Language Teaching (2022), **55**, 116–135
doi:10.1017/S026144482000021X

CAMBRIDGE UNIVERSITY PRESS

REPLICATION RESEARCH

Using elicited imitation to measure global oral proficiency in SLA research: A close replication study

Kevin McManus* and Yingying Liu

Department of Applied Linguistics, The Pennsylvania State University, University Park, USA
*Corresponding author: Email: kmcmanus@psu.edu

Abstract

We closely replicated Wu and Ortega (2013), who found that an elicited imitation test (EIT) reliably distinguished low-level from high-level language abilities among instructed second language (L2) learners of Mandarin Chinese. The original study sampled learners (1) from second-level courses to represent low-level language abilities and (2) from third-, fourth- and graduate-level courses to represent high-level language abilities. Results showed high-level learners outperformed low-level learners on the Mandarin EIT. Our close replication used Wu and Ortega's (2013) materials and procedures in order to understand (1) the extent to which this EIT can additionally distinguish between finer-grained language abilities and (2) the ways in which the broad grouping of language abilities in the high group may have contributed to the findings. Sixty-five instructed L2 learners from four instructional levels were assigned to one of three groups: Beginner (first-level courses), Low (second-level courses), High (third- and fourth-level courses). Consistent with the original study, our results showed clear between-group differences, indicating that the EIT can distinguish between both broad (beginner vs high) and finer-grained (beginner vs low, low vs high) language abilities. These results are discussed in light of the original study's findings with implications for proficiency assessment in second language acquisition (SLA) research.

1. Introduction

The ways in which research design and methodology can influence research outcomes have in recent years become an important focus of attention in our field (Plonsky, 2013, 2015; Hudson & Llosa, 2015; Phakiti, De Costa, Plonsky, & Starfield, 2018). Byrnes (2013) referred to this increased sensitivity as the 'methodological turn' (see also Ellis, 2015), noting that '[…] methodological issues inherently merit a certain level of attention inasmuch as they assure the quality of our work. But it appears that at this point in the development of applied linguistics, they demand a kind of professional scrutiny that goes directly to the core of what we do […]'. One area of applied linguistics research receiving this 'professional scrutiny' is the measurement of L2 proficiency (Pallotti, 2009; Tremblay, 2011; Hulstijn, 2015; Iwashita, 2018), defined in the field of SLA as 'learners' knowledge and automated ability for use of core vocabulary and grammar delivered with reasonably intelligible pronunciation and fluency' (Wu & Ortega, 2013, p. 681; see also Hulstijn, 2011, 2012). Reasons for why the measurement of L2 proficiency is challenging are two-fold, at least: (1) proficiency is a multi-faceted construct that is complex to operationalize and measure, and (2) there is considerable variation not only in the ways in which SLA research has measured proficiency, but whether it is measured at all (Thomas, 1994, 2006; Hulstijn, 2010, 2011, 2012; Leclercq & Edmonds, 2014; Brown, Plonsky, & Teimouri, 2018). There is even considerable variation in research designs that attribute critical roles to proficiency in understanding L2 learning. For example, studies have grouped learners by 'proficiency' using standardized language placement tests (*e.g., Diploma de Español como Lengua Extranjera,* Dussias, Valdés Kroff, Guzzardo Tamargo, & Gerfen, 2013), amounts of time spent learning the L2 (e.g., Slabakova, 2015), course grades (e.g., Laufer & Goldstein, 2004), and self-reports of proficiency (e.g., Dewey,

© The Author(s), 2020. Published by Cambridge University Press

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

Bown, & Eggett, 2012). Each of these measures operationalize and evaluate L2 proficiency in fundamentally different ways. As a field, therefore, our approaches to understanding proficiency and the tools we use for measuring it, if any, are varied. This state of affairs has led to major obstacles in how we appraise research findings, draw between-study comparisons, and advance knowledge and understanding about the role of proficiency in L2 learning.

In light of these concerns, SLA researchers have become increasingly attentive to the difficulties associated with operationalizing and measuring proficiency (Tremblay, 2011; Leclercq & Edmonds, 2014; Brown et al., 2018). One line of research that has sought to address this issue is the development and use of elicited imitation as a proficiency measurement tool (Ortega, Iwashita, Norris, & Rabie, 2002; Wu & Ortega, 2013; Tracy-Ventura, McManus, Norris, & Ortega, 2014; Yan, Maeda, Lv, & Ginther, 2016). As a testing method, EIT usually requires participants to listen to a stimulus and then repeat that stimulus verbatim (Underhill, 1987). This procedure is then repeated for a series of different stimuli that can differ on a number of dimensions, depending on the objective of test (e.g., greater syntactic complexity, increasing syllable length, performing an activity before repetition; for reviews, see Yan et al., 2016; Gass, 2018). An underlying claim of the EIT is that 'in order for the learner to be able to correctly imitate the target language structure, it must be part of the learner's interlanguage system' (Erlam, 2006, p. 467; see also Jessop, Suzuki, & Tomita, 2007; Rebuschat & Mackey, 2013; Tracy-Ventura et al., 2014).

A commonly referenced study on the effectiveness of EIT to measure L2 proficiency is Wu and Ortega (2013). In that study, participants were recruited from different instructional levels and EIT performance was correlated with performance on an oral narrative task. The complexity of the EIT stimuli increased over the course of the test by (1) increasing the number of syllables in each stimulus (from 7 to 19, in order) and (2) incorporating advanced-level or late-acquired syntactic linguistic features (e.g., subordination) and low-frequency lexical items. Results showed that instructed foreign language learners and heritage speakers of Chinese recruited from high-level courses (300-, 400-, and graduate-level) repeated EIT stimuli more accurately than those recruited from low-level courses (200-level). Syllable length provided an important explanation for this performance: 'the longer a sentence was, the harder it was to achieve a high score, and this relationship explained 74% of the variance in the item scores' (Wu & Ortega, 2013, p. 693). Performance in the oral narrative task patterned similarly: high-level learners' performance included higher frequencies of clauses, motion clauses, and motion verbs than low-level learners. Taken together, the Mandarin EIT appeared to successfully distinguish between low-level and high-level language abilities.

The design of Wu and Ortega's Mandarin EIT closely followed a unified set of design features that have been used in EITs for other languages, including English (Ortega et al., 2002), French (Tracy-Ventura et al., 2014), German (Ortega et al., 2002), Japanese (Ortega et al., 2002), Korean (Kim, Tracy-Ventura, & Jung, 2016), and Spanish (Ortega et al., 2002; Bowden, 2016; Solon, Park, Henderson, & Dehghan-Chaleshtori, 2019). These design features include, for example, that stimuli are ordered by syllable length and range from 7 to 19 syllables (or to 18 syllables in the German EIT and 17 syllables in the Spanish EIT, see Ortega et al., 2002). According to Ortega et al. (2002), this is because (1) '8-syllable items discriminate well between L2 grammatical processing abilities and pure repetition strategies', (2) '10-syllable items are too taxing to repeat without accessing grammatical processing', (3) '8- to 14-syllable items can be expected to be repeated with a minimum accuracy of 50% by mixed proficiency and mixed short-term memory capacity learners', and (4) 'most L2 learners reach their ceiling with 18-syllable items or longer' (Ortega et al., 2002, p. 1). As a result, the syllable length of a stimulus is understood to be an important determinant of complexity or processing difficulty (see also Gaillard & Tremblay, 2016). In addition, performance on the EIT is rated using a five-point scoring rubric that assigns scores based on meaning (e.g., complete meaning of stimulus is preserved, half of the meaning is preserved, see 'Method' for more detail with examples). Lastly, all EIT stimuli are merged into a single audio file and played without interruption so that (1) the amount of time between hearing the stimulus and repeating it and (2) the amount of time assigned for repetition

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

is the same for all participants. This contrasts with other EITs in which the researcher manually pauses the recording to allow for repetition (e.g., Erlam, 2006; Li, 2010).

