguistic tasks performed in the web (top) and in the lab (bottom).
Note: **p* < 0.05, ***p* < 0.01, ****p* < 0.001. The diagonal cells plot the standardized score distribution of corresponding tests.

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

were found among most measures (.332 < *r* <. 430, $p < 0.001$), except for the correlations between UGJT and OP in the LB versions ($r = .168$) and the MKT and three implicit knowledge measures, which were either insignificant or weakly correlated (.146 < *r* <. 226, $0.01 < p < 0.05$). Importantly, no statistically significant differences in the strength of correlation were found between LB and WB EI measures with different linguistic tests, with Fisher's *r*-to-*z* transformation with a range of $0.19 < p < 0.72$ (see Figure S1, Online Supplementary Materials).

To explore the relationships more fully, we submitted all measures to a series of CFAs separately for the WB (n = 149) and the LB (n = 151) tasks. As previously mentioned, for each task version, we configured a model with EI indicated as an implicit knowledge construct along with two implicit knowledge measures (i.e., OP and written TGJT). Table 6 presents fit indices for the WB and LB conditions. Both CFA models of the lab and web versions demonstrated excellent fit to the data, with all fit indices (i.e., CFI, SRMR, and RMSEA lower bound) within the acceptable range and χ^2 returning nonsignificant values. The reliability (?) of the implicit factor was .69 (WB) and .67 (LB), and for the explicit factor, .66 (WB) and .66 (LB). Parameter estimates for the models are detailed in Figure 3. All indicators in both models were

Table 6. Model fit indices for the web- and lab-based conditions

	Web-based	Lab-based
Parameters (n)	16	16
χ^2	1.711	5.156
χ^2 *p* (> 0.05)	0.789	0.272
df	4	4
CFI (≥. 95)	1.000	0.986
SRMR (version 0.08)	0.015	0.033
RMSEA	0.000	0.045
RMSEA lower (≤ 0.06)	0.000	0.000
RMSEA upper	0.079	0.163

Note: χ^2 = chi-square; χ^2 *p* = chi-square test *p* value; CFI = comparative fit index; *df* = degree of freedom; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

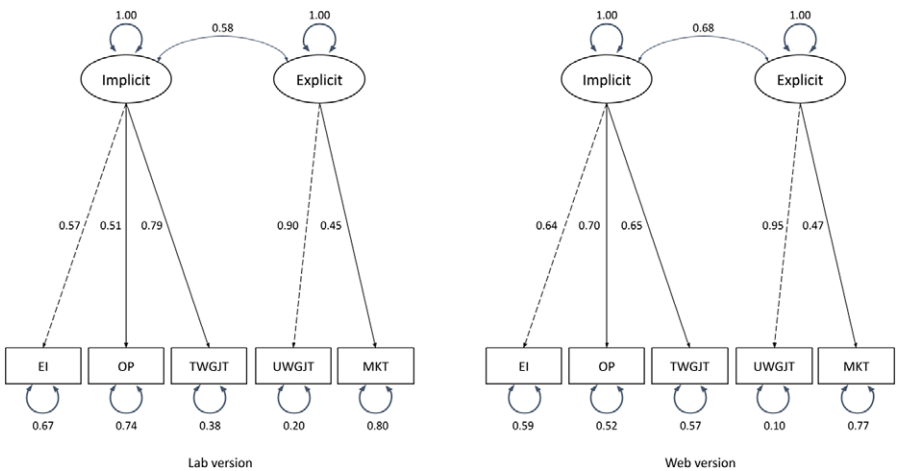

Figure 3. Two-factor CFA model, lab-version (left) and web-version (right).
Note: EI = elicited imitation; EI = EI; MKT = metalinguistic knowledge test; OP = oral production; TWGJT = timed grammaticality judgment test; UWGJT = untimed grammaticality judgment test.

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

significant. Collectively, no significant deviation of the WB EI and LB EI scores was observed in relation to a battery of different explicit and implicit knowledge measures. Such mirroring of the CFA results of the WB and LB measures reaffirms WB EI as a measure sharing a similar underlying construct as the LB EI.

