The LexTALE as a measure of L2 global proficiency: A cautionary tale based on a partial replication of Lemhöfer and Broersma (2012)

Eloi Puig-Mayenco[1], Adel Chaouch-Orozco[2], & Hong Liu[3], Fernando Martín-Villena[4]
King's College London[1], The Hong Kong Polytechnic University[2],
Xi'an Jiaotong-Liverpool University[3], Universidad de Granada[4],

Abstract[1]

The role of proficiency is widely discussed in multilingual language acquisition research, and yet, there is little consensus as to how one should operationalize it in our empirical investigations. The present study assesses the validity of the LexTALE (Lemhöfer and Broersma, 2012) as a 'quick and valid' measure of global proficiency. We first provide an overview review of how the LexTALE has been used since its publication, showing that although the test has gained popularity in the last few years, its reliability has not been thoroughly examined. Thus, herein we present results of a partial replication of Lemhöfer and Broersma (2012), where we empirically assess the validity of the LexTALE as a measure of L2 global proficiency in two groups of learners of English with various degrees of proficiency (L1 Spanish, *n* = 288; L1 Chinese, n = 266). The results indicate that if we are to use LexTALE in our investigations, we should do so with caution as the analyses show that irrespective of the L1 and level of proficiency of the targeted participants, its reliability as a measure of global proficiency is under question evidenced by the low and moderate correlations found with a standardised measure of global proficiency across groups.

Key words: LexTALE, L2 Global Proficiency, Vocabulary Size, Partial Replication

Please, cite as:

Puig-Mayenco, E., Chaouch-Orozco, A., Liu, H. & Martín-Villena, F. (In press). A word of caution for the use of the LexTALE as a measure of global L2 proficiency: a partial replication of Lemhöfer and Broersma (2012). *Linguistic Approaches to Bilingualism*.

[*] We want to thank the King's Undergraduate Research Fellowship programme who funded three undergraduate students to work on the systematic review and data collection of this project. The research fellows were Julia Dziatkowska, Lisa Loy and Yara Chalhoub.

[**] The first author is external member of the project (PID2019-107328GB-I00) funded by XXXXX

1. Introduction

The role of general proficiency in bi-multilingual language acquisition has been extensively discussed across paradigms (see Malovth & Benati 2018, for an overview). And indeed, researchers have been interested in understanding the concept of proficiency in a second language (L2) and have theorised about its operationalisation (e.g., Hulstijn, 2011; Hulstijn, 2012; Lado, 1961; Norris & Ortega, 2003; Norris & Ortega, 2012) with important implications for not only research but also teaching and language learning. Many others within the field of second language acquisition (SLA), however, have used proficiency in their empirical investigations (e.g., to ensure different groups of participants have similar levels of proficiency and/or categorise participants into groups according to those levels, or as a regressor in their analyses). Because, as SLA researchers, we are ultimately interested in informing theories of SLA, our outcomes are inevitably compared to those of related studies to increase the generalisability of the findings. Doing so is not only welcome but very much needed if we want to advance knowledge in the field. However, we also need to make sure that the outcomes of studies are directly comparable. In this light, Norris and Ortega (2003) pointed out that without a valid measure of L2 proficiency, it would be difficult to make meaningful comparisons of results and interpretations across studies. Beyond the inherent challenges of such a task, this is precisely why understanding how L2 proficiency has been operationalised and adopting consistency across studies in the SLA literature is of paramount importance. The focus of this report is, thus, to understand the reliability of the LexTALE (Lemhöfer and Broersma, 2012), which, as shown below, has been extensively used in the literature as a proxy for general proficiency.[2]

[2] The SLA field lacks consensus about what the best definition of proficiency is. Herein, we use general proficiency similarly to Thomas (1994: 330) who defined it as "a person's overall competence and ability to perform in L2 [second language]". The scope of this article is not to provide a once-and-for-all definition of proficiency or a test that taps into it. Our goal is to explore whether the LexTALE, a widely used test assumed to tap into proficiency, correlates