Taken together, this collection of EITs arguably make an important contribution towards systematizing the ways in which proficiency can be measured by providing standardized and more transparent indexes for appraising research findings, drawing between-study comparisons, and advancing knowledge and understanding about the role of proficiency in L2 learning. In light of these potential advantages, it is important that the generalizability (i.e., external validity) of the EIT for measuring L2 proficiency be assessed (i.e., to what extent do these EIT findings apply in different settings, to different populations, and in different contexts, see Brewer & Crano, 2000; Porte & McManus, 2019). While steps can be taken to minimize threats to external validity (see Mackey & Gass, 2016), replication with intentional variable modifications (e.g., participants, contexts) can assess the validity of the EIT for measuring L2 proficiency. To date, no EIT has been replicated with the explicit aim to understand the nature and the stability of the original study's findings. Replication therefore has the potential to increase confidence in this research tool as well as to offer a strong foundation for future research. The present study contributes to this goal through a close replication of Wu and Ortega (2013).

2. Motivation for replication

The decision to closely replicate Wu and Ortega (2013; henceforth W&O) was partly motivated by growing interest in the use of EIT to measure L2 proficiency. According to Porte and McManus (2019, p. 73), a 'close' replication sees 'only one major variable … modified … it can be added or removed and all other are kept as constant as possible to be better able to single out the kind of influences each has on the dependent variable'. Since EITs in multiple languages exist, we specifically focused on the Mandarin EIT for four reasons.

First, only the Mandarin EIT measured L2 proficiency among instructed foreign language learners recruited from different curricular levels. In the other EITs, learners were recruited either from the same curricular level (e.g., Tracy-Ventura et al., 2014), only some of the learners were enrolled in foreign language classes (e.g., Bowden, 2016), or recruitment according to curricular level was unclear (e.g., Kim et al., 2016).

Second, in addition to completing the EIT, W&O correlated a range of lexical and syntactic oral production measures from an oral narrative task to understand 'the extent to which elicited imitation performances yielded reliable and valid evidence that correlated well with an L2 learner's strengths and weaknesses in different oral linguistic dimensions' (Wu & Ortega, 2013, p. 686). In other words, different measures of language ability accepted in the field of SLA were used to understand EIT performance.

Third, as evidenced by new journals and conferences in the field, research into the L2 learning of Chinese is growing (see Yuan & Li, 2019; e.g., *Chinese as a second language* [John Benjamins], *Chinese as a second language research* [De Gruyter], conference of the *Association for Chinese as a Second Language Research*). Evidence-based resources to support growing research in this area are therefore needed, including L2 proficiency measures for Chinese.

Fourth, specific to the original study's design, W&O's participants were recruited from lower-division courses (200-level courses) and upper-division courses (300-level, 400-level, and graduate-level courses). Those in the lower-division courses were assigned to the low group and those in the upper-division courses were assigned to the high group. Compared to the low group, therefore, language ability and classroom learning experience in the high group was more varied because these learners were sampled from three curricular levels. To address this potential influence on the findings, our close replication narrowed sampling in the high group to include students enrolled in 300- and 400-level courses only. This sampling modification allows us to understand the extent to which including graduate students in the high group, arguably the learners with the most advanced language abilities, contributed to W&O's findings (see also Solon et al., 2019, who addressed similar broad sampling

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

concerns in Bowden's (2016) Spanish EIT). In addition, we recruited participants from 100-level courses to create a beginner group (not part of W&O's design). Our rationale for these changes was to understand: (1) the ways in which the broad grouping of language abilities in the high group contributed to W&O's findings and (2) the extent to which W&O's EIT can additionally distinguish between finer-grained language abilities (by comparing beginner, low, and high groups).

3. The replication

In this replication, we note three exceptions from W&O. First, the replication involves foreign language learners only, but W&O examined performance among foreign language learners and heritage speakers. W&O's separate EIT analyses for the foreign language learner and the heritage speaker groups make it possible for the replication to investigate one of these L2 groups only. Second, we excluded graduate students from the high group of foreign language learners. Third, we added a beginner group by recruiting foreign language learners from 100-level courses. All other aspects of the original study remain the same: we used the same test materials, followed the same procedures, and carried out the same analyses as W&O.

4. Research questions

The original study included four research questions (RQ). We used three of W&O's research questions. W&O's RQ2 is not included because it investigated performance differences between foreign language learners and heritage speakers. RQ1 is the same as the original study except for a minor change that reflects our variable modification (underlining in RQ1 indicates our modification). The original study's RQ3 and RQ4 remain the same in this replication:

RQ1: To what extent do Chinese EIT scores distinguish between <u>three institutionally defined high, low, and beginner</u> language ability groups?

RQ2: What is the relationship between participants' performance on the Chinese EIT and their performance on an oral narrative task, evaluated in terms of three CAF-related measures: average numbers of clauses, motion clauses, and motion verb types?

RQ3: What features of the items might influence the outcomes observed for the EIT and help explain sources of varying item difficulty?

5. Method

5.1 Participants

Sixty-five foreign language learners of L2 Mandarin Chinese learners at a large public university in the United States participated in the study, sampled from the entire undergraduate Chinese language programme using two criteria: (1) language background and (2) curricular level. First, using W&O's background questionnaire and classification criteria, all participants were classified as foreign language learners because (1) participants' strongest language before age five was not Mandarin Chinese or another Chinese dialect, (2) participants did not have at least one parent with Mandarin Chinese or another Chinese dialect as their native or dominant languages, and (3) participants did not report exposure to Mandarin Chinese or another Chinese dialect at home. Second, learners were assigned to groups based on curricular level. Participants sampled from 200-level courses constituted the low group ($n = 20$) and learners in 300- and 400-level courses constituted the high group ($n = 25$, 13 participants from 300-level courses and 12 participants from 400-level courses). These closely match the sampling procedures of the original study except in two ways. First, we excluded graduate students from the high group. Second, we recruited learners from 100-level classes to create a beginner group ($n = 20$).

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

Table 1. Participant background information in the replication

		Years learning Chinese	Age started hearing Chinese	Length of time spent in Chinese-speaking country (weeks)
	N	Mean (SD)	Mean (SD)	Mean (SD)
Beginner	20	1.6 (1.9)	16.9 (6)	11.2 (47.9)
Low	20	2.9 (2.8)	14.9 (6.3)	17.7 (59.7)
High	25	5.8 (3.7)	13.2 (5.3)	33.1 (92.6)

Additional background information about the participants is shown in Table 1 (comparable data are not reported in W&O). Between-group effect size (ES) comparisons (Cohen's *d* with 95% CIs for *d*) indicated the following between-group differences:

- the high group had spent more years learning Chinese than the low group (medium ES, $d = 0.87$ [0.24, 1.47])
- the high group had spent more years learning Chinese and started hearing Chinese at a younger age than the beginner group (large ES, $d = 1.38$ [0.71, 2.01]); medium ES, $d = -0.66$ [−1.25, −0.04], respectively)
- no other between-group differences were meaningful (CIs for *d* included zero)[1]

The standard deviations (SDs) for length of time spent in a Chinese-speaking country indicate variation within each group. The ranges of these stays in a Chinese-speaking country are (1) 0–4 years in the beginner group, (2) 0–5 years in the low group, and (3) 0–8 years in the high group. Participants reporting very high amounts of time spent in a Chinese-speaking country were not removed because (1) the original study did not report implementing this procedure and (2) these varied experiences can be typical in Chinese language programmess. Furthermore, in the entire dataset, only five participants reported spending one year or more in a Chinese-speaking country (three participants in the high group, and one participant in each of the beginner and low groups).

5.2 Procedure and instruments

We used the same three instruments (available from IRIS, https://www.iris-database.org) and followed the same procedure as in the original study: oral narrative task, then the background questionnaire, and then the Chinese EIT. Completing these tasks took 15–20 minutes.

