Studies in Second Language Acquisition **41** (2019), 1027–1053
doi:10.1017/S0272263119000342

Research Article

REVISITING THE SPANISH ELICITED IMITATION TASK

A TOOL FOR ASSESSING ADVANCED LANGUAGE LEARNERS?

Megan Solon*

University at Albany, SUNY

Hae In Park

University at Albany, SUNY

Carly Henderson

Augusta University

Marzieh Dehghan-Chaleshtori

University at Albany, SUNY

Abstract

Elicited imitation tasks (EITs) have been shown to be a valid, reliable, and practical method for establishing second-language (L2) oral proficiency across a variety of languages. Nevertheless, research that has validated existing EITs has most often explored a narrow range of learner proficiencies or has examined how well the EIT distinguishes between low- and high-level learners. Bowden's (2016) validation study of the Spanish EIT was the first to expand the range of learner proficiencies by including low, advanced, and very advanced L2 learners; her results suggested that, in its current state, the EIT may not be equipped to distinguish finer-grained levels of proficiency, especially at more advanced levels. The present study revisits the Spanish EIT to examine (a) its ability to distinguish between higher level learners and (b) whether including additional, more challenging EIT items allows for finer-grained distinctions at higher levels.

We would like to thank Cecily Corbett, Pamela Alvarado-McNamara, and Ángel Milla Muñoz for their invaluable assistance in the creation and piloting of the revised EIT instrument. This project benefitted enormously from the feedback provided by attendees to the Evolving Perspectives on Advancedness: A Symposium on Second Language Spanish organized by Mandy Menke and Paul Malovrh in February 2018. We are also grateful to the anonymous reviewers and the SSLA editorial team for their comments, suggestions, and critiques. All remaining errors are our own.

*Correspondence concerning this article should be addressed to Megan Solon, Department of Languages, Literatures, and Cultures, Humanities 235, University at Albany, SUNY, 1400 Washington Avenue, Albany, NY 12222. E-mail: msolon@albany.edu

Copyright © Cambridge University Press 2019

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

Results suggest that the addition of longer, more challenging EIT items can improve the discriminability of the test among more advanced L2 learners.

The ability to reliably and validly establish second-language (L2) proficiency is essential for research in the field of second language acquisition (SLA). Thomas (1994) enumerates what she calls "commonsensical reasons" (p. 310) that SLA research depends upon interpretable L2 proficiency metrics; these reasons include (a) being able to generalize research results beyond the individual learners included in a sample, (b) being able to track developmental sequences, and (c) controlling for proficiency as a variable in comparisons of acquisition of a given L2 by different groups of learners. Ortega, Iwashita, Norris, and Rabie (2002) developed an elicited imitation task (EIT) in four different languages (English, German, Japanese, and Spanish); on this task, participants listen to and are prompted to repeat a series of sentence-length stimuli that are subsequently evaluated for accuracy. Ortega and colleagues proposed that the EIT not only can serve as a reliable and practical tool to measure oral proficiency but also that parallel tasks facilitate comparability across languages. Since the introduction of these four original tasks, parallel EITs have been created in several additional languages (e.g., Chinese: Wu & Ortega, 2013; Zhou & Wu, 2009; French: Tracy-Ventura, McManus, Norris, & Ortega, 2014; Korean: Kim, Tracy-Ventura, & Jung, 2016), and all have been shown to have high internal and external reliability. In particular, learners' EIT performances correlated with other common measures of L2 proficiency (e.g., standardized language test scores; Kim et al., 2016; simulated oral proficiency interview scores [SOPI]; Bowden, 2016) and distinguished well between low- and high-level L2 speakers (e.g., Bowden, 2016; Wu & Ortega, 2013). Similar high validity and reliability have also been reported for EITs of a different design (e.g., Gaillard & Tremblay, 2016, which included 50 French sentences of varying length and complexity in randomized order for self-paced repetition; Graham, Lonsdale, Kennington, Johnson, & McGhee, 2008, which included 60 English sentences and utilized two different scoring methods: rating along a 4-point scale and a syllable count measure), lending further support to the claim that the EIT can be a reliable, valid, and practical tool for L2 proficiency assessment in a research setting.

The EIT's ability to and utility for discriminating between learners with relatively advanced language proficiencies, however, is questionable and has not received sufficient empirical attention to date. Bowden's (2016) validation study with L2 Spanish learners is the only study that includes an examination of this issue. Her findings suggest that, while the current version of the Spanish EIT can discriminate among a broad range of learners (from novice high to superior proficiency as determined by the SOPI), its ability to discriminate among very advanced learners (ranging from advanced mid- to superior proficiency levels) was not supported. The ability to accurately and efficiently distinguish between proficiency levels among more advanced learners (in addition to distinguishing between, e.g., low and high proficiency learners) is important for being able to document and understand advanced language capacities, a crucial goal of SLA research (Byrnes, 2012; Malovrh & Benati, 2018). The present study partially replicates and extends Bowden (2016) by revisiting the EIT's ability to discriminate among advanced learners of Spanish. It is a replication in the sense that it utilized Bowden's

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

(2016) updated version of the Spanish EIT (Ortega et al., 2002) to reexamine its ability to discriminate among advanced learners; and it further extends her work by adding six additional sentence stimuli to the existing 30-item EIT to investigate whether the 36-item EIT would increase the Spanish EIT's ability to discriminate between learners at more advanced levels.

BACKGROUND

PROFICIENCY ASSESSMENT IN SLA

The importance and difficulty of assessing L2 proficiency for research purposes (and doing so reliably, validly, consistently, and efficiently) has been the focus of several reviews and syntheses of SLA research. These works have repeatedly revealed that few studies in SLA assess L2 proficiency adequately. Thomas's (1994, 2004) reviews of proficiency assessment techniques in SLA research revealed, for example, that the most common technique used to specify L2 proficiency level was learner institutional status (e.g., "first year" vs. "third year") and that little changed in terms of L2 proficiency assessment practices between 1994 and 2004. More recently, Tremblay (2011) found that, of 144 SLA studies surveyed from three journals (*Second Language Research*, *Studies in Second Language Acquisition*, and *French Language Studies*), fewer than 40% assessed proficiency independently with a test. Hulstijn (2012) similarly found that a majority (55%) of the 140 empirical papers he reviewed from *Bilingualism: Language and Cognition* did not employ an objective measure of language proficiency. Both Tremblay (2011) and Hulstijn (2012) highlight the urgent need for SLA researchers to assess proficiency independently with some kind of "objective" (to the extent possible) test.[1] The choice of test, however, is not always straightforward. As Hulstijn (2010) notes, "Language proficiency is a complex construct, involving various knowledge and skill components. No test can be, at the same time, maximally efficient, maximally reliable, and measure all LP [language proficiency] components" (p. 197). Instead, researchers must consider the study's goals and research questions and the role of L2 proficiency in those goals and questions while also considering the validity, reliability, and efficiency of proficiency assessment techniques. For instance, while cloze tests have been shown to be reliable and empirically valid tools for measuring L2 proficiency (e.g., Tremblay, 2011), they may not be the most appropriate measure for assessing learners' oral proficiency as cloze tests are conducted in the written modality. Recently, a number of studies have proposed that EITs are an appropriate tool to measure learners' oral proficiency and can do so in an efficient way without compromising the reliability and validity of the test.

THE ELICITED IMITATION TASK

Elicited imitation refers to a psycholinguistic technique whereby participants listen to and repeat a (typically sentence-length) stimulus. Elicited imitation has been widely used as a tool in both first language (L1) and L2 research to assess language performance with regard to a variety of foci (e.g., implicit knowledge: Erlam, 2006; Verhagen, 2011; morphology: West, 2012; listening comprehension: Jensen & Vinther, 2003). As

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

described by Yan, Maeda, Lv, and Ginther (2016), the underlying assumption of elicited imitation is that "if the participant has acquired the grammatical features associated with or displayed in the stimuli, it should be easy to repeat the stimuli. Otherwise, repetition will be difficult" (p. 498). More recently, EITs have been revisited as a potential method for testing global (oral/aural) L2 proficiency. According to Bley-Vroman and Chaudron (1994), when test-takers hear a sentence stimulus, they process the meaning to form a representation at various levels, which is then stored in short-term memory. Based on what they have understood, test-takers will reconstruct its meaning and translate the conceptual representation into a linguistic form. While the required language processing for the task is automatic for native speakers, this is not the case with most L2 learners and, as a result, the EIT is likely to impose greater demands on learners' working memory. However, it is expected that with an increase in L2 proficiency, learners' input and output processing would become more automatized and, as a result, their repetition accuracy would also improve. While some have voiced concerns regarding the possibility of simply parroting in response to stimuli, studies examining the relative contributions of memory and language ability to EIT performances (Dosi, Papadopoulou, & Tsimpli, 2016; Moll, Hulme, Nag, & Snowling, 2015; Park, Solon, Henderson, & Dehghan-Chaleshtori, under review; Riches, 2012) have consistently shown that EITs are best seen as a measure of language skills rather than a measure of memory capacity, despite the observed involvement of memory capacity. These findings are further corroborated by studies that have found no significant correlations between EIT performances and memory capacity (Kim et al., 2016; Okura & Lonsdale, 2012). To minimize the likelihood of simple parroting, most stimuli sentences are long enough to exceed working memory capacity (five to seven words, or six to eight syllables, according to Vinther, 2002), and researchers often insert a period of silence before repetition.

