UMA VERSÃO EM PORTUGUÊS EUROPEU DO *C-TEST*

EUROPEAN-PORTUGUESE VERSION OF THE C-TEST

Masayuki Yamada*
masayu36@miador.nagoya

A avaliação da proficiência é uma questão importante na interpretação dos resultados das investigações sobre a aprendizagem de língua segunda. Dados contraditórios entre os estudos podem decorrer dos diferentes procedimentos empregues para avaliar os aprendentes. De modo geral, são consideradas informações institucionais, estimativas baseadas no perfil linguístico e, mais raramente, são utilizados resultados de testes independentes. No presente estudo, foi desenvolvida uma versão portuguesa do *C-test*, um teste de preenchimento simples utilizado para medir a proficiência geral. Elevados coeficientes de *Cronbach Alfa* e *Omega* revelaram a fiabilidade do teste. A classificação da proficiência dos aprendentes baseada na pontuação do teste demonstrou uma correlação forte com a avaliação pelos professores (r = 0.74), mostrando que o teste é eficiente para avaliar a proficiência geral. Para além disso, outros parâmetros, nomeadamente nível da turma e autoavaliação dos aprendentes também mostraram uma correlação forte com a avaliação por professores (r = 0.79 e 0.77). A utilização futura do teste é discutida.

Palavras-chave: C-test. Avaliação. Proficiência. Aprendizagem de português. Português como língua estrangeira (PLE).

Proficiency assessment is crucial when one interprets the results of SLA studies. Inconsistent results among them can occur from a lack of uniformity in the methods of the proficiency assessment. In general, institutional status and estimates based on learners' linguistic profile are taken into consideration. However, the score of independent tests is rarely utilized. In this article, we developed a European-Portuguese version of the C-test, a simple fill-in-the-blank test used to measure general proficiency of a foreign language. High coefficients of Cronbach Alfa and Omega were found, demonstrating its reliability. Proficiency classification based on the test score showed strong correlation with the assessments made by teachers (r = 0.74), which could imply that the test captures the general proficiency. Moreover, other parameters, namely classroom level and learners' self-assessment, also correlated strongly with the assessment by teachers (r = 0.79 and 0.77). The use of the test for the future is discussed.

* Universidade do Minho, Portugal.

Keywords: C-test. Assessment. Proficiency. Acquisition of Portuguese. Portuguese as a foreign language.

1. Introdução

Há várias décadas que um grande número de estudos (*e.g.* Anderson 1992; Bialystock & Smith 1985; Corder 1967; DeKeyser 1997; Ellis 2015; Firth & Wagner 2007; Isabel 2006; Krashen 1982; Lado 1957; Leiria 1991; Li 2010; Mcdonough & Kim 2009; Pienemann 1998; Pinto 2014; Rodrigues 2015; Selinker 1972; VanPatten 2007) tem sido realizado no que toca à aprendizagem de língua segunda e estrangeira[1] (doravante ALS). Através desses estudos foram revelados, por exemplo, vários fatores que afetam a variabilidade do desempenho e do nível de proficiência dos aprendentes, tais como língua materna, idade de início da aprendizagem da língua-alvo, duração da aprendizagem (em anos), frequência de uso da língua-alvo, entre outros. Estas variáveis têm sido tendencialmente aproveitadas para determinar a proficiência da L2. Sobretudo, o nível da turma e a duração da aprendizagem são as estimativas mais comuns na área da ALS, apesar da inexistência de homogeneidade em proficiência que os aprendentes apresentam nestes grupos (Tremblay 2011). Sendo assim, é necessário avaliar a proficiência de forma mais precisa e cuidadosa, por exemplo, através de um ou mais testes independentes.

Thomas (1994; 2006) examinou os métodos de aferição da proficiência de aprendentes em estudos publicados durante dois intervalos (1988–92 e 2000–04), em quatro revistas de ASL: 1) *Applied Linguistics*, 2) *Language Learning*, 3) *Second Language Research* e 4) *Studies in Second Language Acquisition*. O autor classificou esses métodos em quatro tipos: i) juízo impressionista (*impressionistic judgment*)[2], ii) informações institucionais (*Institutional status*), iii) avaliação interna (*in-house assessment*) e iv) teste padronizado (*standardized test*). A percentagem de utilização das

1 Na área da ALS, o termo "língua segunda" refere-se, frequentemente a uma língua adicional: segunda, terceira, quarta e assim por diante (Ellis 2005). No presente trabalho, utilizam-se os termos língua segunda (L2) e língua estrangeira (LE) como sinónimos.

2 Trata-se de uma aferição subjetiva sem dados de suporte (p. ex., Os participantes deste estudo são principiantes.) ou com base em anos da estadia em locais em que se fala a língua-alvo.

informações institucionais (ii) foi de 40,1% no primeiro período e de 33,2% no segundo período; a de testes independentes (*avaliação interna* e *teste padronizado*) foi de 36,3% e 42,6%, respetivamente.

Também Trembley (2011), numa revisão de estudos sobre L2 publicados entre 2000 e 2008, indica que apenas 37,2% utilizaram um teste independente. Nos 62,8% dos estudos que não utilizaram testes independentes, 60,4% empregaram os parâmetros nível da turma e duração da aprendizagem para classificar o nível de proficiência dos participantes.

Deste modo, apesar de se verificar a utilização de testes independentes, este tipo de aferição ainda não pode ser considerado comum na área. Este cenário é mais saliente relativamente a algumas línguas. Por exemplo, Trembley (2011) relata ainda que, quanto aos estudos sobre o francês como L2/LE, o número é mais reduzido, somando apenas três. Algumas razões possíveis para a utilização reduzida de testes independentes são o facto de serem pagos e demorados e o número reduzido de instrumentos desenvolvidos para o efeito. Além disso, Lee-Ellis (2009) aponta o facto de que, ao contrário de línguas comummente investigadas como o inglês, não existe nenhuma medida "prática" de proficiência para o coreano, por exemplo, e, por conseguinte, testes independentes nos estudos de coreano como L2 são pouco utilizados.