5.2.1 Oral narrative task

The same picture story from W&O was used to elicit information about participants' abilities to express motion events in Chinese. This task included twelve black and white pictures of a boy looking for his missing dog. Each picture illustrates a different motion event in order to elicit a variety of different motion events (e.g., a boy is walking his dog, the boy jumps on to the table, the boy runs out of the classroom).

Participants were given up to three minutes to look over the story before beginning their narration and were provided with the same instructions as in the original study: 'You will see 12 sequential pictures in this task. Your job is to tell a story in Chinese to describe what you see in as much detail as you can' (Wu & Ortega, 2013, p. 704). Each story was narrated one-to-one with the researcher. Participants were allowed to look at the story while narrating. All narrations were audio-recorded.

5.2.2 Chinese EIT

We used the same Chinese EIT (Zhou & Wu, 2009) as the original study (available from IRIS), which was designed following the design features described in Ortega et al. (2002). The Chinese EIT included

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

30 sentences (see Appendix A for stimuli), ranging from 7 to 19 syllables, and contained a variety of vocabulary and grammatical structures that followed as closely as possible Ortega et al.'s English EIT.

The Chinese EIT took ten minutes to complete and was administered by playing the EIT audio file. The audio file was never interrupted. Because the audio file included instructions for completing the task, all participants heard the same instructions as those used in the original study: 'Your task is to try to repeat exactly what you hear. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can.' Before repeating the sentence, there was a short pause (2.5 seconds) followed by a beep. For example, item three was the stimulus 这个城市街道很宽 'the city streets are wide'. The participant first heard the stimulus, which was followed by a pause of 2.5 seconds, then a beep, then the participant repeated the sentence. The 30 Chinese sentences were preceded by four English sentences to familiarize participants with the test format (i.e., waiting for the beep before repeating the sentence). The EIT was time pressured since the stimulus had to be repeated within a specified amount of time. For example, participants had two seconds to repeat item one (seven syllables long) and 2.5 seconds to repeat item three (eight syllables long). The amount of time for repetition increased by 0.5 seconds with every syllable increase (e.g., three seconds for nine-syllable sentences, 3.5 seconds for ten-syllable sentences, see Ortega et al., 2002). The next stimulus played even if the participant did not manage to repeat the previous stimulus in the allotted time.

6. Data analysis and coding

We followed the same data analysis procedures as described in W&O.

First, for the oral narrative task, each learner narrative was orthographically transcribed in CHAT (MacWhinney, 2000). Transcripts were then analysed for three measures: number of clauses, motion clauses, and motion verbs. Following W&O, a clause was defined as 'a unit containing a unified predicate that expresses a single situation (Berman & Slobin, 1994, p. 657)'. For example, the utterance 两只狗跑到学校里 'the two dogs run into the school' was coded as a single clause, but 他离开教室跑下楼 'he leaves the classroom and goes downstairs' was coded as two clauses. Clauses were excluded from the analysis if more than one word or phrase was either unintelligible or in the learner's L1 (e.g., 这个男孩子 *goes up the stairs* 'this boy goes up the stairs'). Motion clauses were identified as those that described 'an entity changing location from one place to another' (Wu & Ortega, 2013, p. 690). For example, 我去了二楼 'I go to the second floor' was coded as a motion clause, but 我觉得我的狗很麻烦 'I think my dog is troublesome' was not because no change in location was specified. Lastly, for the number of motion verbs produced, a motion verb was identified as any verb that indicated movement (e.g., 跑 'run', 追 'chase'). For W&O, these different measures tap into different types of language ability. The total number of clauses 'was thought to provide a task-relevant indication of learners' speaking fluency and communicative adequacy' (Wu & Ortega, 2013, p. 689). The number of motion clauses 'was thought to be a[n] appropriate indicator of communicative effectiveness in terms of degree of task completion' (Wu & Ortega, 2013, p. 689). The number of motion clauses was used as 'a task-relevant indication of lexical diversity [… which] provides a good indicator of learners' general language proficiency' (Wu & Ortega, 2013, p. 690).

As in W&O, the EIT was scored and coded using the five-point rubric (Iwashita & Ortega, 1998; Ortega et al., 2002; Zhou, 2012). Table 2 shows the potential maximum score for each EIT item. As Ortega et al. (2002) discuss, the scoring rubric accords central importance to the integrity of the stimulus's meaning. For example, a score of 3 is given if the whole meaning of the stimulus is preserved with some minor grammatical changes, whereas a score of 2 is given when those grammatical changes influence the meaning. Using this 0–4 rubric, the maximum possible score on the EIT is 120.

In terms of reliability, the internal consistency of the EIT was tested using Cronbach's Alpha, which indicated a very high instrument reliability ($\alpha = 0.96$, $n = 65$), consistent with W&O ($\alpha = 0.97$) and the EITs in other languages (see Tracy-Ventura et al., 2014). For inter-rater reliability, W&O report that 15% of the EIT data and 15% of the oral narrative data were independently scored by two raters. Comparison of these rating showed high reliability (95% agreement). Any disagreements were resolved

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

Table 2. EIT scoring rubric with examples

Score	Definition	Example
4	Perfect repetition	Stimulus: 红色的书在桌子上。 red de book on the table shang 'The red book is on the table' Response: 红色的书在桌子上。 red de book on the table shang 'The red book is on the table' (participant 801)
3	Accurate content repetition with some (ungrammatical or grammatical) changes of form	Stimulus: 他得先把房间打扫干净才可以出去玩。 he have-to first ba room clean up cai can go outside play 'Before he can go outside, he has to finish cleaning his room' Response: 他得先把房间打扫干净然后可以出去玩。 he have-to first ba room clean up then can go outside play 'Before he can go outside, he has to finish cleaning his room' (participant 501)
2	Changes in content or in form that affect meaning	Stimulus: 我喜欢看有快乐结局的电影。 I like watch have happy ending de movie 'I enjoy watching movies that have a happy ending' Response: 我喜欢看…电影。 I like watch…movie 'I enjoy watching movies' (participant 103)
1	Repetition of half of the stimulus or less	Stimulus: 我听说明天可能会下雨。 I hear tomorrow may rain 'It is possible that it will rain tomorrow' Response: 我…明天…听…下雨 I…tomorrow…hear…rain 'I…tomorrow…hear…rain' (participant 302)
0	Silence, unintelligible repetition, or only one word repeated One word = – Only one content word plus function word(s) – Only one content word plus function word(s) plus extraneous words not in the original stimulus – Only function word(s) repeated	Stimulus: 那家饭馆的中国菜应该很不错。 that classifier restaurant de Chinese food should be very good 'That restaurant is supposed to have very good Chinese food' Response: 饭馆 restaurant 'restaurant' (participant 403)

through discussion and one rater coded the remaining data. We followed this inter-rater reliability procedure except that we increased the proportions. For the oral narrative task, two raters independently coded 30% data. Inter-rater reliability was 97% agreement for the number of clauses, 92% agreement for the number of motion clauses, and 93% agreement for the number of motion verbs. The second author coded the remaining files. For the EIT, 100% of files were independently coded by two raters. Inter-rater reliability was 77%, which is lower than W&O's 95% agreement. Any disagreements were resolved through discussion. The two raters were therefore agreed on the final scores used in our EIT analysis (see discussion for potential reasons for inter-rater reliability differences between the studies).

Since all datasets were normally distributed (using visual inspection of Q-Q plots, histograms, boxplots; $p > .05$ for Shapiro-Wilks tests in all datasets), we present the results of parametric tests, specifically one-way ANOVAs for between-group comparisons for EIT and oral narrative scores and pearson's r for correlations between EIT and oral narrative scores. Prior to carrying out these tests,

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

Table 3. Group descriptives for EIT in the replication and Wu and Ortega (2013)

	N	Mean	SD	95% CIs
Beginner				
Wu & Ortega	–	–	–	–
Replication	20	20.9	10.9	15.9, 26.1
Low				
Wu & Ortega	20	38.2	17.1	–
Replication	20	33.8	17.2	25.7, 41.8
High				
Wu & Ortega	20	61.5	23.2	–
Replication	25	51.6	22.1	42.5, 60.7

Note. Maximum EIT score = 120.

all assumptions were tested (linearity, normality of distribution, and no significant outliers for pearson's *r*; normality of distribution, equal variances for ANOVAs).

