CrossMark

Studies in Second Language Acquisition, 2016, **38**, 647–675.
doi:10.1017/S0272263115000443

ASSESSING SECOND-LANGUAGE ORAL PROFICIENCY FOR RESEARCH

The Spanish Elicited Imitation Task

Harriet Wood Bowden
University of Tennessee–Knoxville

Proficiency is a key variable in late second language (L2) learning, but one that is undermeasured in current research. This study investigates whether L2 oral proficiency can be quickly and reliably assessed via the Spanish "elicited imitation task" (EIT; Ortega, Iwashita, Rabie, & Norris, 1999). Thirty-seven L2 learners of Spanish with varied L2 experience (from low to very advanced) were tested. The EIT's external validity was assessed via correlation with the simulated oral proficiency interview. Further analyses examined the EIT's internal reliability, discriminatory power, and usefulness in identifying subgroups (clusters) of learners within the sample. Results strongly support the utility of the Spanish EIT for measurement of oral proficiency in SLA research. Moreover, the availability of parallel EITs in other languages promises to facilitate comparisons across studies. Thus, more precise measurement and analysis of proficiency in SLA research is called for via standardized measurements including the EIT.

Thanks go to Kayla Bobbitt, Laura Gurzynski-Weiss, Jinah Haytko, Stephanie Lukas, and Shadia Kawa for assistance in rating SOPIs and/or EITs. Thanks go to Nina Moreno for assistance in recording the EIT. I would also like to thank Lourdes Ortega for making the EIT available to me and for providing feedback during the project, especially on the rerecording and scoring of the EIT.

© Cambridge University Press 2016

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

INTRODUCTION

Proficiency is a theoretically and pragmatically problematic area in the field of second language acquisition (SLA). Previous research has demonstrated that proficiency influences learners' performance on tasks, affects their activation of prior knowledge in instructional interventions, and limits or enhances their availability of attentional resources during tasks (e.g., Carrell, 1991; Kormos, 2000). Proficiency has also been identified as a key factor (and even the most important factor; Abutalebi, Cappa, & Perani, 2001) affecting the language system of bilinguals in various reviews and meta-analyses of neuroimaging and electrophysiological studies (Abutalebi, 2008; Indefrey, 2006; Kotz, 2009; Steinhauer, White, & Drury, 2009). Moreover, variability in proficiency is a hallmark of second language (L2) acquisition, and therefore, as signaled by Tremblay (2011), this variable should be measured and reported in SLA research as precisely as possible.

Much SLA research does not even attempt to measure proficiency, instead, for example, simply describing learners' current class level in the L2 or years of L2 experience. When proficiency *is* assessed in the research setting, it is often through self-ratings or nonstandardized short-answer tests of grammatical or lexical knowledge. In discussing proficiency assessments, which can range from extensive to shortcut approaches, Norris and Ortega (2012) state that "while the reporting of any L2 proficiency estimate is better than none, it is also the case that one major limitation to the generalizability of SLA research findings is the lack of domain-specific standardized proficiency assessments" (p. 580). This state of affairs may be attributable to the fact that proficiency is difficult to define, because proficiency measurements tend to be difficult and/or time-consuming to administer, and/or simply because there are few reliable, standardized proficiency measurements that are feasible for use in a research setting. Moreover, there is a lack of comparability in measures used across languages, making it difficult to interpret and generalize findings across languages.

This study investigates the elicited imitation task (EIT) developed by Ortega, Iwashita, Rabie, and Norris (1999) for Spanish in terms of its validity, reliability, and feasibility for use in this setting, as well as its ability to discriminate among a relatively broad range of L2 learners. In addition, the study examines to what extent other language background information usually available to researchers through self-report, including self-ratings of proficiency and language background variables, is a reliable proxy for independent proficiency assessments.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

Defining Proficiency

Although L2 learners, practitioners, and researchers have intuitions about what proficiency is, an operational definition is a precursor to assessment. Proficiency has been defined broadly in the field, for example, as "a person's overall competence and ability to perform in L2" (Thomas, 1994, p. 330). Hulstijn (2011, 2012) has recently proposed an alternative conceptualization, stating that language proficiency "should be taken seriously . . . because it is fundamental in the understanding of language acquisition and bilingualism" (2012, p. 429). He suggests making a distinction between basic language cognition (BLC) and higher language cognition (HLC). Basic language cognition is defined as what all native speakers have in common: implicit knowledge of phonology, morphology, and syntax; explicit knowledge in the lexical domain; and the automaticity with which this knowledge can be processed. Basic language cognition is restricted to spoken and aural language (speaking and listening), and to highly frequent items and structures common to all adult speakers of their first language. Higher language cognition is the complement or extension of BLC and includes lower-frequency items and structures, and written as well as spoken utterances. Thus, HLC may show individual differences within (normal) adult native speakers and may be affected by variables including literacy, level of education, profession, and so on. In this study, proficiency is broadly understood as at least potentially reflecting both BLC and HLC, though, as we will see, the elicited imitation task is more closely aligned with BLC, given its aural/oral nature (in that examinees listen to and then repeat back a series of sentences), and with what Mozgalina (2015) terms "oracy" (see subsequently).

Measurement of Proficiency in SLA Research

Many options are available to SLA researchers for measuring proficiency, and a huge variety of these have been included in research. These include standardized measures (such as the American Council on the Teaching of Foreign Languages [ACTFL] oral proficiency interviews or the TOEFL), which are often long, costly, and/or difficult to administer; discrete item and often in-house designed tests such as cloze measures or C-tests; and participant self-ratings, among others. However, along with this wide variety of measures, a number of researchers have highlighted and criticized the lack of consistency in measurement or even the lack of any measurement of proficiency in SLA research. On two occasions, Thomas (1994, 2006) has surveyed published SLA studies and demonstrated that proficiency is often inadequately controlled in SLA research, being

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

measured more often by institutional status or impressionistic judgment than either standardized or in-house assessments. More recently, both Tremblay (2011) and Hulstijn (2012) conducted surveys of published SLA research to assess whether and how proficiency has been measured. These studies highlight the gamut of proficiency measures or estimates being used by SLA researchers as well as the lack of measurement in many cases.

Tremblay (2011) surveyed SLA studies from 2000 to 2008, including 144 studies from three well-regarded journals: *SSLA* (76 studies), *Second Language Research* (61 studies), and *French Language Studies* (7 studies). Of these 144 studies, only about 37% (53 of the 144 studies) measured proficiency via an independent test. In these 53 studies, proficiency measurements that researchers reported using included (a) section(s) of standardized proficiency or placement tests (40%),[1] (b) cloze tests or C-tests (30%), (c) oral interviews or accent ratings (26%), or (d) other (23%).[2] In the remaining 63% (91 of the 144) of studies reviewed, *no independent proficiency assessment* was administered. Instead, proficiency was usually estimated via some other means available to researchers: (a) classroom level or years of instruction (60%); (b) length of residence in the L2 environment (24%); (c) existing proficiency scores (such as TOEFL scores, which might or might not be recent) (19%); (d) self-ratings (8%); or (e) other (15%). In 14% of the studies (13 out of 91), proficiency was not even estimated.

