Signos
Ensino · Humanidades

UNIVATES

Revista Signos, Lajeado, ano 42, n. 2, 2021. ISSN 1983-0378
DOI: http://dx.doi.org/10.22410/issn.1983-0378.v42i2a2021.3027
http://www.univates.br/revistas

A PROPOSAL FOR RECALIBRATION OF A VOCABULARY LEVELS TEST AS DIAGNOSIS OF LATE BILINGUALS' L2 PROFICIENCY

Jesiel Soares-Silva[1]
Luiz Henrique Mendes Brandão[2]
Lara do Nascimento Góes[3]
Brenda Lorraine Grillo Silva[4]
Geovanne Barbosa[5]
Jáliton Luiz Souza Ferreira[6]
Lays Albuquerque Benevides[7]

Abstract: This study explores the recalibration and adequacy of a measure of vocabulary size – the Vocabulary Levels Test (VLT) – as a predictor of Brazilian Portuguese-English speakers' ability to access grammatical representations through their non-dominant language. Such endeavor concerns a specific part of the test (composed majorly by cognates) which has been blurring the results when participants are natives in Latin-derived languages, such as Brazilian Portuguese. A new test (nVLT) was designed, with a novel version of this problematic part (level 4) present in the older test that, now, avoids the proliferation of cognates. Both versions were applied to a number of Brazilian participants and the results were correlated with another proficiency measure, taken from an acceptability judgment task designed according to the model reported in Souza et al (2015). When the low-proficiency participants took the VLT, there were a decreasing pattern in their scores from the first level of the exam all the way to level 3 (because each level is harder than the preceding). But, when they got to level 4, which is "harder" than level 3, their scores increased surprisingly, and then decreased again in level 5. When they performed the nVLT, which has a level 4 recalibrated

[1] Docente do Departamento de Letras (UFOP).

[2] Mestrando em Estudos Linguísticos (UFMG).

[3] Mestranda em Educação (UFVJM).

[4] Graduanda em Letras (UFMG).

[5] Graduando em Letras (UFVJM).

[6] Graduando em Letras (UFVJM).

[7] Graduanda em Letras (UFVJM).

-- Artigo recebido em 24/10/2021. Aceito em 14/12/2021. --

(without latin cognates), the decreasing pattern was maintained evenly through the whole test. These results from nVLT show an internal coherence of the test due to the recalibration.

Keywords: Vocabulary Levels Test; recalibration; bilinguals.

UMA PROPOSTA DE RECALIBRAÇÃO DE UM TESTE DE NÍVEIS DE VOCABULÁRIO COMO DIAGNÓSTICO DE PROFICIÊNCIA L2 DE TARDIÓES

Resumo: Este estudo explora a recalibração e adequação de um instrumento de medida de tamanho de vocabulário – o Vocabulary Levels Test (VLT) – como preditor de habilidade de acesso a representações gramaticais através da língua não dominante por parte de bilíngues brasileiros do par português-inglês. Tal empreendimento está relacionado a uma parte específica do teste (composta majoritariamente por cognatos) que tem poluído os resultados quando participantes são usuários nativos de línguas derivadas do latim, tal como o português brasileiro. Um novo teste (nVLT) foi desenvolvido, com uma nova versão dessa parte problemática (nível 4) presente na versão antiga. Essa nova versão evita a proliferação de cognatos. Ambas as versões foram aplicadas a um número de participantes brasileiros e os resultados foram correlacionados com outra medida de proficiência, abstraída de uma tarefa de julgamento de aceitabilidade desenhada nos moldes estabelecidos por Souza et al (2015). Quando os participantes de baixa proficiência fizeram o teste VLT, houve um declínio em suas pontuações desde o primeiro até o terceiro nível do exame (porque cada nível do exame é mais difícil que o anterior). Mas, quando eles chegam ao nível 4, há uma melhora significante e, portanto, surpreendente neste nível que, teoricamente, é mais difícil que o nível 3, e depois as pontuações voltam a cair quando chegam ao nível 5. Já no nVLT, que tem o nível 4 recalibrado (sem cognatos latinos), o padrão de declínio na pontuação se mantém do primeiro ao último nível. Estes resultados do nVLT mostram uma coerência interna do teste devido à recalibração.

