CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
Programa de Pós-Graduação *Stricto Sensu* em Estudos de Linguagens

Rose Mara Silva

As Novas Tecnologias de Informação e Comunicação na Avaliação de Proficiência em Português como Língua Estrangeira: O Exame Celpe-Bras

Belo Horizonte – MG
Agosto, 2017

Rose Mara Silva

As Novas Tecnologias de Informação e Comunicação na Avaliação de Proficiência em Português como Língua Estrangeira: O Exame Celpe-Bras

Dissertação apresentada ao Programa de Pós-Graduação *Stricto Sensu* em Estudos de Linguagens do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), como requisito parcial à obtenção do título de Mestre em Estudos de Linguagens.

Área de concentração: Linguística aplicada à avaliação de proficiência linguística.

Linha de pesquisa: Linguagens, Ensino e Mediações Tecnológicas.

Orientador: Prof. Dr. Jerônimo Coura-Sobrinho.

Belo Horizonte – MG
Agosto, 2017

Silva, Rose Mara.

S586n As novas tecnologias de informação e comunicação na avaliação de proficiência em português como língua estrangeira: o exame CELPE-BRAS / Rose Mara Silva . – 2017.
150 f.: il.; tabs. ; grafs. –

Orientador: Jerônimo Coura Sobrinho

Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais, Programa de Pós-Graduação em Estudos de Linguagens, Belo Horizonte, 2017.

Bibliografia.

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU* EM ESTUDOS DE LINGUAGENS

ROSE MARA SILVA

"As Novas Tecnologias de Informação e Comunicação na Avaliação de Proficiência em Português como Língua Estrangeira: O Exame Celpe-Bras"

Dissertação apresentada ao Programa de Pós-Graduação em Estudos de Linguagens do Centro Federal de Educação Tecnológica de Minas Gerais em 03 de julho de 2017, como requisito parcial para obtenção do título de Mestre em Estudos de Linguagens, aprovada pela Banca Examinadora constituída pelos professores:

Prof. Jerônimo Coura Sobrinho, Dr. - Orientador
Centro Federal de Educação Tecnológica de Minas Gerais

Prof. José Marcelo Freitas de Luna, Dr.
Universidade do Vale do Itajaí

Prof. Heitor Garcia de Carvalho, Dr.
Centro Federal de Educação Tecnológica de Minas Gerais

Dedico este trabalho a
Deus, por Sua presença em minha vida.

AGRADECIMENTOS

A minha amada família que, apesar de, muitas vezes, não endender o que faço, está sempre ao meu lado, me incentivando a cada passo.

Aos amigos, que são a família que escolhi, pelo carinho e apoio.

Ao meu orientador, Prof. Dr. Jerônimo, por ter me aceito como orientanda e pela oportunidade de aprendizado e crescimento profissional. Verdadeira honra tê-lo como guia durante todo o percurso, nesses últimos dois anos.

Aos meus queridos novos amigos, companheiros de aprendizado neste curso, pelos bons momentos juntos; também foram muitos os desafios compartilhados. Certamente, seria mais difícil sem vocês.

À CAPES, que viabilizou minha permanência no curso como bolsista.

Aos coordenadores, professores e funcionários do Programa de Pós-Graduação em Estudos de Linguagens do CEFET, pela atenção dispensada às minhas solicitações durante todo o curso.

Aos professores que aceitaram o convite para compor a banca de avaliação desta dissertação, pela oportunidade de mais este momento de debate acadêmico.

A minha querida e grande amiga Maria José Moreira (Jô), pelo companheirismo, amizade e incentivo, mesmo estando do outro lado do Atlântico.

Meu muito obrigado. Amo todos vocês!

Seamos realistas,
sueñemos con lo imposible!

Ernesto 'Che' Guevara

Language testing is a field in crisis,
one which is masked by the impressive
appearance of technological advance.

Tim McNamara

RESUMO

O desenvolvimento das Novas Tecnologias de Informação e Comunicação (NTIC) avança em muitas áreas do conhecimento humano, e não poderia ser diferente no campo de Avaliação de Proficiência em Língua Estrangeira (LE). Nesta área, o uso de computadores tem se tornado cada vez mais presente e, por suas especificidades, apresenta características distintas da avaliação impressa. No entanto, poucos trabalhos voltados a essa temática têm sido desenvolvidos, principalmente, no Brasil. Sendo assim, percebe-se a necessidade de mais pesquisas, também, na área de avaliação de proficiência em Língua Portuguesa (LP) como Língua Estrangeira. Devido à crescente demanda de aprendizes, de profissionais e demais interessados pela LP, consequentemente, a procura pelo exame brasileiro Celpe-Bras vem crescendo gradativamente, tornando essencial pensar em meios práticos de se alcançar todos aqueles que se interessem ou necessitem fazer o exame, assim como facilitar a aplicação e a correção do mesmo. Valendo-se do fato de o Celpe-Bras não ser aplicado em versão computadorizada/*online*, este estudo pretende investigar o uso das NTIC como ferramenta para integrar e aprimorar a realização de tarefas destinadas à avaliação de proficiência em Português como Língua Estrangeira, no Exame Celpe-Bras. Para alcançar tal objetivo, buscou-se nesta dissertação: i) identificar as NTIC utilizadas na realização de tarefas avaliativas em testes de proficiência linguística; ii) discutir as qualidades ou critérios de teste e as características de alguns exames de proficiência em meio eletrônico; e iii) Propor alguns cuidados e/ou sugestões para uma versão *online* do Exame Celpe-Bras. Assim, adotou-se para esta pesquisa a linha metodológica descritiva, com a abordagem de análise bibliográfica e documental, de cunho qualitativo. Finalmente, após a análise e discussão dos resultados, a pesquisa apontou para a necessidade da busca permanente pela inovação em testes, e mostrou que as NTIC trazem avanços e alternativas variadas para a área de avaliação em LE, além de colaborar para a ampliação da promoção e difusão da Língua Portuguesa ao redor do mundo.

Palavras-chave: Avaliação. Celpe-Bras. NTIC. Testes de Proficiência em Larga Escala por Computador.

ABSTRACT

The development of New Information and Communication Technologies (NICT) advances in many areas of human knowledge, and it could not be different in the field of Assessment of Foreign Language Proficiency. In this area, using computers has become increasingly present. That way of assessment has its specificities and it is in many aspects quite different from the paper and pencil assessment. However, few studies have been developed on this theme, specially, in Brazil. Thus, there is the need for more research, as well, in the area of assessment of Portuguese as a Foreign Language. Due to the growing demand of learners, professionals and other individuals, interested in Portuguese Language, the demand for the Celpe-Bras exam has increased gradually. Therefore, it is essential to think of practical means to reach all those who are interested or need to take the exam, as well as to facilitate the administration and correction of it. Based on the fact that the Celpe-Bras exam is not applied in a computerized environment, this study aims to investigate the use of NICT as a tool to integrate and improve the tasks used in the assessment of the proficiency in Portuguese as a Foreign Language, in the Celpe-Bras Exam. In order to reach this objective, this master's thesis intends: i) to identify the NICT used in conducting tasks in language proficiency tests; ii) to discuss the qualities of tests usefulness and the characteristics of some language proficiency tests in virtual context; and iii) to propose some suggestions for an online version of the Celpe-Bras Exam. Thereby, as a qualitative research, the descriptive methodological approach was adopted and a bibliographical analysis was done. Finally, after analyzing and discussing the results this research pointed to the need for a constant search for innovation in tests and showed that NICT brings different advances and alternatives to the area of Foreign Language Assessment, as well as contribute to the dissemination and promotion of the Portuguese Language around the world.

Keywords: Assessment. Celpe-Bras. NICT. Large-Scale Computer-Based Testing.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

ACTFL American Council on the Teaching of Foreign Languages

AVA Ambientes Virtuais de Aprendizagem

AWE Automated Writing Evaluation

CALT Computer-Assisted Language Test

CAPLE Centro de Avaliação de Português Língua Estrangeira

CAT Computer- Adaptative Testing

CBMAT Computer-Based Multidimensional Adaptive Testing

CBT Computer-Based Testing

CEBRASPE Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos

CELPE-BRAS Certificado de Proficiência em Língua Portuguesa para Estrangeiros

CELU Certificado de Español Lengua y Uso

CEP Centro de Estudos de Pessoal

CPLP Comunidade de Países de Língua Portuguesa

DELE Diplomas de Español como Lengua Extranjera

EaD Educação a Distância

EP Elementos Provocadores

EPO Entrevista/Exame de Proficiência Oral

EPPLE Exame de Proficiência para Professores de Língua Estrangeira

ESOL Cambridge English Tests

ETS Educational Testing Service

FCE First Certificate in English

iBT Internet Based Test

IELTS International English Language Test System

INEP Instituto Nacional de Estudos e Pesquisas Educacional

L2 Segunda Língua

LE Língua Estrangeira

LI Língua Inglesa

LLT Language Learning & Technology

LP Língua Portuguesa

LTI Language Testing International

LTRC Language Testing Research Colloquium

MCAT Teste Adaptativo Computadorizado Multidimensional

MTRI Teoria de Resposta ao Item Multidimensional

MLA Modern Language Association

NTIC Novas Tecnologias de Informação e Comunicação

OPI Oral Proficiency Interview

OPIc Oral Proficiency Interview Computer Test

PEC-G Programa de Estudantes-Convênio de Graduação

PEC-PG Programa em Nível de Pós-Graduação

PFOL Português para Falantes de Outras Línguas

PLE Português Língua Estrangeira

PLN Processamento de Linguagem Natural

PTE Academic The Pearson Test of English Academic

QECRL Quadro Europeu Comum de Referência para as Línguas

TCF/TEF Test de Connaissance du Français

TEPOLI Teste de Proficiência Oral em Língua Inglesa

TestDaF Der Test Deutsch als Fremdsprache

TIC Tecnologias de Informação e Comunicação

TLMC Testes de Línguas Mediados por Computadores

TOEFL Test of English as a Foreign Language

TOEIC Test of English for International Communication

TRI Teoria de Resposta ao Item

UCAT Teste Adaptativo Computadorizado Unidimensional

UNB Universidade de Brasília

UTRI Teoria de Resposta ao Item Multidimensional

SUMÁRIO

INTRODUÇÃO

O crescente desenvolvimento das Novas Tecnologias de Informação e Comunicação, doravante NTIC[1], e a sua ampla disseminação, quer nos contextos sociais, escolares e/ou profissionais, têm mudado o cenário educativo no que diz respeito às atividades de avaliação em Segunda Língua ou em Língua Estrangeira, doravante L2 e LE, respectivamente.

Segundo Furtoso (2011), o processo avaliativo tem sido muitas vezes reduzido à testagem de determinado conteúdo, supostamente aprendido na escola, ignorando, assim, as interações dentro da escola e fora dela. No que diz respeito às LE, o cenário da avaliação não é diferente, ou seja, há uma tendência de redução do processo avaliativo à aplicação de provas. No entanto, a avaliação não deveria ser abordada como elemento isolado, mas como elemento fundamental em todo o processo de ensino e de aprendizagem. (SCARAMUCCI, 1998. p. 117 *apud* FURTOSO, 2011. p. 106).

Especialmente nos dias atuais, faz-se necessário investigar a avaliação de forma sistemática, como parte do processo de ensino e de aprendizagem, e também considerar as novas práticas pedagógicas por todo esse processo avaliativo, tais como: a Educação a Distância (EaD), o uso das TIC, os Ambientes Virtuais de Aprendizagem (AVA), dentre outras. Segundo Consolo (2004), avaliar o desempenho por meio de interação mediada por recursos eletrônicos, em comparação com as práticas de se avaliar proficiência por meio de testes em versão impressa, caracteriza-se como aspecto inovador na área de avaliação.

Assim, sabendo que as tecnologias digitais e o ambiente *online* surgem como formas, quase naturais, de interligar o Ensino, a Aprendizagem e a Avaliação em L2 ou LE, alguns recursos, cada vez mais utilizados no ensino, como a videoconferência, os *podcasts* e o *Skype*, além de diminuir a distância entre os usuários de uma LE e o contexto real de uso desta língua-alvo, poderiam ser repensados como ferramentas também destinadas à avaliação. Além disso, não podemos negar o fato de que essas NTIC favorecem a interação sociocultural entre as pessoas, o que pode despertar o interesse por línguas estrangeiras até pouco tempo desconhecidas. E entre essas línguas de crescente interesse está a Língua Portuguesa, doravante LP. (FURTOSO, 2011).

[1] A sigla NTIC, ou apenas TIC, é a denominação "utilizada para se referir a uma série de novos meios, como a *Internet*, a multimídia, a TV digital por satélite e a realidade virtual, que giram de maneira interativa em torno das telecomunicações, da informática e dos meios audiovisuais." (SANTOS, 2012).

No que se refere ao Português como Língua Estrangeira (PLE), o movimento de ensino de português para estrangeiros vem crescendo gradativamente pelo mundo, especialmente no Brasil (OLIVEIRA, 2015; DINIZ, 2012; LUNA, 2012), devido ao desenvolvimento econômico por que vivencia o país. Podemos dizer, por exemplo, que, nos últimos cinco anos, o Brasil e seus atrativos turísticos têm sido altamente difundidos pela mídia internacional, além de ter sido escolhido por organizações internacionais e ter sediado os dois maiores eventos esportivos do mundo, a Copa do Mundo de Futebol 2014 e as Olimpíadas Rio 2016. Atualmente, o país é membro do BRICS, bloco econômico composto pelos países emergentes: Brasil, Rússia, Índia, China e África do Sul. Portanto, várias empresas internacionais têm interesse em investir no Brasil.

Outro fator a ser ressaltado é a exigência do Exame Celpe-Bras para o ingresso nas universidades brasileiras, por meio do Programa de Estudantes Convênio de Graduação (PEC–G) e de Pós-Graduação (PEC–PG), bem como para a validação de diplomas de profissionais estrangeiros que pretendem trabalhar no Brasil, e/ou obterem a inscrição profissional, como é o caso dos médicos estrangeiros que precisam realizar o exame, para que seus registros sejam efetivados nos Conselhos Regionais de Medicina. (BRASIL, 2010 *apud* Diniz, 2012). Além disso, "o exame também é pré-requisito para profissionais estrangeiros de algumas empresas multinacionais desempenhando funções no Brasil, bem como para diplomatas argentinos." (DINIZ, 2012. p. 452).

Sendo assim, juntamente com essa crescente demanda de aprendizes, de profissionais e demais interessados pela LP, há a necessidade de mais pesquisas na área de avaliação de proficiência desta língua-alvo e, consequentemente, a aplicação do Exame Celpe-Bras é cada vez mais solicitada.

Com uma maior demanda, torna-se essencial pensar em meios práticos e de baixo custo de se alcançar todos aqueles que se interessem ou necessitem fazer o exame, assim como facilitar a aplicação e a correção do mesmo. Referimo-nos aqui à *praticidade* do teste, um critério que se refere a questões administrativas e de logística, que devem ser levadas em conta na construção, aplicação e correção de um exame avaliativo. Por exemplo, deve-se pensar no custo que este vai ter, no tempo que se levará para fazê-lo e aplicá-lo e se será de fácil correção ao se interpretar os resultados (BROWN, 2010). Bachman e Palmer (1996) destacam que, das seis[2] qualidades observadas no desenvolvimento de um teste - *o construto,*

[2] Essas qualidades ou critérios de testes são explicados, de forma detalhada, na primeira seção do terceiro capítulo desta dissertação.

a validade, a confiabilidade, a praticidade, a autenticidade e *o efeito retroativo* -, a praticidade é a única que se difere das outras em relação à sua natureza, pois todas outras dizem respeito ao uso feito dos resultados dos testes, enquanto que a praticidade "refere-se primeiramente à maneira na qual o teste será executado." (p. 35). Ou seja, apesar de ser ideal alcançar um equilíbrio entre as seis qualidades de teste a praticidade é fundamental, visto que todas as outras qualidades podem ser repensadas de acordo com os recursos necessários e os recursos realmente disponíveis para o *design*, o desenvolvimento e a aplicação dos testes. Neste aspecto, consideramos ser importante identificar, de forma sistemática, as NTIC apropriadas e que também possam colaborar para a praticidade do Exame Celpe-Bras em um possível formato computadorizado ou *online*, na avaliação de desempenho, justificando, assim, a pertinência dessa pesquisa de mestrado.

Definição do problema e pergunta de pesquisa

Esta pesquisa pretende identificar as tecnologias apropriadas para a elaboração de um formato eletrônico e/ou *online* do exame Celpe-Bras e discutir como as NTIC são utilizadas em alguns exames de proficiência linguística. Sabendo que, até o momento, o exame brasileiro possui apenas a versão impressa, consideramos que esse novo formato possibilite maior praticidade na aplicação do mesmo, em vista do aumento da procura pelo teste e da ampliação de postos aplicadores credenciados.

Diante disso, abre-se o seguinte questionamento:

- ✓ Como utilizar as NTIC existentes na área de avaliação de proficiência linguística, para a elaboração e aplicação de uma versão eletrônica/*online* do Celpe-Bras?

Objetivos da pesquisa

Objetivo geral da pesquisa

Investigar o uso das NTIC como ferramenta para integrar e aprimorar a realização de tarefas destinadas à avaliação de proficiência em Português como Língua Estrangeira, no exame Celpe-Bras.

Objetivos específicos da pesquisa

Identificar as NTIC utilizadas na realização de tarefas avaliativas em testes de proficiência linguística.

Discutir as qualidades ou critérios de teste e as características de alguns exames de proficiência em meio eletrônico.

Propor alguns cuidados e/ou sugestões para uma versão *online* do Exame Celpe-Bras.

Organização da dissertação

A dissertação está organizada em cinco capítulos. No primeiro, apresentamos a metodologia da pesquisa, no qual traçamos o caminho percorrido durante todo o processo em busca do recorte teórico e, posteriormente, para a discussão das análises e exposição das propostas.

O segundo capítulo contempla a contextualização do estudo e a descrição do Exame Celpe-Bras na versão impressa.

O terceiro capítulo contempla a revisão de literatura que fundamenta a análise e a discussão das propostas, além de oferecer subsídios para as contribuições desta pesquisa. É relevante destacar que não abordamos ou utilizamos uma teoria específica para embasar esta dissertação. No entanto, percebemos nos trabalhos revisados a presença de conceitos e concepções teóricas que apresentam uma visão sobre a avaliação mediada pela tecnologia. Visão esta que julgamos positiva e promissora. Sendo assim, estas concepções teóricas servem como base para a pesquisa.