Hulstijn (2012) reviewed 140 papers published from 1998 to 2011 in *Bilingualism: Language and Cognition.* These were studies in which, in his judgment, "the assessment of [language proficiency] as an independent or moderating variable is a necessary or preferred requirement" (p. 423). The studies were divided chronologically into an earlier (volumes 1–7; 1998–2004; 37 papers) and later (volumes 8–14; 2005–2011; 103 papers) set. In the earlier set, L2 proficiency was objectively assessed in 19% of the studies; in the later set, 54% used an objective measure. Measurements used varied from proficiency certifications (e.g., TOEFL) to translation tests, panel judgments, and cloze or C-tests. Thus, although the later studies showed progress in terms of percentage using an objective assessment of proficiency, the situation still leaves a good deal of room for improvement. In fact, the recent reviews by both Tremblay and Hulstijn suggest that improvements to measuring proficiency in SLA research are sorely needed, in terms of both assessing proficiency at all and ensuring consistency of measurement across studies and across languages, and researchers have called for the adoption of standardized and validated measurements that can be used for research purposes (Norris & Ortega, 2012; Tremblay, 2011).

In addition to performing the survey of proficiency assessment in L2 research mentioned previously, Tremblay (2011) examined the validity, reliability, and practicality of a cloze test (Tremblay & Garrison, 2010) for assessing French L2 proficiency, observing that "it is much more difficult and time-consuming to train multiple raters to reliably evaluate

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

oral proficiency than to design and administer a cloze test" (p. 346). Although the cloze test was found to have reasonable validity, reliability, and practicality and thus was taken to be a good candidate for providing a measure of L2 proficiency, one acknowledged limitation of the measure is that it does not assess the spoken modality. Thus, Tremblay suggested that an aural/oral test used in conjunction with the cloze test would provide a better picture of participants' proficiency. One promising possibility for such a test seems to be an elicited imitation task.

Elicited Imitation

Elicited imitation tasks require examinees to listen to and then repeat a stimulus, usually at the sentence level. They have a history of use for language testing in both L1 (e.g., Slobin & Welsh, 1968/1973) and L2 (e.g., Naiman, 1974; Savignon, 1982) research. The EIT investigated here was developed by Ortega et al. (1999) and consists of a series of 30 sentences of increasing length to be repeated, meant to tap global (oral) proficiency. The general rationale for using an oral imitation test to measure proficiency is that speakers should only be able to accurately imitate sentences they have comprehended and parsed through their developing grammars (Bley-Vroman & Chaudron, 1994; Vinther, 2002; see Mozgalina, 2015, for additional discussion of the how the underlying constructs of proficiency are claimed to be tapped by elicited imitation tasks).

Vinther (2002) gives a brief review of elicited imitation as a measure of proficiency and offers cautions about the design of such tasks. One of the frequently raised points of concern regarding elicited imitation when used as a proficiency measure is that if stimuli are too short and/or if they are repeated back immediately, they may be "parroted" back rather than processed and understood by the learner and thus may not reflect the learner's proficiency. Thus, sentences should be long enough to be outside the bounds of echoic repetition, and there should be an appropriate time delay before the repetition. In addition, some elicited imitation tasks use ungrammatical utterances in order to examine whether these are imitated as is or corrected, the underlying assumption being that speakers will unconsciously correct such sentences even without being asked to correct them. However, as Vinther points out, interpretation of such results is not clear (especially if task instructions are not clearly reported by researchers), in that ungrammaticalities may be repeated either because participants do not in fact recognize the ungrammaticality or in an effort to reproduce what was heard (which may correspond to instructions).

Recently, a number of SLA researchers have turned to elicited imitation as a test to measure implicit knowledge, especially since R. Ellis (2005) used an oral elicited imitation task (with both grammatical and ungrammatical

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

sentences) as part of a battery of L2 assessments. He found that this task, along with a timed grammaticality judgment and oral narration task, loaded onto a factor that he attributed to implicit knowledge. A metalinguistic knowledge test and the ungrammatical sentences of an untimed grammaticality judgment test, on the other hand, loaded onto a second factor, which he deemed explicit knowledge. In Ellis's elicited imitation task, participants are asked to give a judgment regarding each sentence (agree/disagree/unsure about the content) and then repeat the sentence in correct English. Scoring focuses only on correct provision of the targeted item and is then converted into a percentage correct (out of 34 items). Since then, other researchers (e.g., Bowles, 2011) have used similar tests for such purposes.

Ortega and colleagues (1999) had a different aim in developing their EITs: to develop parallel tasks in several languages that could facilitate global (oral) proficiency measurement and comparability across studies and across languages. They chose an aural/oral measure because it does not require literacy skills, and it engages, in real time, several related aspects of language comprehension: lexical, grammatical, phonological, and prosodic aspects. Finally, it should be noted that although some other elicited imitation tasks add meaning-based judgments to the task, Ortega and colleagues reasoned that this likely adds another cognitive dimension to what is required of the participant, and they chose a simple oral repetition task.

Given that the task examines only aural and oral skills in the repetition of relatively straightforward sentences, it appears to align quite closely with Hulstijn's BLC; this version of the EIT also appears to match his description of "tests assessing participants' ability to process simple oral utterances receptively or productively" (2012, p. 429), taken to tap BLC. Similarly, Mozgalina (2015) argues that this type of EIT reflects a construct she calls oracy, defined as "learners' knowledge of core vocabulary, grammar, and phonology in combination with automated ability for their use in L2 oral perception and production *at the sentence level*" (p. 40). Oracy as so defined is thus a construct very closely related to aural/oral proficiency, but Mozgalina intends that this term will limit the prescriptive use of the EIT beyond the SLA research setting and therefore will not have unintended pedagogical consequences for learners. Thus, the EIT as designed and implemented by Ortega et al. appears well suited to tap L2 aural/oral proficiency; whether the measured construct is closer to BLC or oracy or whether these can be distinguished is addressed in the discussion.

In Ortega and colleagues' (1999) original Spanish EIT, participants were instructed (aurally) in English—the participants' L1—to try to repeat exactly what they heard, and to repeat as much as they could. They were given practice sentences in English, which were ordered by increasing length, from 5 to 17 syllables. The following Spanish assessment stimuli consisted of 30 sentences with 7 to 17 syllables, also in order of increasing length. The sentences included a variety of grammatical structures and

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

levels of complexity. After each sentence, there was a 2- to 3-s pause, followed by a tone that cued the oral repetition of the sentence. The length of the pause was selected based on Chaudron and Russell (1990), who reviewed two available studies (Grigg, 1986; McDade, Simpson, & Lamb, 1982). They concluded that immediate repetition may overestimate participants' linguistic abilities (due to echoic repetition), and a delay of 5 s might allow for more controlled processing strategies to be accessed; thus a delay of about 3 s might be appropriate. The sentences were tape recorded in order to ensure consistent timing across subjects. Thus, given the aural presentation, the brief time delay, and oral imitation, the test should tap both aural comprehension and oral production of the L2. The participants were instructed to repeat as much as they could; this was intended to encourage them to respond even if they did not understand the entire sentence and thus allow for information to be captured about participants' proficiency levels if they were able to repeat some but not all of an item. Thus, the test content and response processes of participants can be argued to support the validity of the EIT for measuring aural/oral proficiency (or oracy).

Results reported in two validation studies with L2 Spanish learners were promising. In Ortega (2000), 16 fifth-semester college learners of L2 Spanish were tested at the beginning and end of the semester. Reliability was measured as Cronbach's alpha of .95 (beginning) and .96 (end). The correlation of the EIT with ratings on the simulated oral proficiency interview (SOPI; see subsequently for more information) administered at the end of the semester (n = 15) was Pearson r = .87. The findings of Ortega, Iwashita, Norris, and Rabie (2002) were similar. In that study, 28 fourth- and sixth-semester college learners of L2 Spanish were tested. Reliability was measured as Cronbach's alpha of .97, and the correlation with SOPI ratings was Pearson r = .88. Moreover, the findings of Ortega et al. (2002) confirmed that increasing length of sentences corresponded to increasing item difficulty, as anticipated. This provides additional evidence for validity of the test based on its internal structure.