Palavras-chave: Vocabulary Levels Test; recalibração; bilíngues.

1. Introduction

When he opened a book chapter by stating "everyone is bilingual," Edwards (2006, p. 7) did not overestimate his appraisal, since his definition of bilingualism encompasses the knowledge of at least one word from a language that is different from the mother tongue. Edwards' description is one among several definitions of bilingualism that reach many aspects from linguistic to political issues (Edwards, 2012).

The last decades have shown the necessity of inserting the bilingualism issue into the psychological, political, and social debate, because the discussion on bilingualism has played a crucial role in constructs, such as ethnicity, communities, minority groups (Edwards, 2012). In this study, we define bilingualism under a psychological aspect rather than social or cultural.

We consider bilinguals to be those who operate in two languages, regardless of their level of proficiency in either language (Grosjean, 1998, 2013). Grosjean (2013, p. 5) characterizes bilingualism (and multilingualism) as "the use of two or more languages (or dialects) in everyday life", and bilinguals as those who "use the two languages—separately or together—for different purposes, in different domains of life, with different people" (Grosjean, 2008, p. 14).

We align ourselves with Grosjean (2008) on the idea that "the bilingual is not the sum of two complete or incomplete monolinguals; rather, he or she has a unique and specific linguistic configuration." Although the proficiency level does not determine whether someone is bilingual or not, an important question to be considered is how to measure one's degree of language competence.

Edwards (2012) mentions types of measurement such as rating scales, tests of speaking fluency, and self-assessment. The author points out that the major part of these measurements is the ability to provide information about a set of one's abilities, but not about all of the facets in which a bilingual is involved. There are several 'labels' to define the highly proficient bilingual, such as balanced bilingual, ambilingual, or equilingual. However, this idea of equilibrium has been overcome, since bilinguals seem 16 to be those who operate two languages in two different ways (Edwards, 2012, Grosjean, 1998).

This study explored the recalibration and adequacy of a measure of vocabulary size in the Vocabulary Levels Test (VLT) as a predictor of Brazilian Portuguese-English speakers' ability to access grammatical representations through their non-dominant language. Such endeavor concerns a specific part of the test (composed majorly of cognates) which has been blurring the results when participants are natives in Latin-derived languages, such as Brazilian Portuguese.

2. L2 proficiency as a construct

Someone who is bilingual is able to use two languages, although proficiency in each language may vary. There are several ranges of fluency in a second language, therefore the level of proficiency may not determine if someone is bilingual or not. For instance, a Spanish native speaker who has moved to the U.S. during her/his teenagehood or early childhood, and who has the need to learn English as a second language through immersion.

In other cases, such as in immigration, the learner can be born in the U.S. and speak English as L1 and Spanish as a heritage language. Bilinguals can learn or even, in some cases, acquire their second language. When the L2 is learned in the beginning of the learner's life it is called early bilingualism and the late bilingualism is when the language is learned after the full development of the cognitive part of the brain (Montrul, 2002)

By all means, a more globalized world has conveyed a need to speak at least a second language and in the academic world this means foremost learning English language. The background and foreground for each learner play a major role in the learning process, so family, job, career and also economic and world events can change a person's second language history. The language field which studies this phenomenon is called bilingualism or multilingualism, if more languages are acquired or learned (Montrul, 2005)

Some reasons for taking proficiency tests are: immigration, citizenship, study or work. In the assessment of languages, tasks are designed to measure learners "productive language skills through performances which allow candidates to demonstrate the kinds of language skills that may be required in a real world context" (Wigglesworth, 2008, p. 111).

Generally, these four skills are split in two major categories, being receptive and productive. The first is related to the skills we do not need to think about before using. Therefore, our first receptive skill is acquired when we are children. We first learn to listen to acquire vocabulary then we start speaking. Speaking fits as productive because we make an effort to formulate what we will say. The same applies for reading and writing. In a nutshell, reading and listening are receptive skills and writing and speaking are productive ones. (Milton, 2013)

Throughout the decades, a necessity to better estimate proficiency among people from different countries who were learning English has emerged. Soares-Silva (2016) states that "the measurement of bilingual speakers' differential proficiency profiles is a matter of absolutely critical importance for the psycholinguistics of bilingualism" (p. 35). Moreover, the creation of the Common European Framework Reference for Languages (CEFR) was a milestone for language appraisal. It was created in 1996 in order to standardize language learning in Europe, it is a standard not only for English, but for all languages.