Neste estudo, entendemos como fundamentais algumas questões, à luz das contribuições já oferecidas por pesquisadores e estudiosos da área de avaliação em âmbito nacional e internacional. São elas: i) a concepção de avaliação e de testes de proficiência linguística, suas qualidades ou critérios (MESSICK, 1989; BACHMAN e PALMER, 1996; McNAMARA, 2000; CONSOLO, 2004; CHAPELLE e DOUGLAS, 2006; TOSTES, 2009; FULCHER, 2010; BROWN, 2010; SALOMÃO, 2010; ANCHIETA, 2010; FURTOSO, 2011;

FURTOSO, GOMES e CONSOLO, 2011; CONSOLO e SILVA, 2014; CONSOLO e FURTOSO, 2015); ii) o histórico, de forma resumida, dos testes de proficiência (BARNWELL, 1996; DUNKEL, 1999; CHALHOUB-DEVILLE 2001; SALOMÃO, 2010); e, o uso das tecnologias em testes de LE ao longo dos últimos vinte anos (DUNKEL, 1999; CHALHOUB-DEVILLE 2001; GODWIN-JONES, 2001; BENNETT, 2001; WALCZAK, 2015; CHAPELLE e VOSS, 2016). Desta forma, a área de avaliação de proficiência em LE é abordada, conjuntamente, com a área das NTIC em testes de alta relevância e/ou em larga escala em meio computadorizado.

Em seguida, no capítulo 4, apresentamos a descrição das NTIC e de alguns testes de larga escala que as utilizam em versões eletrônicas. Assim, as principais características de um *Computer-Based Testing* (CBT) e *Computer-Adaptative Testing* (CAT) são apresentadas, bem como as ferramentas tecnológicas utilizadas por testes como os exames do *Educational Testing Service* (ETS); o exame PTE-*Academic*; o Projeto DIALANG e os testes da *Language Testing International* (LTI). Por fim, apresentarmos as principais mudanças no método de teste de um exame em formato impresso para um teste em versão computadorizada/*online*.

No quinto capítulo, retomamos a pergunta de pesquisa, analisamos o uso das NTIC nos testes descritos e a possibilidade de sua utilização no Exame Celpe-Bras, explicitando algumas propostas para a elaboração e a otimização de uma possível versão *online*.

No último capítulo, discutimos o papel do Exame Celpe-Bras na ampliação da difusão e promoção da LP, e como a praticidade de uma aplicação em meio computadorizado e/ou *online* pode repercurtir no processo de internacionalização da LP. Em seguida, apresentamos as considerações finais e sintetizamos os principais pontos do estudo, a fim de apontar os desafios para a criação de um novo formato do exame, as limitações deste estudo e sugerir direções para pesquisas futuras.

Finalizando, reunimos as referências bibliográficas que fundamentam o estudo, seguidas dos anexos.

CAPÍTULO 1 - METODOLOGIA

Neste capítulo, explicitamos o tipo e o caráter da pesquisa, os procedimentos metodológicos, o instrumento de coleta de dados e as categorias de análise, traçando, por fim, o caminho percorrido até as discussões e as propostas finais do estudo.

1.1 Natureza da pesquisa

Esta pesquisa é de caráter bibliográfico, sendo que, apresentamos um levantamento de estudos acerca do assunto e os resultados alcançados, com o objetivo de "conhecer e analisar as principais contribuições teóricas existentes sobre um determinado tema ou problema" (KÖCHE, 1997. p. 122 *apud* FURTOSO, 2011. p. 33), para que novas pesquisas possam ser feitas com vistas a oferecer contribuições relevantes para o avanço da área de avaliação. Assim, adotamos para esta pesquisa a linha metodológica descritiva, com a abordagem de análise bibliográfica e documental, de natureza aplicada e de cunho qualitativo.

Segundo Dornyei (2007), a pesquisa qualitativa envolve procedimentos de coleta de dados que resultam principalmente em dados não numéricos, analisados principalmente por meio de métodos subjetivos e interpretativos.

Quanto à natureza aplicada, esta se propõe a "gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos." (PRODANOV e FREITAS, 2013. p. 51), aplicando, no caso desta pesquisa, as concepções teóricas encontradas a um recorte da realidade ou *corpus* de pesquisa.

Do ponto de vista do objetivo, classificamos esta pesquisa como descritiva, por descrever os fatos observados sem interferir neles e estabelecer relações entre algumas variáveis, buscando uma nova visão do problema.

Em se tratando da pesquisa documental, Prodanov e Freitas (2013) a caracterizam da seguinte forma:

> A pesquisa documental baseia-se em materiais que não receberam ainda um tratamento analítico ou que podem ser reelaborados de acordo com os objetivos da pesquisa. Assim como a maioria das tipologias, a pesquisa documental pode integrar o rol de pesquisas utilizadas em um mesmo estudo ou se caracterizar como o único delineamento utilizado para tal. (p. 55).

Ainda segundo os autores, o documento é qualquer registro que possa ser usado como fonte de informação, por meio de investigação, que engloba: "observação (crítica dos dados na obra); leitura (crítica da garantia, da interpretação e do valor interno da obra); reflexão (crítica do processo e do conteúdo da obra); crítica (juízo fundamentado sobre o valor do material utilizável para o trabalho científico)." (PRODANOV; FREITAS, 2013. p.56).

Sendo assim, passamos, a seguir, para os procedimentos metodológicos desta pesquisa.

1.2 Procedimentos Metodológicos

Como procedimento inicial, buscamos elencar publicações e documentos que abordem o tema das NTIC na área da avaliação de proficiência em língua(s) dentro da Linguística Aplicada e alguns exames de proficiência em LE que utilizam recursos tecnológicos, a fim de verificar como esses recursos são utilizados. Desta forma, organizamos e interpretamos os documentos, a fim de atingir os objetivos a que nos propusemos anteriormente.

Primeiramente, como base de nossa pesquisa, descrevemos o exame Celpe-Bras em sua versão impressa, para, então, seguir com a revisão bibliográfica de trabalhos na área e dos critérios de qualidade de uso de testes impressos e computadorizados e seus históricos - desde os primeiros testes até os dias atuais. Em seguida, abordamos as tecnologias e seus usos em testes de proficiência linguística em meio eletrônico, descrevendo as ferramentas utilizadas na área e suas principais características. Nesta primeira parte, concentramos nosso estudo em publicações e trabalhos nacionais e internacionais dos últimos vinte anos, realizados sobre o tema. Assim, como dito anteriormente, os conceitos e as concepções teóricas presentes nos documentos revisados fornecem uma visão positiva e promissora sobre a avaliação em meio computadorizado e embasam este estudo.

Em um segundo momento, partimos para a descrição e análise de alguns testes que utilizam as ferramentas tecnológicas, selecionados de acordo com as características que julgamos pertinentes a esta pesquisa, a saber: que sejam testes de proficiência linguística de larga escala, total ou parcialmente aplicados em meio eletrônico, possuindo ou não a versão impressa.

Portanto, alguns exames internacionalmente reconhecidos, como o *International English Language Test System* (IELTS), os *Cambridge English Tests* (ESOL), o *First*

Certificate in English (FCE), o *Test de Connaissance du Français* (TCF/TEF), o *Diplomas de Español como Lengua Extranjera* (DELE), *Certificado de Español Lengua y Uso* (CELU), o *Der Test Deutsch als Fremdsprache* (TestDaF), os testes desenvolvidos pelo Centro de Avaliação de Português Língua Estrangeira (CAPLE) foram descartados para a análise nesta pesquisa por possuírem apenas a versão em papel e caneta, ou por serem testes tradicionais lineares computadorizados[3], utilizando unicamente questões de múltipla-escolha, ou, ainda, por utilizarem entrevistas com interação face a face na parte oral, assim como o Celpe-Bras o faz na versão impressa.

Assim, para a delimitação do estudo, tendo em vista o critério de seleção estabelecido por nós, dentre os inúmeros exames de proficiência existentes no mercado, tais como os citados acima, decidimos abordar os seguintes testes:

i) *Test of English as a Foreign Language* – TOEFL, por apresentar o formato em papel/caneta e a versão *online* (*Internet Based Test* - iBT);

ii) *Test of English for International Communication* – TOEIC, por ter passado por consideráveis mudanças em seu formato, ao longo de trinta anos de existência, e ser configurado em duas partes, uma na versão impressa (compreensão oral e leitura) e outra realizada pela mediação do computador (expressão oral e escrita). Analisamos, portanto, apenas esta segunda parte do exame.

Ambos os exames são americanos, elaborados pelo *Educational Testing Service* (ETS) e reconhecidos internacionalmente no mundo acadêmico e dos negócios.

Além desses, para compor a análise, buscamos outros três exames desenvolvidos para ser aplicados totalmente em formato eletrônico, e encontramos:

iii) *The Pearson Test of English Academic* - PTE *Academic*, que é um teste britânico destinado a falantes não nativos de inglês que desejam estudar no exterior;

iv) O Projeto DIALANG, totalmente gratuito e financiado por instituições e universidades europeias, destinado a avaliar 14 línguas diferentes, podendo ser

[3] Um teste linear computadorizado é um teste padronizado que possui um mesmo número de questões ordenadas na mesma sequência, administrado a todos os examinandos em uma determinada aplicação do teste, o que o torna muito parecido com os testes de papel e caneta, se diferenciando apenas pelo meio de administração - o computador. McNamara (2000, p. 5) explicita a presença de itens ou questões de testes com "formato de respostas fixas", para caracterizar um teste tradicional e sugere que as questões de múltipla-escolha são as mais importantes e utilizadas nesses tipos de testes.

usado em computadores que executem um sistema operacional em um navegador da *Web*;

v) Os testes da *Language Testing International* (LTI), uma instituição americana, antiga na área de avaliação e voltada para a elaboração e desenvolvimento de testes em mais de cem idiomas, em mais de 40 países. Selecionamos, assim, três testes computadorizados dos cinco testes do *American Council on the Teaching of Foreign Language* - ACTFL: o ACTFL *Writing Proficiency* (WPT); o ACTFL *Oral Proficiency Interview* (OPI); e o ACTFL *Oral Proficiency Interview Computer Test* (OPIc).

Embora o interesse pela realização de estudos sobre os testes de proficiência e as NTIC, no Brasil, venha crescendo, ainda há poucos estudos nessa área. Além disso, pouco material é divulgado sobre os exames de proficiência selecionados, logo, os documentos e dados desta pesquisa foram coletados em páginas oficiais na *Internet* das entidades que os elaboram ou são responsáveis pelos mesmos, tais como, os manuais eletrônicos e as pesquisas divulgadas sobre os testes em questão. Assim, as amostragens de exemplificação de tarefas e os modelos dos exames selecionados são apresentados na forma em que foram encontrados nas fontes citadas anteriormente.

Para melhor visualização e organização dos dados, apresentamos alguns quadros, a fim de expor os exames de forma clara e objetiva, colaborando para a identificação e análise das NTIC empregadas em cada seção dos testes.

1.3 Instrumentos de coleta de dados

Como dito anteriormente, as informações que subsidiaram este estudo foram coletadas em documentos e fontes bibliográficas da área das NTIC relativos a testes de larga escala, com vistas a identificar as funcionalidades dessas tecnologias como ferramentas para integrar e aprimorar a realização de tarefas destinadas a avaliação de proficiência linguística. Assim, os dados e os exames selecionados são descritos e/ou analisados e exemplificados, com o intuito de verificar as ferramentas tecnológicas existentes e presentes em alguns testes em LE. No entanto, os exames selecionados para descrição e análise não constituem o cerne desta pesquisa, que pretende se concentrar no uso das NTIC na avaliação de proficiência em LE que

possam servir como subsídios para a criação ou a otimização do exame Celpe-Bras em meio eletrônico ou *online*. Entendemos, assim, que as ferramentas tecnológicas utilizadas por esses testes e as suas características em meio eletrônico propiciam um caminho para alcançar nosso objetivo e traçar algumas respostas à pergunta de pesquisa.

Dessa forma, a presente pesquisa baseia-se na metodologia de coleta de documentos, como material primordial, realizada a partir de documentos contemporâneos (atuais) ou retrospectivos (documentos com mais de dez anos) considerados cientificamente autênticos (não-fraudados, retirados de fontes confiáveis). É necessário destacar que, nessa tipologia de pesquisa, os documentos são classificados em dois tipos principais: as fontes primárias e as fontes secundárias. As fontes primárias são as que não receberam qualquer tratamento analítico, tais como: documentos oficiais, reportagens de jornal, cartas, contratos, diários, filmes, fotografias, gravações etc. As fontes secundárias, por sua vez, são os documentos que, de alguma forma, já foram analisados, que se baseiam em livros, revistas, jornais, publicações avulsas e teses, cuja autoria é identificada. (PRODANOV; FREITAS, 2013).

Portanto, utilizamos documentos dos dois tipos, já que lidamos com publicações sobre o tema da pesquisa e os manuais e as páginas oficiais dos testes de proficiência, além de publicações sobre os mesmos. Optamos por apresentar as fontes de documentos primários utilizados, como os manuais e os tutoriais dos testes, entre outros, ao longo do desenvolvimento do texto nas referências de rodapé, para melhor organização da pesquisa.

1.4 Categorias de análise

O método de análise utilizado para esta pesquisa se enquadra na técnica de análise apresentada por Seliger e Shohamy (1989, p. 205), em que o pesquisador aplica aos dados um sistema de categorias pré-existentes "originados de uma estrutura conceitual ou das perguntas de pesquisa específicas".

Sendo assim, após a investigação das NTIC na avaliação de proficiência, estabelecemos, previamente, algumas categorias para análise dos exames selecionados, lembrando que a partir destas categorias podem surgir subcategorias. Essas categorias são explicitadas em maior ou menor proporção, dependendo do interesse de nossa pesquisa, previamente relatado neste capítulo. São elas: i) Abordagem/ construto do exame (visão de uso de língua contida no teste); ii) Método ou tipo do teste (de desempenho ou tradicional); iii) Linear ou adaptativo; iv) Tarefas/questões: pontos discretos/ integradas/ comunicativas; v)

Critérios de avaliação: analítico/ holístico; vi) Autenticidade/ praticidade; vii) Nível de certificação/ pontuação; e viii) Artefatos tecnológicos utilizados. Essas categorias ou concepções teóricas de avaliação são definidas, explicadas e exemplificadas no terceiro capítulo, e ao longo das descrições das características dos testes computadorizados, no quarto capítulo. Para melhor visualização e resumo dessas categorias, elaboramos uma lista de verificações, ou *checklist*, com as oito categorias aqui mencionadas, referentes aos cinco exames previamente descritos, a fim de identificar a presença ou não dessas categorias e/ou como são apresentadas nos testes.

Em seguida, partimos para a análise dos dados obtidos e apresentamos algumas propostas, a fim de identificar o que as NTIC têm a oferecer aos testes de proficiência linguística elaborados em meios eletrônicos[4], e quais ferramentas poderiam ser também aplicadas ao exame Celpe-Bras, em nova versão computadorizada, visando-se não apenas à praticidade na aplicação e correção do exame, mas respeitando o construto, a validade, a confiabilidade e as demais qualidades inerentes a todos os testes de alta relevância em nível internacional, apresentados no capítulo 3 deste estudo.

Desta forma, a partir dos resultados encontrados, apresentamos alguns critérios fundamentais, que podem servir como um guia para os desenvolvedores de uma possível versão *online* do exame Celpe-Bras, de acordo com os trabalhos de Bachman e Palmer (1996) e Chapelle e Douglas (2006). Além disso, estabelecemos uma lista de questionamentos essenciais, que podem nortear o trabalho dos elaboradores e/ou órgãos responsáveis pelo Celpe-Bras, para assegurar que o exame continue justo, válido e confiável em um formato computadorizado. Os questionamentos por nós apresentados baseiam-se nos estudos de Corbel (1993) e são respondidos de acordo com nosso estudo.

Finalmente, discutimos o alcance ou o impacto da pesquisa e a ampliação de oportunidades para a internacionalização da LP, apresentando as limitações e os desafios para a criação de um novo formato do Exame e a necessidade de pesquisas e investigações futuras neste campo.

No próximo capítulo, abordamos a contextualização da pesquisa, bem como a descrição do exame Celpe-Bras em sua versão impressa. Em seguida, no capítulo 3, revisitamos as concepções teóricas que fundamentam esta pesquisa, dando, então, sequência ao estudo.

[4]Referimo-nos aqui aos dispositivos utilizados em testes mediados por computador não lineares, como o CAT – *Computer Adaptative Tests,* o NLP – *Natural LanguageProcessing,* e a possibilidade de uso de dispositivos *síncronos* ou comunicadores instantâneos via *Internet,* na avaliação do desempenho oral do examinando, por exemplo.

CAPÍTULO 2

CONTEXTUALIZAÇÃO DA PESQUISA

As NTIC têm avançado de forma crescente e, ao que parece, não há limites para o que está por vir. Sua inclusão na sociedade moderna, de uma forma ou de outra, com mais ou com menos vigor, é uma realidade cada vez mais palpável e irreversível. As NTIC estão presentes no cotidiano das pessoas, pelo uso das redes sociais e de modalidades de comunicação síncronas e assíncronas - como o *Facebook, Instagram, Skype, Whatsapp, IMO, Telegram* - e as informações estão sendo disseminadas de forma rápida e global, via *Web World Wide*. Assim, com a evolução da sociedade, percebemos que os meios de comunicação têm sido cada vez mais ligados aos meios eletrônicos e ao uso frequente de celulares, computadores, tablets, iPods, iPads, entre outros. Destarte, o modo de viver e trabalhar vem mudando gradativamente e se mostra, muitas vezes, dependente dos meios eletrônicos e tecnológicos. O desenvolvimento das NTIC avança em diversas áreas do conhecimento humano – científico, social, político - e não poderia ser diferente com a área acadêmica, e, mais precisamente, com o ensino, a aprendizagem e a avaliação em LE. Sendo assim, a área educacional tende a inovar e a acompanhar esse avanço tecnológico, progressivamente.

A avaliação em meios eletrônicos, em função de suas especificidades, é, em vários aspectos, diferente da avaliação em papel e caneta. De acordo com Salomão (2010, p.325), "a área da avaliação em L2 é um dos campos mais jovens da Linguística Aplicada, e a teoria e a prática de testes de proficiência oral representam a subárea mais jovem no campo da avaliação." O que dizer então da prática de testes de proficiência em meios eletrônicos? Tratando de LE e avaliação de proficiência, Furtoso (2011) destaca que adentrar o campo da avaliação nas áreas da LP como LE, doravante PLE, ainda é como entrar em um ambiente quase que inexplorado, com poucas pesquisas sobre o tema. No entanto, é possível notar um progresso no que diz respeito à avaliação de proficiência em PLE, que vem ganhando reconhecida visibilidade a partir de iniciativas de otimização do espaço interacional *online.*

Sobre essa visibilidade, segundo dados da pesquisa de Oliveira (2013), o português é língua oficial em 10 países, além da Guiné Equatorial e da Região Administrativa Especial de Macau, onde é cooficial, junto com o mandarim até o ano de 2049. Ainda, segundo a mesma

pesquisa, há cerca de 7 a 9 milhões de falantes de Língua Portuguesa nas Diásporas, especialmente nos Estados Unidos da América e no Canadá. Interessante abordar o fato que esse interesse pela LP nos EUA, por exemplo, não é recente. Luna (2012) destaca que o crescimento do ensino da LP nos EUA é comprovado por pesquisas, com motivação instrumental, especialmente para fazer negócios, porém, esse interesse vem de uma longa história, de quase quatro séculos, marcada por altos e baixos. Gonçalves (2012) acrescenta que a LP é encarada como uma mais-valia no currículo dos jovens nos EUA, provenientes, na sua maioria, de fora das comunidades de falantes de português.