Note that similar high reliability has been reported for the parallel EITs developed by Ortega et al. for other languages (Cronbach's alpha .92–.97 for German, English, and Japanese; Ortega et al., 1999). Moreover, these EITs were also shown to have some degree of concurrent validity with other measures of proficiency (i.e., with other measures made at the same time purported to tap proficiency) and were able to discriminate among learners at two distinct classroom levels. Additional validation studies have recently been conducted for EITs in French (Tracy-Ventura, McManus, Norris, & Ortega, 2014) and Mandarin (Wu & Ortega, 2013). Note that these studies define global oral proficiency based on Hulstijn's definition of BLC, and this is the construct that their EITs are claimed to measure. These studies establish some degree of correlation of the EIT with other measures that reflect proficiency. Tracy-Ventura et al.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

report a correlation of $r = .78$ with end-of-year grades, and Wu and Ortega report *r*s of between .48 and .53 with various measures of complexity, accuracy, and fluency on an oral narrative task. However, these studies examined a narrow range of learners (in Tracy-Ventura et al., intermediate learners) or only low- vs. high-level learners (Wu & Ortega). Similarly, the Spanish validation studies by Ortega and colleagues tested a limited population of learners (from fourth to sixth semester; see Mozgalina, 2015, for a more thorough review of EIT validation studies).

Thus, there is still a need for studies to examine the validity and reliability of EITs more thoroughly. In particular, converging evidence regarding validity of the instrument from a variety of types of evidence is appropriate to mount a validity argument for this use of the EIT (see Kane, 2006, and Mozgalina, 2015, for discussion of validity). Validity entails both the interpretation and prescriptive uses of the test, and evidence for validity can come from the test content itself, the response process of the participant, the internal structure of the test, the test's relation to other variables, and the consequences of the test.

Motivation for the Current Study

In comparison to previous studies by Ortega and colleagues, which investigated the external validity and internal reliability of the Spanish EIT within a relatively narrow range of L2 Spanish experience (fourth to sixth semester), the current study investigates these same constructs across a broad spectrum of L2 Spanish learners (from beginning to very advanced) and thus examines whether it may be a useful tool for measuring proficiency in SLA research in a broader population of L2 Spanish learners. Moreover, with regard to validity, the study not only examines the degree to which the EIT correlates with another validated and standardized measure of proficiency (a type of evidence for validity known as criterion-related validity evidence) but also explores the EIT's validity through examining its relationship with a number of (more readily available) self-reported variables often used by researchers to gauge proficiency. Thereby, it also examines whether any of these may reliably serve as a proxy for oral proficiency measurements. Further, the study explores the degree to which the EIT discriminates among learners across the broader spectrum as well as within experience-level groups of L2 learners, both to assess its potential usefulness for SLA researchers and to provide additional evidence of the test's validity. The present study therefore offers evidence that can be used to mount a validity claim for the use of EITs in SLA research, both with regard to the test's relation to other variables (especially the SOPI) and with regard to the internal structure of the test.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

Thus, the research questions motivating the current study were the following:

1. What is the external validity of the EIT as a proficiency measure (as measured by correlation with SOPI scores)?
2. What is the external validity of other language background variables and self-ratings as proficiency measures (as reflected by correlation with EIT and SOPI scores)?
3. What is the internal reliability of the EIT?
4. What is the discriminatory power of the EIT among learners, and what is the difficulty of each item?
5. To what extent does the EIT discriminate within a spectrum of L2 experience levels (from low to very advanced)?
6. What is the EIT's utility for clustering L2 learners across these experience levels?

METHOD

Participants

Participants were 37 native English speakers who were L2 learners. They were recruited at three broad levels of experience. The low-experience L2 speakers (n = 15; 8 females; mean age = 19.0, SD = 0.8) were enrolled in second- to fourth-semester university-level Spanish classes and had no immersion experience. The advanced-experience L2 speakers (n = 15; 11 females; mean age = 23.4, SD = 3.4) had several semesters of formal Spanish (mean of 6.7 semesters of university-level Spanish, SD = 1.6) and 1–2 semesters of immersion experience (mean = 9.6 months, SD = 3.5). The very-advanced-experience L2 speakers (n = 7; 4 females; mean age = 32.0 years, SD = 7.3) had at least 19 months of immersion experience (mean = 55 months, SD = 44). Participants were accepted into the aforementioned groups in the study as long as they met the following additional inclusion criteria: they were between the ages of 18 and 45; had at least some college education; were right-handed; had no known developmental, neurological, or psychiatric disorders; and had normal or corrected-to-normal hearing and vision. See Table 1 for additional language background information, including age of exposure, months of immersion, years of formal Spanish, and self-ratings of proficiency. Participants at the low experience level were recruited through their Spanish classes. At the advanced level, participants were recruited through targeted emails to students who had studied abroad in the previous year. Very advanced speakers were recruited through personal contacts at institutions where having a high level of proficiency in a foreign language is common. All gave informed consent and were paid for their participation.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

Table 1. Participants' language background information (by experience group and total)

Group	Age of exposure	Months of Spanish immersion	Years of formal Spanish	Reading self-rating	Writing self-rating	Speaking self-rating	Understanding self-rating
Low (n = 15)	13.6 (3.4)	0.0 (0.0)	5.0 (3.2)	2.7 (0.6)	2.5 (0.7)	2.3 (0.5)	2.7 (0.7)
Advanced (n = 15)	12.3 (0.3)	9.6 (3.5)	8.7 (2.6)	3.7 (0.7)	3.4 (0.9)	3.6 (0.9)	3.8 (0.7)
Very advanced (n = 7)	15.0 (5.2)	55.1 (44.6)	6.1 (3.4)	4.6 (0.5)	4.3 (0.8)	4.7 (0.5)	4.7 (0.5)
Total (n = 37)	13.3 (3.5)	14.3 (27.5)	6.7 (3.4)	3.5 (0.9)	3.2 (1.0)	3.3 (1.1)	3.5 (1.0)

Note. Means are presented with *SD*s in parentheses. Self-ratings: 1 = lowest proficiency (very poor); 5 = highest proficiency (nativelike).

Procedure and Materials

Participants first provided information via a background questionnaire, which elicited demographic, medical, and language background information as well as self-ratings of proficiency. These self-ratings were elicited on a Likert scale from 1 to 5, in which 1 was defined as "lowest proficiency (very poor)" and 5 as "highest proficiency (nativelike)." Self-ratings were given for reading, writing, speaking, and understanding of Spanish. Next, participants performed Spanish production and comprehension tasks not reported here.

Next, they performed the Spanish EIT. The test was virtually identical to the one described previously, developed by Ortega et al. (1999), with only a few changes. First, one verb that could be problematic in some dialects (*coger*) was replaced (with *atrapar*) in consultation with Ortega. Second, the test (including the instructions) was rerecorded digitally, with a female L1 speaker of American English recording the instructions and English practice stimuli, and a female L1 speaker of Ecuadorian Spanish recording the Spanish sentences. In addition, the timing was standardized. The pause after each sentence before repetition was set at 2 s, and the ringtone lasted 0.5 s, for a total of 2.5 s between the end of each sentence and the window for the start of the repetition. This timing was chosen in order to replicate closely the original tape-recorded test while ensuring that timing was consistent across sentences.

Thus the EIT began with aural instructions and 5 practice sentences (with 5–15 syllables, increasing in length) in English. Then 30 unrelated Spanish sentences (with 7–17 syllables, increasing in length) were presented aurally, one at a time. Each sentence was followed by a 2-s pause and then a 0.5-s ringtone to cue the repetition. Participants' responses were digitally recorded and later scored. Scoring followed the Ortega (2000) criteria and was based on the quantity and quality of idea units that were repeated back (4 points maximum per sentence for a perfect repetition; see Table 2 for scoring criteria), with a maximum score on

Table 2. Spanish EIT rating criteria (abbreviated)

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/S0272263115000443 Published online by Cambridge University Press

the test of 120. Each test was independently scored by two raters, and any differences were reconciled.[3] The test takes about 10 min per participant to administer and an approximately equal time to score.