One of the objectives of the CEFR is to help institutions to describe the levels of proficiency required by existing guidelines, tests and exams in order to facilitate the comparison between different certification systems. The CEFR divides language proficiency into six different levels which range from A1 to C2. A corresponds to the basic levels, which is divided into breakthrough (A1) and waystage (A2); B, intermediate, threshold (B1) and vantage (B2); and C, advanced, efective (C1) and mastery (C2).

By using these levels, one can know what the other is capable of communicating in the target language. It is useful in the job market or for institutions, such as colleges. There was a similar initiative from the United States, which generated the American Council on the Teaching of Foreign Languages as the ACTFL framework, created in 1967. This one is divided into eleven levels, from novice low, corresponding A1 in the CEFR, and distinguished, corresponding C2.

Nowadays, the CEFR is the standard reference for the majority of proficiency tests in the market, even the ones not related to English tend to follow this pattern. Tests such as the International English Language Testing System (IELTS) general and academic, the Test of English as a Foreigner Language (TOEFL IBT-Internet based test), TOEFL PBT (paper based test) are able to place the student in one of the six levels of proficiency. Other tests like Cambridge First, C1 advanced and C2 proficiency have only a pass or fail result. The details and characteristics of the proficiency tests will be deepened in the next topic.

2.2 The Vocabulary Level Test

The Vocabulary Levels Test (VLT) (Nation, 1990) was created with the perspective of assessing the vocabulary level of English speakers as a second language. The test works independently of specific contexts, since there is no need to situate words in specific occurrences.

The VLT is divided into five parts – it contains the same exercise structure, although an increase in difficulty in each level advance – each level containing six items (total of 30 items). Participants need to match 3 of the 6 words provided in their respective meanings. Therefore, each level of the VLT produces a total of 18 correct word combinations. The satisfactory achievement of points at each level allows us to identify the vocabulary size of each participant.

Levels 1, 2 and 3 refer to the approximate knowledge of 2000, 3000 and 5000 most frequent words, respectively. Level 4 corresponds to academic and scientific vocabulary; and level 5 corresponds to the knowledge of 10,000 most frequent words. Nation (1990) explains that it would be necessary to get at least 12 out of 18 of the words that are in each level for there to be a satisfactory result.

VLT is based on vocabulary that is already used and frequent in the English language. The British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) provide a *corpus* of 34 billion (BNC) and 1.0 billion words (COCA), being composed by spoken language, fiction, magazines, newspapers and even other hosts that allows its user to expose different contexts of the language. The BNC/COCA also displays a frequency list browser, allowing the user to search a word form or part of a speech providing a large *corpora* that has significant materials for linguistic purposes. VLT tests are composed of these *corpora*.

2.3 The Acceptability Judgment Task

For this study, the design reported in Souza et al (2015) study of a timed Acceptability Judgment Task (AJT) was applied. The sentences in the AJT were displayed on a computer screen using the Ibexfarm (to be explained) software. Task takers were exposed to sentences (presented one-by-one) in the center of the screen. Then, they judged each sentence according to its well-formedness using a 5-point Likert scale which ranges from bad-formed (1) to well-formed (5). Responses were given using the numeric keys of a computer keyboard, and a time limit of 8 seconds was set for the judgment calls (this time limit of 8 seconds was reported in Souza et al, 2015).