Within the past 30 years, there have been four systematic reviews (Park et al., 2022; Thomas, 1994, 2006; Tremblay, 2011) specifically examining how SLA studies have operationalised the construct of L2 proficiency. Thomas (1994) explored 157 studies published in four top journals in the field of SLA from 1988 to 1992. She found that only 36.3% of those studies used an independent measure of global L2 proficiency. Within that 36.3% of studies, there was also some variation as to how proficiency was operationalised; they used institutional status (40.1%), standardised tests (22.3%), impressionistic judgements (21%) and in-house assessments (14%). In a follow-up study, Thomas (2006) replicated her original review on 211 additional studies published between 2000 and 2004. The outcomes were remarkably similar: only 42.6% of the studies used independent measures of proficiency, with, again, some variation regarding its operationalisation: institutional status (33.2%), standardised tests (23.2%), impressionistic judgements (19.4%) and in-house assessments (19%). Similarly, Tremblay (2011) examined 144 studies published in other top journals within the field. Her results aligned with those found by Thomas (1994, 2006) showing that only 36.8% of the inspected studies employed independent measures of proficiency. Out of those that did, variation was again observed as to what specific measures were used: years of instruction (30.9%), length of immersion (12.4%), standardised tests (11.8%) and pre-established proficiency scores (9.5%). In the most recent review, Park et al. (2022) reviewed 500 studies published between 2012 and 2019, finding that although 91.2% of the studies reported the level of proficiency of the participants, only 42% of those used an independent measure to operationalise it. Again, there was significant variation in the measures used: standardised tests (18%), C-tests (8.49%), oral tests (3.4%), vocabulary tests (6.96%) and other independent tests (5.94%).

with a more traditional and standardized test that has been shown to map nicely onto different levels of proficiency from the Common European Framework of Reference (CEFR; Council of Europe, 2011).

When looking at the four systematic reviews together, it is clear that the picture has not changed much since 1988 and that there is still considerable variation in how researchers measure proficiency. The illustration provided by these reviews resonates with the words of Lemhöfer and Broersma (2012), who emphasised that, "[g]iven the central role of proficiency […] in L2 research, it is alarming how little consensus there is on how to measure it" (p. 326).

Notably, although not their original intention, Lemhöfer and Broersma (2012) contributed (indirectly) to alleviate this problem. They proposed the LexTALE (Lexical Test for Advanced Learners of English), a measure of vocabulary size which was shown to correlate with the Quick Placement Test (UCLES, 2001), taken as a more standard measure of L2 proficiency. This test has gained increased popularity in SLA and bilingualism studies. Since 2012, the LexTALE has been used in, at least, 551 studies. As such, the test seems to have effectively achieved a consensual status as a tool to tap into L2 English proficiency.

The present study argues that this practice is unwarranted and calls for caution as per the conclusions that we researchers can directly make about L2 development, or, indirectly, about phenomena assumed to be influenced by it. In what follows, we (a) explore how the LexTALE has been used in the last 10 years since its publication and (b) conduct a partial replication of Lemhöfer and Broersma (2012) to further test its validity and reliability as a measure of L2 global proficiency.[3]

2. The LexTALE: what is it?

[3] We acknowledge that the LexTALE was not proposed as a sole measure of global proficiency. In fact, Lemhöfer and Broersma introduced it as a test of lexical knowledge while also mentioning that, considering they had found moderate correlations with a standardised measure of global proficiency, the LexTALE score could be taken as a adequate measure of global proficiency. Nevertheless, we also acknowledge (see next section) that the LexTALE has been substantially used for assessing L2 proficiency in SLA research—independently from the authors' original intentions.

The LexTALE was originally proposed by Lemhöfer and Broersma (2012) as a 'quick and valid' test of English vocabulary knowledge for advanced English speakers. The test consists of a dichotomous lexical decision task that lasts approximately five minutes. Participants are shown 60 items, of which 40 are real words and 20 are pseudowords, and are asked to decide whether they are English words. The stimuli included in the LexTALE were extracted from a list of 240 items from the unpublished vocabulary test '10K' (Meara, 1996). All items have between 4 and 12 characters and the 40 English words have a mean frequency of 6.3 (range: 1-26) occurrences per million in the CELEX database (Baayen et al., 1995).[4]

Lemhöfer and Broersma (2012) conducted a study to assess the LexTALE's reliability as a measure of vocabulary knowledge and global proficiency. To do so, they tested 72 native Dutch speakers and 87 native Korean speakers on five different measures: (i) LexTALE scores, (ii) L1 to L2 translation, (iii) L2 to L1 translation, (iv) Quick Placement Test (QPT) scores and (v) self-ratings of English proficiency. Note that although they targeted Korean participants with high proficiency based on the TOEIC® (Test of English for International Communication™) scores to ensure the two groups would be as comparable as possible, the Dutch speakers had significantly higher scores in both the LexTALE and QPT, which was taken to mean that they had higher proficiency in English (Lemhöfer and Broersma, 2012: 332). Overall, their results showed that there were strong correlations between the LexTALE score and the other measures for the L1 Dutch group and moderate correlations for the L1 Korean group.[5] In addition, the authors

[4] The reader is referred to Lemhöfer and Broersma (2012, p. 329-330) for a detailed explanation of the items included and the justification for their inclusion.