Example stimuli are presented subsequently (the numbers represent the sentence number of each item in the 30-item test).

1. Quiero cortarme el pelo.
4. Él se ducha cada mañana.
6. Dudo que sepa manejar muy bien.
7. Las calles de esta ciudad son muy anchas.
12. El chico con el que yo salgo es español.
13. Quiero una casa en la que vivan mis animales.
15. Ella sólo bebe cerveza y no come nada.
16. Me gustaría que el precio de las casas bajara.
22. El gato que era negro fue perseguido por el perro.
28. El examen no fue tan difícil como me habían dicho.
29. ¿Serías tan amable de darme el libro que está en la mesa?
30. Hay mucha gente que no toma nada para el desayuno.

Finally, participants were assessed on Spanish oral proficiency via the SOPI (Center for Applied Linguistics, Washington, DC).[4] Although it is acknowledged that neither this nor any proficiency measure is perfect, the SOPI served as the yardstick for measuring oral proficiency in this study both because it is one of few accepted and validated standardized measures[5] and because, although more time consuming, it could feasibly be used in this research setting. Created by test developers at the Center for Applied Linguistics (CAL) and foreign language educators, the SOPI is designed to elicit a representative performance sample of speech in a relatively short time. This measure taps a mixture of both BLC and HLC as defined by Hulstijn, given that to achieve higher levels formal education is evidently required (e.g., some test items targeting the superior level require beginning a formal speech with accuracy and appropriateness). As noted previously, all participants in the study had at least some college education.

The SOPI is based on the ACTFL speaking proficiency guidelines and elicits spoken language in a variety of simulated situations, such as describing activities, narrating a story, apologizing, and convincing others. Each task is presented to participants both aurally and in writing. The SOPI begins with a warm-up and then leads participants through a variety of tasks: picture-based tasks (giving directions, describing activities in a familiar setting, and telling a story), topic-based tasks (describing a procedure, presenting advantages and disadvantages, explaining and defending a point of view, and describing what would happen in a hypothetical situation), and situation-based tasks (giving advice to a friend, apologizing for having offended someone, and making a formal presentation to a group). For example, see the following sample SOPI item,

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

in which participants are given 15 s to plan and 1 min to answer: "Imagine that your family is hosting an exchange student from Mexico. The exchange student, Alma, tells you that she has a toothache and asks you how to make an appointment to see a dentist. After Alma asks her question, *explain to her how to make an appointment to see a dentist*. (15 sec / 1 min)"

Participants' SOPI assessments were digitally recorded and subsequently rated independently by two raters;[6] if these two ratings did not match, the test was independently rated by a third rater to obtain a final rating. The SOPI ratings, which can range from lower-than-novice high to superior, were converted to numerical scores following Henning (1992). See Table 3 for the range of possible SOPI ratings and their numerical conversions. The SOPI takes about 45 min to administer and approximately the same time, or a bit less, to rate.

ANALYSES

Statistical analyses were chosen to answer the aforementioned research questions. First, descriptive statistics for the EIT and SOPI were selected in order to visualize distributions of scores across the different experience levels for the EIT, the SOPI, and language background variables of interest.

Second, Pearson correlations were chosen to address Research Questions 1 and 2:

1. What is the external validity of the EIT as a proficiency measure (as measured by correlation with SOPI scores)?
2. What is the external validity of other language background variables and self-ratings as proficiency measures (as reflected by correlation with EIT and SOPI scores)?

With regard to Research Question 2, in particular, correlations would be performed among the EIT, the SOPI, and (a) self-ratings of proficiency

Table 3. SOPI ratings and numerical conversion (Henning, 1992)

Rating	Numerical conversion
Superior	3.0
Advanced high	2.8
Advanced mid	2.3
Advanced low	2.1
Intermediate high	1.8
Intermediate mid	1.3
Intermediate low	1.1
Novice high	0.8
< novice high	—

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

and (b) three self-reported language background variables—age of initial exposure (with which there might be an inverse relationship), amount of immersion experience, and amount of formal Spanish training.

Third, a series of item analyses were chosen to address Research Questions 3–6:

3. What is the internal reliability of the EIT?
4. What is the discriminatory power of the EIT among learners, and what is the difficulty of each item?
5. To what extent does the EIT discriminate within a spectrum of L2 experience levels (from low to very advanced)?
6. What is the EIT's utility for clustering L2 learners across these experience levels?

Internal reliability (Research Question 3) would be measured via Cronbach's alpha reliability coefficient. Item analyses following Brown (2000, 2003) and Tremblay (2011), explained fully subsequently, were selected to address Questions 4 and 5, and a *k*-means cluster analysis was chosen to explore Question 6. All statistical analyses were performed with SPSS (Version 21) and Excel 2010.

RESULTS

Descriptive statistics for the SOPI and EIT are given in Table 4, both within each experience-level L2 group and across all participants. As can be seen, the range of scores on the EIT is wide (ranging from 13 to 119 out of 120), with some overlap in score ranges across groups with consecutive experience levels. The EIT scores were examined for outliers (via SPSS Explore's boxplot output, as well as by examining whether there were any scores more than 2.5 *SD*s from the mean on the EIT), and no outliers were identified. The SOPI scores also showed a wide range of performance across all participants (from 0.8 to 3.0, out of 3.0—i.e., from novice high to superior) with some overlap across consecutive levels of the three broad L2 groups. Skewness and kurtosis values for the SOPI and EIT, both within and across groups, are within the range of less than 2 *SE* of the measure, indicating relatively normal distributions for the measures (Brown, 1997).

In order to answer Research Question 1 (What is the external validity of the EIT as a proficiency measure [as measured by correlation with SOPI scores]?), correlations were calculated between these two assessments. Results indicated that the SOPI and the EIT were indeed very highly correlated, at Pearson $r = .911$. (See also Figure 1, which plots each participant's SOPI score against his or her EIT.) These results suggest that the EIT has high external validity, comparable to the SOPI.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

Table 4. Descriptive statistics for the EIT and SOPI (by group and total)

	SOPI				EIT			
Group	Mean (*SD*)	Range	Skewness (*SE*)	Kurtosis (*SE*)	Mean (*SD*)	Range	Skewness (*SE*)	Kurtosis (*SE*)
Low (*n* = 15)	1.18 (0.21)	0.8 (NH)–1.8 (IH)	1.552 (0.580)	5.050 (1.121)	47.7 (19.7)	13–85	0.327 (0.580)	–0.180 (1.121)
Advanced (*n* = 15)	2.28 (0.41)	1.3 (IM)–2.8 (AH)	–0.594 (0.580)	0.972 (1.121)	96.5 (16.2)	61–116	–1.150 (0.580)	0.294 (1.121)
Very advanced (*n* = 7)	2.81 (0.24)	2.3 (AM)–3.0 (S)	–1.790 (0.794)	3.693 (1.587)	109.3 (8.2)	99–119	0.018 (0.794)	–2.136 (1.587)
Total	1.94 (0.73)	0.8 (NH)–3.0 (S)	0.035 (0.388)	–1.561 (0.759)	79.1 (31.3)	13–119	–0.488 (0.388)	–1.127 (0.759)

Note. Maximum scores were 120 for the EIT and 3.0 for the SOPI. NH = novice high; IM = intermediate mid; IH = intermediate high; AM = advanced mid; AH = advanced high; S = superior.