No espaço *online*, ou seja, na *Internet*, Oliveira (2013, p. 411-412) afirma que o português "alcançou recentemente a cifra de 83 milhões de usuários, passando a ser, em 2010, a quinta língua mais usada na rede – à frente do japonês." Isso pode ser observado nas redes sociais, como o *Facebook*, *o Twitter*, e no uso de aplicativos como *Whatsapp* e *Skype*, entre outros. Oliveira (*Ibid*, p. 417) traz um panorama do período pós 2004 que, de acordo com ele, tem representado o auge do crescimento da LP:

> O período pós-2004, [...], tem sido um período virtuoso para o crescimento da língua portuguesa, tanto internamente como externamente. Ampliou-se o letramento da população, a inserção dos países na sociedade internacional, o crescimento da classe média, criando uma produção e um consumo cultural mais sofisticado, mais viagens ao exterior e maior acesso à *Internet.* Estes fatores fomentam um interesse maior pelos países de língua portuguesa e, consequentemente, maior disposição para o seu aprendizado como língua estrangeira.

Outro fator que colabora para a difusão da LP é a elaboração e a aplicação de exames de proficiência como o exame Celpe-Bras (Certificado de Proficiência em Língua Portuguesa para Estrangeiros), que é o único exame reconhecido pelo governo brasileiro, aplicado desde 1998. Desde sua primeira aplicação, o exame vem crescendo em procura e possui 29 postos aplicadores credenciados no Brasil e 65 no exterior. (BRASIL, 2016). No entanto, este exame, ainda não apresenta uma versão computadorizada ou *online*. Somente a correção da Parte Escrita é realizada no espaço *online* desde a edição 2015/01.

Por outro lado, há exames internacionais de proficiência linguística, já consolidados, que utilizam as NTIC. Fato é que as pessoas podem fazer testes *online*, viabilizados por entidades governamentais ou particulares responsáveis pela aplicação, visando à praticidade na correção, visto que o resultado pode ser divulgado quinze dias após a realização do mesmo, o que acontece, por exemplo, no exame TOEFL-iBT. Outro fator importante é a questão

mercantilista dos testes de proficiência padronizados e/ou reconhecidos internacionalmente, tais como: *Test of English as a Foreign Language* (TOEFL), *Test of English for International Communication* (TOEIC), *International English Language Test System* (IELTS), *First Certificate in English* (FCE), *Cambridge English Tests* (ESOL), entre outros. Esses testes, diferentemente do Exame Celpe-Bras, são de responsabilidade de instituições que visam ao interesse financeiro, e que têm esses testes como principal negócio. Contribuem para a lucratividade dessas instituições: os diversos tipos de certificação ofertados (normalmente em conformidade com o Quadro Europeu Comum de Referência para o Ensino de Línguas) e a validade da certificação, em torno de dois a três anos. Por outro lado, o Celpe-Bras é uma iniciativa do governo brasileiro, não tem fins lucrativos e, embora seja cobrada uma taxa dos examinandos, o valor dela serve para cobrir as despesas da instituição aplicadora. Nesse sentido, o Celpe-Bras pode ser considerado uma importante ação de política de difusão e divulgação da Língua Portuguesa, ação essa custeada pelo governo brasileiro.

Assim, muitos testes são elaborados visando objetivos distintos dependendo se são controlados, idealizados ou aplicados por órgãos do governo ou entidades privadas/ particulares. Desta forma, eles podem visar ao lucro final, não se preocupando com algumas questões, tais como, onde são utilizados, ou se estão sendo aplicados com um propósito diferente daquele para o qual foram elaborados, comprometendo, assim, a confiabilidade dos mesmos. Quanto maior a praticidade na aplicação e correção de um teste de proficiência, maior cuidado e estudo deve haver para não correr o risco de comprometer a validade do construto[5], por exemplo. Muitos testes pecam em considerar e testar apenas a confiabilidade dos mesmos. No caso dos testes mundialmente conhecidos de Língua Inglesa, doravante LI, eles se sustentam dentro de uma 'norma e ações descentralizadas de promoção via mercado', confirmando o papel que a LI ainda desempenha no sistema mundial, a de 'manutenção das funções de língua hipercentral' com ganho de importância por ser considerada a língua da ciência e da tecnologia em muitos países ao redor do mundo, como a Holanda e alguns países escandinavos, por exemplo. (OLIVEIRA, 2013).

Outra questão que podemos ressaltar diz respeito aos testes de proficiência e seus possíveis formatos. É possível encontrar testes de diferentes línguas, em várias versões: *online*, eletrônica, papel e caneta, entrevistas face a face. Todos esses formatos, teoricamente, tornam-se opções para que o examinando possa escolher, de acordo com seu interesse e/ou

[5] Entende-se aqui por validade do construto se o teste representa precisamente a teoria da variável que ele mede. (SHOHAMY e SELIGER, 1989 p. 190). Validity construct: whether it represents accurately the theory of the variable which it measures. [no original].

finalidade, aquele que melhor lhe convier. No entanto, nem todas as sociedades possuem acesso a diferentes formatos de testes, como é o caso do Celpe-Bras, que tem sido aplicado apenas em versão impressa.

Finalmente, outro fator a ser levado em conta é a diversidade de conteúdos e abordagens utilizadas em cada teste. As opções aumentam gradativamente de acordo, inclusive, com o propósito ou objetivo de cada teste, seja possibilitar a entrada em uma universidade ou empresa, ou servir como parâmetro de desempenho individual para o próprio examinando. Desta forma, acreditamos que a área de estudo de testes de proficiência em ambientes virtuais pode ser de grande ajuda no que se refere à praticidade de testes de proficiência em larga escala, como é o caso do Exame Celpe-Bras, que vem crescendo gradativamente em procura no mundo inteiro. Porém, faz-se necessária a criação de critérios e métodos que possam manter a sua validade, confiabilidade e autenticidade, em formato eletrônico e/ou *online*. Ou seja, é necessária uma investigação acerca dos recursos tecnológicos existentes, aqui tratados como as NTIC, que podem ser utilizados para um novo formato do exame. E é nesse contexto que o presente trabalho se insere.

Prosseguimos, assim, para a próxima seção, com uma descrição do Exame Celpe-Bras em sua versão impressa.

2.1 O Exame Celpe-Bras

O Celpe-Bras foi instituído em 1994 e aplicado pela primeira vez em 1998. Ele é o único certificado de proficiência em LP reconhecido pelo governo brasileiro. É estruturado em duas etapas: a **Parte Escrita**, com duração de 3 horas, composta de quatro tarefas que integram compreensão e produção escrita e compreensão auditiva; e a **Parte Oral**, com duração de 20 minutos, em que o examinando interage com um entrevistador, sendo observado por outro examinador, os quais analisam holística e analiticamente o desempenho do mesmo, de acordo com as grades de correção (ANEXO I e II).

Para realizar o teste, os examinandos devem fazer as inscrições pela página do Instituto Nacional de Estudos e Pesquisas Educacional – INEP, mediante um cadastro *online* e o pagamento de uma taxa. O Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (CEBRASPE) é o responsável pelo apoio técnico e logístico à realização do Exame.

De acordo com o Guia de Capacitação do Examinador[6]:

> O Celpe-Bras é um exame de alta relevância, uma vez que decisões importantes são tomadas com base em seus resultados. É internacionalmente aceito em empresas e instituições de ensino como comprovação de proficiência em Língua Portuguesa. No Brasil, algumas universidades exigem o Celpe-Bras para o ingresso em cursos de graduação, por meio de alguns convênios como o Programa de Estudantes-Convênio de Graduação (PEC-G) e o Programa de Estudantes-Convênio de Pós-Graduação (PEC-PG), bem como para a validação de diplomas de profissionais estrangeiros que pretendem trabalhar no país. É também requisito para inscrição profissional em algumas entidades de classe, a exemplo do Conselho Federal de Medicina (CFM). (BRASIL, 2015).

Os textos que servem de insumo para as partes (Escrita e Oral) são autênticos e retirados de diversas mídias. Os elaboradores do exame selecionam os chamados Elementos Provocadores (EP), usados na Parte Oral, os quais constituem pequenos textos, fotos, cartuns, etc., sobre assuntos do dia-a-dia. O exame é aplicado no Brasil e em outros países em duas edições anuais, geralmente, em abril e outubro; e é oferecido para os examinandos estrangeiros não brasileiros ou lusófonos, cuja língua materna não seja o português.

Os níveis de certificação e pontuação do Exame são apresentados a seguir.

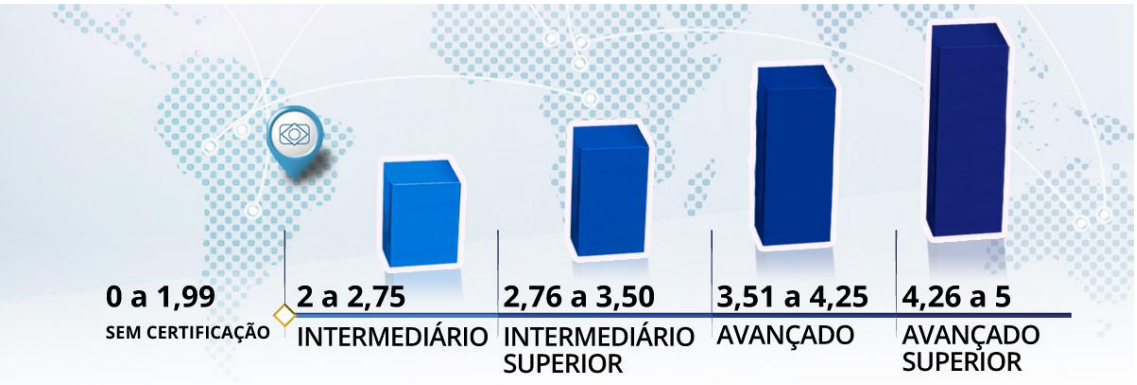
2.1.1 Os níveis de certificação

A certificação ocorre em quatro níveis de proficiência: Intermediário, Intermediário Superior, Avançado e Avançado Superior, sendo que o nível Básico não gera certificação. Sendo assim, a proficiência do examinando é avaliada por meio de seu desempenho nas quatro tarefas da Parte Escrita, com peso igual para cada uma delas, e na Interação Face a Face da Parte Oral. Como, no Exame, as quatro habilidades (falar, ouvir, ler e escrever) são avaliadas de forma integrada, espera-se que o examinando apresente equilíbrio de desempenho em ambas às partes do exame para obter determinada certificação, ou, no caso de níveis diferentes, o nível com nota inferior servirá como base para certificação do examinando.

O gráfico 1 a seguir mostra os níveis de certificação do Exame Celpe-Bras e as respectivas médias de pontuação de cada nível.

[6] Disponível em: < http://portal.inep.gov.br/celpebras-estrutura_exame>. Acesso em: jan/2017.

GRÁFICO 1 – Os níveis de certificação do Exame Celep-Bras



Disponível em: <http://portal.inep.gov.br/web/guest/acoes-internacionais/celpe-bras>. Acesso em: abril/ 2017.

A seguir, são apresentados os postos aplicadores e o crescente número de inscritos desde a primeira aplicação do Exame.

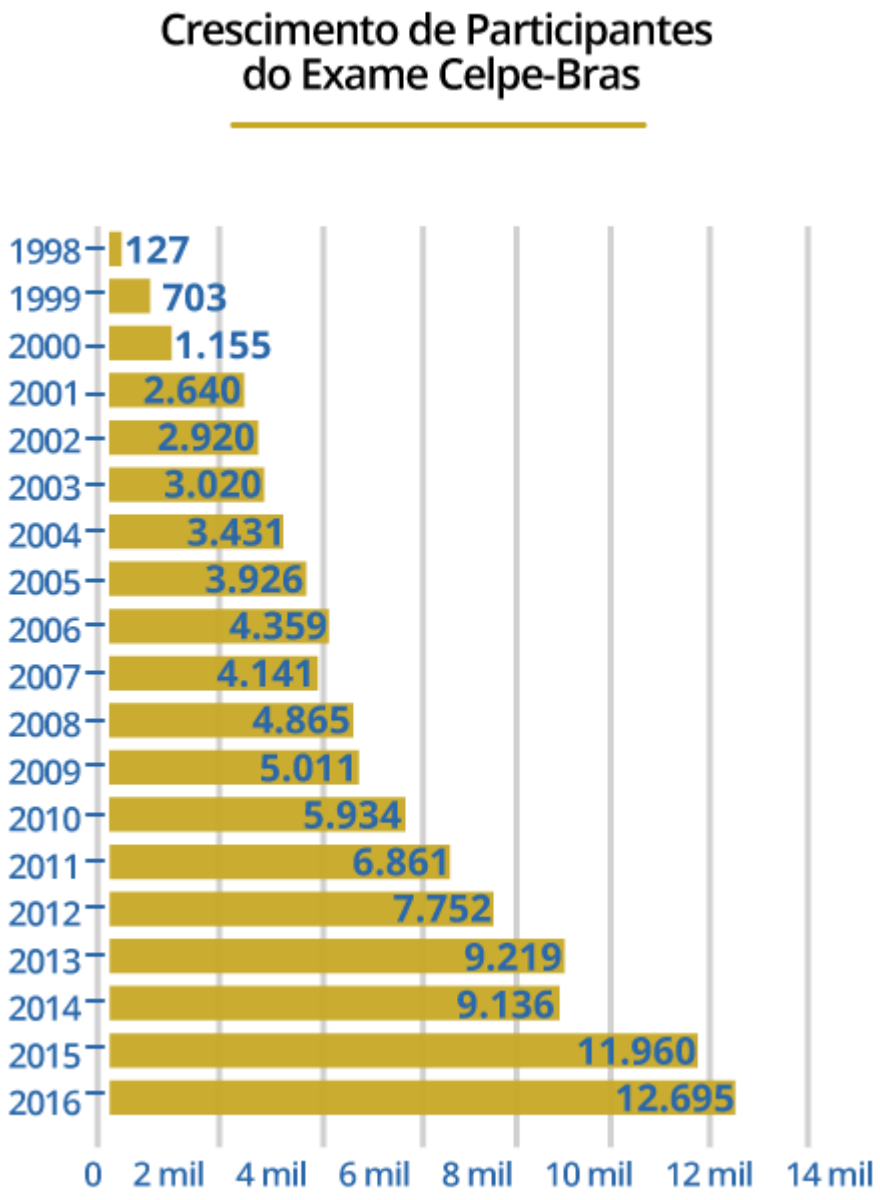
2.1.2 O crescimento pela procura do exame

Desde a sua primeira aplicação, a procura pelo Exame vem aumentando de forma significativa, sendo aplicado em 29 postos aplicadores credenciados no Brasil e 65 postos aplicadores credenciados no exterior[7]. É o que mostram o gráfico e o mapa abaixo, retirados da página oficial do INEP e do Acervo Virtual Celpe-Bras da Universidade Federal do Rio Grande do Sul (UFRGS), respectivamente.

O gráfico 2 abaixo mostra em números o crescimento pela procura do Exame desde sua primeira aplicação até o ano de 2016. E a figura 1, em seguida, mostra a relação dos 65 postos aplicadores existentes no Brasil e ao redor do mundo.

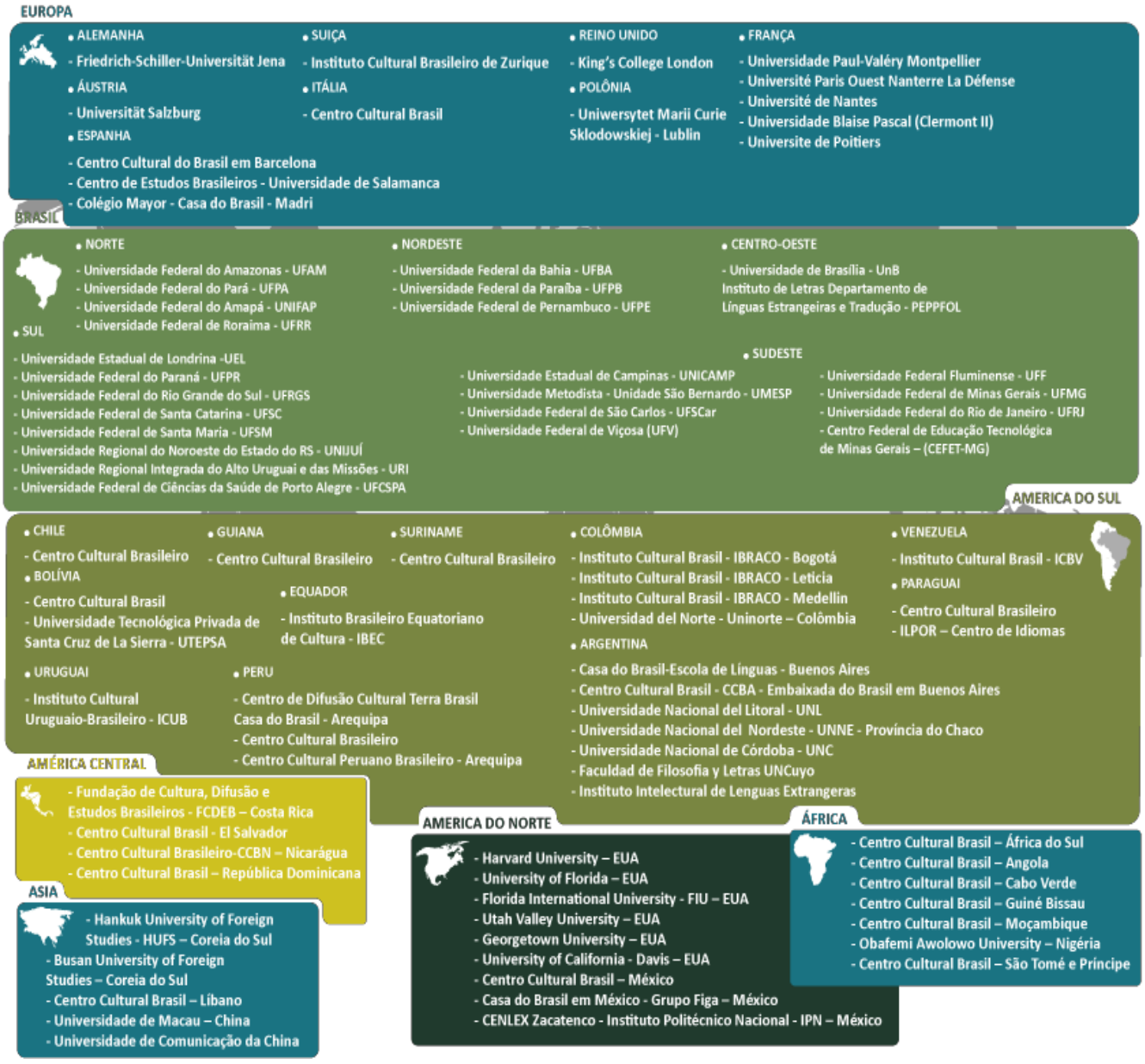
[7] Disponível em <http://portal.inep.gov.br/celpebras>. Acesso em: jan/2017.