[5] The authors reported the Pearson correlation coefficient for the correlations between the LexTALE and the other measures. We direct the reader to Table 4 in Lemhöfer and Broersma (2012: 333) for the exact magnitudes of these correlations.

compared the self-rating results to the translation and QPT tasks, finding less robust correlations.[6] Based on these results, Lemhöfer and Broersma argued that the LexTALE could, indeed, be taken as a useful and valid measure of English vocabulary knowledge *for speakers with advanced proficiencies* (Lemhöfer & Broersma, 2012: p. 340)—this restriction in the use of LexTALE is of crucial relevance for our discussion below on the use of the test. Furthermore, as they found significant correlations between the LexTALE and all the measures except for the self-ratings, they also argued that the LexTALE should be "preferable to self-ratings" (p. 340) as a measure of global proficiency.

As briefly noted above and as will be shown below, the LexTALE has been used in a large proportion of studies. However, its reliability has been scarcely explored. To date, only Nakata et al. (2020) have attempted to further validate the LexTALE in a study with 111 native Japanese speakers who spoke English as a second language. They gathered TOEFL ITP scores and tested them on four different measures: (i) LexTALE score, (ii) translation test score, (iii) vocabulary size test score and (iv) self-rating of English proficiency. Their results replicated those of the L1 Korean group in the original study in that there was a moderate correlation between the LexTALE and TOEFL ITP scores. Nakata and colleagues argued that their findings showed that the LexTALE could be taken as a good measure of vocabulary knowledge, and, importantly, to a lesser extent, L2 proficiency. They did acknowledge, however, that the lower correlation coefficient they found was potentially due to the lower proficiency level of their participants, echoing Lemhöfer and Broersma's (2012) indication about the LexTALE's reliability being dependent on speaker proficiency.

[6] Note that the authors raised some caution against this result, speculating that differences in proficiency between the two groups could potentially account for the observed discrepancy in correlation strength.

To the best of our knowledge, these are the only two studies that have attempted to validate the use of the LexTALE as a measure of L2 global proficiency. Thus, at this point, we must stress that, despite its extensive use—detailed in the next section—the LexTALE's validity has not been tested thoroughly.

3. The LexTALE: how has it been used?

To explore how the LexTALE has been employed to date, we conducted a review of the literature in which we searched for all studies, including peer-reviewed journal articles, book chapters and graduate theses, that mentioned the LexTALE from 2012 to 2022. The search was as exhaustive as possible and was conducted through Google Scholar, ProQuest and Language and Linguistic Behavior Abstracts (LLBA), we last updated it on March 13th 2022. The search led to 814 hits, we then did a first pass look to include only those studies that had an empirical component in it, which led to a database of 732 studies, which were coded with/using the following scheme. We coded for *use* ("citation only" or "actual use"), *L1* and *proficiency level* (pre-advanced to advanced).[7] The review's results indicate that the LexTALE has been used in 551 studies (out of 732 studies reviewed). Further, as shown in Figure 1, there has recently been an upsurge of studies that have utilised it as a measure of vocabulary knowledge and/or proficiency.

[7] We acknowledge that the labels 'pre-advanced' and 'advanced' are not very specific. Authors use many different ways to operationalise proficiency and many different labels. Therefore, we decided to keep this binary coding scheme to capture whether the LexTALE had been used with participants with high proficiency ("advanced") or low proficiency ("pre-advanced"). We included in the latter category all studies with at least one group of speakers with any proficiency below "advanced" (e.g., *ab initio*, beginner, intermediate, upper-intermediate). We took the authors' reporting of proficiency as an indicator of the participants' proficiency.