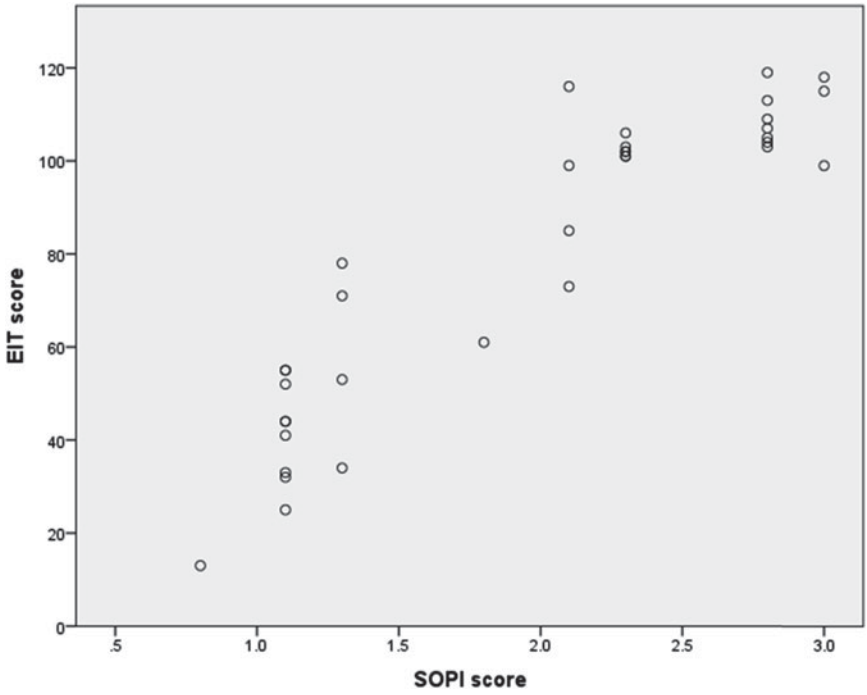


Figure 1. Scatterplot of EIT scores against SOPI scores.

Next, Research Question 2 was addressed: What is the external validity of other language background variables and self-ratings as proficiency measures (as reflected by correlation with EIT and SOPI scores)? In order to assess the external validity of self-ratings of proficiency, correlations were performed for the four self-ratings of proficiency (reading, writing, speaking, and understanding) with both the SOPI and EIT. All four self-ratings were significantly correlated with both the SOPI and the EIT (see Table 5), though the correlations were more modest than those carried out between the SOPI and EIT measures. Unsurprisingly, given that the EIT and the SOPI are largely measures of oral production, self-ratings of speaking had the highest degree of correlation with both the SOPI and the EIT (r = .752 and .737, respectively). Given that this correlation was higher, whereas the correlations with the listening rating (which was next in terms of degree of correlation with the SOPI and EIT) as well as the reading and writing self-ratings were lower, this lends both convergent and discriminant evidence of validation to the EIT.

Then correlations were performed between the SOPI, the EIT, and three language background variables: age of exposure, months of immersion experience, and years of formal Spanish study. The results (see Table 6) revealed significant though weaker correlations between

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

Table 5. Pearson correlations between the SOPI, the EIT, and self-ratings

	SOPI score	EIT score	Reading self-rating	Writing self-rating	Speaking self-rating	Understanding self-rating
SOPI score	1	.911**	.648**	.580**	.737**	.716**
EIT score		1	.624**	.592**	.752**	.734**
Reading self-rating			1	.842**	.811**	.814**
Writing self-rating				1	.851**	.739**
Speaking self-rating					1	.808**
Understanding self-rating						1

** $p < .01$.

the amount of immersion experience and both the SOPI and EIT ($r = .572$ and .433, respectively), as well as between the total number of years of formal Spanish training and the EIT ($r = .371$) but not the SOPI. There was no significant correlation between age of exposure and either the EIT or the SOPI. Thus it was found that with the SOPI as the baseline, the EIT was by far the most strongly correlated measure and thus had the highest degree of external validity for measuring oral proficiency as tapped by the SOPI. In addition, self-ratings were a better proxy for quantifying oral proficiency than either the amount of immersion experience or formal study of the language for this sample of learners. Finally, age of acquisition was not related to proficiency scores for these learners.

Next, a series of item analyses were performed to address Research Questions 3–5. Research Question 3 (What is the internal reliability of the

Table 6. Pearson correlations between the SOPI, the EIT, and language variables

	SOPI score	EIT score	Age of exposure	Months of Spanish immersion	Years of formal Spanish
SOPI score	1	.911**	—	.572**	—
EIT score		1	—	.433**	.371*
Age of exposure			1	.487**	–.821**
Months of Spanish immersion				1	—
Years of formal Spanish					1

Note. The dash indicates correlations that are not significant.
** $p < .01$.
* $p < .05$.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

EIT?) was addressed by calculating Cronbach's alpha reliability coefficient for the EIT, which was .981. This indicates an excellent level of internal consistency for the items in the test. Additional item analyses were performed to answer Research Question 4 (What is the discriminatory power of the EIT among learners, and what is the difficulty of each item?).[7] First, mean scores for each item (out of 4) were calculated across all participants. These are plotted from lowest to highest in Figure 2. As can be seen, the mean scores appear to fall almost entirely within the upper half of the possible range. This is likely due to the fact that more than half of the participants had quite high levels of proficiency (advanced and very advanced).

Following the procedures outlined in Brown (2000, 2003), discrimination indexes were then calculated. Mean scores on each item were calculated for the top (n = 12) and bottom (n = 12) third of participants on the EIT. Next, these mean scores were converted to proportions (given that each item had a possible score of 4), and item discrimination indexes were calculated for each item by subtracting the mean (proportion) accuracy on that item for the bottom third of participants from the mean accuracy on that item for the top third. The Appendix presents the details of this analysis: For each item, the mean score for the top and bottom third of participants as well as the item discrimination index is

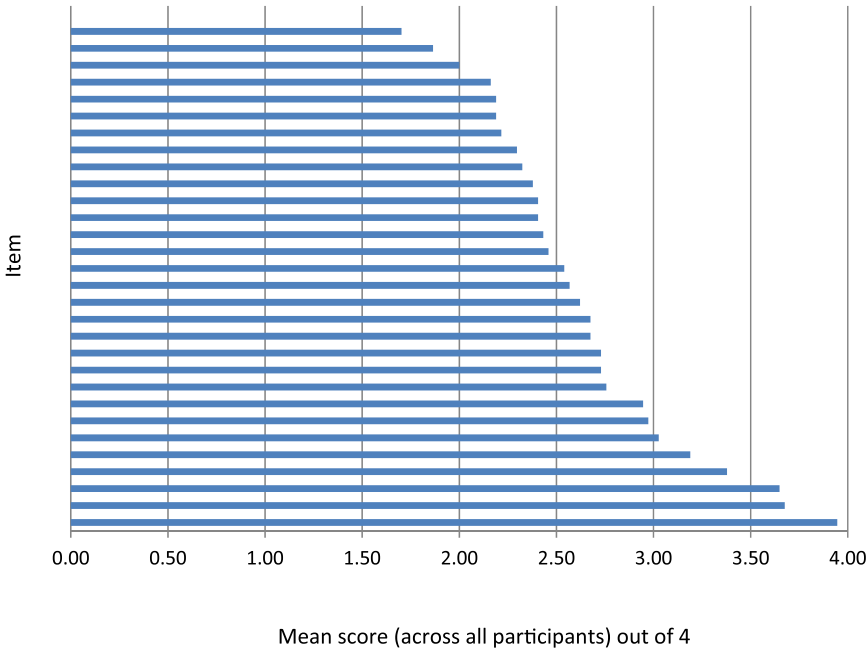


Figure 2. Item difficulty of the 30 EIT items.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

given. Table 7 gives a summary of this analysis, presenting the mean scores and mean discrimination indexes for the top and bottom third of participants, averaged across all items. As can been seen in both Table 7 and the Appendix, there is a range of mean scores, all discrimination index values are positive, and only one item (Item 4) had a discrimination index below .20. This indicates that, with the exception of this item, the EIT items do in fact discriminate among the learners examined here. The correlations and item analyses taken together suggest that the Spanish EIT is a reliable tool that assesses proficiency in a way similar to the SOPI across a broad spectrum of L2 learners such as those examined here, and that it is a tool that is able to discriminate well among a broad range of L2 learners.