Discussion

The goal of this study was to verify the validity of a low-fee, web-programmed EI task that measures English morphosyntactic knowledge. Drawing on theoretical and empirical work in L2 assessment and psycholinguistics, we hypothesized that, first, EI scores will yield significant associations with standardized L2 proficiency test scores as L2 grammar knowledge is a central component of L2 proficiency. Second, linguistic knowledge consists of explicit and implicit knowledge, with the EI task operationalizing as either implicit or automatized explicit knowledge; hence, EI is expected to exhibit more meaningful relationships to time-pressured accuracy measures (implicit or automatized explicit knowledge) than untimed linguistic measures (explicit knowledge). To empirically test the validity of web-programmed EI in relation to these assumptions, we compared the extent to which an EI administered on the web yielded a relationship to English proficiency scores and measures of implicit (i.e., OP and TGJT) and explicit (i.e., UGJT and MKT) grammar knowledge similar to that of EI administered in a lab setting.

In this study, we observed a good amount of shared variance between general English proficiency scores and the two versions of the EI scores. We found relationships of medium to large, comparable associative strength between TOEFL and WB EI (r = .52, 95% CI [0.39, 0.63]) and LB EI (r = .60, 95% CI [0.44, 0.72]), providing initial support for the similar importance of grammar competence in overall language proficiency. Previous research has reported strong relationships between standardized English proficiency and EI scores, demonstrating the use of morphosyntactic knowledge in language proficiency tests (e.g., Elder & R. Ellis, 2009; Erlam, 2006, 2009; Spada et al., 2015). For instance, in Erlam's studies (2006, 2009), scores of EI measuring 17 features of English grammar were strongly linked to the overall IELTS scores (r = .76), with the writing component showing the weakest (r = .46) and the listening component yielding the strongest (r = .72) correlations. Including only one target feature in the EI, Spada and colleagues (2015) used a C-test to measure learners' English proficiency and found one's knowledge of the English passive voice was moderately related to their general English proficiency (r = .50). In a recent meta-analysis (Kostromitina & Plonsky, 2022), studies using binary scoring, as our study, showed a correlation of . 53 (95% CI [0.43, 0.63]) to proficiency scores, which aligns closer to our findings. On par with these findings, our study, which also used only six features of English, yielded a medium correlation to general English proficiency in both versions of the EIs, thus establishing preliminary evidence for the comparable validity of online EI vis-à-vis in-person EI tasks.

Intriguingly, when the EI items were grouped by grammaticality, we observed some evidence of modality effects. A trend of differences emerged in the correlations between TOEFL scores and the ungrammatical EI items (although not reaching statistical significance, p = 0.062), with a weaker correlation for the web version (r = .37, 95% CI [0.22, 0.50]) compared to that of the lab version (r = .57, 95% CI [0.41, 0.70]). This pattern, where the ungrammatical tasks were less consistent but neither improved nor harmed the overall effectiveness of EI, matches the findings of Yan et al. (2016). In our context, such a trend was specific to the WB format, suggesting that the online

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

administrative context might introduce variability in participant responses to ungrammatical items. The lab setting (more specifically, the affordance of direct researcher interaction) allowed for immediate feedback on participants' questions on "correct English" responses when task instructions were given. Although the web version attempted to address this by modeling correction of ungrammatical items, the less controlled condition of the web setting might not have achieved the same effectiveness, thereby potentially affecting the results. We note that the observed trend of modality effect does not necessarily diminish the validity of ungrammatical items as a whole (because the observed effects are not statistically meaningful; and the LB version did show a fairly robust correlation to the TOEFL scores, in fact, numerically slightly better than grammatical items; see Figure 1, right). Instead, our findings caution users of WB EI about the potential for misinterpreting instructions, which could reduce the effectiveness of ungrammatical sentences in EI tasks.

Second, the effectiveness of the online EI was further corroborated by a series of CFA results. Based on a well-established explicit-implicit knowledge model by R. Ellis (2005) and its extensions (Godfroid & Kim, 2021; Godfroid et al., 2018), another goal was to assess the association between the web-programmed EI with different implicit and explicit knowledge measures. To this end, in both WB and LB versions, we configured CFA models with the same structure, anticipating comparable structural validity. As expected, we found excellent model fits for both WB and LB CFA models, with significant loadings for EI scores and other four linguistic measures. Such mirroring of the CFA results adds additional support for both EIs zeroing in on a common construct of L2 knowledge type.