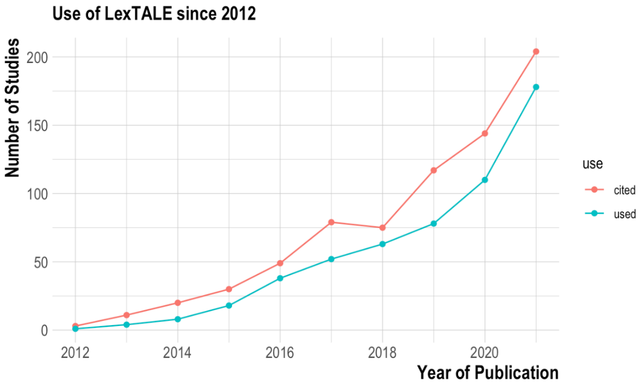


Figure 1. Line graph showing the number of studies that have cited (and used) the LexTALE in the past 10 years.

In addition, the review revealed that the test has been used with 31 different language pairings (e.g., Dutch, Korean, Spanish, Hungarian, Yoruba, Polish, Hindi), and that, out of those studies, 89% included at least one group containing speakers with pre-advanced proficiency.

Importantly, then, there are at least two concerns about how the LexTALE has been employed in SLA research so far. First, as noted, Lemhöfer and Broersma concluded that their findings could be generalised to "most if not all other groups of *advanced* [emphasis added] learners of English" (p.340). However, our review shows that an overwhelming majority of studies (89%) have used it to discriminate among pre-advanced learners too. The second concern relates to the script of the learners' native language. Speakers of different-script languages may have a less entrenched awareness of and sensitivity to phonotactic rules that are highly relevant in English. Analogously, learners whose native language shares the Latin script may more heavily rely on phonotactic constraints to judge the plausibility of an item being an English word or not. Access to this type of knowledge would, then, be a crucial difficulty for learners from different-script languages when completing the test. As such, the native language's script may be a critical factor in determining success in the LexTALE. This is, indeed, something that has not been discussed in the literature. Importantly, we address these two concerns in the present replication.

4. A partial replication of Lemhöfer and Broersma (2012)

Herein, we primarily build upon Lemhöfer and Broersma's (2012) study exploring the reliability of the LexTALE. In their study, they had two interlocked goals. They wanted to investigate the relationships between the LexTALE and translation performance, and the LexTALE and a measure of general proficiency. Here, we further explore their second aim. As such, we investigate whether the LexTALE correlates with a standardised measure of global proficiency in two groups of speakers of L2 English whose L1 shares or does not script with English (L1 Spanish and L1 Chinese, respectively) and whose proficiency varies from intermediate to advanced. Specifically, we entertain the following two research questions:

1. Does the LexTALE equally capture L2 global proficiency in speakers whose L1 shares or does not share script with English?
2. Is the LexTALE a reliable tool to assess L2 global proficiency in intermediate (level B1 to B2 of the Common European Framework of Reference for Languages; CEFR; Council of Europe, 2011), as well as advanced (level C1 to C2 of the CEFR) speakers of English?

5. Methods

5.1. General procedure and instruments

Similarly to Lemhöfer and Broersma's study, our investigation was also web-based and consisted of three parts that participants could complete at the location of their choosing. The first part consisted of a background questionnaire where we gathered information regarding prior experience with English and other languages, which was designed for this study.[8] The second task was the LexTALE. We adopted the same protocols to deliver the test and applied the scoring procedure suggested by Lemhöfer and Broersma (2012).[9] The task is estimated to take between three and five minutes to complete. To increase comparability with the original publication of the LexTALE, we chose the same version of the Quick Placement Test (UCLES, 2001) as our third task to capture English global proficiency in a standardised test. As discussed by Lemhöfer and Broersma (2012), the QPT is "a commercial test that has been validated using thousands of participants and is used by universities and adult education institutions to assign students to English course levels" (p. 328). The QPT consists of 60 multiple-choice questions increasing in difficulty as the participants move through the test. Importantly, it also uses a standardised scoring procedure, in which the range one score falls into corresponds with a level on the CEFR scale. The task is estimated to take between 30 and 40 minutes to complete. A general point to note is that in our replication study we used three of the five tasks used in Lemhöfer and Broersma (2012) (LexTALE test, QPT and self-ratings). We decided not to run the translation tasks as they fell outside the scope of the present study.

The study was conducted in accordance with the recommendations of the ethics committee at King's College London, which gave ethical approval (MRA-20/21-23994). All participants gave informed consent in accordance with the Declaration of Helsinki. Following Open Science Framework practices, all data, scripts and analyses can be found at the first author's OSF repository (https://osf.io/adcr5/?view_only=88cbf349c459448c9bb06f2cb091a93b).