In order to assess Research Question 5 (To what extent does the EIT discriminate within a spectrum of L2 experience levels [from low to very advanced]?), item analyses were performed separately for each a priori experience group (see Table 8), though it should be noted that the very-advanced-experience group especially is rather small (n = 7). Reliability was high within each group, with the lowest Cronbach's alpha value being .899. However, within the advanced and the very advanced groups (recall, as seen in Tables 3 and 4, that the advanced group had an average SOPI rating of around advanced mid, and the very advanced group averaged around advanced high), there were some items that did not discriminate among learners in these groups—2 items for the advanced group and 8 for the very advanced group. As was done across all participants, item discrimination indexes were also calculated within each group, in order to examine the extent to which the EIT appears to be able to distinguish proficiency levels among members of these experience-level groups. A summary of the item discrimination indexes and mean item scores on the EIT for the top and bottom third of each participant group are reported in Table 8. As with all learners, the discrimination indexes were all positive. Compared to discrimination indexes across all participants (Table 7), mean item discrimination indexes within each group (Table 8) were lower, which is to be expected within a narrower range of L2 experience. Nonetheless, the mean discrimination index values were again higher than .20 for both the low and advanced groups

Table 7. Item score and discrimination index summary, across all items

Variable	*M*	*SD*	Range
Mean item score, top third	3.7	0.3	3.0–4.0
Mean item score, bottom third	1.3	0.9	0.5–3.8
Discrimination indexes	.58	.20	.04–.88

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

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

Table 8. Reliability and item discrimination, by experience group

Group	Cronbach's alpha	Mean item score, top third		Mean item score, bottom third		Mean item discrimination index	
		M	*SD*	*M*	*SD*	*M*	*SD*
Low (n = 15)	.950	2.3	0.8	0.9	0.9	.35	.16
Advanced (n = 15)	.942[a]	3.6	0.4	2.6	0.7	.26	.15
Very advanced (n = 7)	.899[b]	4.0	0.2	3.4	0.6	.15	.14

Note. The score range on each item was 0–4.
[a] Two items were excluded from analysis due to zero variation in this group.
[b] Eight items were excluded from analysis due to zero variation in this group.

(but not for the very advanced group), suggesting that the EIT is able to discriminate fairly well among learners sampled at low and advanced levels of experience, but not as well at very advanced levels. As pointed out by an anonymous reviewer, the lack of discrimination at the highest level is not surprising given that there is less to be discriminated at the highest levels of proficiency. Note that three of the speakers in this group were rated as superior and three as advanced high on the SOPI, with the final speaker being rated as advanced mid. Thus, 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.

Finally, in order to address Research Question 6 (What is the EIT's utility for clustering L2 learners across these experience levels?), a *k*-means cluster analysis was run in conjunction with an ANOVA (see Table 9). This procedure offers a systematic and data-driven way to identify subgroups within the sample, and it has been used in previous SLA research (e.g., Tremblay 2011). In this analysis, an increasing number of subgroups are specified—starting with two and adding one on each iteration—until the gain in *F* value produced by the ANOVA begins to decline (Calinski & Harabasz, 1974). In the present data, the analysis revealed that three was the optimal number of subgroups as defined by EIT scores, given that after this point the gains in the variance on the EIT explained by the clusters (quasi-*F* values) began to decrease.

Next, participants were regrouped according to their identified cluster membership, and descriptive statistics were calculated for the regrouped participants. The language background information and the distribution of both EIT and SOPI scores within these three clustered groups are summarized in Tables 10 and 11 (paralleling Tables 1 and 4, which present the data according to experience-level groups). As can be seen in Table 10, the language background variables and self-ratings of proficiency follow the expected direction of the clusters for increasing proficiency level (note that, as in the larger sample, age of exposure appears to be unrelated to proficiency, but all other variables follow the expected direction), further supporting the external validity of the EIT. In Table 11, it can be observed that, as expected from the high correlation of the EIT with the SOPI, SOPI scores also increase with increasing EIT scores; moreover, the ranges of SOPI scores at each level when compared with

Table 9. *K*-means cluster analysis on EIT scores

k (no. of clusters)	*df*	*F*	Gain
2	1, 35	136.192	—
3	2, 34	176.512	40.32s0
4	3, 33	181.235	4.273

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

Table 10. Participants' language background information, by EIT *k*-means cluster groups

Cluster level	Age of exposure	Months of Spanish immersion	Years of formal Spanish	Reading self-rating	Writing self-rating	Speaking self-rating	Understanding self-rating
Level 1 (*n* = 8)	13.3 (4.0)	0.0 (0.0)	5.0 (4.2)	2.8 (0.5)	2.4 (0.7)	2.1 (0.4)	2.4 (0.5)
Level 2 (*n* = 11)	13.1 (1.4)	2.45 (3.9)	6.2 (2.5)	3.1 (0.8)	2.9 (0.7)	2.9 (0.7)	3.3 (0.8)
Level 3 (*n* = 18)	13.8 (4.7)	27.9 (34.7)	7.8 (3.2)	4.1 (0.8)	3.8 (1.0)	4.1 (1.0)	4.2 (0.8)

Note. Means are presented with *SD*s in parentheses. Self-ratings: 1 = lowest proficiency (very poor); 5 = highest proficiency (nativelike).

Table 11. Descriptive statistics for the SOPI and EIT, by *k*-means cluster groups

	SOPI				EIT			
Group	Mean (*SD*)	Range	Skewness (*SE*)	Kurtosis (*SE*)	Mean (*SD*)	Range	Skewness (*SE*)	Kurtosis (*SE*)
Level 1 (*n* = 8)	1.94 (0.73)	0.8–1.3	–1.081 (0.752)	4.091 (1.481)	33.2 (10.5)	13–44	–0.974 (0.752)	0.833 (1.481)
Level 2 (*n* = 11)	1.48 (0.39)	1.1–2.1	0.702 (0.661)	–1.210 (1.279)	67.2 (12.6)	52–85	0.165 (0.661)	–1.558 (1.279)
Level 3 (*n* = 18)	2.59 (0.33)	2.1–3.0	–0.200 (0.536)	–1.696 (1.038)	106.8 (6.6)	99–119	0.719 (0.536)	–0.884 (1.038)

Note. Maximum scores were 120 for the EIT and 3.0 for the SOPI.

those by experience level (cf. Table 4) appear to show less overlap, suggesting that the EIT does a better job of grouping participants by proficiency than the predefined experience levels do. Finally, Figure 3 represents what percentage of each experience level was assigned to the different EIT clusters. As reflected there, the cluster analysis grouped all of the very-advanced-experience participants into the highest cluster, along with the majority of the advanced participants as well. The middle EIT cluster comprised participants from both the advanced- and low-experience groups, whereas the lowest EIT cluster consisted solely of members of the low-experience group. This analysis provides another converging source of evidence for the external validity of the EIT. Thus, for research in which groupings based on aural/oral proficiency or oracy are desirable, the EIT appears to be a good candidate for performing such groupings.