Quanto ao português, o cenário é semelhante ao caso do francês e do coreano. Fizemos uma pesquisa de estudos sobre português em duas revistas, *Second Language Research* e *Studies in Second Language Acquisition*. Seguindo os critérios de Thomas (2006), foram excluídos: i) revisão de literatura ou livro, ii) ensaio, comentário ou outros tipos de estudo, em que se discute uma temática geral e não se realiza um estudo empírico com dados recolhidos. Adicionalmente, como o nosso interesse está voltado para a aprendizagem tardia, ainda foram excluídos estudos referentes à aprendizagem precoce (por crianças) e à de bilingues. Na *Second Language Research*[3] foram encontrados 85 resultados para a palavra "*Portuguese*", sem determinar o período. Apenas o estudo de Montrul *et al.* (2010) foi enquadrado nos critérios de pesquisa estabelecidos.[4] Os autores avaliaram a proficiência de ingleses e espanhóis, aprendentes de português, a partir da autoavaliação e da duração da aprendizagem, ou seja, utilizando o critério do "juízo impressionista" de Thomas (2006). Já na *Studies in Second Language*

3 http://journals.sagepub.com/home/slr (Consultado em: 7 de março de 2019).

4 Dos 85 estudos apenas três continham a palavra em questão, isto é, "*Portuguese*", no seu título. É provável que a palavra tenha sido encontrada no corpo do texto dos artigos.

Acquisition[5], foram encontrados 16 resultados, sendo que nenhum se encaixava nos critérios indicados acima.[6] Consultamos, de seguida, a página da *Internet, Cátedra Português Língua Segunda e Estrangeira*[7], que oferece uma lista de referências sobre aprendizagem e ensino de português L2. Dos 417 estudos listados, 277 foram descarregados a partir da ligação *download*.[8] Considerados os nossos critérios, restaram 55 estudos. A Tabela 1 mostra a utilização dos quatro métodos da classificação de Thomas (2006). As informações institucionais são maioritariamente utilizadas (67%), seguido pelo método juízo impressionista (25%), sendo utilizado algum teste independente apenas em 7% dos estudos.[9]

Tabela 1. Utilização de quatro métodos de avaliação de proficiência

Tipo	Número	%
juízo impressionista	14	25%
informações institucionais	37	67%
avaliação interna	3	5%
teste padronizado	1	2%
total	55	

Obviamente, a utilização de cada método pode ser legítima, dependendo do objetivo de investigação, pelo que não se pode afirmar que todos os estudos apresentem um problema metodológico. Porém, uma vez que i) há, frequentemente, dentro da turma, heterogeneidade em proficiência de

5 https://www.cambridge.org/core/journals/studies-in-second-language-acquisition (Consultado em: 7 de março de.2019)

6 Há alguns estudos que investigam a aprendizagem precoce ou por bilingues, ou que relatam uma temática geral sem recolher dados novos. Entre outros, por exemplo, Major (2007) investigou a identificação de acento de línguas familiares ou não familiares. Neste estudo, participaram americanos com/sem experiência do português. Os participantes com experiência podem ser considerados como aprendentes de português. No entanto, como não foi possível aceder aos textos integrais, apenas o resumo não apresentava informações suficientes para a inclusão na análise.

7 http://catedraportugues.uem.mz/?__target__=bibli&bib=7 (Consultado em: 7 de março de 2019)

8 Embora houvesse a ligação *download*, alguns artigos não foram obtidos por uma questão de direito de acesso.

9 Julgamos que Correia (2011) utilizou "teste padronizado", visto que os textos analisados foram produzidos "num contexto de avaliação/certificação" do nível B2 do CAPLE, exame padronizado de português como língua estrangeira. No entanto, os textos foram produzidos pelos examinandos do exame DIPLE, equivalente ao nível B2, e não fica claro se todos possuem ou já obtiveram esse nível.

aprendentes e ii) cada instituição utiliza exames de posicionamento diferentes (isto é, existe também heterogeneidade em proficiência entre as turmas do mesmo nível), não se pode comparar nem generalizar os resultados dos estudos realizados, o que constitui um problema nos estudos da ASL.

Em português europeu, existe um teste padronizado, CAPLE[10], efetuado nos Centros de Avaliação de Português como Língua Estrangeira para certificação da competência da língua. Este exame não é, no entanto, tal como os exames padronizados de outras línguas, um instrumento prático para estudos da ASL, já que é pago e demorado (normalmente demora-se um ou dois dias). Sendo assim, não temos instrumentos de acesso fácil para avaliar a proficiência de aprendentes para estudos da ASL, e pensamos que é vantajoso desenvolver um que possa ser utilizado de forma geral e económica no país. No presente trabalho, desenvolvemos uma versão em português europeu do *C-test*, um teste de preenchimento simples utilizado em várias línguas para medir a proficiência geral.

2. C-test

O *C-test* é um teste considerado como medida de proficiência geral de língua, que se baseia no princípio de redundância reduzida (Spolsky 1973), correspondendo, por exemplo, aos *noise test* e *cloze test* (Taylor 1953). De acordo com este princípio, considera-se que as mensagens linguísticas transmitidas no dia-a-dia contêm mais informações do que as que são rigorosamente necessárias. Por esta razão, mesmo que uma parte das informações se perca, consegue-se recuperar ou restituir a mensagem a partir das informações que estão intactas (Raatz & Klein-Braley 2002).

O *C-test* foi proposto por Raatz e Klein-Braley (1981) com a intenção de ultrapassar as limitações do *cloze test*. O teste consiste em 4 – 6 textos curtos autênticos. A primeira frase é mantida intacta e, a partir daí, são eliminados caracteres correspondendo a metade de cada palavra, de modo alternado (palavra sim palavra não) (Klein-Braley 1997). Quando o número de caracteres da palavra é ímpar, a maior parte dos caracteres é eliminada. Caso a palavra contenha apenas uma letra como "*a*", "*o*", "*e*", é ignorada na contagem. A parte eliminada é substituída por um sublinhado de tamanho constante, isto é, independentemente do número de caracteres eliminados, conforme o exemplo infra (Baghaei & Tabatabaee 2015).

10 http://caple.letras.ulisboa.pt

If you were to ask most people who Charles Darwin was, many of them would reply that he was the man who said that we were descended from monkeys. They wo___ be wr___. Darwin d___ no mo___ than sug___ the possi___. What h___ said, a___ proved b___ thousands o___ examples, w___ that ov___ millions o___ years ani___ and pla___ have cha___. This he called evolution.
(retirado de Baghaei & Tabatabaee 2015)

Os textos preparados são previamente verificados por falantes nativos da língua alvo. Consideram-se para utilização no teste apenas os textos cujas taxas de acerto sejam superiores a 90%. Os participantes devem inserir a parte eliminada, ou seja, restituir a palavra original. Apenas a restituição completa é considerada como resposta correta, pelo que, os erros ortográficos, por exemplo, podem ser tratados como resposta incorreta.