GRÁFICO 2 – Números de inscritos no exame Celpe-Bras desde a primeira edição em 1998 até a última edição em 2016.



Disponível em: <http://portal.inep.gov.br/web/guest/acoes-internacionais/celpe-bras>. Acesso em: abril/ 2017.

FIGURA 1: Mapa com os Postos Aplicadores – Celpe-Bras 2016.1
Ilustração: Phelippe Lepletier



Disponível em: <http://www.ufrgs.br/acervocelpebras/Manuais/guias>. Acesso em: jan/2017.

Apresentamos, em seguida, a estrutura do Exame Celpe-Bras.

2.1.3 Estrutura do Exame

O Celpe-Bras não apresenta questões de múltipla-escolha ou atividade de *cloze*[8], nem foca apenas em uma habilidade por vez, separadamente. O exame adota o conceito de tarefas comunicativas, possibilitando, assim, interações que simulam atividades da vida real desempenhadas pelo examinando, que por sua vez, usa a língua de forma integrada e com um propósito social.

A seguir, são abordadas as duas partes do Exame.

2.1.4 A Parte Escrita do Exame

De acordo com McNamara (2000, p. 10), o propósito dos testes de proficiência é visualizar a situação futura da língua em uso, sem, necessariamente, referência ao processo de aprendizagem prévio. Portanto, o critério de avaliação baseia-se no comportamento comunicativo do examinando. O autor destaca, também, dois tipos de testes: os chamados 'tradicionais' e os de 'desempenho'.

Os **testes tradicionais** avaliam, de forma separada, as habilidades ou o entendimento receptivo do examinando, seguindo um padrão que verifica o domínio de pontos isolados da LE, tais como aspectos gramaticais e coesão textual. Geralmente, os itens desses tipos de testes têm formato de respostas fixo, tais como: múltipla-escolha, *cloze*, completar lacunas, respostas curtas, entre outros. Desta forma, "por serem de fácil controle, são com frequência, os mais adotados e não se prestam, portanto, a verificar aspectos produtivos da competência comunicativa, como a capacidade de persuasão ou negociação, por exemplo." (TOSTES, 2009. p. 5).

Já nos **testes de desempenho** a avaliação é feita por meio de tarefas autênticas e comunicativas, mais comuns em testes de fala e escrita, sendo necessária a presença de avaliadores treinados, além de um processo previamente definido de avaliação. Trata-se, portanto, do desempenho linguístico do examinando que engloba "não apenas a capacidade de

[8] Um teste de *cloze* é um exercício, teste ou avaliação que consiste em uma parte do texto com certas palavras removidas, em que o participante é convidado a substituir as palavras que faltam. Testes de *cloze* exigem a capacidade de entender o contexto e o vocabulário, a fim de identificar as palavras corretas ou tipo de palavras que pertencem às passagens eliminadas de um texto. Este exercício é comumente administrado em avaliação da aprendizagem das línguas nativas e segunda.

dominar aspectos gramaticais, mas também de empregá-los competentemente de acordo com a situação comunicativa que se apresenta.” (*Ibid*. p. 6).

Desta forma, o Celpe-Bras como um todo se caracteriza como um exame de desempenho, já que, na Parte Escrita, por exemplo, as tarefas comunicativas exigem que os examinandos produzam textos escritos, de acordo com um gênero textual, um propósito e um interlocutor explicitados no enunciado de tarefa, simulando, assim, uma situação real de comunicação. “O critério que está por trás de exames com propósito comunicativo, portanto, é a situação comunicativa real, e se espera que o candidato mobilize conhecimentos e habilidades para que a interação ocorra de maneira satisfatória.” (FERREIRA, 2012. p. 24).

Assim, na Parte Escrita o examinando produz quatro textos interagindo com áudios, vídeos e textos autênticos impressos, que muitas vezes apresentam linguagem verbal e não-verbal. O objetivo, nesta parte do exame, é avaliar a compreensão auditiva e escrita e a produção escrita do examinando.

Os quadros 1 e 2 mostram um exemplo de tarefa da Parte Escrita, da edição 2011/1 do Exame, e apresentam os aspectos discursivos e linguísticos avaliados.

QUADRO 1 – Tarefa 3 – Parte escrita do Exame Celpe-Bras

Enunciado da Tarefa 3 – Cosméticos para Homens – Celpe-Bras 2011-1
Você trabalha no departamento de *marketing* de uma empresa fabricante de cosméticos que pretende ampliar sua produção. Com base nas informações da reportagem abaixo, escreva um texto para a diretoria dessa empresa, salientando o perfil do homem contemporâneo e sugerindo a criação de uma linha completa de produtos para o público masculino.

Disponíveis em <http://portal.inep.gov.br/celpebras-estrutura_exame>. Acesso em: jan/2017.

QUADRO 2 – Aspectos linguísticos e discursivos avaliados na Parte Escrita (Celpe-Bras)

Aspectos discursivos avaliados	Aspectos linguísticos avaliados
Enunciador: Funcionário do departamento de *marketing* da fábrica de cosméticos. **Interlocutor:** Diretor da fábrica de cosméticos. **Propósito comunicativo:** escrever um texto para a diretoria da fábrica de cosméticos sugerindo a criação de uma linha de produtos direcionada ao público masculino e salientando o perfil do homem contemporâneo.	O atendimento aos aspectos linguísticos varia do mais adequado (nota 5) ao menos adequado (notas 1 e 0).

Disponíveis em <http://portal.inep.gov.br/celpebras-estrutura_exame>. Acesso em: jan/2017.

O quadro 3 mostra como as tarefas estão divididas na Parte Escrita.

QUADRO 3 – As quatro tarefas da Parte Escrita do Exame Celpe-Bras

Tarefa	Texto-base	Habilidades envolvidas	Tempo de realização
1	Vídeo	Compreensão oral e Produção escrita	30 minutos
2	Áudio	Comprensão oral e Produção escrita	2 horas e 30 minutos
3	Texto escrito	Compreensão escrita e Produção escrita	
4	Texto escrito	Compreensão escrita e Produção escrita	

Disponível em <http://portal.inep.gov.br/celpebras-estrutura_exame>. Acesso em: jan/2017.

A Parte Oral do Exame e suas características são apresentadas a seguir.

2.1.5 A Parte Oral do Exame

O INEP, órgão responsável pelo exame, em sua página oficial, refere-se à Parte Oral do exame como uma "interação a partir de atividades e interesses mencionados pelo

examinando na ficha de inscrição e conversa sobre tópicos do cotidiano, de interesse geral, com base em elementos provocadores". (BRASIL, 2011).

Desta forma, podemos afirmar que a Parte Oral do exame utiliza uma abordagem direta, avaliando a produção oral do examinando em uma interação face a face com um entrevistador e na presença de um observador por vinte minutos. Nos primeiros cinco minutos são feitas algumas perguntas pessoais baseadas no formulário preenchido pelo examinando no momento da inscrição. Nos quinze minutos restantes, o entrevistador escolhe três EP dentre os vinte possíveis e, baseados em um roteiro de perguntas, inicia-se uma conversa a partir da leitura desse material feita pelo examinando.

Mostramos, a seguir, um exemplo de EP da edição 2015/1 do Exame e o roteiro de questões referente ao mesmo EP, figura 2 e 3, respectivamente.

FIGURA 2 - Exemplo de Elemento Provocador – Celpe-Bras

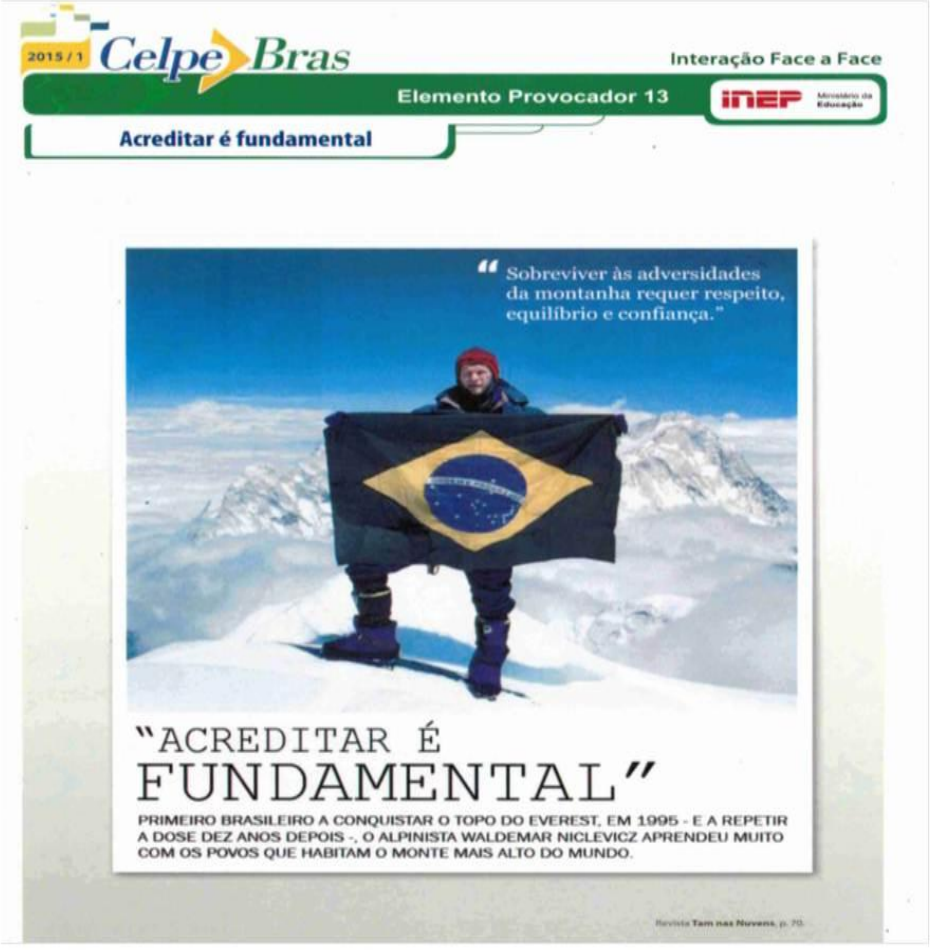


Imagem retirada de <http://www.ufrgs.br/acervocelpebras/acervo/2015-1>. Acesso em: jan/2017.

FIGURA 3 – Exemplo de Roteiro de Interação Face a Face - Celpe-Bras

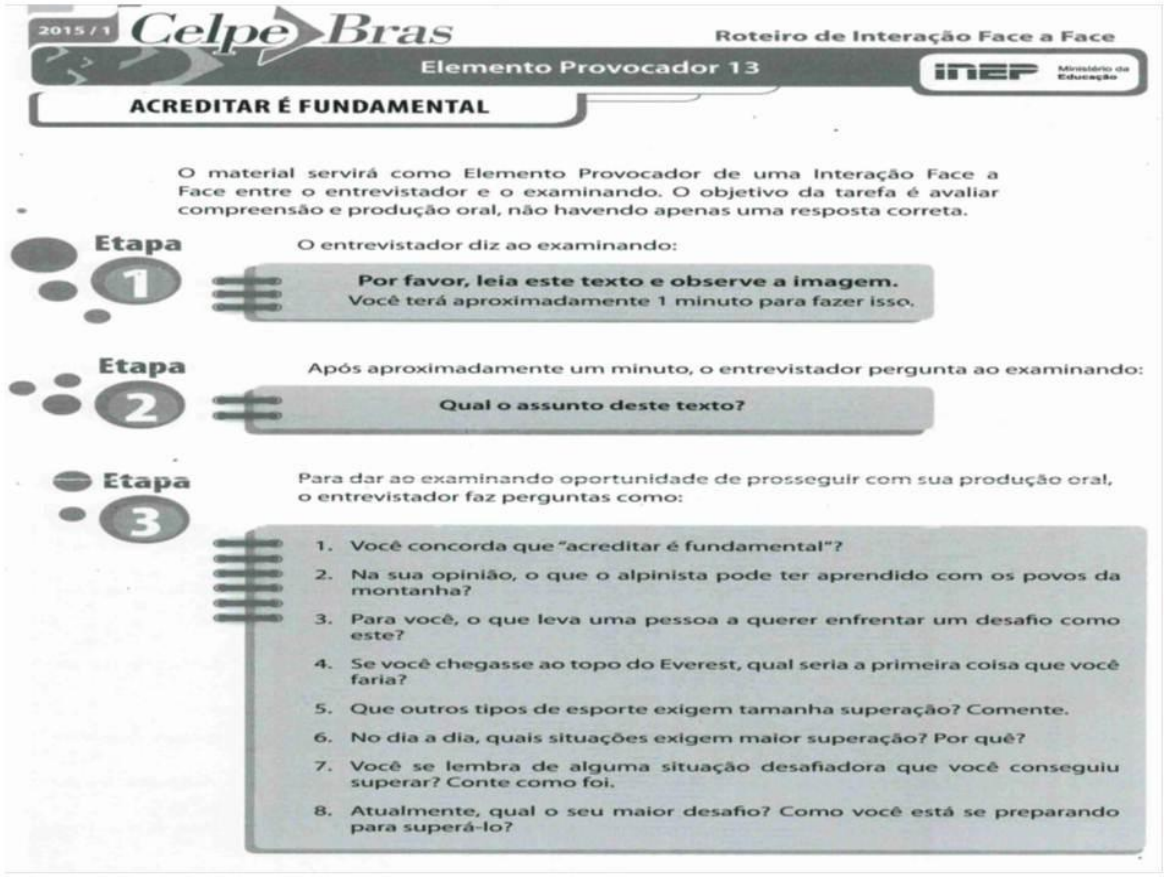


Imagem disponível em <http://www.ufrgs.br/acervocelpebras/acervo/2015-1>. Acesso em: jan/2017.

Ao final da entrevista ou interação face a face, os dois avaliadores atribuem notas com base no desempenho do examinando. Apenas o entrevistador interage verbalmente com o mesmo, cabendo ao observador avaliar o desempenho do examinando de forma analítica. O entrevistador atribui uma única nota ao examinando, de 0 a 5 pontos, de acordo com a grade holística, enquanto que o observador atribui seis notas distintas, de 0 a 5 pontos para cada item, de acordo com os critérios[9] estabelecidos na grade analítica.

Segundo Ferreira (2012), o exame Celpe-Bras se difere de outros testes padronizados em LE, como os testes europeus, por exemplo, os quais se pautam em uma lista de conteúdos, e existe uma prova específica para cada nível de certificação. Nos exames europeus, como o *International English Language Test System* (IELTS), *Cambridge English Tests* (ESOL),

[9] Os critérios avaliados pelo observador são: Compreensão, Competência Interacional, Fluência, Adequação Lexical, Adequação Gramatical e Pronúncia. Ver ANEXO II deste trabalho.

First Certificate in English (FCE), *Test de Connaissance du Français* (TCF/TEF)**,** entre outros, os níveis são pautados pelo Quadro Europeu Comum de Referência para Línguas - (QECR). Já o Celpe-Bras é formatado em uma única prova, com base na qual todos os examinandos inscritos são avaliados e certificados de acordo com o desempenho nas tarefas das duas partes (Oral e Escrita). Portanto, uma vez que o examinando não tenha obtido o nível pretendido, é possível realizar o exame novamente, até alcançar o nível de proficiência e/ou certificação desejada. (FERREIRA, 2012. p. 24).

Finalizamos este capítulo com os quadros 4 e 5, o primeiro quadro mostra como é estruturada a interação face a face da Parte Oral e o segundo quadro resume a estrutura geral do Exame.

QUADRO 4 – Parte Oral do Exame Celpe-Bras

	Conteúdo da Interação	Habilidades envolvidas	Tempo
1	Conversa sobre interesses pessoais do examinando com base nas informações obtidas nos formulários de inscrição.	Compreensão oral e Produção oral	5 minutos
2	Conversa sobre tópicos do cotidiano e de interesse geral com base em três elementos Provocadores.	Compreensão escrita/oral e Produção oral	15 minutos (cinco minutos para cada elemento Provocador)

Imagem retirada sem modificações e disponível em: < http://portal.inep.gov.br/celpebras-estrutura_exame>. Acesso em: jan./2017.

QUADRO 5 – Estrutura Geral do exame Celpe-Bras

O exame Celpe-Bras

Seção do teste	Número de questões	Limite de tempo
Parte Escrita	Composta de quatro tarefas que integram compreensão e produção escrita e compreensão auditiva, sendo permitida anotações em um rascunho que deve ser entregue no fim do exame.	3 horas
Parte Oral	O candidato interage oralmente com um entrevistador sendo observado por outro examinador, os quais analisam holística e analiticamente o desempenho do mesmo.	20 minutos
Total		3 horas e 20 minutos

Elaborado pela autora

Neste capítulo, apresentamos a contextualização do estudo e traçamos um panorama geral do crescimento do uso das tecnologias na sociedade e, consequentemente, na área de educação e avaliação. Destacamos o crescimento de interesse pela LP como LE e a importância de se pensar em meios práticos para avaliar o desempenho dos interessados na língua-alvo. Finalmente, descrevemos o Celpe-Bras, sua estrutura e suas principais características. Além disso, apresentamos os postos aplicadores e o aumento do número de inscrições para o Exame.

A seguir, partimos para o terceiro capítulo da pesquisa, discutindo as concepções de avaliação e de testes de proficiência em LE e apresentando a revisão de literatura nacional e internacional sobre o tema e um breve histórico dos testes e as tecnologias utilizadas no decorrer desse processo histórico.

CAPÍTULO 3

CONCEPÇÕES DE AVALIAÇÃO E DE TESTES DE PROFICIÊNCIA

Every test is vulnerable to good questions, about language and language use, about measurement, about test procedures, and about the uses to which the information in tests is to be put.

Tim McNamara

Ao buscar na literatura da área de avaliação, no âmbito da Linguística Aplicada, o que há de mais recente sobre as NTIC e os ambientes virtuais para elaboração e aplicação de exames de proficiência de LE, nos deparamos com poucos trabalhos, em âmbito nacional, sobre o tema. No entanto, em âmbito internacional, percebemos uma preocupação maior em estudar as tecnologias e os seus usos em avaliações de proficiência. Mesmo assim, o campo ainda carece de muita investigação. Por isso, apesar de alguns destes trabalhos datarem de mais de dez anos atrás, entendemos a importância de destacá-los e relatá-los neste capítulo, visto abordarem questões históricas acerca do percurso inicial do uso da tecnologia em testes de línguas. Estudar o passado dos testes, desde seu início, neste caso, é fundamental para situar a discussão proposta nesta dissertação e, quem sabe, vislumbrar o futuro dos mesmos.