First, a series of one-way ANOVAs with Level (beginner, low, high) as the between-subjects factor were used to compare scores on the EIT and the oral narrative task (number of clauses, number of motion clauses, number of motion verbs). The alpha level was set at 0.05. Second, pearson's *r* correlations were carried out to investigate relationships between performance on EIT and performance on the oral narrative task.

To interpret magnitudes of difference, we present Cohen's *d* ES with 95% confidence intervals (CIs) for *d* for all between-subject comparisons, and *r* values with 95% CIs and R^2 values for all correlations. Between-subject ES were calculated using mean and SDs of two groups. ES (*d*, *r*, R^2) were interpreted using Plonsky and Oswald's (2014) and Plonsky and Ghanbar's (2018) field-specific benchmarks: between-subject: *d*s around .40 are considered small, .70 medium, and 1.00 large; *r* coefficients close to .25 are considered small, .40 medium, and .60 large, and R^2 coefficients close to .18 are considered small, .32 medium, and .51 large. CIs are used to interpret ES reliability (see Field, Miles, & Field, 2012).

In addition, we follow Porte and McManus's (2019) guidelines for interpreting differences and similarities between an original study and a replication. When descriptive results of the original study are provided, these are compared with the replication using between-subject Cohen's *d* ES and CIs for *d* (see also McManus & Marsden, 2018). Although, to date, this method is seldom used in SLA research for understanding reproducibility (see Marsden, Morgan-Short, Thompson, & Abugaber, 2018), Cohen's *d* ES with 95% CIs for *d* can indicate the extent to which any between-study differences are meaningful (Cohen's *d* ES >.40 with 95% CIs for *d* not including 0 are taken to indicate a meaningful difference, see Porte & McManus, 2019) and what the magnitude of those between-study differences might be (e.g., medium, large difference, using Plonsky & Oswald's, 2014 interpretation or more specific indices from meta-analyses).

7. Results

7.1 EIT

Table 3 shows group EIT scores for the replication's beginner, low, and high groups, as well as those for W&O's low and high groups.

Table 3 shows a similar patterning of results between the studies: the high groups perform more accurately than the low groups, and, in the replication, the low group performs more accurately

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

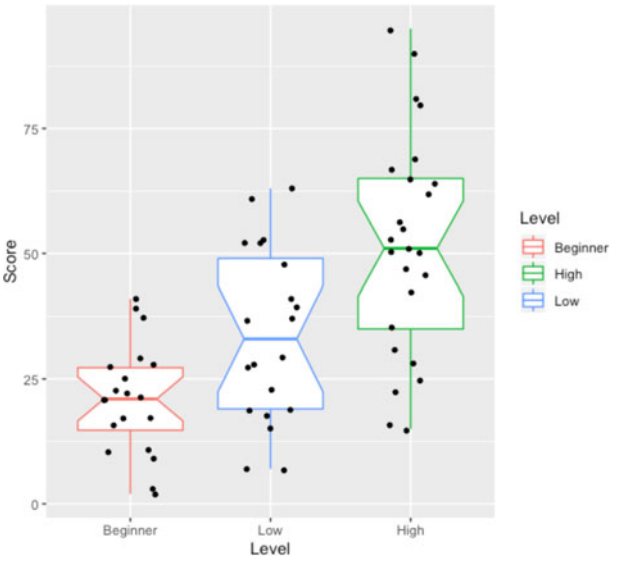


Figure 1. EIT scores in the beginner, low and high groups in the replication

Table 4. Posthoc tests of one-way ANOVA for EIT in the replication

	t	95% CIs	*p*	Cohen's *d*	95% CIs for *d*
High-Beginner	5.79	18.07, 43.63	<.001	1.70	0.99, 2.35
Low-Beginner	2.28	−0.67, 26.27	0.07	0.90	0.23, 1.53
Low-High	−3.39	−30.83, −5.27	0.003	−0.89	−1.49, −0.26

than the beginner group. In addition, the high and low groups in both studies show similar amounts of variation in EIT scores, as indicated by the SDs.

A one-way ANOVA was carried out to test EIT performance between the replication's groups, which indicated statistically significant between-group differences, $F(1.63) = 34.5$, $p < .001$, eta squared = 0.35 (see also Figure 1). Posthoc tests (Table 4) showed that the magnitudes of these differences were large between beginner and high, and medium between beginner and low and low and high.

We compared EIT scores between studies using Cohen's *d* ES comparisons (following McManus & Marsden, 2018; Porte & McManus, 2019), which indicated no reliable differences (all CIs for *d* included zero): between the low groups, $d = 0.26$, CIs = −0.37, 0.87; between the high groups, $d = 0.44$, $d = -0.16$, 1.03. EIT performance at each level (i.e., low and high) was therefore broadly comparable between studies, irrespective of sampling and contextual differences.

Taken together, these results demonstrate that the EIT distinguished between each of the replication's three curricular levels. This result replicates the original study's finding that the EIT is able to distinguish between different language abilities and extends them by additionally distinguishing between more fine-grained language abilities (e.g., between 100-level and 200-level students). The inclusion of more advanced/experienced learners of Chinese in the original study (i.e., graduate students) is one potential explanation for the slightly higher group means in W&O's high group compared to the replication. However, between-study comparisons showed that excluding graduate students from the high group did not change the patterning of results. In sum, this replication indicates that the EIT

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

Table 5. Group descriptives for the oral narrative task in the replication

	Mean	SD	95% CIs
Number of clauses			
Beginner	5.00	3.01	3.59, 6.41
Low	12.25	4.88	9.97, 14.53
High	18.44	6.82	15.63, 21.25
Number of motion clauses			
Beginner	2.30	1.78	1.47, 3.13
Low	4.80	2.55	3.61, 5.99
High	7.04	3.47	5.61, 8.47
Number of motion verb types			
Beginner	1.65	1.09	1.14, 2.16
Low	3.25	1.80	2.41, 4.09
High	5.04	2.37	4.06, 6.02

can be used as an effective proficiency assessment tool for SLA researchers working with L2 learners of Chinese across a broad range of language abilities, including beginning learners of L2 Chinese.

7.2 Oral narrative task

Table 5 shows group descriptives for the number of clauses, the number of motion clauses, and the number of motion verb types in the replication's beginner, low, and high groups (Comparable descriptives are not provided in W&O).

Table 5 shows a similar patterning of results to the EIT in that high group learners produced the most clauses, motion clauses, and motion verb types, while beginner group learners produced the lowest frequencies for these measures. In addition, the SDs and CI intervals show that scores became more dispersed as a function of level, with the greatest heterogeneity found among high group learners (see also Figure 2). That is, performance in the high group was the most varied, perhaps because this group included learners from multiple curricular levels. Performance was most homogenous among beginner group learners. Therefore, even though learners in the high group produced the most clauses, motion clauses, and motion verb types, these larger SDs and wider CIs intervals indicate that language abilities were also the most varied in this group.

A series of one-way ANOVAs were carried out to test between-group performance on these measures, which showed statistically significant differences between the groups: Number of clauses, $F(2, 62) = 35.81$, $p < .001$, eta squared = 0.54; Number of motion clauses, $F(2,62) = 16.4$, $p < .001$, eta squared = 0.35; Number of motion verb types: $F(2,62) = 18.21$, $p < .001$, eta squared = 0.37. Consistent with the EIT, posthoc results (Table 6) showed that the magnitudes of these differences were medium to large because of higher frequencies in the high group compared to the low and beginner groups, and higher frequencies in the low group compared to the beginner group (see also Figure 2). Taken together, these results indicate that, like the EIT, performance on the oral narrative task also distinguished between each of the replication's three curricular levels EIT.