Desde a sua proposta, o teste foi traduzido para mais de 20 línguas (Eckes & Baghaei 2015), contando mais de 500 publicações (Grotjahn 2016). Apesar de haver controvérsia no que respeita à sua validade, muitos estudos apoiam a sua utilização do teste (p. ex., Babaii & Ansary 2001; Babaii & Moghaddam 2006; Eckes & Grotjahn 2006; Katona & Dornyei 1993; Klein-Braley 1997; Lei 2008). Outros autores, no entanto, consideram que o teste é demasiado fácil, revelando valores baixos de discriminação de itens (p. ex., Cleary, 1988; Kamimoto, 1993). Ainda outros julgam que o teste é adequado como medida da competência em nível micro (p.ex., leitura e gramática) mas não como medida da proficiência geral (p. ex., Chapelle & Abraham 1990; Cohen 1984).

De modo geral, a validação do *C-test* tem sido efetuada através da abordagem correlativa.[11] Muitos investigadores relatam correlação moderada ou forte entre o teste e outros tipos de testes considerados válidos. Além disso, apesar de o teste se correlacionar bem com vários componentes linguísticos a nível micro (p. ex., vocabulário, gramática, fala, escrita), apresenta melhor correlação com a pontuação total dos testes, isto é, a nível macro (Babaii & Ansary 2001; Eckes & Grotjahn 2006; Grotjahn & Stemmer 2002; Katona & Dornyei 1993), o que implica que o seu constructo do teste se baseia na avaliação integrativa.

11 Também se investiga através de abordagem fatorial (p. ex., Eckes & Grotjahn 2006; Khodadady 2014; Klein-Braley 1997).

3. Método
3.1. Material

O teste foi criado seguindo os procedimentos propostos por Raatz & Klein-Braley (2002). Foram preparados cinco textos. Para que o teste abarcasse proficiências distintas, os textos foram retirados de manuais didáticos[12] e de jornais. A inteligibilidade[13] dos textos (Curto 2014) foi verificada com recurso à ferramenta LX-CEFR (Curto, Mamede & Baptista 2014), sendo que o valor de cada texto corresponde aos níveis A1, A2, B1, B2 e C1, respetivamente. Os textos são ordenados desde o mais fácil até ao mais difícil, para que os textos mais complexos não desmotivem os alunos de nível baixo logo no início do teste. Cada texto contém 20 lacunas. Cada lacuna foi criada, como habitualmente, pela regra mencionada acima: a partir da segunda frase, elimina-se a metade de palavras alternadas. Quando o número de caracteres da palavra é ímpar, a maior parte dos caracteres é eliminada. A parte eliminada é substituída por um sublinhado de tamanho idêntico. O texto de exemplo pode ser visto abaixo:

> *O Manuel é estudante na Universidade de Lisboa. Ele lev_____-se(1) sempre mu_____(2) cedo. D_____(3) manhã e_____(4) tem au_____(5) das 8h à_____(6) 12h. Dep_____(7) almoça n_____(8) cantina c_____(9) os col_____(10). À ta_____(11), o Manuel pra_____(12) desporto: basqu_____(13). Ele jo_____(14) na equ_____(15) da univer_____(16). À no_____(17), o Manuel ja_____(18) com a fam_____(19). Depois d_____(20) jantar o Manuel gosta de navegar na internet ou falar com os amigos.*

O teste foi aplicado primeiro a 10 falantes nativos de português a fim de verificar se seria fácil para quem tem proficiência alta da língua. Todos os nativos conseguiram preencher corretamente quase todas as lacunas.[14]

---

12 Isto significa que nem todos os textos preparados eram autênticos. No entanto, o autor decidiu extrair os textos de manuais para assegurar que estivessem adequados aos níveis A1 e A2, que também se pretendiam considerar na avaliação do instrumento.

13 Equivalente ao termo inglês "*readability*" (Flesch Reading Ease; Flesch 1948).

14 O teste completo encontra-se em: https://goo.gl/AVmTwv.

3.2. Participantes

Participaram 104 aprendentes de português, que aprendem a língua em universidades no país. A maioria dos aprendentes eram alunos do curso de português para estrangeiros em universidades e alguns eram alunos de mestrado. A Tabela 2 apresenta o resumo do perfil linguístico de acordo com o nível da turma a que pertencem, isto é, uma informação institucional.[15]

Tabela 2. Perfil linguístico de acordo com a informação institucional

Nível da turma	N	Anos de aprendizagem	Média de idade (anos)	Nacionalidade[16]
B1	28	1,83	26,7	CN(10), MO(1), HK(1), VE(7), NL(1), AR(1), FR(2), SY(1), US(1), ES(1), IT(1), TZ(1)
B2	32	4,02	28,6	CN(11), US(4), CA(1), CH(1), JP(1), DE(2), KR(2), FR(1), UA(1), SY(2), EN(1), NL(1), RO(1), ES(1), RU(1), PT(1)
C1	38	6,34	23,3	CN(22), RU(3), US(1), BY(1), IE(1), ET(1), JP(1), DE(1), PT(3), BR(2), MZ(1), ES(1)
M[17]	6	5,88	34,0	CN(4), KR(1), JP(1)

3.3. Procedimento

O teste foi aplicado durante as aulas ou individualmente. Antes do teste, os participantes assinaram o termo de consentimento informado e preencheram

15 Alguns participantes exibiam, em relação ao português, uma proficiência quase nativa. Como se previa que pudessem alcançar pontuações próximas das obtidas pelos falantes nativos, pensou-se na possibilidade de excluir os seus dados da análise. No entanto, desta vez, optou-se por incluí-los, a fim de verificar se o teste demonstra, de facto, um resultado como o esperado. Efetivamente, todos os participantes com este perfil tiveram pontuação acima de 80 (sendo 100 o número máximo de pontos possíveis).

16 A sigla da nacionalidade é baseada na lista de códigos de países usados pela OTAN. O número à direita refere-se ao número de participantes.

17 Refere-se aos alunos de mestrado. Estes alunos não frequentavam um curso de português e não era possível atribuir-lhes um nível da turma (B1, B2 e C1), pelo que foram classificados no grupo M.

uma ficha de informações destinada a caracterizar o perfil da amostra (*cf.* o teste completo em https://goo.gl/AVmTwv). Procedeu-se, de seguida, à explicação do teste e a uma sessão de treino. O teste foi realizado em 30 minutos na presença do investigador para garantir que os participantes não usassem o dicionário ou qualquer material auxiliar.