To our knowledge, this is one of the first studies to validate the effectiveness of EIs deployed on the web. Our findings show promise for utilizing web-programmed EI as an alternative to its LB counterpart bringing potential to language science and education. Notably, the capacity to reliably record utterances via the Internet opens great prospects for research that necessitate elicited production data. Research, such as L2 oral proficiency (e.g., Ortega et al., 2002) and explicit-implicit knowledge development (e.g., R. Ellis, 2005; Erlam, 2006; Suzuki & DeKeyser, 2015; Suzuki et al., 2023) that entails an experimental design similar to that of the current study, may benefit from utilizing web-programmed EI in eliciting speech data. Relatedly, our web-programmed EI may serve as an efficient tool for researchers to conduct equitable and statistically powered research. With increased calls for diverse representation of L2 learners in SLA research (Andringa & Godfroid, 2020; Godfroid & Andringa, 2023), this online task, which enables wider and larger recruitment of participant samples, could hold promise for conducting equitable yet reliable research with unconventional learners (e.g., non-WEIRD populations: Western, Educated, Industrialized, Rich, and Democratic) on a larger scale. Beyond EIs used for research purposes, there is potential for their use in educational applications. With EI demonstrating its effectiveness as a language program placement test (Yan et al., 2016), this WB program, along with the potential for automated scoring of EIs (Isbell et al., 2023), may make the administration and scoring of EIs feasible with limited staff and resources. For researchers and educators interested in web-programming their own EI test, see Appendix S1 (Online Supplementary Material) and Kim et al. (2024.) for the coding guidelines and a tutorial on autoscoring English grammar forms.

Limitations and future directions

As previously mentioned, in the WB condition, the OP task was administered online alongside the EI task. This decision was made deliberately to optimize limited resources in

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

Kim's longitudinal study. However, for the specific focus of this paper (i.e., the validation of online vs. offline EIs), having online OP exclusively in the WB condition and not in the LB condition might introduce variations. To check for potential confounding effects, we looked into how OPs, in both conditions, relate to the three linguistic tests (i.e., TGJT, UGJT, and MKT), all of which were conducted in the lab for both conditions. We then used Fisher's *r*-to-*z* transformation to test for parity between the two conditions. If online modality mediated test performance, the strength of associations between OP and its paired linguistic tasks would yield meaningful differences between conditions. All paired relationships between OP and the linguistic tests were comparable across the conditions, with p values ranging from 0.117 to 0.763 (Figure S2, Online Supplementary Materials). While these results offer some support that the differences in modality between OPs are not meaningful, in future research, it is recommended to eliminate any discrepancies between conditions to facilitate more direct comparisons.

Another important limitation of this study is that participants' English proficiency scores were self-reported. The secondary data we reanalyzed (Godfroid et al., 2018; Kim, 2020) primarily came from university students, many of whom provided the scores achieved at the commencement of their academic programs. Therefore, their entered language proficiency score may not accurately represent their current proficiency levels. To test whether test-taking time changes the associations between EI and standardized proficiency scores, we broke the analysis down by test administration year (Figure S3 and Tables S1–3, Online Supplementary Materials). In essence, there is a gradual decline in correlation strength as the time since taking the test increases. Specifically, among the 139 participants who reported the years of test administration, those who took the English proficiency test within 2 years from the year of experimentation showed a correlation score of $r = .57$ (95% CI [0.33, 0.74], n = 51) to the EI scores. This coefficient is comparable to those who have taken the tests between 3 and 4 years ($r = .55$, 95% CI [0.33, 0.72], n = 63). The correlational strength dropped drastically when the years of test-taking dated after 5 years from 2019 ($r = .30$, 95% CI [–0.11, 0.62], n = 25). Given the strength weakens as the time since test-taking increases, we anticipate a stronger correlation between EI and language proficiency scores when the scores more accurately reflect learners' current proficiency level. Future studies should incorporate language proficiency tests into the experimental design, if not using secondary data.