There were two types of sentence violations applied to the experiment: a) argument structure realization violations involving unergative verbs in transitive syntax, and b) explicit morpho-syntactic violations involving long distance dependencies (Wh-movement) and subject-verb agreement. Argument structure realization violations were adopted because according to White (2003), L2 argument structure may present a challenge to L2 learners, as "interlanguage lexical representations may not correspond to argument structures encoded in the lexicons of native speakers of the L2" (White, 2003, p. 206). Examples of the sentences:

1 *The man laughed the children during the party. *Transitivized unergative verb*

2 *The girl give the cats milk twice a day. *Agreement violation*

3 * What did Steven read the book that Helen talked about? *WH movement violation*

4 The chief ran the boys around the park. *Grammatical sentence (Induced movement)*

5 The girls melted the cheese in the bowl. *Grammatical sentence (change-of-state verbs*

6 The gardener swiped the table clean. *Grammatical sentence (resultatives)*

3. Methodology

For the calibration process, words that would not be perceived by crosslanguage similarities (cognates) were chosen, considering that English has borrowed words from several languages over the course of its history. The Latin language, especially vocabulary, is seen as an important part in academic writings and speeches.

As part of the process of learning English, those words will naturally occur. However in a vocabulary test which has an entire part dedicated to academic English, those words may bias the results. Considering this aspect, through *COCA,* we created a new level 4 for the VLT, not focusing on academic vocabulary, and considering word frequency as a criterion as in the other parts of the test. The main goal was to avoid the proliferation of cognates.

3.1 The Ibexfarm as a software to administer Sentence Judgment Tasks

Ibexfarm is a digital platform created by Alex Drummond and hosted on the Ibex farm website. The tool makes the creation of many online psycholinguistic experiments possible, as acceptability judgments, self-paced reading and many others. Furthermore, the range of possible experiments is steadly growing, according to the website. The name Ibex is an acronym for (Internet Based Experiments).

The system is modular and uses only two types of programming languages, namely Javascript and HTML. It is necessary for the researcher to know a little of those programming languages in order to create the experiments, but there is plenty of information on the internet on how to make use of them. After it is created, the experiments can be sent straight to the participants through a link generated by the website. This link gives instant access to the content of the experiment.

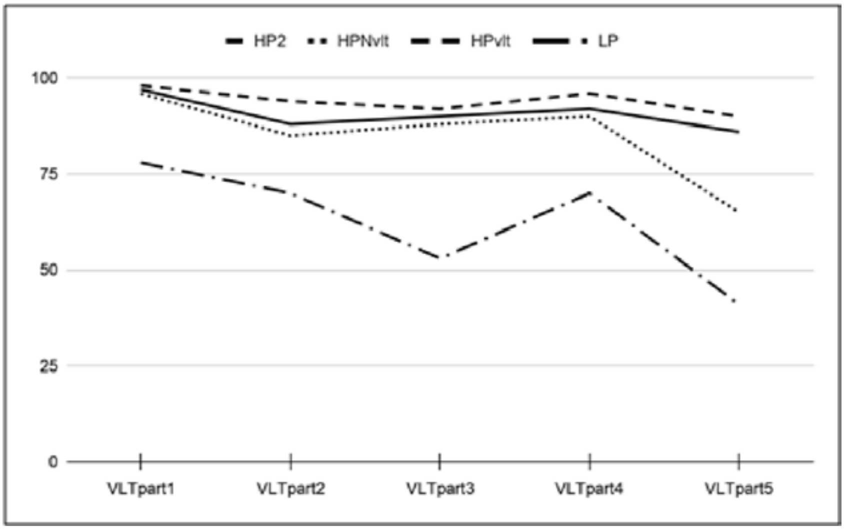
Ibexfarm has been used as a tool on many papers since it was made available online. Chow and Chen (2020), for example, used Ibexfarm to conduct a cloze probability norming study. They analyzed the predictions comprehenders make when processing the input. Kim, Park and Seo (2020) also made use of the tool, conducting a self paced reading with a comprehension question at the end, as well as a cloze task with Korean L2 learners of English. Their goal was to investigate how they predicted upcoming syntactic structure based on newly received input during sentence processing.

4. Data Analysis

As stated, we aimed at investigating how both VLT and its recalibrated version (nVLT) perform within a group composed of different levels of proficiency. Moreover, we intended to correlate the results with another proficiency measure taken from an acceptability judgment test (Soares-silva, 2016).