3.4. Pontuação

Foi dado um ponto para as respostas corretas e 0 para as incorretas. Uma única resposta era aceite, considerando-se como resposta incorreta erros ortográficos e outras palavras que poderiam funcionar gramatical e semanticamente.[18] Assim, a pontuação máxima de cada texto são 20 pontos e o total dos cinco textos 100 pontos.

4. Resultados

A análise foi efetuada com recurso à ferramenta R (v. R 3.3.3) e Microsoft Excel® da forma que se descreve em seguida. Em primeiro lugar, analisou-se a estatística descritiva do teste, a fim de capturar a tendência geral de cada texto e da pontuação de cada grupo. Em segundo lugar, o teste foi analisado em termos da fiabilidade e a discriminação de itens. Em terceiro lugar, foi efetuada uma análise de *cluster* e considerou-se a possibilidade de o teste medir a proficiência geral dos aprendentes, em comparação com a avaliação feita por professores.

4.1. Estatística descritiva

A estatística descritiva do teste é apresentada na Tabela 3. A tabela mostra que a média se vai tornando mais baixa do texto1 (T1) para o texto 5 (T5), o que implica que a dificuldade de cada texto varia conforme a sua inteligibilidade. Tal tendência é consistente entre os grupos, como é demonstrado no Gráfico 1.

---

18 Em algumas questões foram permitidas respostas alternativas devido à variação de uso. Por exemplo, tanto *sande* como *sandes* foram considerados como corretos (2º texto, Nº 8). Além disso, não só *este* como também *esse* foram aceites (5º texto, Nº 7).

Tabela 3. Estatística descritiva do teste

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
T1	1	104	17.07	2.61	18	17.43	1.48	8	20	12	-1.20	0.98	0.26
T2	2	104	15.19	2.93	15	15.36	2.97	6	20	14	-0.61	0.45	0.29
T3	3	104	12.99	4.06	13	13.12	4.45	0	20	20	-0.35	-0.09	0.40
T4	4	104	11.67	4.59	12	11.86	4.45	0	20	20	-0.32	-0.53	0.45
T5	5	104	11.20	4.76	12	11.33	4.45	0	20	20	-0.28	-0.48	0.47

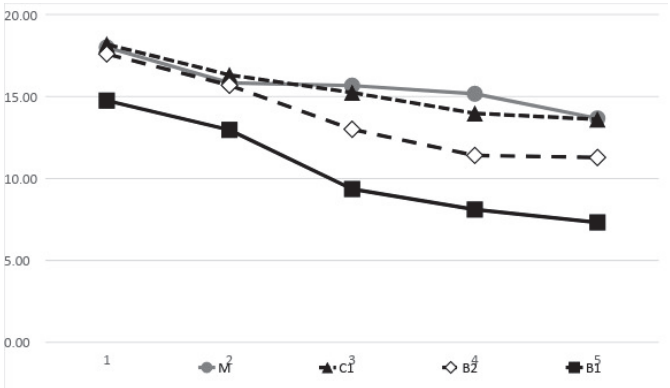

Gráfico 1. Média de pontuação observada na tarefa em cada texto conforme o grupo

4.2. Fiabilidade

A fiabilidade é o critério que torna um teste consistente (Alderson 2005). É um parâmetro que garante que o teste pode apresentar o mesmo resultado independentemente da altura em que se aplica. Embora, tradicionalmente, sejam considerados vários métodos para estimar a fiabilidade, tais como o *test-retest*, *parallel forms*, *split-half* (para uma introdução, *e.g.* Hill & Hill 2008), o procedimento mais comum é o coeficiente *Cronbach Alpha* (Cronbach 1951). Em contrapartida, salienta-se que nos últimos anos alguns investigadores têm posto em causa o *Cronbach Alfa*, sendo que outro parâmetro tem sido recomendado (*e.g.* Okada 2011; 2015): o coeficiente *Omega*. Por estas razões, neste trabalho foram calculados ambos os parâmetros. A Tabela 4 mostra que o teste apresenta os coeficientes elevados. Acrescenta-se que os coeficientes foram calculados, considerando cada texto como um *super-item* ou *testlet* (Wainer & Kiely 1987), visto que se prevê a dependência local das lacunas num texto (Eckes & Grotjahn 2006; Klein-Braley 1985).

Tabela 4. Coeficientes de fiabilidade

Alpha:	0.93
G.6:	0.92
Omega Hierarchical:	0.87
Omega H asymptotic:	0.92
Omega Total	0.95

Para examinar se cada item tinha funcionado corretamente, foram calculadas a correlação item-total (*item-total correlation,* doravante CIT) e a dificuldade de item (doravante DI). A CIT representa o nível de discriminação, isto é, a correlação entre a pontuação de cada item e o total desse mesmo item, tendo o valor desde -1 até 1. O valor fica mais alto e aproxima-se de 1, caso o item tenha mais respostas corretas fornecidas por examinandos de pontuação alta e menos por aqueles de pontuação baixa. Um item cujo valor seja superior a 0,30 é considerado como tendo poder de discriminação (Brown 2005). A DI representa o quão difícil é o item, e calcula-se pelo "número de participantes com resposta correta / o número de participantes". Quanto maior o valor, maior a facilidade do item. Os itens que se enquadram na categoria entre 0,30 e 0,70 são considerados de dificuldade intermédia.

A Tabela 5 apresenta os itens, cujo CIT se encontra abaixo de 0,30, juntamente com a sua DI à direita. A tabela mostra que a maioria dos itens é oriunda dos primeiros dois textos (Q1 – Q40). A baixa CIT destes itens leva-nos a presumir que os itens dos textos menos complexos, cuja inteligibilidade é A1 e A2 (T1 e T2), eram fáceis para todos os grupos (B1, B2, C1 e M). Visto que o teste foi criado para tentar capturar proficiências distintas de aprendentes, é lógico que haja tais itens (*e.g.* Q6, Q4, Q22). No entanto, alguns destes itens representam, ao mesmo tempo, a DI baixa (*e.g.* Q2, Q30, Q36), ou seja, maior dificuldade. Assim, pode-se afirmar que, mesmo nos textos de A1 e A2, havia alguns itens que eram bastante difíceis tanto para os aprendentes de nível intermédio como para os de nível avançado. Por esta razão, pode-se julgar que estes itens devam ser modificados. Em contrapartida, nenhum item demonstrou valor negativo. Por outras palavras, não havia itens problemáticos que fossem fáceis para os aprendentes de nível mais baixo, mas difíceis para os de nível mais avançado, o que nos leva a pensar que o teste foi, em termos gerais, bem construído.