[8] The background questionnaire will be archived on the IRIS database (www.irisdatabase.org) upon acceptance of the manuscript. This questionnaire was not used in Lemhöfer and Broersma's study. We designed it to gather additional information we thought might be useful.

[9] The reader is directed to Lemhöfer and Broersma for a more detailed description of the task itself. Further information can also be found on the LexTALE website (www.lextale.com).

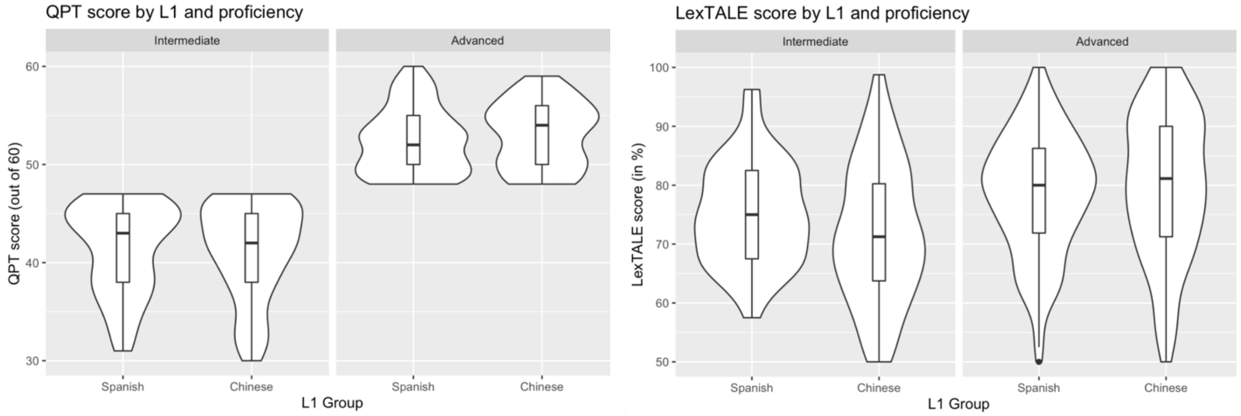
5.2. Participants

Two groups of L2 English speakers took part in our study: the L1 Chinese group (L1 Chinese-L2 English, *n* = 266) and the L1 Spanish group (L1 Spanish-L2 English, *n* = 288). All participants were young adults at the time of testing and reported having begun learning English in their schooling system (primary or secondary education). The mean ages of the participants were 22.86 (*SD* = 3.92) and 24.18 (*SD* = 5.01) for those in the L1 Chinese group and the L1 Spanish group, respectively. Inthe L1 Chinese group, 175 participants were female, 85 were male, and six did not disclose their sex. Inthe L1 Spanish group, 232 participants were female, 55 were male and one did not disclose their sex. We used the scores from the QPT to categorise participants into two groups based on their English proficiency: intermediate (30-47, B1 to B2 of the CEFR) and advanced (48-60, C1 to C2 of the CEFR).[10] It is worth noting that the profile of the participants is fairly similar to those tested in Lemhöfer and Broersma (2012) in that they are all university students in young adulthood. In both our study and theirs, there is a group of speakers of an Indo-European language (Dutch and Spanish) and another one of an east Asian language (Chinese and Korean). However, a crucial difference between both studies is that we included participants who scored at an intermediate level (B1-B2 CEFR) to address our second research question; that is, to examine whether the LexTALE was a reliable test within this proficiency range as well as with advanced participants for which it was originally designed.

6. Results and discussion

[10] We had initially also collected data from beginner (A1 to A2) learners, but decided to exclude them from the analysis due to the low number of participants we had in that group.

Figures 2 and 3 below show the distribution of scores in the QPT and the LexTALE by both L1 and proficiency group, as categorised using the QPT scores. As shown by these figures, the distribution of the data is similar across L1s and proficiency groups.



Figures 2 and 3. Violin plots of the data distribution for the QPT (left figure) and the LexTALE (right figure).