Even though direct between-study comparisons are not possible since descriptives for oral narrative performance are not reported in the original study, ES comparisons reported by W&O show a similar patterning of results to those found in the replication. In both studies, clause, motion clause, and motion verb type frequencies were higher in the high group than in the low group. Also, the magnitudes of low-high group differences were similar: for clauses, 0.68 in W&O and 1.02 in the replication;

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

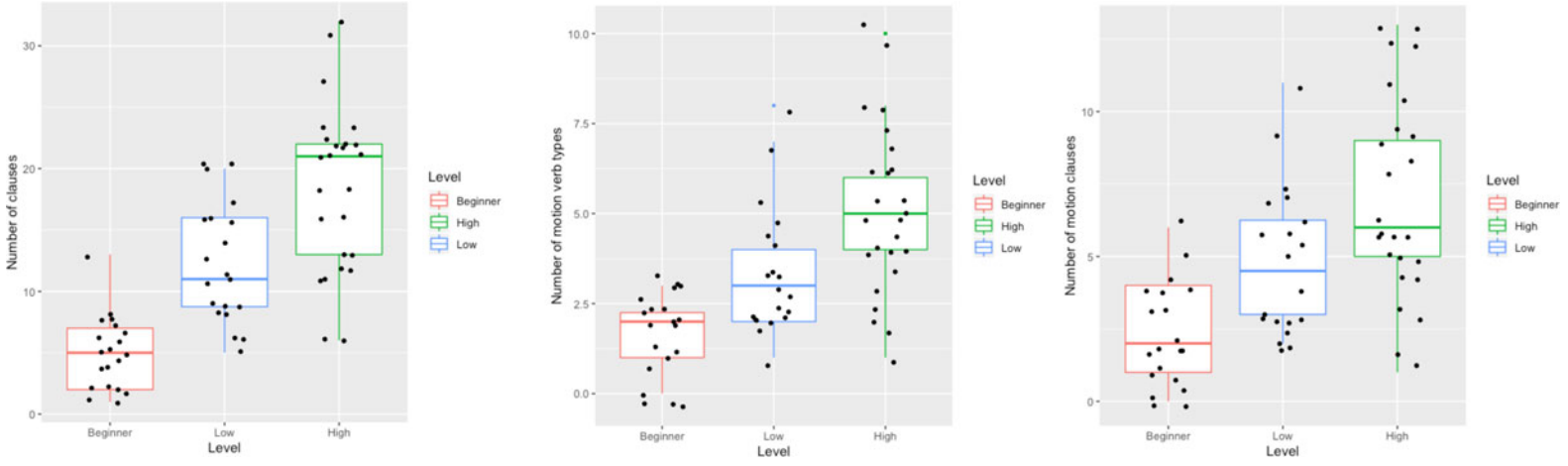


Figure 2. Clause, motion clause, and motion verb type frequencies in the oral narrative task for beginner, low and high groups in the replication

Table 6. Posthoc results for between-group comparisons on clause, motion clause, and motion verb type frequencies in the oral narrative task in the replication

	t	95% CIs	*p*	Cohen's *d*	95% CIs for *d*
Clauses					
High-Beginner	8.46	9.63, 17.25	<.001	2.46	1.64, 3.18
Low-Beginner	4.33	3.23, 11.27	<.001	1.79	1.02, 2.48
Low-High	−3.89	−10.00, −2.38	<.001	−1.02	−1.63, −0.38
Motion clauses					
High-Beginner	5.72	2.75, 6.73	<.001	1.66	0.96, 2.31
Low-Beginner	2.86	0.40, 4.59	0.02	1.14	0.45, 1.78
Low-High	−2.71	−4.23, −0.25	0.02	−0.72	−1.32, −0.10
Motion verb types					
High-Beginner	6.01	2.03, 4.74	<.001	1.77	1.05, 2.43
Low-Beginner	2.69	0.17, 3.03	0.02	1.08	0.39, 1.72
Low-High	−3.17	−3.14, −0.44	0.01	−0.84	−1.43, −0.21

for motion clauses, 0.84 in W&O and 0.72 in the replication; for motion verb types, 0.62 in W&O and 0.84 in the replication.

7.3 Relationships between EIT and oral narrative task performance

Correlations were carried out to investigate potential relationships between performance on the EIT and the oral narrative task. These tests indicated positive, strong, and statistically significant relationships between EIT scores and clause, motion clause, and motion verb frequencies in the oral narrative: EIT-Clauses, $r = 0.74$ [CIs, 0.59, 0.83], $p < .001$, $R^2 = 0.55$; EIT-Motion clauses, $r = 0.75$ [CIs, 0.62, 0.84], $p < .001$, $R^2 = 0.56$; EIT-Motion verbs $r = 0.73$ [CIs, 0.59, 0.83], $p < .001$, $R^2 = 0.53$. See Figure 3 for corresponding plots. Given the relatively high R^2 values (Plonsky & Ghanbar, 2018), these results indicate a high degree of explained variance between performance on the EIT and the oral narrative task. In other words, the higher the EIT score, the higher the frequency of clauses, motion clauses, and motion verb types in the oral narrative task.

Comparisons with W&O indicate a similar patterning of results, albeit with higher degrees of shared variance between EIT and oral narrative performance in the replication (R^2 range = 0.53–0.56) than in W&O (R^2 range = 0.23–0.28). However, it is important to note that W&O's correlations included both heritage speakers and foreign language learners. Nevertheless, despite some differences in the magnitude of shared variance, both studies indicate a significant relationship between EIT and oral narrative performance.

7.4 Linguistic features in the EIT potentially influencing performance

W&O suggested that syllable length explained a large proportion of the variance in the item scores ($R^2 = 0.74$). We also found a negative, strong, and statistically significant relationship between mean EIT score for an item and its length in syllables ($r = -0.87$ [CIs, −0.94, −0.74], $p < 0.001$, $R^2 = 0.76$, see Figure 4). In addition, syllable length accounted for a large proportion of the variation in each group:

- Beginner, $r = -0.82$ [−0.91, −0.65], $p < 0.001$, $R^2 = 0.67$
- Low, $r = -0.85$ [−0.92, −0.70], $p < 0.001$, $R^2 = 0.72$