Tabela 5. Itens de correlação item total baixa e a sua dificuldade

Item	CIT	DI
Q2	0,12	0,23
Q30	0,14	0,28
Q36	0,15	0,28
Q29	0,16	0,31
Q15	0,19	0,35
Q13	0,19	0,32
Q53	0,20	0,37
Q17	0,21	0,31
Q40	0,22	0,36
Q73	0,23	0,29
Q61	0,23	0,30
Q83	0,23	0,33
Q28	0,23	0,29
Q6	0,24	0,85

Item	CIT	DI
Q27	0,24	0,30
Q16	0,25	0,42
Q25	0,25	0,36
Q35	0,26	0,38
Q26	0,27	0,45
Q4	0,27	0,54
Q80	0,27	0,40
Q22	0,28	0,62
Q5	0,29	0,49
Q11	0,29	0,66
Q9	0,29	0,77
Q20	0,29	0,62

4.3. Validade

Foi efetuada uma análise de *clusters* – método hierárquico – a fim de classificar a proficiência de acordo com a pontuação do teste. Foi obtido o dendrograma, apresentado no Gráfico 2. Tendo em conta a forma do dendrograma, foi decidido dividir os dados em quatro grupos, postulando os grupos como *principiante+*, *intermédio*, *intermédio+* e *avançado*. De seguida, uma análise de *cluster* – método *k-means* – foi efetuada com 4 na variável *k*. A estatística descritiva de cada *cluster* encontra-se na Tabela 6, implicando que os *clusters* 1 a 4 representam *intermédio+*, *avançado*, *intermédio* e *principiante+*, respetivamente. Para verificar se o agrupamento foi bem feito, uma ANOVA foi implementada em relação a cada variável (T1 – T5). A significância foi verificada entre todos os *clusters* (T1: $F(3,100) = 51.097$, $p < .001$; T2: $F(3,100) = 31.842$, $p < .001$; T3: $F(3,100) = 112.653$, $p < .001$; T4: $F(3,100) = 114.968$, $p < .001$; T5: $F(3,100) = 105.776$, $p < .001$;).

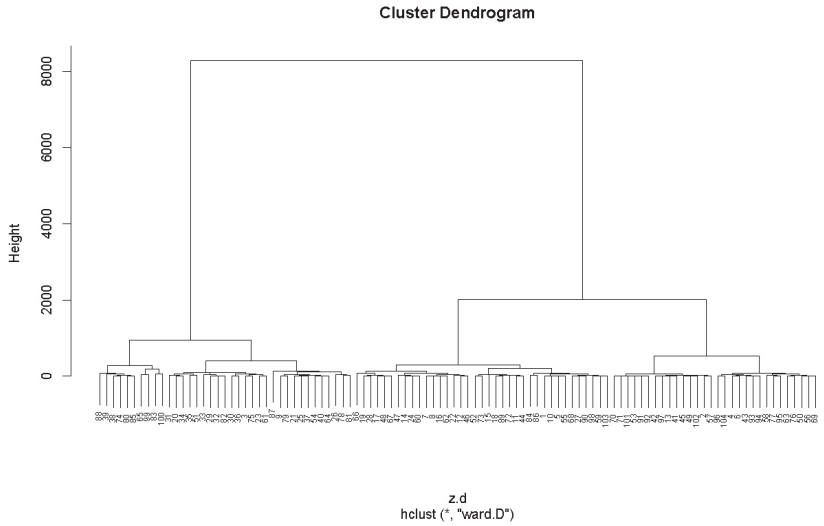

Gráfico 2. Dendrograma de *cluster*

Tabela 6. Estatística descritiva de cada cluster

		T1	T2	T3	T4	T5
cluster 1	Min.	12.0	12.00	10.00	8.00	8.00
	1st Qu.	17.0	14.00	12.00	10.00	11.00
	Median	18.0	16.00	14.00	12.00	12.00
	Mean	17.7	15.49	13.76	11.73	12.27
	3rd Qu.	19.0	17.00	15.00	13.00	13.00
	Max.	20.0	20.00	17.00	15.00	16.00

		T1	T2	T3	T4	T5
cluster 2	Min.	17.00	15.0	12.0	14.00	12.00
	1st Qu.	19.00	16.0	16.0	16.00	14.00
	Median	19.00	17.5	17.5	17.00	16.00
	Mean	19.07	17.6	17.2	17.03	16.07
	3rd Qu.	20.00	19.0	19.0	18.00	18.00
	Max.	20.00	20.0	20.0	20.00	20.00

		T1	T2	T3	T4	T5
cluster 3	Min.	13.00	7.00	6.00	3.00	1.00
	1st Qu.	15.00	13.00	9.00	6.00	6.00
	Median	16.00	14.00	10.00	8.00	7.00
	Mean	15.81	13.78	9.852	8.30	7.15
	3rd Qu.	17.00	15.00	11.00	10.50	9.00
	Max.	19.00	18.00	12.00	13.00	11.00

		T1	T2	T3	T4	T5
cluster 4	Min.	8.0	6.0	0.00	0.00	0.00
	1st Qu.	11.0	10.0	4.50	1.75	0.75
	Median	11.0	11.0	6.50	5.50	4.50
	Mean	12.1	10.7	6.00	4.50	3.60
	3rd Qu.	12.0	12.0	7.75	6.75	5.75
	Max.	18.0	14.0	9.00	9.00	7.00

	cluster4	cluster3	cluster1	cluster2
Min. (total)	39,0	44,0	64,0	75,0
Max. (total)	43,0	65,0	80,0	97,0
Mean (total)	41,7	54,7	72,4	85,5

A Tabela 7 mostra a distribuição dos aprendentes em cada *cluster* conforme a sua informação institucional. Segundo estes dados, tudo indica que há bastante heterogeneidade dentro dos grupos. Por exemplo, nem todos os aprendentes do grupo C1 se enquadram no *cluster* 2 (*avançado*), encontrando-se muitos no *cluster* 1 (*intermédio*+) também. A legitimidade deste agrupamento é discutida de seguida.