The current WB EI program used in the study served as a data collection instrument. Nevertheless, its scalability was limited by the reliance on manual procedures for test item development and response scoring. To address this, a more comprehensive system integrating item development, test administration, and automated scoring is envisioned. With the aid of natural language processing technologies, a system could be created to generate or select sentences from a corpus containing target grammatical constructs, both with and without errors. The system would store participants' responses and employ automatic speech recognition and other natural language processing tools for automated scoring (Isbell et al., 2023; Kim et al., 2024). Such a comprehensive WB EI test system has the potential to greatly benefit the SLA and language learning community.

As also suggested by a reviewer, we encourage future studies to utilize a planned within-subjects design, where all participants engage in both offline and online versions of EIs. The reanalytic nature of our study precluded such a design, but the proposed within-subjects approach offers a rigorous way to investigate the impact of modalities (online vs. offline) on test scores while mitigating potential sources of noise, such as sampling error and participant variability.

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

Conclusion

The aim of this study was to validate the WB EI test by comparing its performance with the traditional LB delivery mode in terms of its correlation with established overall proficiency and linguistic knowledge tests. Using secondary data from Kim (2020) and Godfroid et al. (2018), we found moderate correlations between the EI test results and the TOEFL test for both delivery modes. The factor loadings of the EI test were also similar between the two modes, indicating comparable structural validity in the CFA model. Thus, we show promise for utilizing web-programmed EI as an alternative to its LB counterpart bringing potential to language science and education. The WB delivery mode offers numerous benefits, including logistical convenience, easier participant recruitment, and lower running costs.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263124000214.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford University Press.

Backx, R., Skirrow, C., Dente, P., Barnett, J. H., & Cormack, F. K. (2020). Comparing web-based and lab-based cognitive assessment using the Cambridge Neuropsychological Test Automated Battery: A within-subjects counterbalanced study. *Journal of Medical Internet Research*, *22*, 16792. https://doi.org/10.2196/16792

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., & Naugle, R. I. (2012) Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *The Clinical Neuropsychologist*, *26*, 177–196. https://doi.org/10.1080/13854046.2012.663001

Birnbaum, M. H. (Ed.). (2000). *Psychological experiments on the Internet*. Academic Press. https://doi.org/10.1016/B978-0-12-099980-4.X5000-X

Canale, M., & Swain, M. (1980). heoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*, 1–47. https://doi.org/10.1093/applin/1.1.1

Chapelle, C. (2021). *Argument-based validation in testing and assessment*. SAGE Publishing. https://doi.org/10.4135/9781071878811

Cumming, G., & Finch, S. (2005). Inference by Eye, Confidence intervals and how to read pictures of data. *American Psychologist*, *60*, 170–180. https://doi.org/10.1037/0003-066X.60.2.170

Elder, C. & Ellis, R. (2009). Implicit and explicit knowledge of an L2 and language proficiency. In R. Ellis, S. Loewen, C. Elder, H. Reinders, R. Erlam & J. Philip (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 167–193). Multilingual Matters. https://doi.org/10.21832/9781847691767-009

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172. http://doi.org/10.1017/S0272263105050096

Ellis, R., & Loewen, S. (2007). Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in Second Language Acquisition*, *29*, 119–126. https://doi.org/10.1017/S0272263107070052

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*, 464–491. https://doi.org/10.1093/applin/aml001

Erlam, R. (2009). The elicited oral imitation test as a measure of implicit knowledge. In R. Ellis, S. Loewen, C. Elder, H. Reinders, R. Erlam & J. Philip (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 65–93). Multilingual Matters. https://doi.org/10.21832/9781847691767-005

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

Erlam, R., & Akakura, M. (2015). New developments in the use of elicited imitation. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 93–111). Routledge.