To begin with, we will have a look at the performance of those participants who are highly proficient in both VLT and nVLT (HP2, henceforth), those who were tested high proficient only in the nVLT (HPNvlt, henceforth), those who tested high proficient only in the VLT (HPvlt, henceforth) and those who were tested low proficient (LP, henceforth). We displayed their performance in each test for it to be able to show the details of each part. First of all, all groups' performance in the VLT in its original version:

Graphic 1: Groups' performance in the VLT



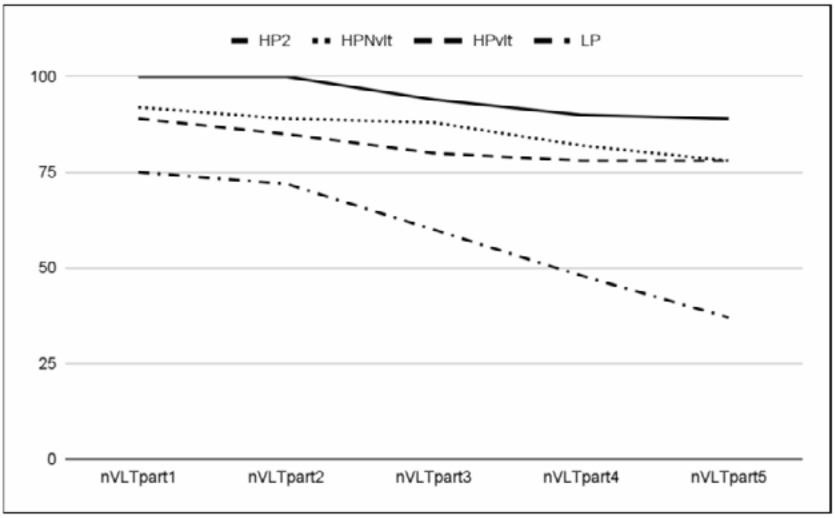
As we can observe in Graphic 1, the performance of HP2 in the VLT is quite as expected, despite a slight increase from par 2 to part 4, which can be associated with the Latin words presented in the level 4 of VLT. Likewise, HPNvlt and HPvlt are similar throughout the parts of the test. However, we call the attention to the LP group. As can be easily noted, as expected, LP's performance decreases along the test from part 1, to part 2 to part 3. However from part 3 to part 4, we notice a considerable increase in their performance, which goes against the expected in the test, since part 4 is composed of less frequent vocabulary.

As we stated in our hypothesis, this may be due to the fact that part 4 of VLT (academic part) is composed majorly of words derived from Latin, which make them cognates for Brazilian Portuguese speakers. This way, even at a more difficult level, they performed better.

The meaning of words such as *computer* and *exercise* can be taken through Portuguese cognates (*computador* and *exercício*). This fact can leave it blurred if the speaker is actually performing well in a less-frequent set of words in the VLT or just relying on cognates. Even considering that knowing cognates is obviously part of knowing the language, one can observe that this part 4 can bias the test in its totality.

Differently, when we tested all groups' performance in the nVLT, in which we recalibrated the part 4, avoiding the cognates and following the frequency criteria from the corpus, results present themselves differently:

Graphic 2: Group's performance in the nVLT

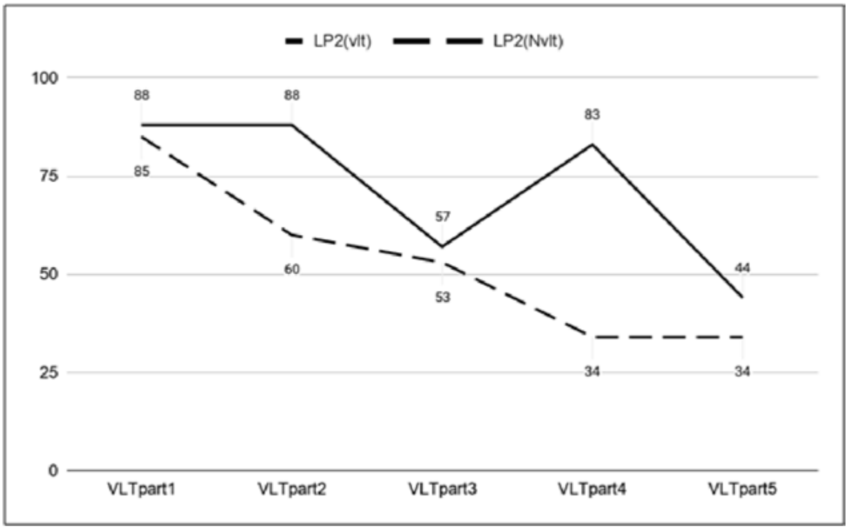


In graphic 2, it is noticeable that the performance of the high proficient groups (HP2, HPNvlt, HPvlt) are similar to the extent of decreasing as the test goes by from more to less frequent vocabulary. It is shown that all groups, including LP, seemed to maintain the same performance from part 1 to part 2, and from the part 3 on, it starts to decrease.