Tabela 7. Distribuição dos aprendentes (cluster x grupo)

Grupo	cluster1 (intermédio+)	cluster2 (avançado)	cluster3 (intermédio)	cluster4 (iniciante+)
B1	4	2	14	8
B2	13	8	9	2
C1	20	15	3	0
M	0	5	1	0

Recorde-se que a validação do *C-test* é feita principalmente através da abordagem correlativa, ou seja, a verificação de correlação com outros testes. Porém, em relação ao português, temos poucos instrumentos que poderiam servir de base de comparação. A maioria dos participantes não possuía o certificado do CAPLE, exame padronizado para certificação da competência de português, pelo que não havia nenhum parâmetro exterior que representasse a proficiência geral dos aprendentes. Assim sendo, no presente estudo, além da autoavaliação de aprendentes e do nível de turma (informação institucional), foi decidido utilizar a avaliação pelos professores. Os professores eram especialistas em PLE (português como língua estrangeira) e davam aulas aos aprendentes há algum tempo, pelo que as suas avaliações poderiam ser consideradas, até certo ponto, como um parâmetro confiável. A avaliação pelos professores foi obtida, considerando-se nove níveis, com base no QECR (Quadro Europeu Comum de Referência), mas de forma mais minuciosa para uma melhor captação da heterogeneidade de proficiência: A1, A1+, A2, A2+, B1, B1+, B2, B2+, C1, C1+, C2. O nível da turma corresponde na verdade a três níveis: B1, B2 e C1. A autoavaliação de alunos compreende 6 níveis: A1, A2, B1, B2, C1 e C2. A Tabela 8 apresenta a distribuição dos alunos conforme o nível obtido a partir da avaliação pelos professores. Pode-se considerar que há bastante heterogeneidade da proficiência dentro dos grupos. Se o *C-test* mede a proficiência geral, a sua pontuação também deve capturar tal heterogeneidade e correlacionar-se bem com a avaliação pelos professores. Os níveis de cada avaliação foram convertidos em escala numérica (ordinal) e foi calculada a correlação com a pontuação do teste.

Tabela 8. Classificação da proficiência dos aprendentes com base na avaliação pelos professores

Grupo[19]	A1	A1+	A2	A2+	B1	B1+	B2	B2+	C1	C1+	C2
B1	0	0	1	2	14	8	3	0	0	0	0
B2	0	0	0	0	3	5	10	8	5	0	1
C1	0	0	0	0	0	0	5	6	14	4	9

Tabela 9. Correlação entre os quatro parâmetros examinados

	C-test	Aval. Prof.	Autoaval.	N. Turma
C-test	1			
Aval. Prof.	0,74	1		
Autoaval.	0,71	0,79	1	
N. Turma	0,60	0,79	0,77	1

Como mostra a Tabela 9, a pontuação do *C-test* correlaciona-se melhor com a avaliação pelos professores do que com outros parâmetros, i.e., a autoavaliação e o nível da turma. Partindo do pressuposto de que a avaliação pelos professores é uma medida confiável de proficiência dos aprendentes, este valor de correlação suporta, até certo ponto, a validade do teste. Porém, note-se que, ao mesmo tempo, tanto a autoavaliação como o nível da turma apresentaram correlação forte com a avaliação pelos professores.

Para capturar mais detalhadamente a relação entre as avaliações, foi calculado o coeficiente de *Kappa*, que mede o grau de concordância de avaliações nominais. Para tal, devido à discrepância de escala, cada avaliação foi convertida em quatro escalas nominais. A Tabela 10, abaixo, apresenta o coeficiente de *Kappa* entre a avaliação pelos professores e outras avaliações, juntamente com a proporção de concordância simples. O coeficiente de *Kendall*, que mede a proporção de concordância de escalas ordinais, também foi calculado, convertendo cada avaliação em escala ordinal. Os resultados de ambos os testes mostraram que os três parâmetros apresentaram correlação moderada e forte com a avaliação pelos professores.

19 Para os alunos de mestrado não foi possível obter a avaliação pelos professores, pelo que foram excluídos da análise.

Tabela 10. Coeficiente de *Kappa* e a proporção de concordância simples

	C-test	N. Turma	Autoavaliação
avaliação p/ prof.	0.408 ***	0.532 ***	0.459 ***
proporção de conc.	58.2%	68.4%	62.2%

Tabela 11. Coeficiente de concordância de Kendall

	C-test	N. Turma	Autoavaliação
avaliação p/ prof.	0.838 ***	0.897 ***	0.877 ***

5. Discussão

No que toca à fiabilidade do *C-test*, foram obtidos valores elevados dos coeficientes de *Cronbach Alpha* e *Omega*. Apenas com estes parâmetros, não se pode afirmar com segurança que o teste é confiável, mas eles sustentam a sua fiabilidade. Em relação à análise da discriminação de item, alguns itens foram considerados como não apropriados, daí ser melhor ter em consideração a sua modificação (*cf.* Jafarpur 1999). Em termos de dificuldade, não houve muitos itens demasiado difíceis. No entanto, como há aprendentes cuja pontuação dos últimos dois textos foi de 0 (*cf.* T4 e T5 na Tabela 6), é bem provável que tais itens estivessem a levantar bastantes dificuldades aos alunos de nível mais baixo, sobretudo os do grupo B1. Eventualmente, esta versão do *C-test* talvez não seja apropriada para os grupos de nível baixo e intermédio.

Quanto à sua validade, a interpretação é difícil. Aparentemente, o teste assinalou a heterogeneidade em proficiência que podia existir dentro dos grupos classificados de acordo com o nível da turma. Simultaneamente, a sua correlação com a avaliação dos professores é forte bem como os outros parâmetros. Este resultado implica que o teste funciona até certo ponto, mas, ao mesmo tempo, não garante a sua superioridade em relação aos outros parâmetros: autoavaliação e o nível da turma. Além disso, o valor de correlação do teste é ligeiramente mais baixo do que os valores dos demais parâmetros, o que nos leva a considerar algumas limitações possíveis.

Em primeiro lugar, é provável que o teste não esteja a medir a proficiência geral, mas sim alguns componentes micro e específicos, como, por exemplo, o conhecimento do vocabulário ou o conhecimento gramatical. A Tabela 12 apresenta a correlação da pontuação do *C-test* com a autoavaliação dos alunos e avaliação pelos professores em termos de quatro

competências linguísticas: ouvir, falar, ler e escrever. Apesar de não ter sido verificada uma tendência saliente, a modalidade de leitura, em particular, parece apresentar uma melhor correlação com o teste.