Gagné, N., & Franzen, L. (2023). How to run behavioural experiments online: Best practice suggestions for cognitive psychology and neuroscience. *Swiss Psychology Open*, *3*, 1. https://doi.org/10.5334/spo.34

Godfroid, A., & Andringa, S. (2023). Uncovering sampling biases, advancing inclusivity, and rethinking theoretical accounts in second language acquisition: Introduction to the special issue SLA for all? *Language Learning*, *73*, 981–1002.

Godfroid, A., Kim, K., Hui, B., & Isbell, D. (2018, October). *Validation research on implicit and explicit knowledge: A research synthesis*. Conference presentation at the 2018 Second Language Research Forum, Montreal, Canada.

Godfroid, A., & Kim, K. (2021). The contributions of implicit-statistical learning aptitude to implicit second-language knowledge. *Studies in Second Language Acquisition*, *43*, 606–634. https://doi.org/10.1017/S0272263121000085

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

Houben, K., & Wiers, R. W. (2008). Implicitly positive about alcohol? Implicit positive associations predict drinking behavior. *Addictive Behaviors*, *33*, 979–986. https://doi.org/10.1016/j.addbeh.2008.03.002

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Hulstijn, J.H. (2011) Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*, 229–249. https://doi.org/10.1080/15434303.2011.565844

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research* (Vol. *41*). John Benjamins. https://doi.org/10.1075/lllt.41

Isbell, D. R., Kim, K. M., & Chen, X. (2023). Exploring the potential of automated speech recognition for scoring the Korean Elicited Imitation Test. *Research Methods in Applied Linguistics*, *2*, 100076.

Isbell, D. R., & Rogers, J. (2020). Measuring implicit and explicit learning and knowledge. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (1st ed., pp. 305–315). Routledge. https://doi.org/10.4324/9781351034784-33

Kerz, E., Wiechmann, D., & Riedel, F. (2017). Implicit learning in the crowd: Investigating the role of awareness in the acquisition of L2 knowledge. *Studies in Second Language Acquisition*, *39*, 711–734. https://doi.org/10.1017/S027226311700002X

Kim, J., & Nam, H. (2017). Measures of implicit knowledge revisited: Processing modes, time pressure, and modality. *Studies in Second Language Acquisition*, *39*, 431–457. https://doi.org/10.1017/S0272263115000510

Kim, K. (2020). *Exploring the interface of explicit and implicit second-language knowledge: A longitudinal perspective*. (Doctoral dissertation). Michigan State University.

Kim, Y., Tracy-Ventura, N., & Jung, Y. (2016). A measure of proficiency or short-term memory? Validation of an Elicited Imitation Test for SLA research. *The Modern Language Journal*, *100*(3), 655–673.

Kim, K., Chen, X., & Liu, X. (2024). Accuracy scoring of English elicited imitation: A tutorial. https://doi.org/10.31219/osf.io/8uzn9

Kim, K. M., & Godfroid, A. (2023). The interface of explicit and implicit second-language knowledge: A longitudinal study. *Bilingualism: Language and Cognition*, *26*, 709–723. https://doi.org/10.1017/S1366728922000773

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, *44*, 886–911. https://doi.org/10.1017/S0272263121000395

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

Lee, E., & Phillips, C. (2022). Why non-native speakers sometimes outperform native speakers in agreement processing. *Bilingualism: Language and Cognition*, *26*, 152–164. https://doi.org/10.1017/S1366728922000414

Loukina, A., Lopez, M., Evanini, K., Suendermann-Oeft, D., Ivanov, A. V., & Zechner, K. (2015). Pronunciation accuracy and intelligibility of non-native speech. *Interspeech, 2015,* 1917–1921. https://doi.org/10.21437/Interspeech.2015-423

Markman, B. R., Spilka, I. V., & Tucker, G. R. (1975). The use of elicited imitation in search of an interim French grammar. *Language Learning*, *25*, 31–41. https://doi.org/10.1111/j.1467-1770.1975.tb00107.x

McManus, K. (2021). Replication and open science in applied linguistics research. In Plonsky, L. (Ed.) Open science in applied linguistics. *John Benjamins*. https://doi.org/10.31219/osf.io/bqr9w

Nagle, C. (2019). Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk. *Journal of Second Language Pronunciation*, *5*, 294–323. https://doi.org/10.1075/jslp.18016.nag

Nagle, C., & Huensch, A. (2020). Expanding the scope of L2 intelligibility research: Intelligibility, comprehensibility,and accentedness in L2 Spanish. *Journal of Second Language Pronunciation*, *6*, 329–351. https://doi.org/10.1075/jslp.20009.nag

Nagle, C. L., & Rehman, I. (2021). Doing L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition*, *43*, 916–939. https://doi.org/10.1017/S0272263121000292

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October). *An investigation of elicited imitation tasks in crosslinguistic SLA research*. Paper presented at the Second Language Research Forum, Toronto, Canada.