Differently from graphic 1, we can see that the low proficient group keeps decreasing from levels, which we expect from a test. These results suggest the problem with cognates may have been fixed to Brazilian Portuguese-English learners who take VLT as a proficiency measure.

After reaching the fact that the High Proficient groups and mainly LP groups for both VLT and nVLT gives us elements to sustain our hypothesis, we decided to consider only the LP group and analyze their performance in part 4 of both VLT and nVLT to separate the result and have a clearer look at our assumption. The results can be seen in Graphic 3:

Graphic 3: Low Proficient groups' performance in the parts of the test



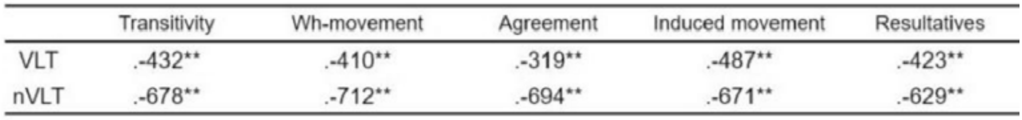
Graphic 3 demonstrates that the LP group in the VLT performs as expected in part 1 (88%), part 2 (88%) and part 3 (57%). However in part 4, their performance goes up to 83%, which is unexpected since those are less frequent vocabularies. Finally, they go down to 44% in part 5, returning to the expected. Differently, the LP group when performing in the nVLT decreases continuously from part 1 (85%) to part 2 (60%) to part 3 (53%) to part 4 (34%) and keeping the same value at part 5. These results from nVLT show an internal coherence of the test due to the re-calibration.

In order to establish a criterion for our findings, we calculated Cronbach's alpha for both tests to check the internal consistency. Our hypothesis was that nVLT would present a more reliable internal consistency than the VLT. For the VLT, consisting in 5 parts of 6 items per part, we had α = .46. For the nVLT, with the same amount of parts and items, but with the recalibration of the fourth part, we had α = .63. As we can see, there is a difference between the internal consistency between the tests. Although the literature claims that a Cronbach's alpha of .63 is not excellent, it is clearly higher than .46.

We then proceeded to the confirmatory investigation of the correlation of VLT and nVLT scores and the Acceptability Judgment Task. A Pearson product moment correlation coefficient (r) was computed to assess the relationship among the two tests and the task in order to verify the degree of correlation between the scores each one produces as diagnosis of L2 proficiency.

For VLT and the AJT, we had *r*(30)=-0.39, p< .05 and for nVLT and the AJT we had *r*(30)= -0.63, p< .05. Both results were significant at a p-value inferior to 0.5. However, the correlation index for the nVLT with the AJT was higher, showing that the recalibration was useful to make the test more cohesive when compared to another proficiency measure (AJT). In order to gather more detail about the correlation, we performed the same test correlation considering both tests and each structured in the AJT:

Table 1 - Correlation of VLT and nVLT with the the Acceptability Judgment Task



	Transitivity	Wh-movement	Agreement	Induced movement	Resultatives
VLT	.-432**	.-410**	.-319**	.-487**	.-423**
nVLT	.-678**	.-712**	.-694**	.-671**	.-629**

**Correlation is significant at the 0.05 level (2-tailed)

Table 1 demonstrates that in all types of sentences, correlation between AJT and nVLT was higher than with VLT. For the transitivity-violation sentence type (.678 > .432), the Wh-movement violation (.712 > .410), the agreement violation (.694 > .319), the induced movement (.671 > .487) and resultatives (.629 > .423) the nVLT presented a higher correlation than the VLT considering the AJT as reference. These results reinforce our assumption in this study.