Tabela 12. Correlação de quatro modalidades da autoavaliação e da avaliação pelos professores com o *C-test*

	autoavaliação					aval. prof.				
	ouvir	falar	ler	escrever	total	ouvir	falar	ler	escrever	total
c-test	0,68	0,62	0,69	0,60	0,69	0,75	0,75	0,78	0,77	0,74

Em segundo lugar, é possível que o agrupamento pela análise de *cluster* não tenha sido bem feito. Como mostra a Tabela 6, o *cluster* 1, que é considerado como *intermédio+*, tem um traço diferente dos outros: a pontuação do T5 é ligeiramente melhor do que a do T4. Assim, alguns alunos que tiveram pontuação elevada (mais de 80 pontos) foram classificados no *cluster1* (*intermédio+*) e não no *cluster* 2 (*avançado*), provavelmente, por causa da distribuição das suas pontuações. Da mesma maneira, outros aprendentes com pontuação mais baixa do que a daqueles foram classificados no *cluster* 2 (*avançado*). De modo geral, na análise de *cluster* são usados vários métodos em termos da distância entre os dados e do agrupamento de *cluster,* e não há um método absolutamente fiável. No presente estudo, usamos o quadrado da distância Euclidiana e o método *Ward*. Talvez pudéssemos ter obtido outro resultado com outros métodos. Por outro lado, visto que a maioria dos aprendentes que tiveram mais de 80 pontos foi considerada pelos professores como estando no nível C1, C1+ ou C2, talvez seja melhor traçar linhas de referência para classificar a proficiência sem depender do agrupamento por *clusters*.

Por fim, deve-se considerar a questão do peso pontual. O grupo incluía muitos aprendentes classificados em níveis diferentes devido a uma diferença de apenas alguns pontos. Supondo que todos os que obtiveram mais de 80 pontos podiam ser considerados como estando no nível *avançado*, será que os que obtiveram 78 ou 79 correspondem realmente ao nível *intermédio+*? Provavelmente, o teste pode distinguir os aprendentes de nível avançado dos aprendentes de nível mais baixo, mas dificilmente poderá distinguir os aprendentes que estejam próximos da fronteira entre dois níveis. Recorde-se que, muito provavelmente, o teste foi bastante difícil para os aprendentes de nível mais baixo. Sendo assim, talvez seja mais apropriado que o teste seja utilizado para verificar se os aprendentes exibem ou não proficiência a um nível avançado.

6. Considerações finais

O presente estudo examinou a fiabilidade e a validade de uma versão em português europeu do C-test. Foi constatada a fiabilidade elevada através da Teoria de Teste Clássica. Por outro lado, é de salientar que, nos últimos vinte anos, os investigadores tentaram examinar o *C-test* através da Teoria de Resposta ao Item (TRI; *Item response theory*), mais concretamente através de *Polytomous Rasch Models* (Masters 1982; Samejima 1968) e *Testlet Response Theory* (Wang & Wilson 2005). Para tais análises, no entanto, são necessários muitos dados. Por esta razão, neste trabalho, que contém como amostra cerca de 100 participantes, não foi efetuada a análise do ponto de vista da TRI, o que é uma das limitações do estudo. Com a recolha de mais dados e uma análise baseada na TRI poderíamos obter mais informações em relação à fiabilidade e aos itens ou textos.

Em relação à validade, esta versão do teste ainda não pode ser considerada para avaliar a proficiência geral, visto que se implicou uma maior relação com a competência da leitura, bem como a impossibilidade de agrupamento minucioso. Sendo assim, chegamos à conclusão de que seria mais seguro utilizar a avaliação feita pelos professores de PLE. Naturalmente, esta nem sempre estará disponível, por exemplo, no caso de os aprendentes serem alunos de licenciatura ou mestrado, que não frequentam um curso de PLE. Nesses casos, pode-se utilizar o teste com algumas limitações como, por exemplo, apenas para verificar a proficiência elevada.

O *C-test* tem sido controverso desde a sua proposta. Muito embora existam vários estudos a suportar a sua validade, é importante não sobrestimar o seu alcance. Pelo seu formato, evidentemente, não se pode considerar um teste suficientemente robusto para capturar todo o conhecimento dos aprendentes. É provável que o teste funcione para um grupo de aprendentes, mas não para outro (Tremblay 2011). Será importante, por isso, interpretar os resultados e o potencial do teste com cautela, de vários pontos de vista e com várias amostras, inclusive, porque a validação é um processo constante. Por outro lado, também se pode afirmar que, quando devidamente validado, o *C-test* será um instrumento útil, uma vez que é relativamente fácil de desenvolver e é aplicável em tempo curto. Espera-se que sejam realizados mais estudos no que toca à avaliação de proficiência de aprendentes, que é indispensável para estudos empíricos em ALS.

Referências

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Londres & Nova Iorque: Continuum.

Anderson, J. (1992). Automaticity and the ACT theory. *The American Journal of Psychology*, *105*(2), 165–180.

Babaii, E. & Ansary, H. (2001). The C-test: A valid operationalization of reduced redundancy principle? *System*, *29*(2), 209–219.

Babaii, E. & Moghaddam, M. J. (2006). On the interplay between test task difficulty and macro-level processing in the C-test. *System*, *34*(4), 586–600. https://doi.org/10.1016/j.system.2006.09.002.

Baghaei, P. & Tabatabaee, M. (2015). The C-Test: An integrative measure of crystallized intelligence. *Journal of Intelligence*, 3(2), 46-58. https://doi.org/10.3390/jintelligence3020046.

Bialystock, E. & Smith, M. S. (1985). Interlanguage is not a state of mind: An evaluation of the construct for second-language acquisition. *Applied Linguistics*, *6*(2), 101–117.

Brown, J. (2005). *Testing in Language Programs*. McGraw-Hill Companies.

Chapelle, C. A. & Abraham, R. (1990). Cloze method: what difference does it make? *Language Testing*, *7*(2), 121–146.

Cleary, C. (1988). The C-test in English. *RELC Journal*, *19*(2), 26–37.

Cohen, A. D., Segal, M., & Bar-Siman-To, R. (1984). The C-Test in Hebrew. *Language Testing*, *1*(2), 221–225. https://doi.org/10.1177/026553228400100206.

Corder, P. (1967). The significance of learner's errors. *International Review of Applied Linguistic*, *5*(1), 161–170.

Correia, R. D. Z. (2011). *Os erros no discurso escrito de Hispano-Falantes de nível B2* (Dissertação de Mestrado, Universidade de Lisboa).

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Curto, P. (2014). *Classificador de textos para o ensino de português como segunda língua* (Dissertação de Mestrado, Universidade de Lisboa).