Motivated by a need to assess proficiency across four L2s included in a crosslinguistic study of syntactic complexity measures, Ortega et al. (2002) explored, as previously mentioned, parallel EITs in four languages—English, German, Japanese, and Spanish. These EITs each included 30 items ranging in length from 7 to 19 syllables. A five-point scoring rubric was also developed to facilitate parallel coding across L2s. The rubric, which guides the assignation of a score from 0 to 4 to each item, focuses on content and, to a lesser degree, formal accuracy. Ortega et al. (2002) as well as a follow-up study (Bowden, 2016; focused just on the Spanish EIT) established the external validity and internal reliability of these four tests. As previously mentioned, subsequent EITs in additional L2s have been developed to parallel the original four, and several studies have corroborated the internal reliability and external validity of this overall version of the EIT (across languages) as a measure of oral/aural proficiency (i.e., Chinese: Wu & Ortega, 2013; Zhou & Wu, 2009; French: Tracy-Ventura et al., 2014; Korean: Kim et al., 2016).

It is worth noting, however, that research that has validated existing EITs has most often explored a narrow range of learner proficiencies or has examined how well the EIT distinguishes between high- and low-level learners. For example, Tracy-Ventura et al.'s (2014) validation of the French EIT explored the performance of 29 French-degree students all recruited from the same curricular level. Wu and Ortega (2013) sampled Chinese language students from lower-division and upper-division classes to purposefully represent low and high linguistic abilities, respectively.[2] Bowden's (2016) validation study of the Spanish EIT was notable in its inclusion of L2 learners from

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

a broad range of experience levels, from low (second- to fourth-semester learners with no immersion experience) to very advanced (at least 19 months of immersion experience). Given the present study's focus on the Spanish EIT and, in particular, its ability to discriminate among learners at higher proficiency levels, in the next section, we review Bowden's (2016) exploration of the Spanish EIT in greater detail.

PREVIOUS ANALYSES OF THE SPANISH EIT: BOWDEN (2016)

Bowden (2016) revisited Ortega and colleagues' Spanish EIT to examine its external validity and internal reliability "across a broad spectrum of L2 Spanish learners (from beginning to very advanced)" (p. 654). Specifically, she explored the external validity of the Spanish EIT as compared to Simulated Oral Proficiency Interview (SOPI) scores, the internal reliability of the Spanish EIT, the discriminatory power and difficulty of each item, the extent to which the Spanish EIT discriminates within a spectrum of L2 experience, and the Spanish EIT's utility for clustering L2 learners across experience levels.

The study included 37 L2 learners of Spanish who were recruited at three broad experience levels: those in the "low" level ($n = 15$) were enrolled in second- to fourth-semester university-level Spanish and had no immersion experience; "advanced" level participants ($n = 15$) had an average of 6.7 semesters of university-level Spanish study and 1–2 semesters of immersion experience; and "very advanced" participants ($n = 7$) had "at least 19 months of immersion experience" ($M = 55$ months; p. 655). Results of Bowden's (2016) analyses suggested high external validity of the Spanish EIT as related to participants' SOPI scores (Pearson $r = .911$); a high internal consistency of items (Cronbach's alpha = .98); high discriminatory power of the EIT among learners (discrimination indices above .20 for all items except one); and high discriminatory power among learners in the low (M discrimination index = .35) and advanced (M discrimination index = .26) groups but not among learners in the very advanced group (M discrimination index = .15). While the lack of discrimination at the very advanced level may be attributed to the fact that there is less to be discriminated at that proficiency level, SOPI scores of the seven learners in this group suggested that their proficiency levels varied considerably. Based on these findings, Bowden suggested that "Additional—more difficult—items might be able to be constructed for the Spanish EIT to allow the test to discriminate among learners at such high levels of proficiency" (p. 667).

With regard to its utility for clustering L2 learners across experience levels, a k-means cluster analysis resulted in three learner groups that exhibited, according to Bowden, less overlap in SOPI scores between groups than did the predetermined experience levels. Nevertheless, it should be noted that all "very advanced" and most "advanced" participants were clustered together in the highest group ($n = 18$), suggesting that the Spanish EIT in its current form may not be very useful for discriminating between learners at higher levels of proficiency.

MOTIVATION FOR AND OBJECTIVES OF THE PRESENT STUDY

In the introduction to their volume *The Handbook of Advanced Proficiency in Second Language Acquisition*, Malovrh and Benati (2018) stress the increasing importance of

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

understanding advanced L2 capacities given the growing "social demands for multilingual actors" (p. 1) in our globalized world. They echo Byrnes's (2012) call for greater attention to probing advanced L2 capacities both to provide a more comprehensive theoretical basis for our understanding of advanced L2 proficiency and ultimate attainment and to inform instructional and curricular decisions so as to guide learners toward such capacities in a principled fashion. Nevertheless, the potential inability of established research tools, such as the EIT, to distinguish between learners at higher levels of proficiency inhibits our ability to describe and understand advanced L2 proficiency and performance. The present study aims to partially replicate and expand on Bowden (2016) to further explore the ability of the existing Spanish EIT to distinguish between higher level L2 learners with varied experiences and to examine whether additional EIT items increase the utility of the EIT for making finer-grained distinctions between learners, especially at higher proficiency levels. Following Bowden (2016), the present study is guided by the following research questions:

1. What is the external validity of the EIT as a proficiency measure (as compared to scores on a modified Diploma of Spanish as a Foreign Language [DELE] test)?
2. What is the internal reliability of the EIT?
3. What is the discriminatory power of the EIT among learners, and what is the difficulty of each item?
4. To what extent does the EIT discriminate among a wide range of L2 experience levels?
5. What is the EIT's utility for clustering L2 learners across a wide range of L2 experience levels?

For each question, the current study compares findings for the original 30-item EIT and those for a modified 36-item EIT to investigate whether additional, longer EIT items increase the validity, reliability, and/or discriminability of the EIT, especially when considering participants at higher levels.

METHOD

INSTRUMENT

The EIT used in the current study was based on the Spanish version developed by Ortega et al. (2002) and adapted by Bowden (2016). Bowden's (2016) version consisted of six practice sentences in English increasing in length from 5 to 17 syllables followed by 30 unrelated Spanish sentences increasing from 7 to 18 syllables. Each sentence is presented aurally, one by one, followed by a 2-second pause and then a 0.5-second tone prompting the participant to repeat the sentence. For the present study, two modifications were made to the sentences on Bowden's version to attempt to limit dialect-specific terminology. First, in item 14, the subject and indirect object pronouns *vosotros* and *os* (second-person plural, informal, specific to Peninsular Spanish) were changed to *ustedes* and *les* (second-person plural, no dialect specificity), respectively. Second, in item 17, the adverb *recto* "straight" (more common in Peninsular Spanish) was changed to its more dialectally universal synonym, *derecho*.

In addition to these modifications to the original 30-item EIT, for the present study, six additional sentences were added to the end of the original task. The additional sentences

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

were adapted from the 50-item French EIT created by Gaillard and Tremblay (2016) and increased in length from 17 to 27 syllables. In previous research, item length has repeatedly been shown to be the strongest predictor of item difficulty (e.g., Graham, McGhee, & Millard, 2010; Hendrickson, Aitken, McGhee, & Johnson, 2010; Ortega et al., 2002). For example, Graham et al.'s (2010) analysis concluded that item length accounted for nearly 73% of the variability in item scores. For this reason, the selection of these six items focused mainly on sentence length as determined by number of syllables and followed Gaillard and Tremblay in terms of maximum item length.[3] In addition to item length, research has also suggested a smaller but still influential role of other characteristics such as the frequency of lexical items included in EI sentences (e.g., accounting for 8% of the variation in item difficulty in Graham et al., 2010), lexical familiarity and syntactic complexity (Ortega et al., 2002), the presence of a preposition or a modal (e.g., Hendrickson et al., 2010), as well as morphosyntactic complexity and issues of prosody and register (e.g., analyzed qualitatively in Tracy-Ventura et al., 2014). Thus, the additional sentences in the present study were also selected to vary in and expand the range of lexical, morphological, and syntactic complexity of the existing EIT. The sentences were piloted with a native speaker of Spanish to ensure naturalness. The list of the 36 sentence stimuli is included in the Appendix. The new instrument—including the original 30 items with modifications and the additional six sentences—was recorded by a female native speaker of Mexican Spanish and is available on IRIS (www.iris-database.org). The addition of these sentences to the end of the existing instrument permits a direct comparison between results on the original 30-item EIT and on the new 36-item version for the same learners, thus allowing us to explore whether additional items increase the utility of the EIT for making finer-grained distinctions between learners, especially among those with more advanced language proficiency.