DISCUSSION AND CONCLUSIONS

The data analyzed here strongly suggest that the Spanish EIT is both internally reliable (as reflected in the Cronbach's alpha of .981) and an externally valid measure of aural/oral proficiency (or oracy). External validity is supported not only by the strong correlation with the SOPI (Pearson's *r* of .911 between the EIT and SOPI) but also by virtue of the correlations with self-ratings of proficiency, including the fact that these were stronger for speaking and understanding than reading and writing. Further evidence of validity is found in the EIT's (weaker) correlations

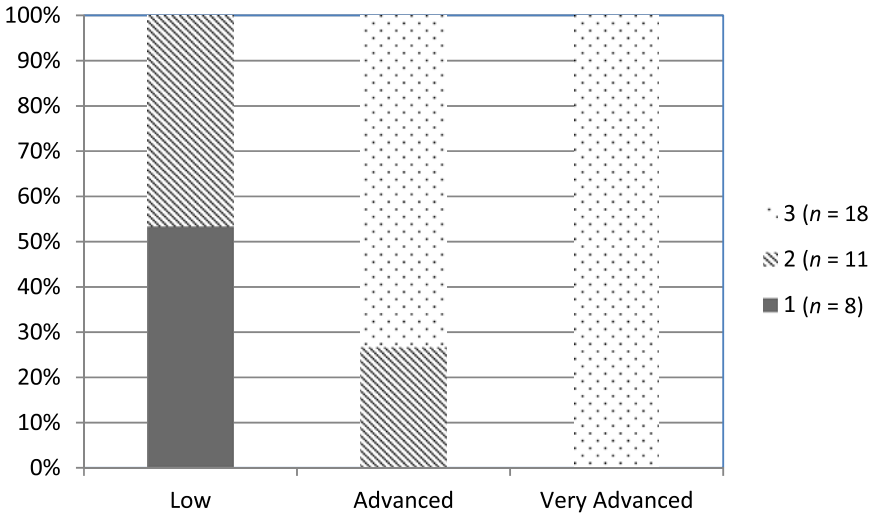


Figure 3. Percentage of participants in each experience group assigned to EIT cluster groups.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

with amount of immersion experience and amount of formal study. Likewise, clustering participants into groups via the EIT and examining resultant descriptive statistics of the language background data offer converging evidence that broadly supports the external validity of the EIT.

Notably, the Spanish EIT correlates with the SOPI much more strongly than do either self-ratings of proficiency or the three examined language background variables (age of exposure, amount of immersion, and amount of formal exposure). It is nonetheless interesting to note that self-ratings of speaking proficiency correlate with both the SOPI and the EIT (as reported previously, $r = .752$ and .737, respectively) better than might be expected, and better than the other language background variables. This suggests that self-ratings may in fact be a better proxy for measuring proficiency than researchers often assume, and that they are certainly better than amount of formal exposure (as reflected in the much weaker correlations). Clearly, however, the use of the EIT is more strongly supported given its much higher correlation with the SOPI.

The Spanish EIT also appears to serve to discriminate among and/or cluster the broad range of learners examined here—from novice high to superior—as well as to discriminate *within* the low (learners rated from novice high to intermediate high) and advanced (intermediate mid to advanced high) L2 learner groups examined here. The Spanish EIT's ability to discriminate among very advanced learners (ranging from advanced mid to superior proficiency levels), however, was not supported in the current analysis, though the sample size in that group was particularly small. These findings provide additional evidence for the validity of the Spanish EIT for SLA research.

Thus, based on the sample investigated here, the Spanish EIT appears to be a valid, reliable, and feasible instrument for measuring aural/oral proficiency in the L2 research setting at least for populations similar to the one investigated here. As to whether the evidence here suggests that the EIT taps BLC or oracy more specifically, the very high correlation between the Spanish EIT and the SOPI should be considered. Given that the SOPI involves production beyond the sentence level and the Spanish EIT correlates extremely highly with it, this suggests that the EIT may tap not only oracy—limited to the sentence level—but also more global aural/oral proficiency, or BLC as defined by Hulstijn (2011, 2012). This would make sense if the proficiency that allows learners to perform the EIT is associated with their abilities not only at the sentence level but also beyond.

The EIT may be appropriate for use in a research setting not only in order to more precisely characterize learners' oral proficiency, but also to group learners objectively and reliably, to identify outliers in preestablished groups, and/or to include proficiency as a factor or covariate in analyses (except possibly at very advanced levels of

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

proficiency, at which additional, more difficult items might better discriminate among learners). It is also important to point out the potential usefulness of including this measure in SLA research that employs more sophisticated techniques of analysis beyond simple groupings of participants—for example, in research using mixed models, growth curve analyses, and so on. Having more precise measurements of proficiency promises to allow for the analysis of the effects of proficiency (or to control this variable) in such research in a much more fine-grained manner than is available via such coarse groupings as class level.

An additional and important advantage of the EIT is that it is currently available in several languages. In addition to the original tests developed initially in German, English, Spanish, and Japanese, EITs have recently been developed for Mandarin Chinese (Wu & Ortega, 2013; Zhou, 2011), French (Tracy-Ventura et al., 2014), and Russian (Mozgalina, 2015); moreover, all of these are now available on IRIS (http://www.iris-database.org/). Elicited imitation tasks have also been developed in Korean and Portuguese, and these are expected to be available in IRIS shortly as well (Ortega, personal communication). The use of such parallel, reliable proficiency assessments in a variety of languages offers an important means for researchers to better understand and compare SLA studies conducted in different languages and to better understand the role of proficiency in SLA.

As always, the limitations of the current study should be kept in mind. First, the EIT measures only aural/oral proficiency. Thus its use for measuring global proficiency would be complemented by use of a measure in the written modality, such as a cloze measure, as suggested by Tremblay (Tremblay, 2011; Tremblay & Garrison, 2010). Second, in the current study, the EIT was examined for external validity primarily by using the SOPI as the yardstick; thus the tested external validity is largely limited to the validity of the SOPI. However, see Malone (2000) for a review of the SOPI, including research on the SOPI and validation studies, and note the converging sources of evidence for external validity detailed previously. Third, as always, the study is limited by the population sample that was obtained here. In particular, it is possible that participant recruitment especially above the low level of experience may have included a self-selection bias in terms of those learners who decide to continue studying the language (and studying abroad) and thus may limit the generalizability of the study if these learners are not representative of the larger population.

The current state of research in SLA has been shown to be lacking in terms of independent assessment of the crucial variable of L2 proficiency (e.g., Hulstijn, 2012; Tremblay, 2011). The present study sought to examine the potential usefulness for research of the Spanish EIT for assessing aural/oral proficiency in the L2 research setting. The results strongly supported

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

its reliability, validity, and feasibility for such use. Especially in complement with a validated and reliable written assessment such as Tremblay's cloze measure, EITs appear to offer researchers a valuable tool for assessing learners' proficiency, potentially across languages. Thus SLA researchers are called on both to continue validating these measures and to employ them in L2 research in order to facilitate comparisons across studies and to improve the methodological rigor and interpretation of SLA studies.

Received 22 May 2015
Accepted 3 November 2015
Final Version Received 22 October 2015

NOTES

1. These included both original and simplified versions of test sections. Examples given by Tremblay (2011) include the "Greek Language Proficiency Test, Michigan Test, Oxford Proficiency Test, Test of Adolescent and Adult Language, TOEFL—e.g., the reading comprehension section of the practice version—[and] Woodcock Language Proficiency Battery" (p. 342).

2. Note that, here and subsequently, the percentages add up to more than 100% because some studies used more than one measure.