Parsons, T. D., McMahan, T., & Kane, R. (2018). Practice parameters facilitating adoption of advanced technologies for enhancing neuropsychological assessment paradigms. *The Clinical Neuropsychologist*, *32*, 16–41. https://doi.org/10.1080/13854046.2017.1337932

Park, H. I., Solon, M., Henderson, C., & Dehghan-Chaleshtori, M. (2020). The roles of working memory and oral language abilities in elicited imitation performance. *The Modern Language Journal*, *104*, 133–151. https://doi.org/10.1111/modl.12618

Preacher, K. J. (2002). Calculation for the Test of the Difference Between Two Independent Correlation Coefficients [Computer software]. Vanderbilt University: Quantpsy.

Révész, A., & Brunfaut, T. (2021). Validating assessments for research purposes. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 21–32). Routledge.

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*, 226217. https://doi.org/10.1371/journal.pone.0226217

Saito, K., Hanzawa, K., Petrova, K., Kachlicka, M., Suzukida, Y., & Tierney, A. (2022). Incidental and multimodal high variability phonetic training: Potential, limits, and future directions. *Language Learning*, *72*, 1049–1091. https://doi.org/10.1111/lang.12503

Spada, N., Shiu, J. L.-J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies: validating an elicited imitation task. *Language Learning*, *65*, 723–751. https://doi.org/10.1111/lang.12129

Sudina, E. (2021). Study and scale quality in second language survey research, 2009-2019: The case of anxiety and motivation. *Language Learning*, *71*, 1149–1193. https://doi.org/10.1111/lang.12468

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229–1261. https://doi.org/10.1017/S014271641700011X

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge: Elicited imitation and word monitoring. *Language Learning*, *65*, 860–895. https://doi.org/10.1111/lang.12138

Suzuki, Y., Jeong, H., Cui, H., Okamoto, K., Kawashima, R., & Sugiura, M. (2023). fMRI reveals the dynamic interface between explicit and implicit knowledge recruited during elicited imitation task. *Research Methods in Applied Linguistics*, *2*, 100051. https://doi.org/10.1016/j.rmal.2023.100051

Van Teil, B., & Kissine, M. (2018). Quantity-based reasoning in the broader autism phenotype: A web-based study. *Applied Psycholinguistics*, *39*, 1373–1403. https://doi.org/10.1017/S014271641800036X

Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*, 54–73. https://doi.org/10.1111/1473-4192.00024

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press

Welch, N., & Krantz, J. H. (1996). The World-Wide Web as a medium for psychoacoustical demonstrations and experiments: Experience and results. *Behavior Research Methods, Instruments & Computers*, *28*, 192–196. https://doi.org/10.3758/BF03204764

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*, 497–528. https://doi.org/10.1177/0265532215594643

Zumbo, B. D. (2021). *A novel multimethod approach to investigate whether tests delivered at a test centre are concordant with those delivered remotely online: An investigation of the concordance of the CAEL: Research monograph*. Paragon Testing Enterprises/UBC Paragon Research Initiative, University of British Columbia. https://doi.org/10.14288/1.0400581

Cite this article: Kim, K. M., Liu, X., Isbell, D. R., & Chen, X. (2024). A comparison of Lab- and Web-based elicited imitation: Insights from explicit-implicit L2 grammar knowledge and L2 proficiency. *Studies in Second Language Acquisition,* 1–22. https://doi.org/10.1017/S0272263124000214

https://doi.org/10.1017/S0272263124000214 Published online by Cambridge University Press