PARTICIPANTS

Eighty-eight L2 learners of Spanish enrolled in or teaching Spanish classes at one of several universities in the United States participated in this study. Learners were grouped into experience levels using a hierarchical cluster analysis (HCA) that included course level,[4] months of Spanish immersion, and years of formal Spanish study as variables. Following the recommendations of Staples and Biber (2015), the HCA resulted in four distinct experience levels, which are described in Table 1.

It should be noted that the high and very high experience levels exhibited very similar experience profiles with regard to course level ($M = 5.71$; i.e., graduate level) and years of formal study of Spanish ($M > 14$ years). The two groups were distinguished in the HCA due to differences in time spent abroad: learners in the high experience group had, on average, almost 10 months of immersion experience; learners in the very high experience group had an average of more than 2 years of immersion experience.

Five native speakers of Spanish also participated in the present study and were included to provide a baseline for EIT performance and to confirm that native speakers perform at ceiling on the task. These participants—three women and two men—were living in the United States, were graduate students or university lecturers, and were from a variety of countries and regions of origin (i.e., Chile, Argentina, Spain, Puerto Rico, and Colombia).

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

TABLE 1. Summary of learner participant experience levels

Experience level	*n*	*M* Course level[a] (*SD*)	*M* Immersion experience in months (*SD*)	Years of formal study (*SD*)
Low	31	1.55 (0.5)	.28 (0.9)	4.02 (2.3)
Intermediate	26	3.69 (0.7)	1.59 (2.3)	8.06 (2.2)
High	24	5.71 (0.6)	9.74 (5.6)	14.4 (3.1)
Very high	7	5.71 (0.5)	25.36 (3.2)	14.6 (3.0)
Total L2 learners	88	3.65 (1.9)	5.24 (7.9)	8.90 (5.11)

[a]For course level, 1 = 100 level; 2 = 200 level; 3 = 300 level; 4 = 400 level; 5 = graduate level in coursework; 6 = graduate level beyond coursework.

TASKS AND PROCEDURE

Each participant met individually with a researcher to complete a series of tasks including a background questionnaire, a nonword repetition task, the EIT, a three-part narration task, and a modified version of the *Diploma de español como lengua extranjera* "Diploma of Spanish as a Foreign Language" (DELE) exam (Montrul, 2012). The results from the nonword repetition task and the narration tasks are described in a separate study (Park et al., under review) and will not be discussed further here.

The background questionnaire, administered using Google Forms, elicited information about participant demographics; native and additional language background; experience and contact with Spanish; and self-rated proficiency in English, Spanish, and other languages. Information from the background questionnaire was used to group participants into experience levels as reported in Table 1. The EIT was administered using PowerPoint. Participants listened to instructions in English recorded by a female native speaker of American English. Participants were told to listen to several sentences in English and repeat back as much of the sentence as they could after hearing a tone sound. After the practice sentences, participants were given the same instructions for the Spanish sentences and proceeded with the task. All responses were audio recorded. Finally, the modified version of the DELE (Montrul, 2012) was administered and scored using Qualtrics (2018). This 50-item test includes a 30-item multiple-choice vocabulary test followed by a 20-item cloze grammar test. The DELE is "the official accreditation degree of fluency in the Spanish language, issued and recognized by the Ministry of Education, Culture and Sports of Spain" (Montrul, 2016, p. 187). Although scant empirical information is available regarding the reliability or validity of the DELE, the test follows criteria established by the Common European Framework of Reference for Languages, and its results, supported by royal decree, are used to "gain access to education and professional institutions in Spain" (Garvida, 2009, p. 4). Additionally, this standardized test has been widely used in linguistics research to determine L2 proficiency levels and to predict linguistic performance (see discussion in Montrul, 2016). Because the 30-item EIT tested in this study had previously been validated against the Spanish SOPI (Bowden, 2016), we elected to compare EIT results here with the DELE as an additional, commonly employed and widely recognized (albeit written) proficiency measure. It should be noted that oral language samples were also collected from all

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

participants and an analysis of their properties in relation to participants' EIT scores is presented in Park et al. (under review).

CODING AND ANALYSIS

Following a norming session conducted between the first and third authors, the EIT was scored by the first author using the 5-point rubric developed by Ortega et al. (2002). An abbreviated version of the rubric, reproduced from Bowden (2016, p. 657, Table 2), is presented in Table 2.

According to the rubric, each item receives a score from 0 to 4 resulting in a total possible EIT score of 120 on the 30-item EIT and 144 on the 36-item EIT. The third author coded 15% of the EIT data; interrater reliability was calculated and indicated a high level of agreement between coders (Cronbach's alpha = .98).

A series of analyses were performed to explore the research questions detailed previously. Here, we briefly describe our analytical approach to each research question; each analysis will be described in more detail alongside its corresponding results in the next section. To determine the external validity of the EIT as a proficiency measure (Research Question 1), the relationship between EIT and DELE scores was calculated using descriptive statistics and Pearson correlations. Cronbach's alpha reliability coefficients were used to explore the internal reliability of the EIT (Research Question 2). Item analyses were used to explore item difficulty and the discriminatory power of the EIT among learners (Research Questions 3 and 4). Finally, the utility of the EIT for clustering L2 learners across a wide range of L2 experience levels was explored using HCAs. Within each analysis, we investigated differences between the 30-item and the 36-item EIT versions to explore the potential effect of additional EIT items. Additionally, given previous findings regarding potential shortfalls of the EIT for distinguishing learners at more advanced levels (e.g., Bowden, 2016), special attention was paid to the results corresponding to the present study's highly experienced learners.

RESULTS

The descriptive statistics for participants' scores on both versions of the EIT (30 items and 36 items) and on the DELE and are presented by experience level in Table 3.

TABLE 2. Spanish EIT rating criteria (abbreviated)

Item score	Criteria
4	Perfect repetition
3	Meaning preserved; use of synonyms or (grammatical or ungrammatical) changes in grammar that do not affect meaning
2	More than half of content preserved; slight changes in content that make content inexact, incomplete, or ambiguous
1	Half or less of content repeated; important content is left out; meaning may be unrelated or opposed to stimulus
0	Silence, unintelligible content, or only one content word

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

TABLE 3. Descriptive statistics for DELE and EIT scores by experience level

	DELE				EIT (30 Items)				EIT (36 Items)			
Experience Level	Mean (*SD*)	Range	Skewness (*SE*)	Kurtosis (*SE*)	Mean (*SD*)	Range	Skewness (*SE*)	Kurtosis (*SE*)	Mean (*SD*)	Range	Skewness (*SE*)	Kurtosis (*SE*)
Low (n = 31)	16.97 (4.41)	10–24	–0.15 (0.42)	–1.26 (0.82)	31.06 (19.60)	3–77	0.75 (0.42)	–0.15 (0.82)	33.48 (21.24)	5–84	0.79 (0.42)	–0.11 (0.82)
Intermediate (n = 26)	29.77 (7.86)	16–44	0.29 (0.45)	–.97 (0.89)	80.15 (15.98)	55–117	0.31 (0.47)	–0.40 (0.89)	89.27 (19.03)	60–134	0.47 (0.46)	–0.28 (0.89)
High (n = 24)	42.88 (3.34)	35–48	–.32 (0.47)	0.17 (0.92)	108.21 (6.88)	94–120	–0.21 (0.47)	–0.67 0(.92)	125.13 (9.91)	107–140	–0.24 (0.47)	–1.23 (0.92)
Very High (n = 7)	44.43 (4.47)	37–48	–1.23 (0.79)	–0.46 (1.59)	108.71 (9.67)	90–120	–1.27 (0.79)	2.11 (1.59)	124.71 (13.34)	102–143	–0.61 (0.79)	0.43 (1.59)
Native Speakers (n = 5)	47.00 (2.34)	43–49	–1.74 (0.91)	3.32 (2.00)	119.80 (0.44)	119–120	–2.24 (0.91)	5.00 (2.00)	143 (1.00)	142–144	0.00 (0.91)	–3.00 (2.00)
Total with Native Speakers	30.91 (12.60)	9–49	–.10 (.25)	–1.50 (.50)	75.31 (36.79)	3–120	–.50 (.25)	–1.14 (.50)	85.48 (43.63)	5–144	–.38 (.25)	–1.22 (.50)
Total without Native Speakers	30.00 (12.32)	9–48	–.02 (.26)	–1.48 (.51)	72.78 (36.21)	3–120	–.44 (.26)	–1.19 (.51)	82.22 (42.57)	5–143	–.34 (.26)	–1.25 (.51)

Note: Maximum scores were 50 for the DELE, 120 for the 30-item EIT, and 144 for the 36-item EIT.