Comecemos, assim, por Tostes (2009) e seu artigo sobre o Exame de Proficiência Oral (EPO), realizado pelo Centro de Estudos de Pessoal (CEP) e destinado a avaliar a proficiência oral no Exército Brasileiro. Nesta pesquisa, a autora admite que avaliar a fluência oral em uma LE é algo complexo e uma tarefa árdua, porém, entende que é possível obter uma visão ponderada da fluência e proficiência linguística do examinando pela "inserção de descritores linguísticos[10] no modelo de verificação e a adoção de situações que visem ao uso da língua para realizar funções comunicativas". (*Ibid*. p. 1). Por se tratar de um modelo de exame com mais de um avaliador, o EPO se utiliza do padrão de conceituação analítica, segundo o qual

[10] Tem-se adotado como critérios de avaliação os descritores linguísticos. Esses permitem que, de acordo com a produção oral evidenciada em interações comunicativas, oexaminando seja classificado num nível que retrate seu perfil linguístico, como os exames de *Cambridge*, por exemplo, que se valem dos parâmetros de descritores linguísticos do QECR. (TOSTES, 2009. p. 3)

todos os examinadores avaliam variados aspectos da comunicação que são, então, enquadrados em escalas de comunicação estabelecidas para cada um dos aspectos, isoladamente. Sendo assim, a autora defende que o teste oral do EPO serve como um meio para verificar os aspectos da conversação em LE (LI e Espanhol) à distância e em tempo real, por meio do uso de videoconferência. Segundo a autora, a videoconferência é uma ferramenta conhecida na área de educação à distância, porém, pouco utilizada em avaliações de proficiência linguística. O CEP é o único estabelecimento de ensino que atende aos militares brasileiros de todas as regiões do Brasil. Portanto, "devido a custos operacionais com deslocamentos de pessoal, a videoconferência apresentou-se como a solução mais viável encontrada." (*Ibid*. p. 4). Dessa forma, o examinando necessita apenas se dirigir para o Comando Militar de Área, sede do exame na sua região, na data e hora marcadas para a entrevista que é realizada de maneira síncrona.

O artigo de Salomão (2010), por sua vez, traz uma retrospectiva da origem e do desenvolvimento das provas de proficiência oral para falantes de LE, além de discutir os aspectos a serem levados em consideração para a formulação deste tipo de teste em ambiente virtual. Desta forma, a autora organiza o artigo em três seções: i) recapitula a história dos testes de proficiência oral, desde sua criação, focando na avaliação por intermédio da entrevista oral; ii) discorre sobre as TIC e seu impacto nos testes de proficiência oral; e iii) responde às perguntas iniciais elaboradas para discussão, a saber:

> Que fatores devem ser levados em consideração ao desenvolver-se um teste de proficiência oral em contexto virtual, mediado por computador? a) Que aspectos devem ser adequados para este meio? b) Quais as possíveis vantagens e limitações de um teste de proficiência oral em tal contexto? c) O uso de computador neste caso pode ser inovador e ao mesmo tempo manter a validade que um teste de proficiência deve ter? (SALOMÃO, 2010.p. 3).

Ao final de seu trabalho, a autora destaca que "as ameaças e preocupações que permeiam a pesquisa e uso dos Testes de Línguas Mediados por Computadores (TLMC), principalmente para testes de proficiência oral, são as mesmas existentes em outros formatos de testes, como os de papel e caneta, por exemplo." (*Ibid*. p. 18). Sobre as limitações dos fatores a serem levados em consideração na elaboração de um teste de proficiência oral em ambiente virtual, Salomão afirma que o questionamento principal deve ser se e como os

elementos tecnológicos podem afetar o uso da língua e o desempenho do examinando no teste. Por fim, a autora reconhece a carência de pesquisas na área.

Seguindo uma perspectiva de pesquisa parecida a de Salomão (2010), a dissertação de Anchieta (2010), desenvolvida na UNESP/SJRP, apresenta resultados sobre um levantamento de dados a respeito de cinco testes de proficiência em LI, mediados por computadores, existentes no mercado - um deles o TOEFL - com o intuito de contribuir para a elaboração e implementação do Exame de Proficiência para Professores de Língua Estrangeira (EPPLE), em ambiente virtual. Além da análise dos testes, Anchieta utiliza as respostas dadas aos questionários elaborados no intuito de mapear as opiniões de profissionais, voltados à formação de professores de LI, a respeito dos exames de proficiência selecionados. Ao final de sua pesquisa, a autora sugere possíveis características a serem incorporadas na elaboração do EPPLE em ambiente eletrônico, ainda em andamento na época. Entre essas características, Anchieta (2010) sugere que o EPPLE: i) siga o exemplo dos testes analisados nos aspectos espaciais e temporais, sendo estipulado o local onde a prova será realizada e o tempo limite para a realização de suas tarefas; ii) utilize diferentes tipos de tarefas que envolvam habilidades desenvolvidas por professores de LI em contextos de sala de aula; e iii) apresente insumos e questões contextualizadas, referentes ao contexto de sala de aula de LI, abordando temas específicos e fazendo uso de diferentes artefatos tecnológicos, além de permitir anotações durante o teste e oferecer simulados do exame no *site* oficial de divulgação. (ANCHIETA, 2010. p. 170-171).

Ambas as autoras, Salomão (2010) e Anchieta (2010), tomam como base teórica os trabalhos de Chapelle e Douglas (2006), Consolo (2004) e McNamara (2000). Ao discutirem sobre os efeitos positivos e negativos de um teste realizado em meio eletrônico, e as características peculiares a esses tipos de testes, Salomão destaca tanto os pontos positivos, quanto os negativos e/ou que merecem maior estudo, sem, no entanto, procurar dar solução para tais eventuais questionamentos, enquanto, Anchieta procura definir os pontos positivos que deveriam ser usados no EPPLE.

Outra pesquisa mais recente, ainda sobre o EPPLE, o artigo de Consolo e Silva (2014), apresenta as justificativas para a elaboração e a aplicação desse exame voltado, especificamente, para professores de LI no cenário brasileiro, com formação em Letras. Após relatar o trajeto histórico percorrido ao longo do processo de implementação do exame em maior escala, os autores esclarecem que existem dois formatos do teste, o presencial (papel e caneta) e a versão eletrônica, sendo esta última a mais utilizada para as pesquisas. Segundo os

autores, a decisão de se elaborar uma versão computadorizada do exame se deve à viabilidade de se aplicar um exame aos profissionais de diversas regiões do país de maneira não presencial e, principalmente, ao custo de deslocamento de examinadores ou de examinandos que poderia ser gerado. O exame é aplicado em algumas universidades brasileiras desde 2009, e os dados coletados nessas aplicações estavam, à época, sendo analisados, para aprimoramento do mesmo, por um grupo de pesquisa da universidade onde os autores atuam. A versão eletrônica foi aplicada pela primeira vez em 2011 e, de acordo com a página oficial do exame[11], a aplicação mais recente ocorreu em agosto de 2016, na Universidade de Brasília (UNB). O teste é dividido em duas partes: Oral e Escrita, e as competências linguística e pedagógica em LI do professor-examinando são avaliadas por meio de tarefas descritas pelos autores da seguinte forma:

> O teste escrito contém questões de compreensão de textos de interesse de professores de línguas, e questões nas quais o candidato deve realizar tarefas verossímeis à produção escrita de professores [...] O teste oral contém três partes. Na primeira, solicita-se que o examinando fale sobre si mesmo, na perspectiva de aluno e falante-usuário da língua estrangeira [...]. Na segunda parte do teste oral apresenta-se um breve excerto de vídeo, baseado no qual o candidato responde questões sobre seu tema e conteúdo. Na terceira parte do teste os candidatos devem demonstrar sua proficiência no uso [...] da linguagem específica para atuação pedagógica. Apresentam-se problemas de cunho realista enfrentados por alunos, sobre aspectos da língua estrangeira em questão, para os quais os examinandos devem oferecer explicações e soluções. A proficiência é avaliada por meio de descritores que constituem níveis representados em uma escala, com variação de A (faixa superior) a E (faixa inferior). Propõe-se o estabelecimento de escalas de caráter mais analítico para cada um dos dois testes e de uma escala de caráter holístico para classificação do desempenho global do candidato no EPPLE. (CONSOLO e SILVA, 2014. p. 13).

Em suas considerações finais, Consolo e Silva esperam que os resultados da pesquisa e das aplicações do exame sirvam como norte para o aprimoramento da formação de professores de LI nos cursos de Letras no Brasil e contribuam para as pesquisas na área de avaliação linguística.

Vejamos agora alguns trabalhos voltados ao estudo do Projeto Teletandem Brasil[12]. Neste projeto, o ensino e a aprendizagem da LE são realizados totalmente em ambiente *online*, fazendo uso simultâneo da produção e compreensão oral, além da escrita e de

[11] Informações disponíveis em <http://www.epplebrasil.org/index.php>. Acesso em: jan/2017.
[12] Para maiores informações sobre o Teletandem, vide a página oficial do projeto disponível em <http://www.teletandembrasil.org/>. Acesso em: jan/2017.

imagens, por meio da interação entre duplas de parceiros (alunos), um brasileiro aprendendo LE e um estrangeiro aprendendo português como LE.

Primeiramente, Furtoso (2011), em sua tese de doutorado, apresenta uma pesquisa sobre a avaliação da aprendizagem em contexto *online*, com ênfase no desempenho oral em Português para falantes de outras línguas (PFOL), pelo Projeto Teletandem Brasil que foca no uso da LE pela comunicação síncrona entre os pares e analisa a avaliação como componente da sessão do Teletandem. O objetivo da pesquisa foi analisar a avaliação como elemento integrador entre o ensino e a aprendizagem, e defender o uso de *podcastings*[13] como meio de realizar essa avaliação, servindo para aprimoramento do *feedback* e da autoavaliação, entre os pares. A autora procurou subsídios para definição de critérios para a avaliação do português falado em contexto *online* e a inclusão das TIC na avaliação da oralidade.

Outro trabalho sobre o Teletandem é o de Furtoso, Gomes e Consolo (2011), apresentado na VII Conferência Internacional de TIC na Educação, em Braga, Portugal, em que continuam a análise do uso de *podcasts* como estratégia e instrumento de avaliação das aprendizagens em LE. A partir dos resultados da revisão de literatura, de pesquisas anteriores dos próprios autores, sobre o uso de *podcasts* no contexto de aprendizagem em LE, eles apresentam uma proposta de inserção do *podcasting* como atividades complementares de comunicação assíncrona em um contexto *online* de aprendizagem de LE, no caso, o Teletandem. Os autores destacam que o projeto é um tandem à distância, que faz "uso do aspecto oral (ouvir e falar) e do aspecto escrito (escrever e ler) das línguas envolvidas por meio de áudio e videoconferências, utilizando textos falados e escritos e imagens de vídeo por meio de uma *Webcam* - em tempo real". Tudo isso é possível utilizando-se o *Skype*, que dispõe de recursos que permitem a comunicação em LE entre os aprendizes, de maneira síncrona. Desta forma, em suas considerações finais, os autores concordam que o uso de *podcastings* é de grande ajuda não apenas na aprendizagem, mas também como complemento na avaliação recíproca entre os pares e tutores-alunos, servindo, também, como *feedback* oral.

Finalmente, em um trabalho mais recente, Consolo e Furtoso (2015) apresentam um artigo sobre a avaliação do ensino e aprendizagem em contexto *online*. Com o objetivo de identificar as características da avaliação de proficiência oral do PLE, presentes na interação entre os parceiros no Projeto Teletandem Brasil, os autores sugerem modos de otimizar o

[13] ***Podcasts*** são programas de áudio ou vídeo, cuja principal característica é um formato de distribuição chamado *podcasting*. ***Podcasting*** é um meio de publicação de arquivos de mídia digital por meio de *feed RSS*, o que permite aos seus assinantes o acompanhamento ou download automático do conteúdo à medida que é atualizado. Retirado de: <http://cursodepodcast.com.br/o-que-e-podcast/>. Acesso em: junho/2017.

processo de avaliação durante o processo de aprendizagem. Além de considerar a avaliação como parte do processo de ensino e de aprendizagem, os autores a veem como elemento fundamental para o efeito retroativo das interações entre os pares. Em suas considerações finais, os autores destacam o mecanismo de *feedback* imediato da produção linguística entre os parceiros e a autoavaliação como partes necessárias ao sucesso do projeto, fazendo com que a conversa flua naturalmente, de modo contextualizado. No entanto, segundo os autores, alguns cuidados devem ser tomados, como: a preparação para a avaliação por parte dos parceiros e a existência contínua de reciprocidade entre os aprendizes ao se empenharem na construção da aprendizagem uns dos outros.

Prossigamos com a exposição das principais concepções de avaliação e as qualidades primordiais em testes de proficiência linguística de larga escala ou de alta relevância, presentes na próxima seção.

3.1 Critérios ou qualidades de testes de proficiência linguística

Levando em consideração a importância das concepções de testes de proficiência linguística, esta seção procura explicitar as concepções teóricas clássicas do campo da Linguística Aplicada à avaliação em LE.

Comecemos com Tim McNamara, pesquisador amplamente reconhecido na área de avaliação. McNamara (2000) aborda a importância dos testes de línguas logo no primeiro capítulo de seu livro intitulado *Language Testing*, afirmando que os "testes de línguas exercem um papel poderoso na vida de muitas pessoas, agindo como porta de entrada em momentos cruciais de transição na educação, no emprego, e na mudança de um país para o outro." (p. 4). [tradução nossa][14]. Por este motivo, torna-se imprescindível o estudo e o esforço para se avaliar de forma válida e justa o desempenho linguístico dos examinandos na língua alvo, pois é por meio dos resultados obtidos em um determinado teste que se tem acesso a amostras da aprendizagem dos examinandos, o que torna possível fazer julgamentos acerca de sua proficiência. Entretanto, julgamentos são inferências abstratas, por isso é preciso definir quem fará esses julgamentos e decidir o quão válido são essas amostras ou evidências de aprendizado. (McNAMARA, 2000).

[14] [...] language tests play a powerful role in many people´s lives, acting as gateways at important transitional moments in education, in employment, and in moving from one country to another.

Sendo assim, alguns critérios devem ser considerados ao discutirmos sobre testes de proficiência em LE, sejam estes realizados em formato impresso ou *online*, tais como: *o construto, a validade, a confiabilidade, a praticidade, a autenticidade* e *o efeito retroativo.* Além desses critérios, Chapelle e Douglas (2006) discutem outros aspectos que devem ser considerados juntamente com as mudanças trazidas pelas novas tecnologias e suas vantagens e desvantagens, ao se desenvolver um teste computadorizado, tais como: as circunstâncias físicas, temporais e de segurança dos testes, as instruções e procedimentos para as respostas do teste, o formato do teste, os participantes, o local, a interação entre o input e a resposta esperada, e, por fim, as características da avaliação – definição do construto, critério de correção e procedimento de pontuação[15].

A depender dos tópicos abordados no teste, McNamara (2000) define o domínio do construto como **operacional** (de desempenho), ou seja, aquele que se utiliza de tarefas práticas do mundo real para, por meio de situações comunicativas, avaliar a habilidade linguística do examinando; ou como **abstrato** (tradicional), quando se faz uso apenas de estruturas gramaticais ou itens de vocabulário em um nível apropriado, como questões de múltipla-escolha, avaliando separadamente as habilidades linguísticas do examinando. (McNAMARA, 2000).

Segundo Fulcher (2010. p. 96), definir *construto* é uma tarefa árdua devido ao seu caráter abstrato, e está intrinsicamente ligado ao propósito e ao contexto do teste. As habilidades ou conhecimentos do examinando servem como base para seu desempenho no teste, mas não é possível observar tais habilidades diretamente, por isso a necessidade de se definir o construto do teste. De acordo com McNamara (2000), o termo *construto* refere-se à filosofia do teste, ou melhor, à visão de língua e língua em uso contida no teste. E esta visão exerce um impacto significativo não apenas no *design* do teste, mas também na forma como a pontuação é dada e na interpretação do desempenho do examinando, ou seja, visões diferentes resultam em testes diferentes. Assim, "definir o construto de teste envolve entender de que consiste o conhecimento de língua, e como esse conhecimento é empregado no desempenho (língua em uso)" do examinando. [tradução nossa][16]. (McNAMARA, 2000. p. 13). Por exemplo, se os elaboradores ou desenvolvedores de testes possuem uma visão estruturalista da língua, ou entendem a língua como um sistema formal, o teste será projetado para avaliar um domínio específico das características da língua, uma por vez, de maneira isolada, como o

[15] Para mais informações acerca dessas características, limitações, vantagens e desvantagens do CALT em testes, ver o segundo capítulo de Chapelle e Douglas (2006) e o ANEXO VIII desta pesquisa.
[16] Defining the test construct involves being clear about what Knowledge of language consists of, and how that knowledge is deployed in actual performance (language use).

conhecimento gramatical, o léxico e os aspectos de pronúncia. McNamara define esse tipo de construto como abstrato. Nesse tipo de teste, as questões de múltipla-escolha são uma opção adequada. Entretanto, se as características de língua forem entendidas como um sistema de conhecimento integrado (pragmático), ou utilizadas para fins de comunicação, dentro de um contexto social de uso de língua, o teste a ser projetado, provavelmente, utilizará tarefas destinadas a medir o desempenho do examinando por entrevistas orais, *role plays*, simulação de tarefas cotidianas, redação de textos e questões envolvendo a compreensão de discursos orais e/ou escritos prolongados. No primeiro tipo de teste, a entrega de resultados e a pontuação são mais rápidas, objetivas e práticas, enquanto que no segundo tipo de teste a interpretação dos resultados é mais complexa e subjetiva, podendo gerar aumento de custos e tempo empregado, exigindo a contratação de avaliadores treinados.

No que diz respeito à *validade*, muitos autores concordam que o termo é a característica mais importante de um teste em LE e refere-se ao significado e à relevância que atribuímos ao mesmo. (BACHMAN e PALMER, 1996; MCNAMARA, 2000; FULCHER, 2010; BROWN, 2010). Messick (1989) define validade como "um julgamento integrado valorativo do grau em que as evidências empíricas e teóricas apoiam a adequação das inferências e ações baseadas na pontuação dos testes ou outros modos de avaliação." (p. 11). [tradução nossa][17]. McNamara (2000) e Brown (2010) destacam quatro tipos de evidências da *validade* de um teste, são elas: i) *validade do conteúdo* (o conteúdo do teste está de acordo com o que se pretende avaliar?), ii) *validade do critério ou método* (como o examinando se relaciona ou entende as tarefas do teste e como suas respostas serão avaliadas?), iii) *validade do construto* (quais habilidades básicas do examinando estão sendo captadas pelo teste? Elas estão de acordo com o construto do exame?) e, iv) *validade consequencial ou impacto* (quais as consequências ou impacto o teste pode causar na sociedade ou no examinando?).