3. Initial ratings between the two raters correlated at Pearson *r* of .984, although only 2 of the 37 ratings (5.4%) were an exact match. Note that after the initial ratings were made, the raters realized that they had not dealt with hesitations, false starts, and self-corrections in the same manner. Thus, after consulting with Ortega regarding the manner in which these were scored by Ortega and colleagues, a third rater checked for treatment of these—in short, participants were not penalized for them—with the rationale being that "memory is not perfect, but their grammatical knowledge to parse a given stimulus is in place if they can reconstruct meaning and/or form or both, even with difficulty (which is traceable in self-corrections, hesitations, or false starts)" (Ortega, personal communication). Discrepancies were thus resolved, and a final rating was determined.

4. A computer-adapted version of the test was administered (this adaptation was developed in the Brain and Language Lab at Georgetown University). It exactly replicated the SOPI but presented the visual and audio prompts via a self-timed PowerPoint presentation rather than via a booklet and master tape.

5. See Stansfield and Kenyon (1992a, 1992b).

6. Raters were trained via CAL's Multi-Media Rater Training Program and/or in CAL SOPI rating workshops.

7. An alternative means for assessing item quality would have been the use of item response theory analyses. See Ackerman, Gierl, and Walker (2003) and Thissen and Steinberg (2009), for examples.

REFERENCES

Abutalebi, J. (2008). Neural aspects of second language representation and language control. *Acta Psychologica*, *128*, 466–478.

Abutalebi, J., Cappa, S. F., & Perani, D. (2001). The bilingual brain as revealed by functional neuroimaging. *Bilingualism: Language and Cognition*, *4*, 179–190.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37–53.

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

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

Bowles, M. A. (2011). Measuring implicit and explicit knowledge: What can heritage learners contribute? *Studies in Second Language Acquisition*, *33*, 247–271.
Brown, J. D. (1997). Statistics Corner. Questions and answers about language testing statistics (skewness and kurtosis). *Shiken: JALT Testing and Evaluation SIG Newsletter, 1*(1), 20–23.
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 Evaluation SIG Newsletter*, *3*(2), 19–21.
Brown, J. D. (2003). Norm-referenced item analysis (item facilitation and item discrimination). *Shiken: JALT Testing and Evaluation SIG Newsletter*, *7*, 16–19.
Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*, 1–27.
Carrell, P. (1991). Second language reading: Reading ability or language proficiency? *Applied Linguistics*, *12*(2), 159–179.
Chaudron, C., & Russell, G. (1990). *The validity of elicited imitation as a measure of second language competence*. Unpublished manuscript. Manoa, HI: University of Hawai'i.
Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172.
Grigg, T. (1986). The effects of task, time, and rule knowledge on grammar performance for three English structures. *University of Hawai'i Working Papers in ESL*, *5*, 37–60.
Henning, G. (1992). The ACTFL oral proficiency interview: Validity evidence. *System*, *20*(3), 365–372.
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.
Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, *15*(2), 422–433.
Indefrey, P. (2006). A meta-analysis of hemodynamic studies on first and second language processing: Which suggested differences can we trust and what do they mean? *Language Learning*, *56*, 279–304. doi:10.1111/j.1467-9922.2006.00365.x
Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.
Kormos, J. (2000). The role of attention in monitoring second language speech production. *Language Learning*, *50*, 343–384.
Kotz, S. A. (2009). A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain and Language*, *109*, 68–74.
Malone, M. E. (2000). The simulated oral proficiency interview: Recent developments. In ERIC, *ERIC Digest*. Washington, DC: ERIC Clearinghouse on Languages and Linguistics. Retrieved from http://files.eric.ed.gov/fulltext/ED447729.pdf.
McDade, H. L., Simpson, M. S., & Lamb, D. E. (1982). The use of elicited imitation as a measure of expressive grammar. *Journal of Speech and Hearing Disorders*, *47*, 19–24.
Mozgalina, A. (2015). *Applying an argument-based approach for validating language proficiency assessments in second language acquisition research: The elicited imitation test for Russian* (Unpublished doctoral dissertation). Washington, DC: Georgetown University.
Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism*, *2*, 1–37.
Norris, J. M., & Ortega, L. (2012). Assessing learner knowledge. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 573–589). New York: Routledge.
Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners* (Unpublished doctoral dissertation). University of Hawai'i at Manoa.
Ortega, L., Iwashita, N., Norris, J., & Rabie, S. (2002, October). *An investigation of elicited imitation tasks in crosslinguistic SLA research*. Paper presented at the Second Language Research Forum Conference, Toronto.
Ortega, L., Iwashita, N., Rabie, S., & Norris, J. M. (1999). *A multilanguage comparison of measures of syntactic complexity* [Funded project]. Honolulu: University of Hawai'i, National Foreign Language Resource Center.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

Savignon, S. J. (1982). Dictation as a measure of communicative competence in French as a second language. *Language Learning*, *32*(1), 33–51.
Slobin, D. I., & Welsh, C. A. (1973). Elicited imitation as a research tool in developmental psycholinguistics. In C. Ferguson & D. I. Slobin (Eds.), *Studies of child language development* (pp. 485–497). New York: Holt, Rinehart & Winston. (Reprinted from *Working Paper No 10, Language Behavior Research Laboratory*, pp. 485–497, 1968, University of California, Berkeley.)
Stansfield, C. W., & Kenyon, D. M. (1992a). The development and validation of a simulated oral proficiency interview. *Modern Language Journal*, *76*(2), 129–141.
Stansfield, C. W., & Kenyon, D. M. (1992b). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, *20*(3), 347–364.
Steinhauer, K., White, E. J., & Drury, J. E. (2009). Temporal dynamics of late second language acquisition: Evidence from event-related brain potentials. *Second Language Research*, *25*(1), 13–41.
Thissen, D., & Steinberg, L. (2009). Item response theory. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The Sage handbook of quantitative methods in psychology* (pp. 148–177). Thousand Oaks, CA: Sage.
Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–336.
Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279–298). Amsterdam: John 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: 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.
Vinther, T. (2002). Elicited imitation: A brief review. *International Journal of Applied Linguistics 12*, 54–73.
Wu, S., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*(4), 680–704.
Zhou, S. (2011). *Learning to express motion events in L2 Chinese: A cognitive linguistic perspective* (Unpublished doctoral dissertation). Manoa, HI: University of Hawaii.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press

APPENDIX

MEAN SCORES AND DISCRIMINATION INDEX FOR EACH ITEM

	Mean score				Mean score		
Item	Top third	Bottom third	DI	Item	Top third	Bottom third	DI
1	4.00	2.67	.33	16	3.67	0.50	.79
2	4.00	2.92	.27	17	3.50	1.08	.60
3	3.33	2.42	.23	18	3.50	1.83	.42
4	4.00	3.83	.04	19	3.42	0.75	.67
5	3.92	1.42	.63	20	3.08	1.58	.38
6	3.92	1.33	.65	21	3.25	0.58	.67
7	3.92	1.00	.73	22	3.67	0.58	.77
8	3.42	0.67	.69	23	3.75	0.50	.81
9	4.00	3.17	.21	24	3.58	1.08	.63
10	4.00	2.17	.46	25	3.00	0.58	.60
11	3.58	1.08	.63	26	3.25	0.58	.67
12	4.00	0.50	.88	27	3.92	1.17	.69
13	3.75	1.33	.60	28	3.75	1.25	.63
14	3.08	0.50	.65	29	3.75	0.50	.81
15	3.83	1.08	.69	30	4.00	1.42	.65

Note. DI = discrimination index.

https://doi.org/10.1017/S0272263115000443 Published online by Cambridge University Press