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

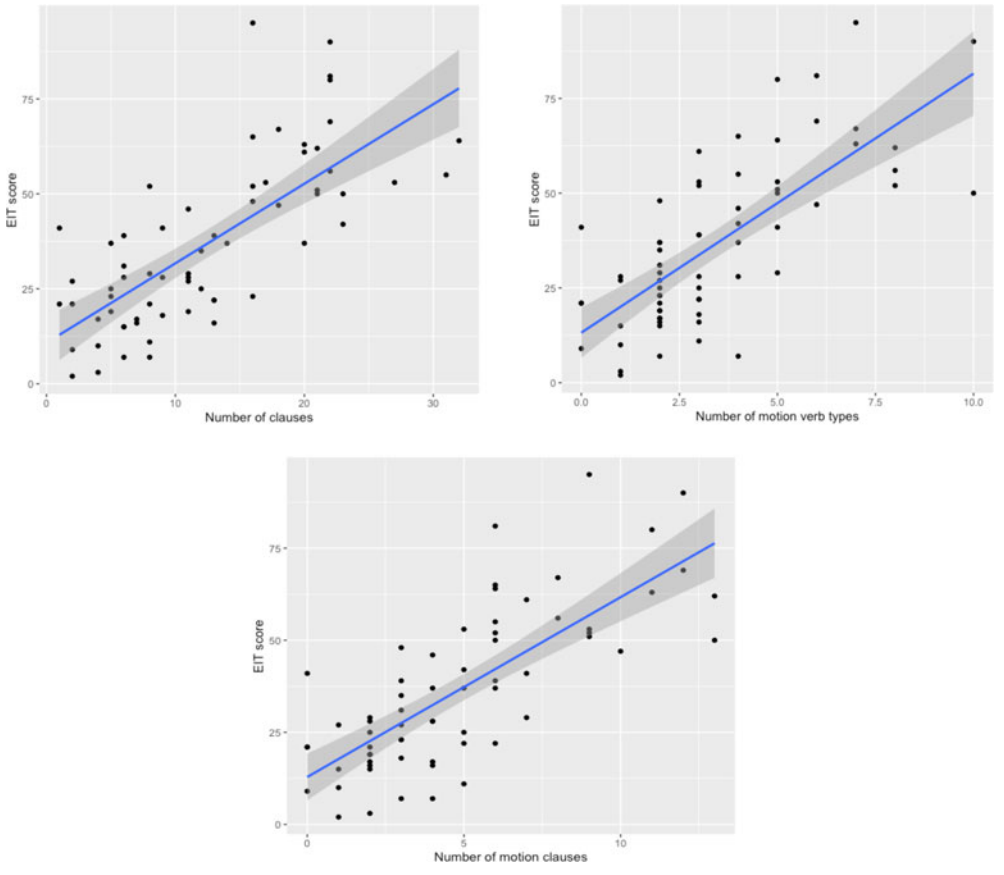


Figure 3. Scatter plots of relationships between EIT and oral narrative task performance

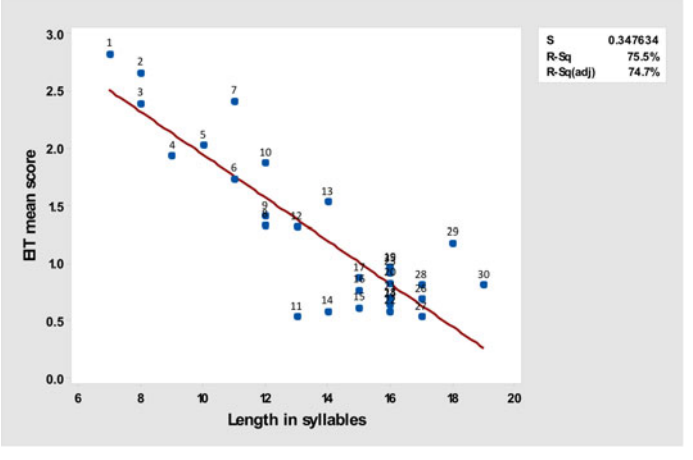


Figure 4. Correlation between EIT mean score and length in syllables for all learners

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

- High, $r = -0.87$ [−0.94, −0.74], $p < 0.001$, $R^2 = 0.75$

In line with W&O, analysis of individual EIT items also showed that EIT items 11, 14, 17, 18, 21, 22, 23, 24, and 27 received the lowest mean scores and items 1, 2, 3, and 7 received the highest scores. In addition to syllable length of the EIT item, W&O proposed that syntactic complexity, familiarity with the construction and vocabulary could also contribute to performance on the EIT. For example, W&O reported that the mean score for item 11 (1.47) was lower than item 12 (2.22), even though both items are 13 syllables long. They suggest that the presence of an attributive clause in item 11 (but not in item 12) may have induced this extra difficulty:

Item 11
[昨天 死了 小猫 的] 小男孩 很伤心。(13)
yesterday die-perfective kitten de little boy very sad
'The little boy [whose kitten died yesterday] is sad'

Item 12
那 家 饭馆 的 中国菜 应该 很 不错。(13)
that classifier restaurant de Chinese food should be very good
'That restaurant is supposed to have very good food'

Our findings patterned similarly: the mean EIT score for item 11 (0.52) was lower than that of item 12 (1.31), with a medium ES (Cohen's $d = 0.78$ [0.42, 1.14]). Furthermore, of the nine lowest scoring EIT items, seven included an attributive clause (i.e., items 11, 17, 18, 22, 23, 24, 27). To examine the extent to which the presence of an attributive clause contributed to item difficulty in the replication, we fitted a linear regression model using syllable length and the presence of an attributive clause on mean EIT scores (see Table 7). This analysis indicated that the syllable length and presence of an attributive clause were significant predictors of EIT difficulty, both individually and collectively ($F(2, 27) = 50.97$, $p < .001$, $R^2 = 79.1$). In addition, this analysis showed that when syllable length remains constant, the predicted mean score of a test item is equal to 3.74 when there is no attributive clause, but equal to 3.45 when there is an attributive clause. In sum, the presence of an attributive clause appears to play a role in understanding EIT performance.

An additional factor influencing EIT outcomes suggested by W&O was learners' familiarity with the constructions, as supported by the comparatively longer, but easy-to-repeat item 7. They attributed the comparatively high EIT score for item 7 to learners' familiarity with the VERB-COPYING CONSTRUCTION of 开车开得很好*/kai-che-kai-de-hen-hao*, literally 'drive-car-drive-particle de-very-well'. In the replication, item 7 was also among the highest scoring items. Based on Figure 4, items 10, 13, 29 and 30 could also be interpreted as easier to repeat than expected given that mean scores for these items are located above the regression line. Similar to item 7, these items contain a grammatical construction to which learners may be frequently exposed in the classroom. Item 10 has the conceding construction of …是…, 就是 '…is…, but it is…', item 13 contains the 对不对 'right or wrong' construction, item 29 includes 我不知道 *'I don't know'* and 是不是 'whether or not' chunks, and item 30 contains the grammatical construction of 什么…都(不)… 'every…all (no)…'. It is possible that familiarity with these constructions could provide some explanation for the comparative difficulty of these items in comparison to other items with the same length of syllables.

8. Discussion

The current study closely replicated Wu and Ortega (2013), who found that a Mandarin Chinese EIT reliably distinguished between low and high language abilities, operationalized as learners enrolled in 200-level classes (the low group) and learners enrolled in 300-, 400-, and graduate-level classes (the high group). Performance on an oral narrative task, analysed in terms of clause, motion clause, and motion verb frequencies, patterned similarly to the EIT. Based on these results, Wu and Ortega

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

Table 7. Summary of linear regression analysis for predicting EIT item difficulty from syllable length and presence of an attributive clause

	Coefficient	Standard error	*t*	*p*
Constant	3.74	0.27	13.72	<0.001
Syllable length	−0.18	0.02	−8.83	<0.001
Presence of attributive clause	−0.29	0.14	−2.13	0.04

(2013, p. 697) concluded that the EIT can be 'an effective research tool in gauging global oral proficiency in L2 Chinese'.

In this close replication, we followed the same procedures, used the same tests, and carried out the same analyses as W&O in order to better understand the original study's findings as well as the external validity of the Mandarin EIT as a proficiency measurement tool. To achieve this, foreign language learners of L2 Mandarin were sampled and assigned to language ability groups following W&O's selection criteria, with three modifications. First, W&O examined performance among foreign language learners and heritage learners, but this replication included foreign language learners only. Second, W&O included learners from three curricular levels in the high group (300-, 400-, graduate-level) and from one curricular level in the low group (200-level only). The replication excluded graduate-level students from the high group to understand the extent to which including these potentially more experienced and more proficient users of L2 Chinese may have elevated mean scores in the high group. Third, we recruited learners from 100-level classes to a beginner group (in contrast to W&O) to understand the extent to which the EIT might also be able to (1) distinguish among finer-grained language abilities (e.g., beginner-low-high) and (2) function as a proficiency tool among relatively inexperienced L2 learners (i.e., beginners). All other aspects of the original study remained the same in the replication.

Overall, our results patterned very similarly to those in the original study. We found no reliable differences between studies for mean EIT performance in the low and high groups (assessed using Cohen's *d* ES and 95% CIs for *d*, see McManus & Marsden, 2018; Porte & McManus, 2019). Comparisons between the replication's groups showed that performance was more accurate as a function of language ability grouping, consistent with W&O. That is, mean EIT scores were highest in the high group, followed by the low group, and lowest in the beginner group. ES comparisons showed that the magnitudes of these between-group differences were medium to large. In terms of performance on the oral narrative, similar patterns of results were found: medium to large differences between the replication's different language ability groups for number of clauses, number of motion clauses, and number of motion verb types.