Brown (2010), por sua vez, enumera algumas características peculiares da *validade*, afirmando que um teste válido: i) deve medir exatamente o que ele propõe medir; ii) não mede variáveis irrelevantes; iii) apoia-se o máximo possível em evidências empíricas (desempenho); iv) envolve o desempenho que exemplifica o critério do teste (objetivo); v) oferece informações úteis e significativas sobre a habilidade do examinando; vi) e, finalmente, é apoiado por um argumento ou fundamentação teórica. (*Ibid*. p. 30). Segundo Wielewicki (1997 *apud* ANCHIETA, 2010. p. 57), "considera-se que um teste apresenta

[17] [Validity is] an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or others modes of assessment.

validade de conteúdo se este constituir uma amostra representativa e relevante das habilidades e estruturas de língua que o teste busca especificamente avaliar". Bachman e Palmer (1996) utilizam-se apenas do termo *validade do construto* como uma das qualidades que devem ser consideradas no desenvolvimento de um teste de língua, e considera que o termo é "usado para se referir à medida que podemos usar para interpretar uma determinada pontuação no teste como um indicador da(s) habilidade(s), ou construto(s) os quais queremos medir." (p. 21). [tradução nossa].[18] Ou seja, se o teste é de leitura, suas tarefas ou questões devem medir a proficiência da leitura e o entendimento de texto demonstrado pelo examinando, e não sua fluência na língua falada, por exemplo.

Em se tratando do critério *confiabilidade*, Weir (1993) afirma que, para ser confiável, um teste precisa: i) apresentar consistência no formato, conteúdo das questões e duração do exame; ii) ter grau de familiaridade dos alunos com o formato do teste; iii) apresentar fatores favoráveis, como a boa administração do teste, as condições do local do teste (iluminação, assentos, isolamento acústico, dentre outros) e o modo como o professor ou responsável administra a aplicação; iv) e ter um bom corretor ou equipe de corretores. No caso de existirem diversos corretores, são necessários medidas ou critérios que garantam a consistência da correção, uma vez que, quanto maior a concordância entre os corretores, maior será a confiabilidade dos resultados. (WEIR, 1993 *apud* ANCHIETA, 2010. p. 54).

Brown (2010) destaca que um teste confiável tem que ser consistente e dependente, ou seja, mesmo que um teste seja aplicado duas vezes aos mesmos examinandos, em diferentes ocasiões, os resultados obtidos devem ser similares, trata-se, portanto, da consistência dos resultados. Bachman e Palmer (1996) reforçam o fato de que a confiabilidade é uma qualidade essencial nos resultados dos testes de línguas, no entanto, reconhecem que "não é possível eliminar inconsistências inteiramente", daí a necessidade de tentar minimizar os efeitos causados por determinada inconsistência que estão fora de nosso controle, por meio de cuidados específicos como, por exemplo, conhecer as características gerais dos examinandos e fazer a pilotagem do teste durante o processo de desenvolvimento e *design* do mesmo. (*Ibid*. p. 20).

A *praticidade*, por sua vez, refere-se à "maneira pela qual o teste será executado em uma determinada situação" (BACHMAN e PALMER, 1996. p. 39). Um teste pode ser chamado de *prático* se não gerar um alto custo e puder ser aplicado em um tempo limite estipulado, suficiente para os examinandos o concluírem, além de ser de fácil administração e

[18] The term *construct validity* is therefore used to refer to the extent to which we can interpret a given test score as an indicator of the ability (ies), or construct (s), we want to measure.

correção. A praticidade de um teste está relacionada aos recursos necessários no *design*, desenvolvimento, aplicação e correção do teste e os recursos realmente disponíveis para seu pleno desenvolvimento e execução. Se o teste exigir mais recursos que os disponíveis ou gerar altos custos, ele não é prático. Esses recursos podem ser humanos, materiais, relativos a aspectos de segurança, espaço e tempo. Essa é a única qualidade presente em todos os estágios durante o processo de desenvolvimento de um teste, podendo este ser redefinido a depender dos recursos necessários e disponíveis. Os outros critérios ou especificações de teste podem, portanto, ser revisados ou reconsiderados em um movimento cíclico, de acordo com a praticidade almejada.

Sobre a *autenticidade* de um teste, Brown (2010) afirma que uma determinada tarefa para avaliação é autêntica quando pode ser executada no mundo real, e, para que isso aconteça, a amostra de língua em uso desta tarefa deve ser a mais natural possível, contextualizada, significativa, relevante e com tópicos interessantes. Bachman e Palmer (2010) definem autenticidade como o "grau de correspondência das características de uma determinada tarefa de teste com as características de uma tarefa cotidiana da vida real, na língua-alvo." (p. 23). [tradução nossa].[19] A autenticidade, neste sentido, está intimamente ligada com a validade do construto do teste, visto que pode servir como referência no momento da interpretação dos resultados gerais do teste, no que diz respeito ao uso efetivo da língua-alvo pelo examinando, em situação de avaliação, em comparação com o uso real da língua. McNamara (2000) trata da autenticidade de resposta e acrescenta que é preciso não somente definir os tipos de textos que servirão de insumo para as tarefas do teste, mas também ponderar sobre como eles podem ser utilizados no teste e como os examinandos poderão interpretá-los. Segundo o autor, a autenticidade de um teste está controversamente ligada à praticidade, pois "quando uma avaliação se torna mais autêntica, ela também se torna mais cara, complexa e potencialmente pesada." (*Ibid*. p.29).

Finalmente, os exames de proficiência em LE podem exercer um *impacto* (positivo ou negativo) no ensino e na aprendizagem, como, por exemplo, no contexto de preparação dos examinandos e na formação e atuação de professores, o que se convencionou chamar de *efeito retroativo*. Bachman e Palmer (1996) vão além desse conceito e declaram que o efeito retroativo ou impacto dos testes é muito mais complexo, podendo refletir no comportamento dos indivíduos, nos elaboradores ou entidades responsáveis pelo teste, e nos professores, de maneiras e intensidades diferentes e em áreas específicas, tais como: "conteúdo e metodologia

[19] [Authenticity is] the degree of correspondence of the characteristics of a given language test task to the features of a TLU task.

do ensino e a maneira de se avaliar o aprendizado adquirido.” (p. 31). McNamara (2000, p. 74), por sua vez, destaca que as autoridades responsáveis pelos testes de larga escala podem, muitas vezes, reavaliar seus testes em uma tentativa de gerar reforma no currículo, o que pode ter um efeito positivo. No entanto, esse efeito ou impacto é, geralmente, algo imprevisível, dependente das tradições de ensino, das condições e motivação dos aprendizes, etc., e isso poderá ser verificado somente após a aplicação do teste, baseado nas informações anteriores à reforma. Neste sentido, Agossa (2017) amplia a discussão sobre efeito retroativo e impacto, diferenciando-as uma da outra, e situa essa discussão no espaço da validade consequencial. O autor concorda com Bachman e Palmer (1996) e adere ao conceito de que “o efeito retroativo limita-se àquilo que acontece na sala de aula no que diz respeito à forma como os professores ensinam e como os aprendizes aprendem, e que impacto representa aquilo que pode acontecer nas próprias pessoas, dentro ou fora da sala de aula”, no que se refere às consequências do teste em suas vidas, por exemplo, no sistema educacional e na sociedade, em geral. (AGOSSA, 2017. p. 19). Desta forma, Agossa (2017) concorda com os estudos que consideram a validade consequencial como um conceito dependente do impacto ou das consequências do teste, pois a atenção dada a tais consequências promovem a validade consequencial do teste. Ou seja, segundo o autor, os testes podem ter consequências nos indivíduos e na sociedade, de acordo com as inferências sobre o uso dos resultados, o construto, as implicações de valor, a utilidade posterior e as consequências sociais do teste. Sendo assim, a validação de um teste deve ser realizada antes e depois de sua aplicação, prevendo tanto as consequências desejáveis quanto as indesejáveis.

Após discorrer sobre esses conceitos, entendidos por estudiosos da área como qualidades ou critérios primordiais em testes de proficiência linguística, independente de seu formato ou versão, apresentamos, a seguir, um breve histórico da evolução dos testes de proficiência até a utilização das novas tecnologias, como forma de introduzirmos o funcionamento e a utilização das NTIC em testes de avaliação linguística.

3.2 Breve histórico dos testes de proficiência

Nos Estados Unidos, o início do uso de testes de língua aconteceu nos anos finais do século XIX. Em 1875, Harvard College começou a aplicar os primeiros testes de francês e em 1896 a *Modern Language Association* (MLA) criou um comitê especial para estudar as

condições do ensino das línguas modernas e, consequentemente, mais tarde, dos testes de proficiência. (BARNWELL, 1996. p. 2). Nessa época, os testes eram baseados nas traduções de trechos de textos, com foco na gramática e na habilidade de verbalizar as suas regras e exceções, em uma visão formal da língua estrangeira, nos três tipos diferentes de exames: elementar, intermediário e avançado. Testes de leitura e escrita começaram a surgir em 1914 nos EUA, sendo que, em 1915 a *State University of New York* criou o que podemos considerar o primeiro teste oral, aplicado aos futuros professores de língua do corpo docente daquela universidade. No entanto, estes testes avaliavam apenas os conhecimentos de fonética e pronúncia de palavras avulsas escolhidas em determinados textos, e, além disso, o termo "teste oral" se referia ao que hoje poderíamos chamar de testes de "escuta ou audição". (*Ibid*. p. 16-19). Somente nos anos 70 as noções de competência comunicativa ganharam espaço para estudo, alavancado pelo prestígio da sociolinguística e de outras disciplinas que relacionam língua e sociedade, dando lugar a testagem de habilidades integradas, sem, no entanto, abandonar a herança dos pontos discretos[20].

No que se refere às entrevistas em testes orais de proficiência em LE, podemos dizer que essa prática também é antiga. As entrevistas tiveram origem na necessidade do Governo norte-americano preparar e avaliar pessoas para lidar com a comunicação oral em línguas consideradas estratégicas durante a Guerra Fria. (SALOMÃO, 2010). No entanto, segundo Fulcher (2003 *apud* SALOMÃO, 2010. p. 326), naquela época, o seu uso inicial pretendia "avaliar apenas questões de pronúncia e acuidade gramatical, por meio de tarefas de leitura em voz alta". Com o passar do tempo, os testes passaram a utilizar estilos menos estruturados e mais conversacionais, conhecido internacionalmente como O*ral Proficiency Interview* (OPI). (SALOMÃO, 2010).

Sendo assim, podemos, resumidamente, afirmar que o desenvolvimento e uso de testes de língua estrangeira seguiu o modelo de aquisição de línguas ou abordagem de ensino dominante em cada época. (BROWN, 2010. p. 7 e 8). Passando por vários estágios: o *pré-científico* (até a década de 1950) com foco na forma e tradução em testes orais; o *psicométrico-estruturalista* (entre as décadas de 1950 e 60) que se caracterizou pelo "uso de elementos isolados que fragmentavam a língua em categorias fonológicas, lexicais e

[20] Pontos discretos (*discrete points testing*) é a prática de se testar as quatro habilidades (leitura, escrita, fala e escuta) por meio de pequenos itens ou partes, como a estrutura gramatical, o conhecimento léxico, morfológico, sintático e discursivo, de maneira isolada, totalmente separadas umas das outras. Para testar essas habilidades de forma individual, as questões de múltipla-escolha ainda são as mais adequadas. Para maiores esclarecimentos sobre o assunto, ver os trabalhos de McNamara (2000, p. 13-14) e Brown (2010, p. 8).

sintáticas, mensurando o conhecimento de tais categorias sem levar em conta as correlações entre essas categorias no discurso" (MARTINS, 2005 *apud* SALOMÃO. 2010. p. 329); e o *sociolinguístico integrativo*, "no qual a língua passou a ser vista como um código inserido em determinado contexto, usado em situações reais de comunicação." (SALOMÃO, 2010. p. 330). E é dentro dessa perspectiva que alguns testes tendem a seguir, com a elaboração baseada em tarefas e maior ênfase no desempenho do examinando, substituindo os testes de múltipla-escolha ou por testes de *cloze,* com respostas curtas ou abertas, como acontece no TOEIC e TOEFL, ou baseado em tarefas comunicativas, simulando situações reais do cotidiano e, portanto, cedendo espaço para correções mais complexas e subjetivas, como acontece no Celpe-Bras e no Celu, por exemplo.

No que se refere a testes eletrônicos, o uso desses testes ocorre em meados dos anos sessenta, sendo amplamente disponível para uso após o advento do computador pessoal, no final dos anos setenta e início dos anos oitenta, nos Estados Unidos. O trabalho de Godwin-Jones (2001) apresenta um histórico técnico sobre as primeiras tecnologias de teste de línguas e suas ferramentas de autoria (ou criação) e os aplicativos de *Internet*, utilizados e desenvolvidos até o ano de 2001, destinados à avaliação. De acordo com esse histórico, o autor afirma que a criação do CD-ROM e a chegada do *World Wide Web* trouxeram as vantagens do armazenamento e da recuperação de pontuação baseada no servidor. No entanto, inicialmente, a experiência do usuário com testes baseados na *Web* não era diferente das versões papel e caneta, com relativamente pouca interatividade ou *feedback* para o usuário. Com a chegada e o aprimoramento do *JavaScript*[21] a partir de 1995, os testes de língua começaram a expandir as opções de formato, adicionando a resposta curta e a correspondência ou ordenação de sentenças, por exemplo. Godwin-Jones (2001) já previa que o desenvolvimento acelerado de tecnologias de reconhecimento de fala forneceria novas opções para testes de proficiência computadorizados que envolvessem a escuta e a fala. Além de apostar em testes personalizados, com opções para serem adaptados a cada aluno e as suas necessidades específicas.

Segundo Dunkel (1999), a quantidade de CBT e CAT desenvolvida por organizações acadêmicas continuava aumentando na época em que seu trabalho foi escrito. Entretanto,

[21] Segundo a Wikipédia, ***JavaScript*** é uma linguagem de programação interpretada. Foi originalmente implementada como parte dos navegadores *web* para que *scripts* pudessem ser executados do lado do cliente e interagissem com o usuário sem a necessidade desse *script* passar pelo servidor, controlando o navegador, realizando comunicação assíncrona e alterando o conteúdo do documento exibido. É a principal linguagem para programação *client-side* em navegadores web. Disponível em https://pt.wikipedia.org/wiki/JavaScript. Acesso em: jan/2017.

apesar de muitas universidades e, praticamente, todas as principais organizações de testes linguísticos estarem envolvidas no desenvolvimento de um CBT ou CAT, muita pesquisa ainda é necessária para que os desenvolvedores e usuários aproveitem o máximo do potencial do CAT a partir do funcionamento de seu modelo psicométrico, em termos da Teoria de Resposta ao Item (TRI), uni ou multidimensional[22], e, acima de tudo, as pesquisas podem auxiliar na compreensão do que é necessário para implementar um CAT confiável. (DUNKEL, 1999. p.89).

Finalmente, podemos citar algumas instituições e seus investimentos em testes computadorizados, tais como: i) a *Brigham Young University*, pioneira no desenvolvimento de instrumentos para CAT franceses, alemães e espanhóis, utilizados para dar acesso a universidades; ii) o ETS, com o desenvolvimento do TOEIC, do TOEFL-CBT, em 1998 e, posteriormente, o TOEFL-iBT, em 2000, nos Estados Unidos e em vários países ao redor do mundo; iii) a *University of Cambridge Local Examinations Syndicate* (UCLES), lançando alguns CAT em várias línguas e para vários fins, como por exemplo, o *CommuniCAT*, destinado a programas de idiomas em ambientes acadêmicos, e o *Business Language Testing Service* (BULATS), com versões em inglês, francês, alemão e espanhol, destinado ao setor corporativo; e, por fim, iv) o projeto DIALANG, patrocinado pelo Conselho da Europa, que prevê uma avaliação diagnóstica em catorze línguas. (CHALHOUB-DEVILLE, 2001).

Na seção seguinte, abordamos a revisão de literatura internacional na área dos testes em meio eletrônico e virtual.

3.3 O uso das tecnologias em testes de LE: desenvolvimento e progresso

[22] Segundo Piton-Gonçalves e Aluísio (2015), um item pode conter opções de resposta dicotômica (em que apenas uma opção de resposta é a correta e as demais, incorretas), politômicas (por exemplo, quando são consideradas mais de duas opções de resposta como corretas) e dissertativas. Neste caso, itens que apresentam resposta dissertativa devem ser corrigidos de acordo com alguma graduação ou critérios, o que implica em uma resposta dicotômica ou politômica. Do ponto de vista do número de habilidades consideradas, a TRI pode ser dividida em duas classes:

1. Teoria de Resposta ao Item Unidimensional (ou simplesmente TRI): define uma única habilidade associada ao examinado e, consequentemente, os modelos e os métodos envolvidos também são unidimensionais (ou univariados). Do ponto de vista educacional, a habilidade **x** pode representar, por exemplo, a proficiência de Escrita em LE.
2. Teoria de Resposta ao Item Multidimensional (MIRT): define um vetor **x** de habilidades associado ao examinado. Do ponto de vista educacional, pode-se, por exemplo, interpretar um vetor habilidade **x** tridimensional, como a proficiência em escrita, leitura e escuta no espaço conhecimento da LE.

(PITON-GONÇALVES, 2012 *apud* PITON-GONÇALVES e ALUÍSIO, 2015. p. 395-396).

.

As novas tecnologias e os testes de proficiência linguística têm sido temas de debates em congressos e despertado o interesse e discussão de pesquisadores da área de avaliação ao longo do século XX e início do século XXI. Pesquisadores afirmam que os computadores foram, primeiramente, utilizados para atribuir a pontuação aos acertos dos examinandos e para a emissão do relatório de resultados em testes de língua. Assim, testes computadorizados começaram a ser aplicados somente a partir dos anos 80 e, finalmente, no início dos anos 2000, quando a nova tecnologia tornou-se mais generalizada na avaliação educacional, e devido à maior disponibilidade e baixo custo dos computadores, esses testes foram aperfeiçoados, passando a ter uma maior relevância na área de avaliação. (DUNKEL, 1999; CHALHOUB-DEVILLE 2001; GODWIN-JONES, 2001; BENNETT, 2001; WALCZAK, 2015; CHAPELLE e VOSS, 2016).

Vejamos alguns trabalhos voltados ao tema, no âmbito internacional, que datam entre os anos de 2001 até 2016.