As can be seen in Table 3, the range of scores among L2 learners on both versions of the EIT was wide, varying from 3 to 120 out of 120 for the 30-item EIT and from 5 to 143 out of 144 for the 36-item EIT. It should be noted that two learners scored at ceiling (i.e., 120) on the 30-item EIT, whereas no learner received a perfect score (i.e., 144) on the 36-item EIT. As was also found by Bowden (2016), in the current study, there was some overlap in EIT scores across consecutive experience levels; in fact, learners at the high and very high experience levels overlapped almost entirely in EIT scores, as measured by either version (i.e., 30 item or 36 item). Native Spanish speakers, as expected, scored at near-perfect levels on both versions of the EIT; for this reason, in all subsequent analyses, the native speaker data were removed and only L2 learner data were explored.

To explore the external validity of the EIT as compared to scores on a modified DELE exam (Research Question 1), correlations were calculated between these two assessment scores. Results indicated that the DELE and the EIT were indeed very highly correlated (Pearson $r = .92$ for the 30-item EIT; $r = .93$ for 36-item EIT). Figure 1 includes a scatterplot of each participant's DELE score against his or her 30- and 36-item EIT scores, respectively. These results suggest that both versions of the EIT have high external validity as compared to the DELE.

Next, to answer Research Question 2, the internal reliability of both the 30- and 36-item EITs was explored by calculating Cronbach alpha coefficients. Results indicated high internal consistency for both task versions, with Cronbach alpha values of .99 for both the 30-item and 36-item EIT.

To address Research Questions 3 (i.e., What is the discriminatory power of the EIT between learners, and what is the difficulty of each item?) and 4 (i.e., To what extent does the EIT discriminate across a wide range of L2 experience levels?), a series of item analyses were performed following Bowden (2016). First, mean scores for each item were calculated across all participants. These are plotted, by item, from highest mean (least difficult) score at the top to lowest mean score (most difficult) at the bottom in Figure 2. As shown in Figure 2, mean scores on EIT items range from just above 1 to just above 3.5 with the majority of mean item scores falling between 1.5 and 2.5. Notably, the six sentences added for the current study's 36-item EIT version (items 31–36) received the lowest mean scores, suggesting that these items do, in fact, represent more difficult items than the existing items on the original 30-item EIT version.

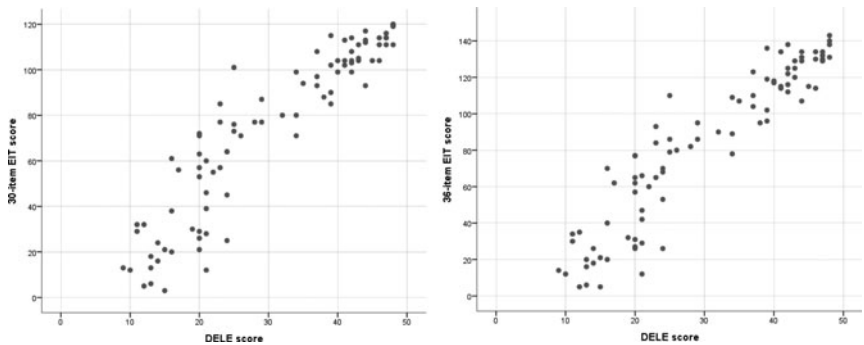


FIGURE 1. Scatterplot of EIT (30-item version on left, 36-item version on right) scores against DELE scores.

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

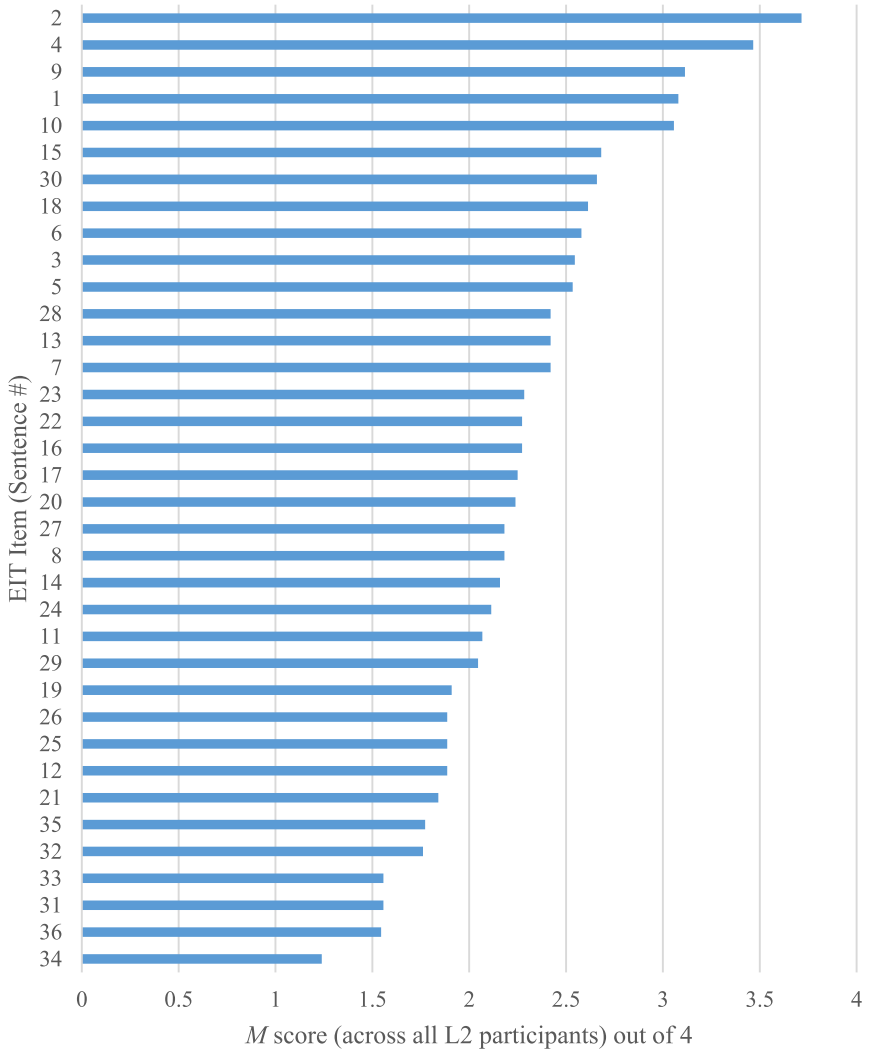


FIGURE 2. Mean item score by EIT item (labeled by sentence number corresponding to Appendix).

Next, discrimination indices were calculated for each item. To do so, mean item scores were converted to accuracy proportions (given that each item had a possible score of 4). Participants were divided to yield an upper group (participants earning the highest 27% of scores; $n = 25$) and a lower group (participants earning the lowest 27% of scores; $n = 24$).[5] Mean proportion scores on each item for participants within the lower group were subtracted from the scores of the upper group to yield an item discrimination index. Mean item scores and overall mean discrimination indices for both EIT versions are presented in Table 4. Table 5 presents the discrimination index for each individual item.

The present analysis employs Ebel's (1979) guidelines in the interpretation of discrimination indices: A discrimination index below 0.19 indicates poor discriminability of

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

TABLE 4. Summary of item scores and discrimination indices on both EIT versions

	30-item EIT			36-item EIT		
Variable	*M*	*SD*	*Range*	*M*	*SD*	*Range*
Mean item score, Upper group (n = 25)	3.71	0.30	2.80–4.00	3.59	0.42	2.52–4.00
Mean item score, Lower group (n = 24)	0.75	0.63	0.21–3.08	0.67	0.30	0.08–3.08
Discrimination indices	0.74	0.15	0.23–0.92	0.73	0.15	0.23–0.92

Note: The score range on each item was 0–4.

an item; indices between 0.20 and 0.29 indicate marginal discriminability; levels of 0.30 to 0.39 indicate reasonably good discriminability, and 0.40 and above are very good. As can be seen in Table 4, both the 30-item and the 36-item EIT versions exhibited high overall discriminatory power with mean item discrimination indices of 0.74 and 0.73, respectively. As for the discriminatory power of individual items, Sentence 2 exhibited the lowest discrimination index (0.23), indicating marginal discriminability. All other items exhibited discrimination indices above .40, indicating good discriminability. Item discrimination index results for the first 30 items were similar to those found by Bowden (2016), although it is interesting to note that, in her analysis, Sentence 4 exhibited the lowest discrimination index (falling below .20). Finally, it is worth mentioning that, although all the six added items exhibited relatively high discrimination indices (ranging from 0.53 for Sentence 35 to 0.82 for Sentence 31), they did not exhibit the highest overall discrimination indices. Several of the items on the existing 30-item EIT exhibited markedly higher discrimination indices (e.g., Sentence 7: 0.93; Sentence 14: 0.92; Sentence 5: 0.90).