To explore whether there were correlations between the QPT and LexTALE scores, we ran a series of correlation analyses, the results of which we convey in the scatterplots below. Our first research question asked whether the use of similar script between the L1 and L2 would affect the reliability of the LexTALE as a measure of global L2 proficiency. Recall that in the original publication, Lemhöfer and Broersma found a higher correlation between the QPT and LexTALE for the L1 Dutch group (*r* = .60) than for the L1 Korean group (*r* = .30). Thus, we wanted to explore if these results would be replicated with speakers of L1 Chinese and L1 Spanish. Figure 4 below contains a scatterplot with regression lines for the two groups. The results of both correlational analyses show that the LexTALE score is significantly correlated with the QPT score for both groups. However, while the correlation was moderate for the L1 Chinese group, *r* (276) = .36, *p* < .001, 95% CI [.26; .46], it was only a low one for the L1 Spanish group, *r* (286) = .26, *p* < .001, 95 % CI [.15; .36].[11] Note, however, that this first analysis collapsed the intermediate and advanced groups encompassing the full range of proficiency scores found in our study, whereas in Lemhöfer and Broersma (2012), only advanced learners were included.

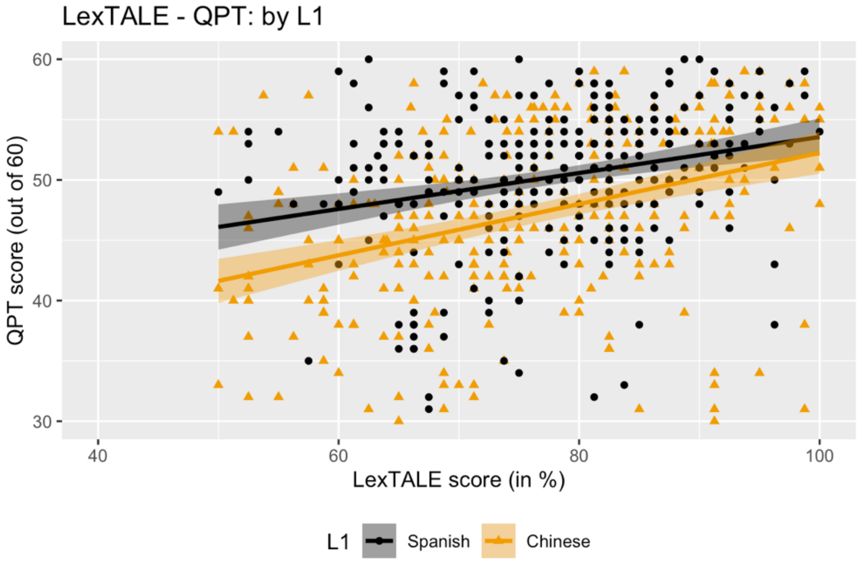


Figure 4. Scatterplot and regression lines of QPT scores (out of 60) and LexTALE scores (in %) for the two L1 groups.

Following up on the previous analysis, our second research question asked whether the LexTALE was an appropriate measure for learners with different proficiency levels. Recall that the LexTALE was initially proposed as a test to be conducted on advanced speakers only. Crucially, however, as our review clearly indicates, many studies to date have used it to assess learners with so-called

[11] Note we employ Cohen's (1988) commonly used benchmarks for interpretation of correlation analyses, which can be interpreted as follows. An *r* coefficient of 0.2 reflects a low correlation; a coefficient value of ~0.5, a moderate correlation; and a coefficient value of ~0.8, a large one. However, we also acknowledge that these are arbitrary numbers and we should be cautious when interpreting them (Thomas, 2007).

lower proficiencies. Thus, we sought to investigate whether the LexTALE was a reliable measure of proficiency in both intermediate and advanced speakers for each of the L1 groups. Figures 5 below contain scatterplots with regression lines for the two proficiency groups in each of the L1 groups.