Em 2001, Chalhoub-Deville publicou um artigo para a revista *Language Learning & Technology*, em que aponta a tecnologia como grande aliada da educação e, consequentemente, da avaliação de proficiência linguística. De acordo com a autora, a Conferência Anual sobre Testes Linguísticos nos Estados Unidos da América, *Language Testing Research Colloquium* (LTRC), nas edições de 1985 a 2001, destacou o crescente uso da tecnologia na área de proficiência linguística, bem como no desenvolvimento dos testes em larga escala, as medições gerais e os contextos educacionais. Temas referentes à aplicação e aos modelos de banco de itens, seleção de itens e testes adaptativos por computador foram os mais discutidos nessas conferências, segundo a autora. Naquele ano de 2001, Chalhoub-Deville lamentava o pouco avanço nos estudos sobre os CAT e os testes computadorizados alertando que a área de avaliação linguística não estava, à época, acompanhando o ritmo de desenvolvimento tecnológico mundial. Em suas palavras:

> Talvez o principal motivo pelo qual o campo de L2 está atrasado nesta área é porque há muito promoveu a avaliação baseada no desempenho, uma forma de avaliação que não se presta tão facilmente à administração computadorizada quanto os formatos de teste mais tradicionais. [...] Assim, enquanto os pesquisadores de medidas gerais, especialmente aqueles que trabalham com CAT, têm se preocupado mais com os tipos de item de resposta selecionada, o campo de L2 continuou a promover a avaliação baseada no desempenho. Ainda hoje [2001], a avaliação baseada no desempenho em testes computadorizados continua a ser um desafio. (CHALHOUB-DEVILLE, 2001. p.95). [tradução nossa][23].

[23] Perhaps the main reason the L2 field has lagged behind in this area is because it has long promoted performance-based assessment, a form of assessment that does not lend itself as readily to computerized

Outros autores como Stansfield (1986) e Dunkel (1991, 1997, 1999) são retratados pela autora como precursores da discussão sobre o uso de computadores nos mais variados tipos de testes e da crescente evolução dos CAT. Na época em que foi escrito, o trabalho de Chalhoub-Deville destacava as principais vantagens do uso da tecnologia nos testes linguísticos, tais como: i) ajuda a tornar a avaliação mais eficiente, flexível e individualizada e mais precisa, no caso dos CAT; ii) monitora o desempenho do aluno, por meio de *feedback* imediato; iii) oferece segurança e aprimoramento de novos tipos de itens e/ou tarefas; iv) permite a adaptação de item difícil, adequando-o de acordo com a habilidade do examinado; e v) facilita a entrega e administração de testes. No entanto, a autora reconhecia, na época, a necessidade de mais pesquisas sobre como a tecnologia poderia gerar mudanças e inovações no empreendimento dos testes computadorizados, pois "o construto de um teste de proficiência linguística é multidimensional e envolve uma variedade de componentes e processos que interagem entre si". (*Ibid*, p. 96). Na visão da autora, a demanda de testes computadorizados, que avaliam e diagnosticam o uso da língua e o desenvolvimento de habilidades de uma pessoa por meio de computadores, aumentaria com o passar dos anos, destinados a suprir as necessidades dos alunos em ambientes não convencionais. Portanto, desenvolvedores e elaboradores de testes devem fazer uso da tecnologia para uma transformação radical das práticas de avaliação, por meio de ações inovadoras.

Prosseguindo em nosso estudo, encontramos Bennett (2001) e seu trabalho intitulado '*How the Internet Will Help Large-Scale Assessment Reinvent Itself',* em que analisa como o avanço da tecnologia, especialmente o uso da *Internet*, a medição e a ciência cognitiva podem ajudar a mudar a concepção dos testes de larga escala. Segundo o autor, as mudanças necessárias devem focar em alguns problemas a serem resolvidos na "base cognitivo-científica desatualizada para o *design* do teste; no desajuste com o currículo; no diferente desempenho dos grupos populacionais; na falta de informação para ajudar os indivíduos a melhorar; e na ineficiência de alguns testes". [tradução nossa][24]. (*Ibid*. p.1). Para ele "as contribuições da ciência cognitiva e de medição são, em muitos aspectos, mais fundamentais do que as das novas tecnologias, embora a nova tecnologia permeie nossa sociedade".

---

administration as do more traditional test formats. […] So, whereas general measurement researchers, especially those working with CAT, have concerned themselves more with selected-response item types, the L2 field has continued to promote performance-based assessment. Even today, CBT performance based assessment continues to be a challenge.

[24] […] an outmoded cognitive-scientific basis for test design; a mismatch with curriculum; the differential performance of population groups; a lack of information to help individuals improve; and inefficiency.

[tradução nossa][25]. (*Ibid.* p.14). Desta forma, Bennett (2001) afirma que a revolução tecnológica, presente no comércio, na educação e nas interações sociais, deve ocorrer de maneira sistemática também na área de avaliação de larga escala. Na visão do autor, estes avanços tecnológicos giram em torno, principalmente, da *Internet*, dadas as suas características de poder ser interativa, de banda larga, oferecendo benefícios para a área de avaliação, tais como, o potencial para: i) fornecer, eficientemente, um conteúdo altamente envolvente para quase qualquer *desktop*; ii) obter *feedback* imediato; iii) processar dados; e iv) disponibilizar informações em qualquer lugar do mundo, a qualquer hora do dia ou da noite, alcançando, portanto, um maior número de pessoas e uma maior interatividade entre elas. (BENNETT, 2001. p.4).

Em um estudo mais recente, um artigo de revisão, Chapelle e Voss (2016) apresentam uma visão geral das maneiras como o processo avaliativo de competência em língua foi impactado pelo desenvolvimento tecnológico ao longo dos anos, nas últimas duas décadas. No artigo, os autores analisam o tema da tecnologia a favor da eficiência e da tecnologia a favor da inovação nos testes de língua, em 25 documentos criteriosamente selecionados dentre os 198 encontrados, entre artigos de revisão, artigos de pesquisa e revisões e comentários de livros publicados na revista eletrônica *Language Learning & Technology* (LLT), desde 1997, ano de seu lançamento, até o ano de 2015.

Em se tratando dos documentos voltados à tecnologia a favor da eficiência dos testes, os autores destacaram o crescimento significativo do uso da tecnologia em testes de larga escala desde os anos 60 e a euforia por parte dos elaboradores e desenvolvedores de testes, acerca dos avanços psicométricos que deram origem a testes adaptativos por computador naquela época. Muitos pesquisadores enxergavam a possibilidade de um teste mais eficiente, rápido e menos oneroso que os testes de papel e caneta. Além da possibilidade da presença de multimídia, como imagens, vídeos e áudios, podendo representar situações autênticas de uso da língua e, ainda, possibilitar uma aceleração do processo de correção, *feedback* e divulgação do resultado em testes de autocorreção, em inglês *Automated Writing Evaluation* (AWE). Essas características representam, de acordo com os autores, um benefício financeiro significativo para as organizações responsáveis pelos testes de larga escala que se encontram preparadas e com a infraestrutura de *hardware* e *software* necessária no desenvolvimento desses tipos de testes. Segundo Chapelle e Voss (2016):

[25] Whereas the contributions of cognitive and measurement science are in many ways more fundamental than those of new technology, it is new technology that is pervading our society.

> Cronologicamente, o primeiro grande desenvolvimento foi o teste adaptativo por computador, que forneceu um meio de adaptar um teste para cada aluno individualmente durante a realização do mesmo. A eficiência também foi evidente a partir dos primeiros usos da avaliação automatizada da escrita (AWE), cuja promessa foi, e em parte permanece sendo, o uso da tecnologia para realizar, ao menos, uma parte do trabalho intensivo de avaliar os ensaios dos alunos. [tradução nossa][26]. (p. 117).

Desta forma, a maioria dos documentos analisados traziam pesquisas sobre os CAT e os CBT e a comparação entre a interpretação dos resultados ou pontuação dos CBT e os testes impressos ou fornecidos por outros mecanismos. Dentro desse último tema, vários trabalhos foram desenvolvidos referentes à 'comparabilidade do conteúdo da tarefa e as condições de administração, o critério psicométrico de estabilidade das estimativas do parâmetro do item, e as características do examinando e das condições de teste', entre outros. No entanto, os autores reconhecem que os estudos comparativos entre o "antigo e o novo" podem ser indícios de perda de oportunidades de inovação na área.

Outra questão destacada no mesmo trabalho refere-se à leitura de textos na tela do computador, em que se verificou não ser mais motivo de estranheza entre os examinandos a partir de 2001, além disso, segundo os autores, existem mais pessoas familiarizadas com a leitura em tela do que em papel. Outros estudos, por sua vez, avaliavam as atitudes e percepções dos examinandos, após completarem uma avaliação, depois de feitas algumas adaptações na versão computadorizada como, por exemplo, permitir que eles desempenhassem um papel na seleção de tarefas, a fim de ajudá-los a não se sentirem desconfortáveis usando a tecnologia.

No que se refere aos documentos voltados para a tecnologia a favor da inovação, Chapelle e Voss (2016) compreendem que a inovação em testes de proficiência linguística 'vai além do objetivo de fazer testes mais eficientes para expandir os usos da avaliação e suas utilidades'. (*Ibid*. p. 120). Assim, a introdução de testes na rede *Web* foi tema constante da LLT (2001), nos documentos analisados pelos autores. Nesses trabalhos, a tecnologia aparece como principal recurso para o aprimoramento dos métodos e uso dos testes:

[26] Chronologically, the first big development was computer-adaptive testing, which provided a means of tailoring a test to each student interactively during test taking. Efficiency was also evident from the first uses of automated writing evaluation (AWE), whose promise was, and in part remains, the use of technology to perform at least some of the time intensive work of evaluating students' essays.

> No decurso dos últimos anos, foram levantadas na LLT várias questões relacionadas à inovação, incluindo a utilização de avaliações para aumentar e melhorar as oportunidades de aprendizagem, a importância de repensar as construções linguísticas a serrem mensuradas, a necessidade de investigar os impactos potenciais das inovações e o envolvimento da profissão em novas avaliações linguísticas, por meio do acesso a ferramentas de autoria para a avaliação. [tradução nossa][27]. (CHAPELLE e VOSS, 2016. p. 120-121).

Ainda de acordo com Chapelle e Voss (2016), a LLT publicou vários artigos entre os anos de 2007 a 2013 cujos temas lidam com questões de definição do construto dos testes e das maneiras de interpretar o desempenho do examinando, a elaboração de novos tipos de itens, tarefas e *designs* dos testes mediados por computador e, principalmente, os baseados na *Internet*. Questões como o tipo de itens a serem criados, em que base os itens devem ser projetados e como devem ser avaliados com o uso da tecnologia foram amplamente discutidas. Outro tema relevante tratado nas publicações analisadas por Chapelle e Voss diz respeito ao *design* de tarefas e ao uso de multimídia em um teste de compreensão auditiva. Além desses temas, foram também considerados os impactos da inovação pela tecnologia na validação do resultado dos testes, o comportamento para a preparação dos testes e o quanto os examinandos apreciam realizar o exame, tema ainda carente de pesquisas nos dias atuais.

Além de todas essas questões levantadas, os autores reconhecem a necessidade da criação e do desenvolvimento de ferramentas de *software* que permitam "aos linguistas participarem sem que eles se tornem programadores de computador". (*Ibid*. p. 124). Os autores não explicam como isso pode ser feito. Entretanto, apesar de ser pouco discutido na área, esse é um tema fundamental e digno de mais estudo, pois, com a implementação das NTIC, o envolvimento e o trabalho conjunto de profissionais da linguística aplicada e da avaliação, assim como a colaboração de programadores, técnicos de informática e manutenção, é imprescindível. Aos linguistas cabem as decisões sobre a elaboração e a escolha das tarefas ou tipos de questões a serem utilizadas, como serão apresentadas e executadas pelo examinando, o conteúdo, o modo de correção e pontuação. Aos técnicos de informática e programadores cabe entender o propósito do teste e decidir quais as ferramentas tecnológicas que melhor atendem a demanda do teste, qual tipo de teste atende a esse propósito, adaptativo ou linear, e pensar como e quais dispositivos ou ferramentas podem ser

[27] A number of issues in innovation have been raised in *LLT* over the past years including the use of assessment to increase and improve opportunities for learning, the importance of rethinking the language constructs to be measured, the need to investigate potential impacts of innovations, and the engagement of the profession in new language assessments through access to authoring tools for assessment.

utilizadas, por exemplo. Para que isso aconteça, é necessário o diálogo entre ambas as partes, linguistas e programadores, em uma troca de conhecimentos, para que ambos profissionais consigam reconhecer as limitações, as possibilidades e os recursos necessários e disponíveis de cada área. Na opinião de Chapelle e Voss (2016), o mercado possui ferramentas de autoria ou "uma gama de *softwares* projetados para diferentes propósitos, que podem, inclusive, ser utilizados no desenvolvimento de avaliações na área de línguas". (*Ibid*. p. 124). Muitos desses *softwares* são projetados primeiramente para o ensino e adaptados à avaliação posteriormente. No entanto, a tendência é que os *softwares* utilizados em testes sejam criados especificamente para a avaliação e desenvolvidos tendo em vista a praticidade e a facilidade de uso e manuseio, tanto por parte do examinando, quanto por parte de quem aplica e avalia o teste, sem comprometer o seu construto e a sua validade.

Por fim, Chapelle e Voss (2016) afirmam que a inovação deve acontecer tanto no âmbito das instituições organizadoras e/ou partes interessadas, como no conteúdo da avaliação, pois ambos estão interligados. Usar a nova tecnologia para fazer a avaliação mais rápida e mais barata pode liberar os recursos necessários para desenvolver uma avaliação melhor, mais eficiente e, por que não, também inovadora. (BENNETT, 2001).

...

Neste capítulo, apresentamos uma revisão de literatura da área de testes de proficiência linguística e as NTIC utilizadas nesses testes. Primeiramente, buscamos expor trabalhos de âmbito nacional que tratam sobre avaliações em meio computadorizado para diferentes fins. Em seguida, discutimos as concepções de avaliação, as qualidades ou critérios essenciais, para o pleno desenvolvimento de testes e, então, partimos para um breve histórico da evolução dos testes de proficiência e a utilização das novas tecnologias ao longo desse percurso. Por fim, abordamos a revisão de literatura internacional na área dos testes em meio eletrônico e virtual.

Destacamos que, ao observar a revisão da literatura cinzenta, não encontramos trabalhos que abordassem o objeto de estudo desta pesquisa, que consiste em investigar o uso das NTIC como ferramenta para integrar e aprimorar a realização de tarefas destinadas à avaliação de proficiência em Português como Língua Estrangeira, no Exame Celpe-Bras. Muito tem sido discutido sobre as ferramentas tecnológicas e suas contribuições para testes eletrônicos ou *online*, especialmente em testes como o EPPLE, Teletandem, EPO, em âmbito nacional, no entanto, estudos voltados ao uso dessas ferramentas no Celpe-Bras e a

possibilidade de uma nova versão computadorizada do Exame não foram ainda discutidas sistematicamente. Este fato torna este estudo justificável e válido dentro da área de avaliação de testes de proficiência linguística, podendo este estudo proporcionar a abertura para novas pesquisas e questionamentos que possam levar ao desenvolvimento e elaboração de pesquisas experimentais. Além disso, o desenvolvimento de uma versão *online* ou computadorizada do Exame Celpe-Bras colabora para a ampliação da internacionalização da LP, ou seja, sua difusão e promoção.

Apesar desta pesquisa não ser embasada por uma teoria ou teorias específicas, buscamos explicitar aqui os conceitos sobre avaliação mediada pela tecnologia, com o intuito de oferecer subsídios para as contribuições deste estudo. Desta forma, percebemos que os trabalhos revisados apresentam uma visão promissora acerca desses conceitos ou concepções teóricas de teste. Consideramos promissora e temos uma visão otimista da área, pois, apesar dos poucos estudos em âmbito nacional, os testes de proficiência computadorizados têm despertado um crescente interesse, principalmente, por oferecerem a praticidade tão almejada pelos desenvolvedores de testes. O uso de ferramentas tecnológicas, como *podcasting*, *chats*, *webcam*, vídeoconferência, multimídia audiovisual, programas de comunicação síncrona, são recursos cada vez mais experimentados na elaboração de testes, em busca do aprimoramento da avaliação à distância e presencial, em exames diagnósticos e de larga escala, como o EPO e o EPPLE. Projetos como o Teletandem têm fornecido bons resultados e mostram-se inovadores ao utilizar a tecnologia a favor da aprendizagem e da avaliação, concomitantemente, para autoavaliação e *feedback* contínuo entre os pares, por meio dessas ferramentas.

Ao discorrermos sobre a história dos testes e a introdução da tecnologia, percebemos a busca constante pelo teste ideal, ou seja, a busca por testes mais justos, válidos e práticos. A evolução histórica das teorias linguísticas e dos testes mostra que o critério de praticidade torna-se uma constante nas pesquisas e, nesse sentido, os recursos tecnológicos oferecem a oportunidade de desenvolvimento de testes práticos, especialmente na aplicação, na correção e na entrega de resultados. Desta forma, as NTIC surgem com a promessa de oportunidade de inovação na área de avaliação, em busca de um teste mais eficiente. No início, os testes computadorizados pouco se diferenciavam dos impressos, seguindo modelos psicométricos e avaliando habilidades não-integradas, com questões de múltipla-escolha, muitas vezes, descontextualizadas. Assim, a autenticidade das tarefas e conteúdos é amplamente discutida apesar da tensão entre essa qualidade e a praticidade. O surgimento de multimídias

favoreceram os testes de escuta e o teste adaptativo, recebido com euforia pelos estudiosos da área no fim do século XX, sendo, ainda, uma opção em testes de proficiência linguística. Com o advento da *Internet*, os testes *online* passaram a lidar com questões de segurança e autenticidade do usuário, tema ainda oportuno na área. Discussões sobre a elaboração de tarefas e a percepção do examinando, assim como o letramento digital[28] ou a diferença de leitura de um texto na tela e impresso, são temas menos discutidos, por causa do aumento do uso do computador e a disseminação do ensino em ambiente virtual.