5. Final remarks

In this study, we advanced the work reported by Souza et al (2015) aiming at validating a measure of vocabulary size — VLT (Nation, 1990) and Soares-Siva (2016) — as a diagnostic instrument to evaluate bilinguals. In Souza, Duarte & Berg (2015), the bilingual group was composed of Brazilian Portuguese-English bilinguals at college level, similarly to participants of this study. In Soares-Silva (2016), the part 4 of VLT was taken off in order to avoid the cognate bias, making the test more cohesive, although consequently incomplete. By proposing a recalibration of the test, we presented here an advance in this useful tool.

Moreover, the results of this study expands the reach of VLT by correlating it with a psycholinguistic task in which the knowledge type involved is from a different nature. In other words, we assumed VLT requires explicit knowledge, since after taking the test, a participant would be able to verbalize and explain his choices, to talk about the knowledge being measured. In addition, the questions allow the test-takers to rely on cognitive strategies such as association, elimination.

On the other hand, the AJT with those specific structures, and with a time pressure, requires a more implicit knowledge from the individuals. It is important to mention that there are several proficiency tests, including tests of vocabulary size, that produce more than two levels of proficiency, including intermediate levels. Depending on the way scores are computed, even VLT is able to produce different levels of proficiency.

In this study, by following the same procedures of Soares-Silva (2016) and advancing in the part 4 problem, our purpose was to administer the nVLT as a diagnostic test. In other words, we were looking for the detection of high-proficient individuals from non-high proficient individuals (low). After the recalibration, nVLT demonstrated to be adequate in generating two groups of proficiency, being the high proficiency group significantly different from the low-proficiency group.

References

BEGLAR, D. A Rasch-based validation of the Vocabulary Size Test. Language Testing, Japan: Temple University, 2010.

CHOW, W. Y.; CHEN, D. Predicting (in)correctly: listeners rapidly use unexpected information to revise their predictions. Language, cognition and neuroscience, 35(9), 2020, p. 1149-1161.

EDWARDS, J.. Bilingualism and Multilingualism: Some Central Concepts. In: In: BHATIA, T.; RITCHIE, W. (Eds.) The Handbook of Bilingualism and Multilingualism, Chichester: John Wiley & Sons, 2012.

GROSJEAN, F. Studying bilinguals: methodological and conceptual issues. Bilingualism: Language and Cognition, 1 (2), 1998, p. 131-149.

______. Studying Bilinguals. Oxford: Oxford University Press, 2008.

______. Bilingualism: A short introduction. In: GROSJEAN, F. & Li, P. (eds.) The Psycholinguistics of Bilingualism. Oxford: Wiley-Blackwell, 2013.

KIM, B.; PARK, M. K.; SEO, H. J. L2ers' predictions of syntactic structure and reaction times during sentence processing. Linguistic research, 37(special edition), 2020, p. 189-218.

MILTON, J. Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In: BARDEL, C.; LINDQVIST, C. & LAUFER, B. (Eds.) L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis. European Second Language Association, 2013.

MONTRUL, S. Incomplete acquisition and attrition of Spanish tense/ aspect distinctions in adult bilinguals Bilingualism: Language and Cognition, 5 (1), 2002.

______. Second language acquisition and first language loss in adult early bilinguals: exploring some differences and similarities. Second Language Research, 21 (3), 2005, p.

199-249.

NATION. P. Teaching and Learning Vocabulary. Boston, MA: Heinle & Heinle, 1990.

SOARES-SILVA, J. Exploring a vocabulary test and a judgment task as diagnoses of early and late bilinguals' L2 proficiency - Doctorate Thesis, UFMG, 2016.

SOUZA. R.; OLIVEIRA, C.; SOARES-SILVA, J.; PENZIN, A. & SANTOS, A. Estudo sobre um Parâmetro de Tarefa e um Parâmetro Amostral para Experimentos com Julgamentos de Aceitabilidade Temporalizados. Revista Estudos da Linguagem, 23(1), 2015, p. 211-244.

WHITE, L. Second language acquisition and Universal Grammar. Cambridge: Cambridge University Press, 2003.

WIGGLESWORTH, G. Task and Performance Based Assessment. Encylopedia of Language and Education (pp.209-226) Springer, 2008 (pp.209-226)