Taken together, this replication demonstrated that performance in the EIT mirrored that in the oral narrative by distinguishing between the different language ability groups. Correlations between the EIT and the oral narrative task indicated a high degree of shared variance. We therefore consider that this replication provides strong evidence in support of the original study's claim that the Mandarin EIT can function as an effective L2 proficiency measurement tool. In addition, this replication's findings for beginning learners indicate that the EIT can be used with inexperienced and low proficiency L2 learners. This is an important contribution to our understanding of the EIT since most previous research has assessed its use among intermediate-to-advanced L2 speakers.

8.1 Design features of the EIT

Both studies were agreed in how specific linguistic attributes of the stimuli could have shaped performance (e.g., construction frequency, syntactic complexity). As previously noted, one major source of agreement was that mean EIT score and the syllable length of the stimulus shared a high degree of variance ($R^2 = 0.76$ [replication], 0.74 [original]), indicating that mean EIT scores tended to be higher

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

for stimuli with fewer syllables, which has also been shown for EITs in other languages (e.g., Tracy-Ventura et al., 2014; Gaillard & Tremblay, 2016).

One aspect of the EIT that has received little discussion so far, however, is the ease and reliability of the coding. This plays a central role in its use as a reliable proficiency assessment tool. As previously noted, inter-rater reliability for the EIT was higher in W&O (95%) than in this replication (77%), but was very similar between studies for the oral narrative (95%). In both cases, the replication followed the same coding protocols as reported by W&O. In order to understand this between-study difference for EIT inter-rater reliability, we examined each of the five coding options in the replication (see Table 2 and Iwashita & Ortega, 1998 for full definitions of coding options).

- 0 = silence, unintelligible repetition, or only one word repeated (81% agreement)
- 1 = half or less of stimulus repeated (71% agreement)
- 2 = change to stimulus meaning (71% agreement)
- 3 = accurate content with minor changes to form (78% agreement)
- 4 = perfect repetition (98% agreement)

Agreement was therefore highest for coding options 0 and 4, while codes 1, 2, and 3 represented more inter-rater disagreement. Further examination showed that the most common discrepancies were between scores 0 and 1 (49% of all discrepancies), 1 and 2 (32%), and 2 and 3 (14%).

One source of inter-rater disagreement included distinguishing between content words and function words in Chinese. For instance, one response to item 28 (participant #401) was 递给我吗 'pass-*gei*-I-question particle', an exact repetition of the last four words of the test item. The syntactic status of 给 'gei' is controversial as some researchers consider it to be a verb (a content word 'to give') while others consider it to be a preposition (a function word 'to') (Li, 2015). Consequently, this response can either be seen as two content words (i.e., 递, 给) plus two function words (i.e., 我, 吗), or as one content word (i.e., 递) plus three function words. The first interpretation verges more towards a score of 1, while in the second interpretation this response should be scored with 0 because there is only one content word plus function words. The raters finally settled on the first interpretation and the score of 1.

In addition, potential pronunciation difficulties contributed to disagreements with codes 2 and 3. For example, a participant (participant #309) repeated item 8 as 挖饭以后，我好好睡觉了 'After (ambiguous word), I have a good nap'. It was unclear whether the participant tried to say 晚饭/*wanfan* '*dinner*' and instead uttered 挖饭/*wafan*. Such ambiguity led to two possible ratings, score 3 for the first interpretation or 2 for the second interpretation. After discussion, the raters agreed on the second interpretation to avoid inflating the scores.

Taken together, these rater disagreements suggest that the EIT may benefit from more rigorous coding protocols and the development of rater training packages to ensure greater between-study consistency. While it is important to remain consistent across EITs (e.g., Spanish, French, Chinese) to enhance their comparability and use in crosslinguistic research, it is possible that some of the coding descriptors might not work equally well across all languages. For example, operationalizing 'word' is a well-known difficulty in Chinese (Qian & Pan, 2002; Ng & Low, 2004), which has been addressed in Chinese SLA research by using characters as the unit of analysis (see Du, 2013; Di Silvio, Diao, & Donovan, 2016). One important area of future research will likely include investigations of EIT rater protocols to understand how specific coding rubrics contribute to proficiency assessment, including the development of online training packages for EIT raters.

9. Conclusions

The current study's close replication of Wu and Ortega (2013) provided supporting evidence that the Mandarin Chinese EIT can function as an effective proficiency measurement tool for SLA research involving foreign language learners. In so doing, it contributes to a growing body of research indicating

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

that EITs can provide reliable and straightforward means to measure proficiency in SLA research. In these ways, the EIT can offer a more reliable index of language ability than other indexes used in the field (e.g., course grades, see Brown et al., 2018). We also note that replication research offers a powerful means for understanding the generalizability of individual studies and, in particular, for assessing the external validity of research tools.

Acknowledgements. A version of this article was presented at the 2019 annual conference of the American Association for Applied Linguistics in Atlanta, GA, USA. Many thanks to the learners who participated in this study, to Chunyuan Di, Ying Feng, Shuishui Long, Jia Yu, Alex Magnuson, and Ben Pin-Yun Wang for help with recruitment, and to Lu Yu and Jingyuan Zhuang for research assistance. Thank you also to the two reviewers and the editor for their valuable feedback that greatly improved this article.

Note

[1] All differences between beginner and low were negligible: years learning Chinese (d [CIs] = −0.54 [−1.16 0.10]), age started learning Chinese (d[CIs] = 0.33 [−0.31, 0.94]), length of time spent in a Chinese-speaking country (d[CIs] = −0.12 [−0.74, 0.50]). No reliable differences were found between beginner and high for length of time spent in a Chinese-speaking country (d[CIs] = 0.29 [−0.31, 0.87]). No reliable differences were found between low and high for age started learning Chinese (d[CIs] = 0.29 [−0.30, 0.88]), length of time spent in a Chinese-speaking country (d[CIs] = −0.19 [−0.78, 0.40]).

References

Berman, R., & Slobin, D. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Mahwah, NJ: Erlbaum.

Bowden, H. W. (2016). Assessing second-language oral proficiency for research: The Spanish elicited imitation task. *Studies in Second Language Acquisition*, *38*(4), 647–675.

Brewer, M. B., & Crano, W. D. (2000). Research design and issues of validity. In H. T. Reis, & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3–16). Cambridge, UK: Cambridge University Press.

Brown, A. V., Plonsky, L., & Teimouri, Y. (2018). The use of course grades as metrics in L2 Research: A systematic review. *Foreign Language Annals*, *51*(4), 763–778.

Byrnes, H. (2013). Editor's introduction. *Modern Language Journal*, *97*(1), 105–108.

Dewey, D. P., Bown, J., & Eggett, D. (2012). Japanese language proficiency, social networking, and language use during study abroad: Learners' perspectives. *The Canadian Modern Language Review*, *68*(2), 111–137.

Di Silvio, F., Diao, W., & Donovan, A. (2016). The development of L2 fluency during study abroad: A cross-language study. *Modern Language Journal*, *100*(3), 610–624.

Du, H. (2013). The development of Chinese fluency during study abroad in China. *Modern Language Journal*, *97*(1), 131–143.

Dussias, P. E., Valdés Kroff, J. R., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When gender and looking go hand in hand. *Studies in Second Language Acquisition*, *35*(2), 353–387.

Ellis, N. C. (2015). Foreword. *Language Learning*, *65*(S1), v–vi.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*(3), 464–491.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Los Angeles, CA: Sage Publications.

Gaillard, S., & Tremblay, A. (2016). Oral proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, *66*(2), 419–447.

Gass, S. (2018). SLA elicitation tasks. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 313–337). London, UK: Palgrave Macmillan.