Curto, P., Mamede, N. & Baptista, J. (2014). Automatic readability classifier for European Portuguese. *System*, 5, 6.

DeKeyser, R. (1997). Beyond explicit rule learning. *Studies in Second Language Acquisition*, *19*(2), 195–222.

Eckes, T. & Baghaei, P. (2015). Using Testlet Response Theory to Examine Local Dependence in C-Tests. *Applied Measurement in Education*, 28(2), 85-98. https://doi.org/10.1080/08957347.2014.1002919.

Eckes, T. & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*(3), 290–325. https://doi.org/10.1191/0265532206lt330oa

Ellis, N. (2015). Implicit and explicit learning.pdf. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 3–23). Amsterdão: John Benjamins.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition*, *27*(2), 141–172.

Firth, A. & Wagner, J. (2007). Second/Foreign Language Learning as a Social Accomplishment: Elaborations on a Reconceptualized SLA. *The Modern Language Journal*, *91*(s1), 800–819. https://doi.org/10.1111/j.1540-4781.2007.00670.x.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233.

Grotjahn, R. (2016). *The electronic C-Test bibliography: Version 2015*. Disponível em: <http://www.c-test.de/deutsch/ctest/pdf/C%20Test%20Bibliography/Grotjahn_Electronic_Ctest_Bibliography.pdf>.

Grotjahn, R. & Stemmer, B. (2002). C-Tests and language processing.pdf. In J. Coleman, R. Grotjahn, & U. Raatz (Eds.), *University language testing and the C-Test* (pp. 115–130). Bochum: AKS-Verlag.

Hill, M. & Hill, A. (2008). *Investigação por questinário*. Edições Sílabo.

Isabel, L. (2006). *Léxico, aquisição e leitura do português europeu língua não materna*. Lisboa, Portugal: Fundação Calouste Gulbenkian.

Jafarpur, A. (1999). Can the C-test be improved with classical item analysis? *System*, *27*(1), 79–89.

Kamimoto, T. (1993). Tailoring the test to fit the students : Improvement of the C-test through classical item analysis. *Fukuoka Women's Junior College Studies*, *30*(11), 47–61.

Katona, L. & Dornyei, Z. (1993). The C-test. *FORUM*, *31*(2), 35–38.

Khodadady, E. (2014). Construct Validity of C-tests: A Factorial Approach. *Journal of Language Teaching and Research*, *5*(6). https://doi.org/10.4304/jltr.5.6.1353-1362.

Klein-Braley, C. (1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing*, 2(1), 76-104.

Klein-Braley, C. (1997). C-Tests in the context of reduced redundancy testing: An appraisal. *Language Testing*, *14*(1), 47–84.

Krashen, S. (1982). *Principles and Practice in Second Language Acquisition*. Pergamon Press: Longman.

Lado, R. (1957). *Linguistics across cultures*. Estados Unidos da América: The University of Michigan Press.

Lee-Ellis, S. (2009). The development and validation of a Korean C-Test using Rasch Analysis. *Language Testing*, *26*(2), 245–274. https://doi.org/10.1177/0265532208101007.

Lei, L. (2008). Validation of the C-Test amongst Chinese ESL Learners. *THE JOURNAL OF ASIA TEFL*, *5*(2), 117–140.

Leiria, I. (1991). *A aquisição por falantes de Português Europeu língua não materna dos aspectos verbais expressos pelos Pretéritos Perfeito e Imperfeito* (Dissertação de Mestrado, Universidade de Lisboa).

Li, S. (2010). The Effectiveness of Corrective Feedback in SLA: A Meta-Analysis: Meta-Analysis of Corrective Feedback. *Language Learning*, *60*(2), 309–365. https://doi.org/10.1111/j.1467-9922.2010.00561.x.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Mcdonough, K. & Kim, Y. (2009). Syntactic Priming, Type Frequency, and EFL Learners' Production of *Wh*- Questions. *The Modern Language Journal*, *93*(3), 386–398. https://doi.org/10.1111/j.1540-4781.2009.00897.x.

Montrul, S., Dias, R. & Santos, H. (2010). Clitics and object expression in the L3 acquisition of Brazilian Portuguese. *Second Language Research*, *27*(1), 21–58.

Okada, K. (2011). Beyond Cronbach's alpha: a comparison of recent methods for estimating reliability. *The Japanese Journal for Research on Testing*, *7*(1), 38–50.

Okada, K. (2015). Reliability in psychology and psychological measurement, with focus on Cronbach's Alpha. *The Annual Report of Educational Psychology in Japan*, *54*(1), 71–83.

Pienemann, M. (1998). *Language processing and second language development: Processability Theory*. Amsterdão: John Benjamins.

Pinto, J. (2014). A aquisição do género e da concordância de género em português língua terceira ou língua adicional. In P. Osório & F. Bertinetti (Eds.), *Teorias e Usos Linguísticos*. Lisboa: Lidel.

Raatz, U. & Klein-Braley, C. (1981). The C-testa – modification of the cloze procedure. *Practice and Problems in Language Testing, University of Essex Depart-ment of Language and Linguistics Occasional Papers No. 26*.

Raatz, U. & Klein-Braley, C. (2002). Introduction to language and to C-Tests. In James Coleman, R. Grotjahn, & U. Raatz (Eds.), *University Language Testing and the C-Test*. AKS-Verlag Bochum.

Rodrigues, E. (2015). Concordância de número e gênero em estruturas predicativas no português brasileiro. *LinguíStica*, *11*(1), 135-152.

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Report Series*, 1968(1), i-169.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*(1-4), 209-241.

Spolsky, B. (1973). What does it mean to know a language; or how do you get somebody to perform his competence? In J. Oller & J. Richards (Eds.), *Focus on the learner* (pp. 164-176). Newbury House Pub.

Taylor, W. (1953). Cloze procedure - a new tool for measuring readability. *Journalism Quarterly*, *30*(4), 414-438.

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. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279-298). Amsterdão & Philadelphia: John Benjamins Publishing Company.

Tremblay, A. (2011). Proficiency assessment standard in second language. *Studies in Second Language Acquisition*, 33(3), 339-372. https://doi.org/10.1017/S0272263111000015.

VanPatten, B. (2007). Input processing in adult SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 115-135).

Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185-201.

Wang, W.-C. & Wilson, M. (2005). The Rasch Testlet Model. *Applied Psychological Measurement*, *29*(2), 126-149. https://doi.org/10.1177/0146621604271053.

[recebido em 27 de abril de 2018 e aceite para publicação em 20 de dezembro de 2018]