Next, to assess the ability of each version of the EIT to discriminate within a spectrum of different L2 experience levels (low to high), item analyses were carried out for each participant experience level. As with the overall item analyses, the analyses by level included determining upper and lower groups of participants in each level by identifying the top and bottom 27% of scorers on each EIT version in that level. It should be noted that the very high experience level is small (n = 7) resulting in upper and lower groups consisting of two participants each. Table 6 presents item reliability and mean discrimination indices for each learner participant level as described in Table 1.

The rightmost column of Table 6 presents the mean discrimination indices for each version of the EIT (30 and 36 item) by the current study's learner experience levels. The discrimination indices for each experience level were, as expected, much lower than the mean overall discrimination indices for each EIT, given the narrower range of L2 experience within each level. For the two lower learner experience levels, both versions of the EIT exhibited discrimination indices above the 0.20 level. At the low experience level, the 30-item EIT exhibited a higher mean discrimination index (0.41) than the 36-item EIT (0.37), whereas, at the intermediate level, the two versions of the EIT exhibited identical mean discrimination indices (0.33). At the higher experience levels (high and very high), discrimination indices were below the 0.20 threshold except for the 36-item EIT among very highly experienced learners (0.21). For both the high and the very high experience groups, the 36-item item exhibited higher discrimination indices than the 30-

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

TABLE 5. Mean proportion scores and discrimination index for each item

	Mean Score		
Item	Upper group	Lower group	DI
1	0.99	0.38	0.62
2	1	0.77	0.23
3	0.97	0.24	0.73
4	1	0.55	0.45
5	1	0.10	0.90
6	0.97	0.08	0.89
7	1	0.07	0.93
8	0.95	0.13	0.83
9	1	0.38	0.63
10	0.99	0.38	0.70
11	0.88	0.29	0.72
12	0.84	0.15	0.69
13	0.89	0.22	0.67
14	0.99	0.07	0.92
15	1	0.13	0.88
16	0.93	0.09	0.84
17	0.89	0.11	0.76
18	0.96	0.19	0.77
19	0.88	0.09	0.79
20	0.88	0.15	0.73
21	0.77	0.03	0.74
22	0.93	0.21	0.72
23	0.96	0.13	0.84
24	0.93	0.05	0.88
25	0.70	0.18	0.52
26	0.80	0.10	0.70
27	0.93	0.09	0.84
28	0.95	0.24	0.71
29	0.90	0.12	0.79
30	0.98	0.10	0.88
31	0.86	0.04	0.82
32	0.80	0.05	0.75
33	0.79	0.02	0.77
34	0.63	0.05	0.58
35	0.69	0.16	0.53
36	0.67	0.13	0.55

Note: DI = discrimination index.

item EIT (high experience group: 0.16 [36 item] as compared to 0.13 [30 item]; very high experience group: 0.21 [36 item] as compared to 0.17 [30 item]). These results suggest that the modified, 36-item EIT is, in fact, better able to discriminate among learners at higher levels of proficiency than is the 30-item EIT. It is also noteworthy that Bowden (2016) found a poor mean item discrimination index (0.15) for the original 30-item EIT for her most advanced experience group, similar to the 0.13 and 0.17 found for the present study's high and very high experience groups, respectively.

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

TABLE 6. Reliability and item discrimination by experience group

Learner experience level	EIT version	Cronbach's alpha	*M* item score, Upper group		*M* item score, Lower group		*M* item discrimination index	
			M	*SD*	*M*	*SD*	*M*	*SD*
Low	30 item	0.97	1.96	0.77	0.33	0.49	0.41	0.16
	36 item	0.97	1.77	0.83	0.31	0.46	0.37	0.17
Intermediate	30 item	0.94	3.35	0.44	2.05	0.84	0.33	0.17
	36 item	0.95	3.18	0.60	1.87	0.89	0.33	0.17
High	30 item	0.78[a]	3.86	0.21	3.35	0.55	0.13	0.11
	36 item	0.84[a]	3.79	0.29	3.13	0.73	0.16	0.14
Very high	30 item	0.91[b]	3.90	0.24	3.23	0.76	0.17	0.19
	36 item	0.93[b]	3.82	0.38	3.00	0.89	0.21	0.20

Note: Low experience level: Upper group, $n = 8$, Lower group $n = 8$; Intermediate experience level: Upper group, $n = 7$, Lower group, $n = 7$; High experience level: Upper group, $n = 7$, Lower group, $n = 7$; Very high experience level: Upper group, $n = 2$, Lower group, $n = 2$.

[a]Five items were excluded from analysis due to zero variation across participants at this level.

[b]Eight items were excluded from analysis due to zero variation across participants at this level.

Individual item mean discrimination indices by level, presented in Table 7, reveal that the six added items exhibited higher mean discrimination indices for the more experienced learners (high and very high experience groups) as compared both to the discrimination indices for the other two learner groups for these items and to the discrimination indices for preceding items for these two groups. These added items discriminate less well among lower experience level learners (which is to be expected given their difficulty), thus explaining the overall lower mean discrimination indices observed in Table 5 and reproduced in the second column of Table 7.

TABLE 7. Mean discrimination indices for each item overall and by learner experience level

Item	Overall	Low	Int	High	Very High
1	0.62	0.72	0.25	0	0.25
2	0.23	0.31	0.04	0	0
3	0.73	0.31	0.54	0.14	0
4	0.45	0.59	0.04	0	0
5	0.90	0.50	0.07	0.07	0
6	0.89	0.53	0.25	0.07	0
7	0.93	0.47	0.50	0	0
8	0.83	0.28	0.39	0.14	-0.13
9	0.63	0.63	0	0.04	0
10	0.70	0.59	0.11	0	0
11	0.72	0.23	0.43	0.18	0.38
12	0.69	0.31	0.36	0.32	0.50
13	0.67	0.44	0.46	0.29	0
14	0.92	0.22	0.46	0.11	0.25
15	0.88	0.50	0.32	0	0.13
16	0.84	0.38	0.46	0.21	0.13
17	0.76	0.56	0.32	0.18	0.13
18	0.77	0.66	0.25	0.07	0.25
19	0.79	0.28	0.46	0.39	0.50
20	0.73	0.47	0.25	0.18	0.38
21	0.74	0.28	0.43	0	0.63
22	0.72	0.13	0.50	0.11	0
23	0.84	0.38	0.32	0.21	0.13
24	0.88	0.28	0.64	0.14	0.25
25	0.52	0.47	0.18	0.21	0.25
26	0.70	0.22	0.39	0.25	0.25
27	0.84	0.44	0.21	0.07	0.38
28	0.71	0.34	0.36	0.11	0.25
29	0.79	0.13	0.57	0.32	0.13
30	0.88	0.59	0.21	0.04	0
31	0.82	0.16	0.57	0.46	0.13
32	0.75	0.31	0.39	0.25	0.50
33	0.77	0.13	0.39	0.29	0.63
34	0.58	0.09	0.29	0.39	0.38
35	0.53	0.16	0.11	0.18	0.25
36	0.55	0.16	0.18	0.46	0.50

Note: Int = intermediate.

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

These results suggest that the addition of these more difficult items does increase the discriminability of the EIT specifically for more experienced learners.

Finally, to explore the utility of the 30- and 36-item EIT versions for clustering learners across a wide range of experience levels (Research Question 5), two separate HCAs were performed—one utilizing the results on the 30-item EIT, the other utilizing the results of the 36-item EIT. It should be noted that the decision to use HCA instead of *k*-means clustering also differs from Bowden (2016). In the present study, HCA was chosen following the recommendations of Staples and Biber (2015) given that the data are presumed to be hierarchical (as opposed to disjoint) in nature; that is, the clusters are presumed to represent groups of speakers that can be ranked as more or less proficient. Furthermore, HCA does not require an *a priori* specified number of clusters while the nonhierarchical clustering method (i.e., *k*-means) does. Thus, with HCA, each observation starts as its own cluster, and clusters with the greatest similarity or shortest distance between them are gradually combined until all samples are ultimately gathered into one category. As HCA does not automatically yield one optimal number of clusters (there is no singular measure to determine the number of clusters), the results are, to some degree, subjective. As Staples and Biber (2015) emphasize, "HCA is an exploratory technique" (p. 266), and there "is no single correct solution resulting from cluster analysis" (p. 268). Rather, the goal is to explore the data so as to determine a cluster solution that is maximally informative. This is done using multiple channels including by examining dendrogram plots that illustrate the hierarchical nature of the clustering, by investigating fusion coefficients (or the "numerical value[s] at which cases merge to form a cluster" Staples & Biber, 2015, p. 255) to explore when the difference between coefficients begins to decline, and by exploring the composition of each cluster in various cluster solutions.