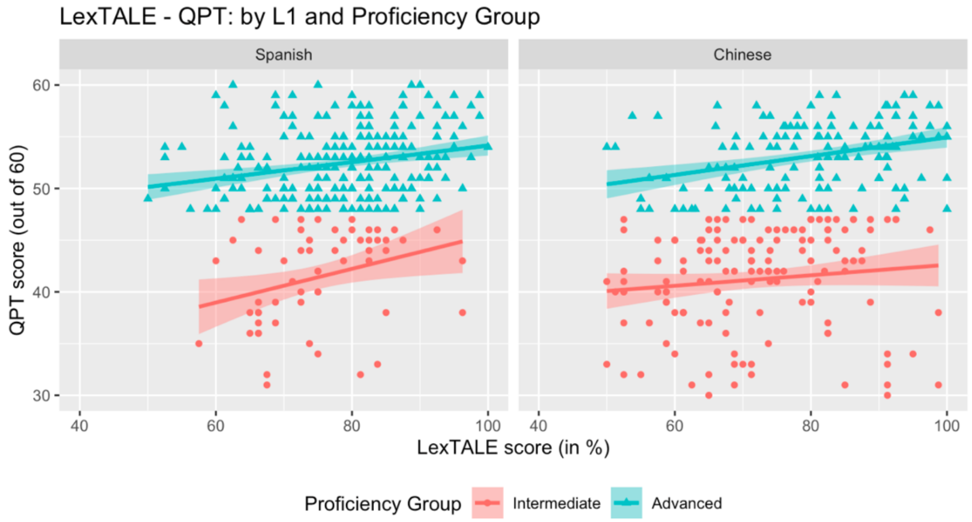


Figure 5. Scatterplots and regression lines of QPT score (out of 60) and LexTALE score (in %) for the two proficiency groups in the two L1 groups.

Again, the results of the correlation analyses show that the scores of the LexTALE and the QPT are significantly correlated, albeit with lower strengths. There was a low correlation for the L1 Spanish advanced group, *r* (229) = .25, *p* < .001, 95% CI [.12; .36], and a slightly higher and moderate one for the L1 Chinese advanced group, *r* (136) = .34, *p* < .001, 95% CI [.18; .48]. Moreover, there was a low correlation for the L1 Spanish intermediate group, *r* (55) = .31; *p* = .018, 95 % CI [.05; .52]. The correlation for the L1 Chinese intermediate group was, in fact, much lower and not significant, *r* (138) = .12; *p* = .148, 95 % CI [-.04; .28].

Although the results indicate that LexTALE scores are positively correlated with scores on the QPT, a standardised measure of general global proficiency, the strength of the relationship is weaker than reported in the original publication for both speakers of L1 Spanish (same script) and L1 Chinese (different script). Recall that, in the original study, the correlation coefficient was relatively high for the L1 Dutch participants(*r* = .60) and close to moderate for the L1 Korean participants (*r* = .29). In light of these divergences, we must note, however, that the number of participants in the present study is considerably larger than in the only two previous studies. More importantly, one needs not to rely exclusively on *r* coefficients to establish the reliability of the LexTALE for measuring global L2 proficiency. We can also look at the distribution of the data. A quick glimpse (as shown in Figure 3 above) shows that there is great—certainly greater than ideal—dispersion within each proficiency group, indicating that the LexTALE is not quite capturing the same construct as that the QPT, a standardised measure of global proficiency, is sensitive to.

Further, it is important to note that the strong correlation found by Lemhöfer and Broersma (2012) was observed in the L1 Dutch speakers, which is typologically closer to English than Spanish is—English and Dutch being West Germanic languages and Spanish a Romance one. Thus, typology may explain the higher correlation in the L1 Dutch speakers in Lemhöfer and Broersma (2012). Nevertheless, if this conclusion is on the right track, it is not clear why correlations are lower in L1 Spanish speakers than in L1 Chinese speakers, given that Spanish is certainly overall typologically closer to English than Chinese is. As the LexTALE has been used with speakers of 31 different L1s, it is of paramount importance that more studies like the present one are conducted with speakers of different native languages. In addition, more fine-grained qualitative studies on strategies and cues used by learners in their lexical judgement may also shed light on whether speakers of different native languages draw onto their L1 for cues of identification. Finally, the use of different strategies accrued from different learning experience might also contribute to LexTALE results in addition to their actual lexical knowledge and global language proficiency.

7. A concerning state of affairs and implications for future research

Herein, we maintain that the LexTALE might serve its purpose when examining vocabulary knowledge, *but* the test does not seem to work as well as a measure of global L2 proficiency. Although our data shows that, for advanced groups, significant correlations between the LexTALE scores and the QPT are obtained, the strength of these correlations is less than ideal. Even more importantly, since its publication, the LexTALE has been widely used to capture L2 proficiency in learners with proficiency levels way below advanced. Our results are clear with regards to this practice: trying to measure L2 proficiency in intermediate L2 learners is unwarranted. The LexTALE is not a reliable measure of global L2 proficiency with intermediate learners and the test should not be used as such.