Hudson, T., & Llosa, L. (2015). Design issues and inference in experimental L2 research. *Language Learning*, *65*(S1), 76–96.

Hulstijn, J. H. (2010). Measuring second language proficiency. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 185–200). Amsterdam, Netherlands: John Benjamins.

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*(3), 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.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research* (Vol. 41). Amsterdam, Netherlands: John Benjamins.

Iwashita, N. (2018). Grammar and language proficiency. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching*. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118784235.eelt0069

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

Iwashita, N., & Ortega, L. (1998). Scoring guidelines for elicited imitation task L2 English. Retrieved from https://www.iris-database.org/iris/app/home/download?id=york:852670&ip=73.187.219.252&placeValuesBefore%20, https://www.iris-database.org/iris/api/resource/york:852670/asset/

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *The Canadian Modern Language Review*, *64*(1), 215–238.

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.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*(3), 399–436.

Leclercq, P., & Edmonds, A. (2014). How to assess L2 proficiency? An overview of proficiency assessment research. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 3–23). Bristol, UK, Buffalo, NY, and Toronto, Canada: Multilingual Matters.

Li, S. (2010). *Corrective feedback in perspective: The interface between feedback type, proficiency, the choice of target structure, and learners' individual differences in working memory and language analytic ability*. Ph.D. dissertation, Michigan State University.

Li, X. (2015). Ditransitive constructions in Gan Chinese: A case study of the Yichun dialect. In B. Nolan, G. Rawoens, & E. Diedrichsen (Eds.), *Causation, permission and transfer: Argument realization in get, take, put, give and let verbs* (pp. 177–193). Philadelphia, PA: John Benjamins.

Mackey, A., & Gass, S. (2016). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*(2), 321–391.

McManus, K., & Marsden, E. (2018). Online and offline effects of L1 practice in L2 grammar learning: A partial replication. *Studies in Second Language Acquisition*, *40*(2), 459–475.

Ng, H. T., & Low, J. K. (2004). Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing 9, 277–284.

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.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601.

Phakiti, A., De Costa, P., Plonsky, L., & Starfield, S. (2018). Applied linguistics research: Current issues, methods, and trends. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 5–29). London, UK: Palgrave Macmillan.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655–687.

Plonsky, L. (2015). *Advancing quantitative methods in second language research*. New York, NY: Routledge.

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

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.

Porte, G., & McManus, K. (2019). *Doing replication research in applied linguistics*. New York, NY: Routledge.

Qian, Y., & Pan, W. (2002). Prosodic word: The lowest constituent in the Mandarin prosody processing. *Proceedings of the ISCA International Conference on Speech Prosody*, *2002*, 591–594.

Rebuschat, P., & Mackey, A. (2013). Prompted production. In C. A. Chappelle (Ed.), *The encyclopedia of applied linguistics* (Vol. 5). Oxford, UK: Wiley-Blackwell.

Slabakova, R. (2015). The effect of construction frequency and native transfer on second language knowledge of the syntax-discourse interface. *Applied Psycholinguistics*, *36*(3), 671–699.

Solon, M., Park, H. I., Henderson, C., & Dehghan-Chaleshtori, M. (2019). Revisiting the Spanish elicited imitation task: A tool for assessing advanced language learners? *Studies in Second Language Acquisition*, *41*(5), 1027–1053.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*(2), 307–336.

Thomas, M. (2006). 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). Philadelphia, PA: John Benjamins.

Tracy-Ventura, N., McManus, K., Norris, J. M., & Ortega, L. (2014). 'Repeat as much as you can': Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds (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*(3), 339–372.

Underhill, N. (1987). *Testing spoken language*. Cambridge, UK: Cambridge University Press.

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*(4), 680–704.

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*(4), 497–528.
Yuan, F., & Li, S. (2019). *Classroom research on Chinese as a second language*. New York, NY/London, UK: Routledge.
Zhou, Y. (2012). *Willingness to communicate in learning Mandarin as a foreign and heritage language*. Ph.D. dissertation, University of Hawai'i at Mānoa, Honolulu.
Zhou, Y., & Wu, S. L. (2009). *Can elicited imitation be used for the measurement of oral proficiency in L2 Chinese? A pilot study*. Unpublished manuscript, University of Hawai'i at Mānoa.

Appendix A. Mandarin EIT

1	我得去剪头发了。(7) I have to get a haircut.
2	红色的书在桌子上。 (8) The red book is on the table.
3	这个城市街道很宽。(8) The streets in this city are wide.
4	他每天早上都要洗澡。(9) He takes a shower every morning.
5	我听说明天可能会下雨。(10) It is possible that it will rain tomorrow.
6	你刚才说你今天在做什么？(11) What did you say you were doing today?
7	我不觉得他开车开得很好。(11) I doubt that he knows how to drive that well.
8	晚饭以后，我好好地睡了一觉。(12) After dinner I had a long, peaceful nap.
9	我喜欢看有快乐结局的电影。(12) I enjoy movies that have a happy ending.
10	这些房子好是好，就是太贵了。(12) The houses are very nice but too expensive.
11	昨天死了小猫的小男孩很伤心。(13) The little boy whose kitten died yesterday is sad.
12	那家饭馆的中国菜应该很不错。(13) That restaurant is supposed to have very good food.
13	你真的很喜欢听流行音乐，对不对？(14) You really enjoy listening to country music, don't you?
14	她刚把公寓所有的的房间都漆完了。(14) She just finished painting the inside of her apartment.
15	在红绿灯那过马路，然后一直往前走。(15) Cross the street at the light and then just continue straight ahead.
16	我希望别墅能便宜一些，我才买得起。(15) I wish the price of town houses would become affordable.
17	我现在交往的那个人非常有幽默感。(15) The person I'm dating has a sense of humor.
18	我想要一个我的宠物可以住的大房子。(16) I want a nice, big house in which my animals can live.
19	我希望今年的天气会比去年暖和一点。(16) I hope it will get warmer sooner this year than it did last year.
20	我的一个好朋友老帮他的邻居看孩子。(16) A good friend of mine always takes care of my neighbor's three children.

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press

21	他得先把房间打扫干净才可以出去玩。(16) Before he can go outside, he has to finish cleaning his room.
22	我最快乐的回忆就是那次跟你去看戏。(16) The most fun I've ever had was when we went to the opera.
23	那个被警察抓到的小偷长得又高又瘦。(16) The terrible thief whom the police caught was very tall and thin.
24	根据统计，每年吸烟的人数越来越多了。(16) The number of people who smoke cigars is increasing every year.
25	次考试根本没有你跟我说的那么难。(16) The exam wasn't nearly as difficult as you told me it would be.
26	她点菜的时候只点有肉的，从来不点青菜。(17) She only orders meat dishes and never eats vegetables.
27	昨天喂的那只黑猫就是被狗追的那只。(17) The black cat that you fed yesterday was the one chased by the dog.
28	可以麻烦您把桌子上的那本书递给我吗？(17) Would you be so kind as to hand me the book that is on the table?
29	我不知道十点半的火车是不是已经开走了。(18) I don't know if the 11:30 train has left the station yet.
30	为什么还是有很多人早上什么东西都不吃呢？ (19) There are a lot of people who don't eat anything at all in the morning.

Kevin McManus is Watz Early Career Professor in Language and Linguistics, Director of the Center for Language Acquisition, and Associate Professor in the Department of Applied Linguistics at The Pennsylvania State University, USA. His research specializations include second language learning, psycholinguistics, and research methodology, especially replication research.

Yingying Liu is a Ph.D. candidate in the Department of Applied Linguistics at The Pennsylvania State University, USA. Her research interests include corpus linguistics, English for academic purposes, and lexicography.

Cite this article: McManus, K., & Liu, Y. (2022). Using elicited imitation to measure global oral proficiency in SLA research: A close replication study. *Language Teaching*, *55*(1), 116–135. https://doi.org/10.1017/S026144482000021X

https://doi.org/10.1017/S026144482000021X Published online by Cambridge University Press