The results of the two HCAs (one on the 30-item EIT results; one on the 36-item EIT results) were largely similar. Inspection of the dendrogram plots and of the fusion coefficients suggested that the optimal number of clusters in both HCAs was likely three. Table 8 presents the distance between fusion coefficients and the relative gain in the *F* value of the corresponding ANOVAs for both HCAs.

As can be seen in Table 8, both HCAs show gains in *F* values with the three-cluster solutions over the two-cluster solutions, suggesting that the three-cluster solutions for both are more informative. Both HCAs also show a decrease in *F* values with a four-cluster solution, again suggesting that the three-cluster solution for both may be optimal.

TABLE 8. Hierarchical cluster analysis on 30- and 36-item EIT scores

EIT version	Number of clusters	*df*	*F*	Gain
30 item	2	1, 86	256.57	–
	3	2, 85	519.22	262.65
	4	3, 84	484.64	–34.58
36 item	2	1, 86	211.97	–
	3	2, 85	523.95	311.98
	4	3, 84	505.88	–18.07

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

When following the three-cluster solution, the clustering of participants according to scores on either EIT version resulted in largely identical learner clusters: the three clusters differed only in the cluster membership of one participant, who resided in the middle group according to the HCA corresponding to the 30-item EIT and in the highest cluster according to the HCA corresponding to the 36-item test. Table 9 presents the descriptive statistics for the three clusters as grouped by the HCA for the 30-item EIT with the information for the one participant whose grouping differed by EIT version provided in the table note.

Table 10 presents the EIT, DELE, and *a priori* experience level breakdown of the three groups as determined by the cluster analysis of the 30-item EIT; once again, the table note provides the individual data for the participant whose level membership changed with the clustering by 36-item EIT results.

Similar to Bowden's (2016) findings with regard to mean SOPI scores of the EIT-determined clusters, the clustering by EIT in the present study appears to result in three groups whose corresponding DELE scores overlap less than did the *a priori* groups established on the basis of experience level.

Nevertheless, as Staples and Biber (2015) discuss, the determination of the optimal number of clusters resulting from an HCA is not an exact science and, instead, largely depends upon the goals of the research underway. When the four-cluster solutions for each HCA were explored, important differences emerged between the results for the 30-item and 36-item EITs. First, the four clusters from scores on the 30-item EIT resulted from dividing the middle cluster of the three-cluster HCA, whereas, for the 36-item EIT, four clusters resulted from dividing the highest cluster. Thus, the HCA corresponding to the 36-item EIT divided higher-level learners into two distinct groups in a way that the 30-item EIT was not able to. Table 11 presents the four-cluster solution for the 30- and 36-item EITs.

Additionally, as can be seen in Table 11, the four-cluster solution resulting from the 36-item EIT exhibits less overlap between clusters in DELE scores than does the 30-item HCA's four-cluster solution and more closely aligns with expectations regarding the distribution of learners by experience level. Considering previous recognition of the EIT's relatively poor ability to discriminate among high-level learners, this distinction between clustering results is important: The additional EIT items appear to improve the EIT's ability to form clusters that distinguish between learners at higher proficiency levels.

One final analysis further emphasizes this distinction. Considering that HCAs cluster existing cases, the number of optimal clusters depends entirely on the data structure—in this case, the characteristics of the participants included in the study. If we imagine a study that aims to investigate a particular linguistic structure that is expected to appear only in later stages of acquisition, we could imagine a participant sample that resembles only a subset of the current dataset. For example, perhaps researchers elect to focus their recruitment on learners registered in 300-level classes and above.[6] From that sample (n = 56 from the current dataset), we might ask, can the EIT discriminate levels of proficiency between these (higher level) learners? Two additional HCAs (one exploring results from the 30-item EIT and one the 36-item EIT) suggest that the 36-item EIT is better able to discriminate among proficiency levels of this more experienced subset of learners. Table

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

TABLE 9. Participant descriptive statistics as grouped by HCA results for 30-item EIT

Cluster level	Course level (*SD*)	Age of onset of Spanish study (*SD*)	Length of immersion (months) in Spanish-speaking region (*SD*)	Years of formal Spanish (*SD*)	Self-rated reading (*SD*)	Self-rated writing (*SD*)	Self-rated speaking (*SD*)	Self-rated listening (*SD*)
Level 1 (n = 25)	1.48 (.51)	15.32 (9.71)	0.05 (.10)	3.88 (2.40)	2.28 (.98)	2.24 (1.05)	1.92 (1.12)	1.84 (.99)
Level 2 (n = 26)	3.29 (1.12)	12.42 (3.049)	1.81 (4.55)	7.60 (2.85)	3.62 (.75)	3.81 (.75)	3.73 (.83)	3.65 (.80)
Level 3 (n = 37)	5.38 (.86)	12.32 (3.49)	11.16 (8.44)	13.20 (3.97)	4.62 (.59)	4.60 (.55)	4.73 (.51)	4.59 (.50)

Note: The HCA for the 36-item EIT resulted in identical clusters except one participant in this table's Level 2 moved to Level 3. Their information includes: Course level = 6; Age of onset = 13; Length of immersion = 23 months; Years of formal Spanish = 14; Self-rated reading = 5; Self-rated writing = 5; Self-rated speaking = 5; Self-rated listening = 5.

TABLE 10. Participant EIT scores, experience levels, and DELE scores as grouped by HCA for 30-item EIT

	EIT (30 items)		Experience Level	DELE	
Cluster Level	Mean (*SD*)	Range		Mean (*SD*)	Range
Level 1 (n = 25)	23.32 (11.59)	3–46	Low (25)	16.20 (4.42)	9–24
Level 2 (n = 26)	71.15 (11.27)	53–90	Low (6), Int (19), Very High (1)	25.81 (6.65)	16–39
Level 3 (n = 37)	107.35 (7.63)	93–120	Int (7), High (24), Very High (6)	42.27 (4.78)	25–48

Note: Data for the participant whose cluster membership changed according to the HCA for the 36-item EIT: 30-item EIT score = 90; Experience level = very high; DELE score = 39.

12 presents the distance between fusion coefficients and the relative gain in the F value of the corresponding ANOVAs for both HCAs.

As shown in Table 12, after two clusters, the informativeness of successive clustering with the 30-item results declines, as indicated by the drop in F value between the two- and the three-cluster solutions. For the 36-item EIT, in contrast, the F value increases with each successive cluster through five clusters, although the gain in F decreases between four and five clusters as compared to the gain made between three and four clusters (suggesting that the four-cluster solution may be the most informative and/or optimal solution). Thus, this analysis suggests that the 36-item EIT may provide researchers, especially those aiming to study more advanced learners, a tool to better identify and discriminate between higher proficiency levels.

DISCUSSION

The present findings largely corroborate those of Bowden (2016), confirming the validity, reliability, and practicality of the Spanish EIT as an independent global

TABLE 11. Participant EIT and DELE scores as grouped by four-cluster HCAs for 30- and 36-item EITs

	EIT		Experience Level	DELE	
Cluster Level	Mean (*SD*)	Range		Mean (*SD*)	Range
30-item EIT					
Level 1 (n = 25)	23.32 (11.59)	3–46	Low (25)	16.20 (4.42)	9–24
Level 2 (n = 10)	59 (3.94)	53–64	Low (4), Int (6)	20.70 (2.71)	16–24
Level 3 (n = 16)	78.75 (6.31)	71–90	Low (2) Int (13), Very high (1)	29.00 (6.21)	20–39
Level 4 (n = 37)	107.35 (7.63)	93–120	Int (7), High (24), Very High (6)	42.27 (4.78)	25–48
36-item EIT					
Level 1 (n = 24)	23.92 (11.61)	5–47	Low (24)	15.88 (4.30)	9–24
Level 2 (n = 26)	76.35 (12.84)	53–96	Low (7), Int (19)	25.23 (6.09)	16–38
Level 3 (n = 17)	112.29 (5.28)	102–120	Int (6), High (9), Very high (2)	39.41 (4.98)	25–46
Level 4 (n = 21)	131.76 (4.42)	122–143	Int (1), High 15), Very high (5),	44.43 (3.17)	37–48

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

TABLE 12. Hierarchical cluster analysis on 30- and 36-item EIT scores on subset of participants (n = 56)

EIT version	Number of clusters	*df*	*F*	Gain
30 item	2	1, 54	177.29	–
	3	2, 53	155.98	–21.31
36 item	2	1, 54	171.14	–
	3	2, 53	200.32	29.18
	4	3, 52	278.75	78.43
	5	4, 51	283.03	4.28

proficiency measure. Specifically, in the present study, as in Bowden (2016), the 30-item EIT demonstrated high external validity as a general proficiency measure ($r = .92$, $p < .01$) when compared to results on the DELE, high internal reliability (Cronbach's alpha = .99), and a mean item discrimination index of .74 (compared to Bowden's .58; keeping in mind differences in calculation decisions mentioned previously). Finally, like Bowden (2016), the present study also found that the 30-item EIT was able to cluster learner participants into three groups that appeared to better distinguish learners based on their proficiency (as measured by the DELE) than did *a priori* experience levels determined according to years of Spanish study, course level, and months of Spanish immersion.