It is surprising that even though (at least) 551 studies have used the LexTALE within the past 10 years, there has only been one investigation (Nakata et al., 2020) that has partially replicated the original study where the LexTALE was proposed (Lemhöfer & Broersma, 2012). Worryingly enough, this seems to be a mere reflection of the lack of replication practices in the SLA literature in general (see Marsden et al., 2018; McManus, 2021 for overviews). McManus (2021) raises an important concern where he states that new studies may in fact be built on ‘unverified and unconfirmed’ results. We make this claim ours, too. We further stress that this is even more worrisome in the present case because the field is using a tool that has not been adequately validated to operationalize a key construct in our investigations. Thus, we call for caution when using the LexTALE, but, more importantly, we join many others in asking for increased replication practices in the SLA literature (e.g., Marsden et al., 2018; Porte & McManus, 2019), with particular emphasis on studies proposing widely used tests of important constructs, such as proficiency.

Finally, we strongly agree with Lemhöfer and Broersma (2012) and others who have raised concerns regarding how little consensus exists in the SLA literature about how proficiency is operationalised. Replicating their words: '[t]he field urgently needs a quick, valid and reliable measure of L2 global proficiency'—but let us add: We, as a field, also need to join forces in further validating and making responsible use of the ones that we currently have at our disposal. Until the field comes up with more specific and reliable measures of L2 proficiency, we consider that we can only make strong claims about such a crucial factor if our investigations employ an array of tests/tasks to measure it in all (or a large extent of) its complexity—always depending on the type of grammatical phenomena we are interested in.

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database (Release 2)*. University of Pennsylvania, Linguistic Data Consortium.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

Council of Europe (2011). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe.

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. https://doi.org/10.1080/15434303.2011.565844

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. https://doi.org/10.1017/S1366728911000678

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. McGraw Hill.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Malovth, P., & Benati, A. (2018). *The Handbook of Advanced Proficiency in Second Language Acquisition* (P. A. Malovrh & A. G. Benati (Eds.)). Wiley-Blackwell. https://doi.org/10.1002/9781119261650

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. https://doi.org/10.1111/lang.12286

Meara, P. (1996). *English Vocabulary Tests: 10 k. Unpublished manuscript.* Center for Applied Language Studies.

Nakata, T., Tamura, Y., & Aubrey, S. (2020). Examining the Validity of the LexTALE Test for Japanese College Students. *The Journal of AsiaTEFL*, *17*(2), 335–348. https://doi.org/10.18823/asiatefl.2020.17.2.2.335

Norris, J., & Ortega, L. (2003). Defining and Measuring SLA. In C. Doughty & M. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 716–761). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470756492.ch21

Norris, J., & Ortega, L. (2012). Assessing learner knowledge. In S. Gass & A. Mackey (Eds.), *The Routledge Handbook of Second Language Acquisition* (pp. 573–589). Routledge.

Park, H. I., Solon, M., Dehghan-Chaleshtori, M., & Ghanbar, H. (2022). Proficiency Reporting Practices in Research on Second Language Acquisition: Have We Made any Progress? *Language Learning*, *72*(1), 198–236. https://doi.org/10.1111/lang.12475

Porte, G., & McManus, K. (2019). *Doing Replication Research in Applied Linguistics*. Routledge.

Thomas, M. (1994). Assessment of L2 Proficiency in Second Language Acquisition Research. *Language Learning*, *44*(2), 307–336. https://doi.org/10.1111/j.1467-1770.1994.tb01104.x

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). John Benjamins.

Thompson B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44, 423–432. https://doi.org/10.1002/pits.20234

Tremblay, A. (2011). Proficiency Assessment Standards in Second Language Acquisition Research. *Studies in Second Language Acquisition*, *33*(3), 339–372. https://doi.org/10.1017/S0272263111000015

Eloi Puig-Mayenco *Corresponding author*
School of Education, Communication and Society
Faculty of Social Sciences and Public Policy
King's College London
Waterloo Campus
London SE1 9NH, United Kingdom
e.puig-mayenco@kcl.ac.uk

Adel Chaouch-Orozco
Department of Chinese and Bilingual Studies
Faculty of Humanities
The Hong Kong Polytechnic University
Hong Kong SAR, China
adel.chaouchorozco@polyu.edu.hk

Hong Liu
Department of Applied Linguistics
Xi'an Jiaotong-Liverpool University
Suzhou Industrial Park – Dushu Lake Higher Education Town
Ren'ai Road 111
Suzhou, China
Hong.Liu@xjtlu.edu.cn

Fernando Martín-Villena
Department of English and German Philology
Faculty of Arts and Humanities
Universidad de Granada
Campus de la Cartuja
Calle del Prof. Clavera S/N
18011 Granada, Spain
fmartinvillena@ugr.es