It should be highlighted that, with regard to the measure of external validity, in the present study, the EIT was compared to a written measure of general language proficiency (the modified DELE exam), whereas Bowden (2016) compared EIT scores to learners' scores on a SOPI (an oral measure). Although such a high correlation between an oral-aural (EIT) and a written (DELE) test of general language proficiency may be somewhat surprising, the present results point to the EIT's ability to capture language proficiency in ways similar to those of other established measures, including those that extend across different modes of presentation and language use. These findings of high external validity as compared to the DELE thus extend existing evidence of the EIT's validity as related to a variety of widely and commonly used measures of language proficiency.[7]

In addition to the overall corroboration of Bowden's (2016) findings of validity, reliability, and discriminability of the 30-item EIT, the present study's analysis of an extended 36-item EIT provides evidence that the addition of these six items improved the EIT's ability to discriminate between learners at higher experience levels. Specifically, overall mean item discrimination indices for the high and very high experience learner groups were higher for the 36-item EIT than for the 30-item EIT, and the individual-item mean discrimination indices of these six added items were especially high among more experienced learners, indicating their greater ability to discriminate among more advanced learners. Additionally, scores on the 36-item EIT permitted the clustering of learner participants into four groups that specifically distinguished between learners at the highest levels, and analysis of a subset of more experienced learners demonstrated the increased ability of the 36-item EIT to cluster more advanced learners into distinct levels. These findings suggest that the modified version of the Spanish EIT (with 36 items) is, in

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

fact, better able to discriminate among a broad range of learners, especially when the sample includes learners at higher levels. These improvements in discriminability did not appear to affect the EIT's high validity and reliability: Correlations with the DELE as well as internal reliability measures were very similar (and even slightly higher) for the 36-item EIT as compared to the 30-item test. Additionally, the six items added less than two minutes of time to the approximately 10-minute task, ensuring that the EIT remains an efficient proficiency assessment tool.

The present analysis did not explore the specific characteristics of the added items that contributed to their increased discriminability. Instead, as mentioned in the "Method" section and following previous research, items were selected mainly by considering their length. An in-depth analysis of the item characteristics and their contribution to discriminability on this corpus is currently underway. Nevertheless, it is worthwhile to point out that item length was clearly not the sole contributor to item difficulty. For example, item 36 was, by several syllables, the longest item (27 syllables) on the modified EIT. Nevertheless, as shown in Figure 2, it was item 34 that received the lowest mean score overall, despite being substantially shorter in length (i.e., 21 syllables). A cursory qualitative review of participant repetitions of this item suggest that difficulty was presented in its lexical components (e.g., the phrase *tener prisa* "to be in a hurry" [literally, "to have urgency"] and the word *calidad* "quality" both presented challenges for numerous learners) as well as in its syntactic complexity (i.e., the sentence begins with an adverbial clause used in a rather infrequent comparative construction). The two L2 learners who scored the highest overall on the 36-item EIT (with scores of 143/144 and 140/144), both scored a 3 on this item due to a lack of gender agreement between the adjective pronoun *cuanta* "how much" and the noun *prisa* "hurry/urgency." Other high-scoring learners replaced *calidid* "quality" with similar words like *cualidad* "quality [as in characteristic]" or *cantidad* "quantity." Likewise, sentence 21 was the most difficult item on the 30-item EIT, despite not being the longest sentence. Issues related to lexical frequency or familiarity (e.g., difficulties with words such as *vecino* "neighbor" and *cuida* "take care of (3sg.)" as well as syntactic complexity (e.g., in use of the phrase *una amiga mía* "a friend of mine" instead of the more common *mi amiga* "my friend") seemed to contribute to learners' challenges in repetition. Detailed analysis of the numerous potential factors that influence item difficulty can further our understanding of the specific characteristics of the added items that increase their discriminability for higher level learners.

The development and validation of the new 36-item version of the Spanish EIT should, in no way, detract from the utility of the established 30-item version. In fact, use of the 36-item EIT clearly eliminates a central benefit of the established 30-item version: the existence of comparable EITs in other languages. Nevertheless, depending on the research goals, sample characteristics, and empirical necessity of establishing L2 proficiency, the extended, 36-item EIT may very well present itself as a useful tool for researchers particularly interested in measuring and distinguishing advanced language capacities. The ability to track developmental sequences among learners across a spectrum of experience and proficiency levels—not just to separate high-level learners from low-level learners—is central to research in SLA. As such, the present study's 36-item EIT offers researchers an important additional tool for making finer-grained

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

proficiency distinctions to facilitate more consistent and reliable analyses of L2 development.

CONCLUSION

The present study corroborates the existing and growing evidence that the EITs developed by Ortega and colleagues (2002) and extended to other languages (e.g., Tracy-Ventura et al., 2014; Wu & Ortega, 2013, etc.) indeed offer a valid, reliable, and efficient L2 proficiency assessment tool for SLA research purposes. In the present study, scores on the original 30-item Spanish EIT (with a few modifications) correlated highly with an external written proficiency measure (DELE; Montrul, 2012) and exhibited high internal reliability and relatively high discriminability. Nevertheless, the addition of six more difficult items improved the EIT's ability to discriminate among learners at more experienced levels and altered the EIT's ability to cluster participants according to proficiency level. These results both substantiate the utility of this proficiency assessment tool and suggest that modifications to this tool can provide researchers with a similarly valid, reliable, and practical assessment technique with the ability to capture participant details particularly relevant to a study's research questions or goals. The ability to independently establish proficiency, to do so efficiently and reliably, and to do so in a manner that better captures details relevant to the acquisition processes explored will only improve our ability to rigorously perform, describe, and interpret research in SLA.

NOTES

[1]The second half of Tremblay's (2011) study explores the validity, reliability, and discriminability of a specific cloze test designed by Tremblay and Garrison (2010) to measure the language proficiency of adult L2 learners of French.

[2]Wu and Ortega (2013) also explored participant language background, recruiting second and heritage language learners at both curricular levels.

[3]Maximum item length on Gaillard and Tremblay's (2016) EIT was 28 syllables. Due to a lexical item change after piloting, maximum item on the present 36-item EIT was 27 syllables.

[4]At universities from which participants were recruited, Spanish classes are identified numerically by level such that 100-level courses correspond to introductory language courses, 200-level courses to intermediate language courses, and so on through graduate-level coursework in the 500, 600, and 700 levels. Some participants were graduate students who had completed all coursework except that corresponding to the writing of the dissertation and others had graduated from graduate school and were working as faculty in university Spanish programs; these participants were considered to be at the "graduate beyond coursework" level as described in Table 1.

[5]The use of 27% to yield upper and lower groups for item analyses follows the commonly accepted recommendation of Kelley (1939). Nevertheless, it should be pointed out that our methods in this calculation differed slightly from those of Bowden (2016), who states having compared the highest third and lowest third of participants following Brown (2000, 2003). Thus, any comparison between the discrimination indices reported here and those reported in Bowden (2016) should bear these differences in mind. The number of participants pertaining to high and low groups differ due to duplications in scores at the upper group cutoff point.

[6]This level was chosen arbitrarily but also purposefully given that studies often focus their recruitment efforts on particular populations defined by a concrete characteristic (e.g., course level) with the aim of encountering participants with certain experiential or proficiency profiles that meet the goals of the study. For research that focuses on some aspect of advanced language capacity or development, initiating recruitment at a 300-level (i.e., third year) population is a reasonable scenario for purposes of this hypothetical illustration.

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

[7]The DELE is an official diploma granted by the Ministry of Education, Culture, and Sport of Spain. It is used extensively for language proficiency testing and certification, especially outside of the United States. The OPI, developed by the American Council on the Teaching of Foreign Languages, and SOPI are primarily used within the United States.

REFERENCES

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. Tarone, S. Gass, & A. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 245–261). Mahwah, NJ: Lawrence Erlbaum.

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

Brown, J. D. (2000). Statistics corner. Questions and answers about language testing statistics (how can we calculate item statistics for weighted items?). *Shiken: JALT Testing and Education SIG Newsletter*, *3*, 19–21.

Brown, J. D. (2003). Norm-referenced item analyses (item facilitation and item discrimination). *Shiken: JALT Testing and Evaluation SIG Newsletter*, *7*, 16–19.

Byrnes, H. (2012). Advanced language proficiency. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 506–520). New York, NY: Routledge.

Dosi, I., Papadopoulou, D., & Tsimpli, I. M. (2016). Linguistic and cognitive factors in elicited imitation tasks: A study with mono-and biliterate Greek-Albanian bilingual children. In J. Scott & D. Waughtal (Eds.), *Proceedings of the 40th Annual Boston University Conference on Language Development* (pp. 101–115). Somerville, MA: Cascadilla Press.

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, *66*, 419–447.

Garvida, M. M. (2009). A review of Diplomas de Español como Lengua Extranjera (DELE). *International Journal of Pedagogies and Learning*, *5*, 4–10.

Graham, C. R., McGhee, J., & Millard, B. (2010). The role of lexical choice in elicited imitation item difficulty. In M. T. Prior, Y. Watanabe, & S. Lee (Eds.), *Selected Proceedings of the 2008 Second Language Research Forum* (pp. 57–72). Somerville, MA: Cascadilla Proceedings Project.

Graham, C. R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. In *Proceedings of the International Conference on Language Resources and Evaluation*. Retrieved from www.lrec-conf.org/proceedings/lrec2008/index.html.

Hendrickson, R., Aitken, M., McGhee, J., & Johnson, A. (2010). What makes an item difficult? A syntactic, lexical, and morphological study of elicited imitation test items. In M. T. Prior, Y. Watanabe, & S. Lee (Eds.), *Selected Proceedings of the 2008 Second Language Research Forum* (pp. 48–56). Somerville, MA: Cascadilla Proceedings Project.

Hulstijn, J. H. (2010). Measuring second language proficiency. In S. Unsworth & E. Blom (Eds.), *Experimental methods in language acquisition research* (pp. 185–199). Amsterdam, The Netherlands: Benjamins.

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

Jensen, E. D., & Vinther, T. (2003). Exact repetition as input enhancement in second language acquisition. *Language Learning*, *53*, 373–428.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, *30*, 17–24.

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. *Modern Language Journal*, *100*, 655–673.

Malovrh, P. A., & Benati, A. G. (2018). Introduction. In P. A. Malovrh & A. G. Benati (Eds.), *The Handbook of advanced proficiency in second language acquisition* (pp. 1–5). Oxford, UK: Wiley Blackwell.

Moll, K., Hulme, C., Nag, S., & Snowling, M. J. (2015). Sentence repetition as a marker of language skills in children with dyslexia. *Applied Psycholinguistics*, *36*, 203–221.

Montrul, S. (2012). DELE Proficiency Test. Unpublished instrument. Retrieved from http://international.ucla.edu/nhlrc/data/example.

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

Montrul, S. (2016). *The acquisition of heritage languages*. Cambridge, UK: Cambridge University Press.
Okura, E., & Lonsdale, D. (2012). Working memory's meager involvement in sentence repetition tests. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 2132–2137). Hillsdale, NJ: Lawrence Erlbaum.
Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002). *An investigation of elicited imitation tasks in crosslinguistic SLA research*. Paper presented at the Second Language Research Forum, Toronto, Canada, October.
Park, H. I., Solon, M., Henderson, C., & Dehghan-Chaleshtori, M. (under review). Revisiting an elicited imitation task as a proficiency measure: The role of working memory in EIT performance.
Qualtrics. (2018). Provo, UT: Qualtrics [computer software].
Riches, N. G. (2012). Sentence repetition in children with specific language impairment: An investigation of underlying mechanisms. *International Journal of Language & Communication Disorders*, *47*, 499–510.
Staples, S., & Biber, D. (2015). Cluster analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 243–274). New York, NY: Routledge.
Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–336.
Thomas, M. (2004). Research synthesis and historiography: The case of assessment of second language proficiency. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279–298). Amsterdam, The Netherlands: Benjamins.
Tracy-Ventura, N., McManus, K., Norris, J., & Ortega, L. (2014). "Repeat as much as you can": Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 143–166). Bristol, UK: Multilingual Matters.
Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition*, *33*, 339–372.
Tremblay, A., & Garrison, M. D. (2010). Cloze tests: A tool for proficiency assessment in research on L2 French. In M. T. Prior, Y. Watanabe, & S.-K. Lee (Eds.), *Selected proceedings of the Second Language Research Forum 2008* (pp. 73–88). Somerville, MA: Cascadilla Press.
Verhagen, J. (2011). Verb placement in second language acquisition: Experimental evidence for the different behavior of auxiliary and lexical verbs. *Applied Psycholinguistics*, *32*, 821–858.
Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*, 54–73.
West, D. E. (2012). Elicited imitation as a measure of morphemic accuracy: Evidence from L2 Spanish. *Language and Cognition*, *4*, 203–222.
Wu, S.-L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*, 680–704.
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 Teaching*, *33*, 497–528.
Zhou, Y., & Wu, S.-L. (2009). *Development and pilot of a Mandarin L2 elicited imitation task*. Honolulu: University of Hawai'i at Mānoa. Unpublished manuscript.

Appendix

Sentences included in EIT with their English translations.

1. *Quiero cortarme el pelo*. (7 syllables)
"I want to get my hair cut."

2. *El libro está en la mesa.* (7 syllables)
"The book is on the table."

3. *El carro lo tiene Pedro.* (8 syllables)
"Pedro has the car."

4. *Él se ducha cada mañana.* (9 syllables)
"He showers every morning."

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

5. *¿Qué dice usted que va a hacer hoy?* (9 syllables)
"What do you say that you are going to do today?"

6. *Dudo que sepa manejar muy bien.* (10 syllables)
"I doubt that he knows how to drive very well."

7. *Las calles de esta ciudad son muy anchas.* (11 syllables)
"The streets of this city are very wide."

8. *Puede que llueva mañana todo el día.* (12 syllables)
"It is possible that it will rain tomorrow all day."

9. *Las casas son muy bonitas pero caras.* (12 syllables)
"The houses are very pretty but expensive."

10. *Me gustan las películas que acaban bien.* (12 syllables)
"I like movies that end well."

11. *El chico con el que yo salgo es español.* (13 syllables)
"The boy that I am dating is Spanish."

12. *Después de cenar me fui a dormir tranquilo.* (13 syllables)
"After eating dinner, I went to sleep peacefully."

13. *Quiero una casa en la que vivan mis animales.* (14 syllables)
"I want a house in which my animals can live."

14. *A ustedes les fascinan las fiestas grandiosas.* (14 syllables)
"You [plural] are fascinated by grandiose parties."

15. *Ella sólo bebe cerveza y no come nada.* (15 syllables)
"She only drinks beer and does not eat anything."

16. *Me gustaría que el precio de las casas bajara.* (15 syllables)
"I would like for the price of the houses to come down."

17. *Cruza a la derecha y después sigue todo derecho.* (15 syllables)
"Cross to the right and then continue straight."

18. *Ella ha terminado de pintar su apartamento.* (14 syllables)
"She has finished painting her apartment."

19. *Me gustaría que empezara a hacer más calor pronto.* (15 syllables)
"I would like for it to start being warmer soon."

20. *El niño al que se le murió el gato está triste.* (15 syllables)
"The boy whose cat died is sad."

21. *Una amiga mía cuida a los niños de mi vecino.* (16 syllables)
"A friend of mine takes care of my neighbor's children."

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press

22. *El gato que era negro fue perseguido por el perro.* (16 syllables)
"The cat that was black was followed by the dog."

23. *Antes de poder salir él tiene que limpiar su cuarto.* (16 syllables)
"Before being able to leave he has to clean his room."

24. *La cantidad de personas que fuman ha disminuido.* (16 syllables)
"The number of people who smoke has decreased."

25. *Después de llegar a casa del trabajo tomé la cena.* (17 syllables)
"After arriving home from work, I ate dinner."

26. *El ladrón al que atrapó la policía era famoso.* (17 syllables)
"The thief that the police caught was famous."

27. *Le pedí a un amigo que me ayudara con la tarea.* (17 syllables)
"I asked a friend to help me with the homework."

28. *El examen no fue tan difícil como me habían dicho.* (17 syllables)
"The exam was not as hard as they had told me."

29. *¿Serías tan amable de darme el libro que está en la mesa?* (18 syllables)
"Would you be so kind as to give me to book that is on the table?"

30. *Hay mucha gente que no toma nada para el desayuno.* (18 syllables)
"There are many people who do not eat anything for breakfast."

31. *Son ellas las que acaban de decorar la sala de espera.* (19 syllables)
"They [fem.] are the ones who just finished decorating the waiting room."

32. *¿Sabe usted si el tren de las once y media ya ha salido de la estación?* (20 syllables)
"Do you [form.] know if the 11:30 train has already left the station?"

33. *Nunca me divertí tanto como cuando fui a la pista de hielo.* (20 syllables)
"I have never had as much fun as when I went to the ice rink."

34. *Cuanta más prisa tenía en su trabajo, menos calidad producía.* (21 syllables)
"The more rushed he was in his work, the lower quality he produced."

35. *Ellos lo organizaron el año pasado en la universidad cercana.* (23 syllables)
"They organized it last year at the university nearby."

36. *Acabamos de volver del supermercado donde las ofertas eran muy interesantes.* (27 syllables)
"We just got back from the supermarket where the offers were very interesting."

https://doi.org/10.1017/S0272263119000342 Published online by Cambridge University Press