Finalmente, os trabalhos internacionais revisados apresentam a evolução dos testes e as tecnologias utilizadas durante esse processo, pontuando os cuidados na elaboração e os aspectos positivos e negativos de testes eletrônicos, reafirmando, ainda, a necessidade de estudos sistemáticos e criteriosos a respeito, principalmente, das falas mais longas em testes orais mediados por computador, que percebemos ser um tema pouco explorado ao longo da evolução tecnológica dos testes, se apresentando, ainda, como o maior desafio da área. Talvez isso aconteça porque avaliar automaticamente o desempenho oral por meio do reconhecimento computadorizado da fala é tarefa complexa. A natureza e a complexidade da fala humana ainda não permitem que todos os níveis de fala do examinando sejam reconhecidos, apenas respostas restritas e breves, com algum progresso recente na geração de pontuações para amostras de fala mais longas (ZECHNER, HIGGINS, XI, e WILLIAMSON, 2009 *apud* CHAPELLE e VOSS, 2016. p. 125). Por isso, na maioria das vezes, prioriza-se a elaboração de testes eletrônicos de Leitura e Escuta, deixando a parte oral para testes humanizados por meio de entrevistas, apesar da menor praticidade e maior custo com avaliadores, por exemplo. Essa é uma área em progresso contínuo e avanços tecnológicos futuros podem trazer as inovações necessárias para um reconhecimento automatizado da fala mais acurado, entretanto, o quadro da evolução tecnológica para testes de proficiência oral tornam os sistemas síncronos de comunicação uma alternativa para suprir essa deficiência na área que ainda necessita de pesquisas que viabilizem seu uso e praticidade em quaisquer circunstâncias ou locais. Voltamos a essa discussão nos capítulos seguintes deste estudo.

---

[28] Entendemos o termo letramento digital e multimodal, "da perspectiva da prática social, [...] como o que as pessoas fazem com a língua por meio de da tecnologia e não o que precisam saber sobre o idioma ou as estratégias necessárias para usar a língua por meio da tecnologia". [tradução nossa][28]. (Chapelle e Douglas, 2006. pp. 1195-1196).
From the perspective of social practice, electronic literacy and multimodal literacy, for example, are seen as what people do with language guage through technology rather than what they need to know about language guage and the strategies they need to use language through technology.[original].

No próximo capítulo, prosseguimos com a descrição das tecnologias e dos tipos de testes em meio eletrônico, além de alguns testes computadorizados e/ou *online* que utilizam as NTIC, suas principais características e abordagens teóricas.

CAPÍTULO 4

DESCRIÇÃO DAS NTIC & CBT

Technology applied to the service of understanding the learning we want will help us fix the presently unfixable - the deep validity problem at the heart of our testing system.

Chapelle & Douglas

Neste capítulo, buscamos explicitar os principais tipos de testes computadorizados existentes no mercado e suas características. Desta forma, os CBT e os CAT são aqui descritos da mesma forma que algumas ferramentas utilizadas na elaboração de testes diagnósticos ou de nivelamento por universidades norte-americanas e europeias, e alguns testes *online* de alto impacto social, em inglês *high-staketests*, como o TOEIC, o TOEFL, o PTE *Academic* e os testes do ACTFL desenvolvidos por organizações especializadas em avaliação em LE. Os dois primeiros possuem formatos computadorizados e impressos e os dois últimos são realizados totalmente em versão *online*. São descritos, também, o Projeto DIALANG, um tipo de CAT diagnóstico, multi-línguas e virtual. Após as descrições, apresentamos as considerações parciais referentes ao tema de pesquisa.

4.1 Testes, testar, testagem: os artefatos tecnológicos

Pensar na construção de um teste em LE significa considerar o seu propósito e definir as questões práticas como: o custo, o pessoal qualificado e/ou disponível e a tecnologia a ser utilizada. Segundo Chapelle e Douglas (2006) "o que temos hoje em dia, em termos de tecnologia desenvolvida para a área de avaliação, são: sistemas modestos, planos de sistemas ideais e trabalhos em progresso". Estes autores destacam que, à época, ainda eram desconhecidos testes de proficiência oral, realizados por meio de recursos de áudio e voz ou por dispositivos de comunicação síncrona e, ao mesmo tempo, oferecidos pela *Internet*, o que

poderia, na opinião deles, modificar as formas de avaliação, trazendo possibilidades de interação quase face a face, devido à utilização de vídeo e de áudio em tempo real. No entanto, sabemos que ainda são necessárias mais pesquisas sobre esse tipo de interação. Nesse âmbito, os autores (*Ibid*) relatam a tecnologia incorporada a este tipo de teste por meio de telefone, usando o Processamento de Linguagem Natural (PLN). O teste feito por telefone não tem o formato de entrevista, como na maioria dos testes de proficiência oral, mas contém uma variedade de tipos de tarefas, incluindo leitura e repetição de sentenças, pronúncia de palavras. A avaliação é feita pelo próprio programa de computador, de acordo com critérios que avaliam a acuidade na repetição de palavras, pronúncia, fluência de leitura e de repetição. Este tipo de teste é baseado em um grande *corpus* de língua falada de falantes nativos de diferentes regiões e dialetos sociais ingleses, com um sistema de reconhecimento de fala pelo computador.

Além desses recursos, segundo Chapelle e Douglas (2006), algumas instituições, tais como, *Ordinate Coorporation, Purdue University e a Center of Applied Linguistics* em Hong Kong, vem desenvolvendo outros dispositivos de avaliação eletrônica que trabalham sem intervenção humana. Sem dúvida, o potencial deste tipo de dispositivo está ligado ao aumento da praticidade na aplicação dos testes, porém, reforçamos a ideia de que mais pesquisas são necessárias "para se entender os usos e as limitações da tecnologia do PLN", no intuito de "compreender sob quais circunstâncias esses dispositivos são mais úteis do que os humanos para avaliação da produção linguística dos candidatos." (SALOMÃO, 2010. p. 331), visto que:

> Em termos de avaliação mediada por computador, o que se viu mais frequentemente como advento do uso da tecnologia para fins de avaliação foram testes de múltipla escolha, de ordem estruturalista, oferecidos em CD-ROM, e o desenvolvimento de CAT – *computer adaptative tests*, ou seja, testes com flexibilidade e adaptabilidade, nos quais os computadores podem adaptar o insumo subsequente de acordo com as respostas anteriores dos candidatos, possibilitando a criação de um teste sob medida para o candidato. (ALDERSON; CLAPHAM; WALL, 1998; CHAPELLE; DOUGLAS, 2006 *apud* SALOMÃO, 2010. p.330).

Em relação às inúmeras ferramentas existentes que podem ser utilizadas na elaboração de um teste de proficiência linguística, Chapelle e Douglas (2006) destacam alguns sistemas desenvolvidos por universidades norte-americanas, no intuito de servir como parâmetro formativo, diagnóstico e de nivelamento dentro do campus. Algumas dessas ferramentas ou

programas são gratuitos e de fácil manuseio, podendo o próprio professor responsável pela aplicação criar seu teste, de acordo com a necessidade da instituição. Outros são mais elaborados e com mais recursos, porém, pagos ou privados. Geralmente, todos eles utilizam ou estão localizados em um servidor *Web CT*[29]. Entre eles: i) *Respondus*, um teste privado, em que podem ser criadas questões de múltipla-escolha, verdadeiro ou falso, escrita de um parágrafo, ligar sentenças, respostas curtas. Este programa permite ao elaborador decidir como serão as instruções e a aparência do teste; ii) *Hot Potatoes*, um sistema grátis para criação de testes e mais simples de manuseio. Permite a elaboração de questões de múltipla-escolha, respostas curtas, frases desordenadas, palavras cruzadas e preenchimento de lacunas; iii) *Discovery School Website*, sistema gratuito de criação de testes, simplificado e de fácil manuseio. Permite a construção de questões de múltipla-escolha, respostas curtas, verdadeiro ou falso, escrita de ensaio e compilação de informações; e iv) *Blackboard e Questiomark*, programas privados que oferecem vários procedimentos para a criação de tarefas e de respostas. Permitem comandos de 'arrastar e soltar', uso de vídeos e insumo multimodal, pelos sistemas *Java* e *Macromedia Flash*, além de todos os outros recursos citados anteriormente. No entanto, reconhecemos que esses sistemas ou programas, abordados acima, não são recomendados para a construção de um teste de alta relevância ou de larga escala, servindo apenas para testes diagnósticos ou de nivelamento, devido ao formato dos itens e questões de segurança e confiabilidade do teste.

Finalmente, além das questões apontadas, outro fator a ser considerado é como a comunicação mediada por computador afeta o uso da língua, pois, segundo McNamara (2000. p. 10), "os testes são baseados nas teorias da natureza do uso da língua alvo e a maneira na qual isso é entendido será refletido no *design* dos testes." [tradução nossa].[30] Ou seja, a diferença de formatos entre testes impressos e testes de desempenho (computadorizados ou não-computadorizados), por exemplo, não é apenas uma coincidência, pois reflete uma diferença implícita entre visões de língua e de uso da língua.

Assim, apresentamos, a seguir, as especificações voltadas aos testes computadorizados e aos testes adaptativos.

---

[29] O WebCT (*Web Course Tools*) teve sua origem em um projeto da University of British Columbia - Canadá, com o objetivo de desenvolver uma ferramenta que permitisse aos professores a construção de cursos a distância, sem a necessidade e utilizar recursos ou técnicas de manuseio. Entre os recursos oferecidos estão o *chat*, grupo de discussão, e-mail, criação de testes e avaliações, e vários recursos específicos para o professor acompanhar o andamento do curso. Permite instalação para teste. Disponível em: <http://www.webct.com>. Acesso em: julho/2017.

[30] Tests are based on theories of the nature of language use in the target setting and the way in which this is understood will be reflected in test design.

4.2 CBT versus CAT

Do ponto de vista de Piton-Gonçalves e Aluísio (2015), os CBT são uma possibilidade inovadora para as avaliações educacionais, possibilitando correções automatizadas, com a produção de estatísticas. Walczak (2015) entende que os CBT podem ser divididos entre testes **lineares** e testes **adaptativos**. Um teste linear é um teste padronizado que possui um mesmo número de questões ordenadas na mesma sequência, administrado a todos os examinandos em uma determinada aplicação do teste, o que o torna parecido com os testes de papel e caneta, se diferenciando apenas pelo meio de administração - o computador. Em um CAT, por sua vez, cada examinando realiza diferentes tarefas ou questões que se adaptam ao seu nível de proficiência, demonstrado por meio de seu desempenho durante o teste, de acordo com as respostas corretas ou incorretas fornecidas. Neste caso, é o computador que seleciona as questões a partir de um banco de itens definido para o teste. "Se um candidato respondeu corretamente à pergunta anterior, a próxima pergunta administrada será mais difícil. Por sua vez, se um candidato respondeu a pergunta anterior incorretamente, a próxima pergunta administrada será menos difícil". [tradução nossa][31]. (*Ibid*. p. 37). Desta forma, a primeira questão tende a ser de nível médio ou intermediário, e a partir da resposta do examinando, o computador estima o seu nível de habilidade na primeira fase do teste, e em seguida, o algoritmo seleciona uma pergunta ou item em um nível de dificuldade apropriado para a capacidade demonstrada pelo examinando, e assim, sucessivamente até o término do teste. O processo de administrar os itens com base nas respostas do examinando continua até que o teste estabeleça o nível de habilidade do mesmo. O fim do teste, portanto, pode ocorrer por esgotamento das questões no banco de dados, pelo término do tempo estimado pelo exame ou pela exaustão do examinando.

De acordo com Walczak (2015), na *Cambridge English Language Assessment*, os testes lineares mediados por computador têm sido amplamente administrados desde o início do século XXI, juntamente com a versão em papel e caneta dos mesmos, enquanto que a primeira aplicação de um CAT pela instituição data do início dos anos 90, com a criação de testes para negócios e agências de empregos. Porém, como a autora destaca, esses testes adaptativos avaliam as habilidades de leitura e escuta, ao contrário dos testes lineares que avaliam as quatro habilidades da língua: leitura, escrita, audição e oralidade. De acordo com

[31] If a candidate answered the preceding question correctly, the next question administered will be more difficult. In turn, if a candidate answered the preceding question incorrectly, the next question administered will be less difficult.

Walczak, a administração e produção dos CAT diferem dos CBT lineares e dos de papel e caneta. Nos testes da *Cambridge*, por exemplo, o resultado de cálculo de medida do desempenho do examinando é o mesmo em testes lineares e adaptativos, mas o processo para se chegar ao resultado é diferente.

Além desses fatores, os CAT exigem uma preocupação com a especificação do teste e a construção do banco de itens ou tarefas, ou seja, os desenvolvedores e elaboradores de testes devem pensar sobre o que querem exatamente medir, qual habilidade será avaliada (leitura, escrita, habilidades integradas, etc) e como ela será medida. Isso engloba pensar nos tipos de tarefas ou questões (múltipla-escolha, *quiz*, *cloze*, etc), no conteúdo e na quantidade de tarefas, pois um banco de itens deve permitir que o teste seja administrado sem correr o risco de repetição frequente das questões. (*Ibid*. p. 36). Essas questões estão relacionadas à validade e confiabilidade dos testes. No entanto, a construção de um banco de itens é uma tarefa árdua, visto que é aconselhável que as tarefas passem por um pré-teste para estar adequadas para a administração por computador. Além disso, os materiais selecionados para um CAT devem abranger todos os níveis de dificuldade, o que depende do propósito para o qual o teste é desenvolvido e para que tipo de examinando ele é destinado. Uma vez disponível, o teste passa pelo processo de verificação e edição, para então ser administrado.

4.2.1 Definição de CAT

Os testes adaptativos por computador ou CAT utilizam a tecnologia para avaliar o desempenho do examinando de forma personalizada, adaptando cada item ou questão de resposta de acordo com o nível de habilidade demonstrado pelo examinando. Esses testes caracterizam-se por adaptar os itens ou tarefas de acordo com as respostas do examinando as questões. Por exemplo, se um examinando acerta ou executa um item de dificuldade intermediária de forma apropriada, a próxima questão será mais difícil, se ele errar é apresentado uma questão ou tarefa mais fácil, e assim, sucessivamente, até o final do exame ou até que seu nível seja estabelecido. Tal adaptação é baseada em um algoritmo especificado pelo desenvolvedor, por meio de um modelo psicométrico, em termos da TRI. Segundo Piton-Gonçalves e Aluísio (2015):

> A TRI está ligada a aspectos psicométricos do teste e propõe uma modelagem estatístico-matemática para as características latentes do indivíduo e para os parâmetros associados aos itens. O CAT baseado na TRI (unidimensional) pode ser denominado Teste Adaptativo Computadorizado Unidimensional (UCAT). Geralmente, a literatura utiliza o termo CAT quando se refere a um UCAT. (p. 390).

Alguns testes educacionais, especialmente os de proficiência linguística em larga escala, como o Celpe-Bras e o CELU, por exemplo, têm apontado para avaliações de desempenho do examinado em múltiplas competências e habilidades integradas, baseado em tarefas que remetem à vida cotidiana. Essa mudança progressiva de foco avaliativo deve-se ao reconhecimento dos estudiosos da área ao fato de que o conhecimento humano é plural e multifacetado. Sendo assim, Piton-Gonçalves e Aluísio (2015) advertem que o CAT ou UCAT não considera essas premissas, avaliando apenas uma habilidade ou competência, separadamente. Desta forma, os autores defendem a criação e uso do Teste Adaptativo Computadorizado Multidimensional (MCAT) para fins educacionais como uma resposta a uma avaliação que considera simultaneamente os múltiplos traços latentes ou as múltiplas habilidades dos examinandos. Este modelo 'supõe que o examinado possua mais de uma habilidade estimada pela Teoria de Resposta ao Item Multidimensional (MIRT)'. (*Ibid*. p. 408). No entanto, os autores concordam que tanto a pesquisa quanto o desenvolvimento de um MCAT, operacional em meio educacional, ainda são incipientes nos cenários nacional e internacional, conforme a revisão da literatura feita pelos mesmos:

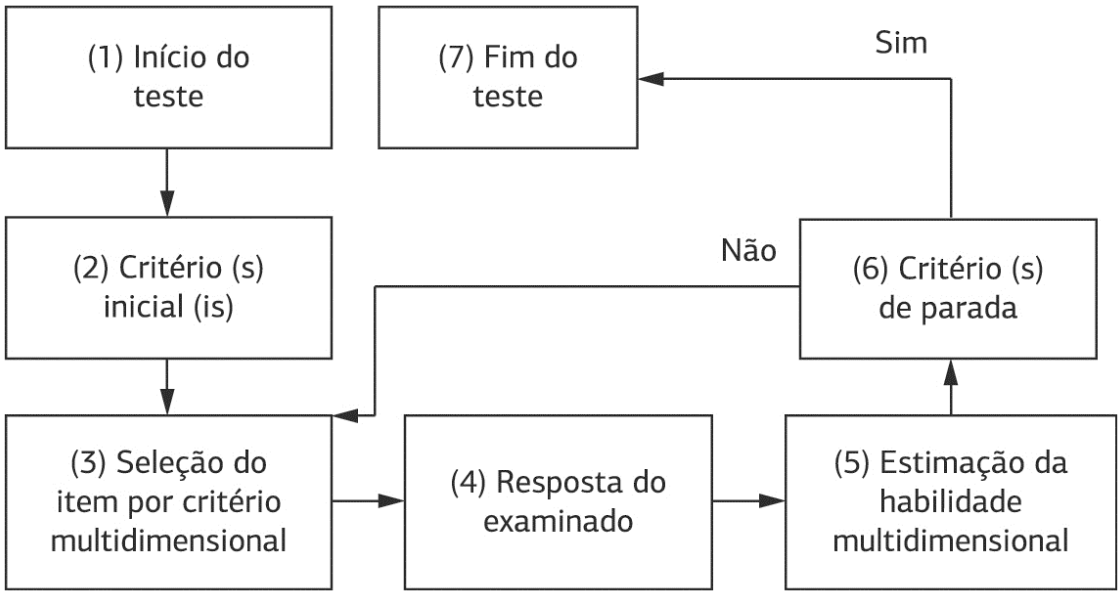
> Na literatura, o termo MCAT geralmente está associado a um conjunto de teorias, métodos e modelos psicométricos da área de estatística e de estatístico-matemática multivariada, e não necessariamente a um sistema computacional que permite a utilização com usuários reais. Para suprir esta lacuna, Piton-Gonçalves e Aluísio (2012) desenvolveram a abordagem *Computer-based Multidimensional Adaptive Testing* (CBMAT), que norteia o projeto e o desenvolvimento computacional de testes multidimensionais. Dessa forma, um CBT poderá contemplar um MCAT, realizando automaticamente a montagem, a aplicação e a correção de testes adaptativos. (PITON-GONÇALVES e ALUÍSIO, 2015. p. 391).

Esse modelo multidimensional proposto e em construção parece ideal para a área de avaliação em LE, um caminho para a aplicação de testes computadorizados que mais se aproxima das múltiplas habilidades e competências que o examinado possui; acurado e mais curto do que os testes lineares tradicionais. Entretanto, existem limitações e dificuldades

relacionadas ao MCAT, dentre elas, o alto custo operacional, de desenvolvimento e de manutenção, pois dependendo da metodologia ou construto adotado para o teste, a inserção e a remoção de itens do banco tornam-se medidas onerosas, envolvendo uma análise criteriosa dos itens e a necessidade da presença de especialistas de conteúdo e métodos estatísticos multivariados para avaliação dos mesmos.

Abaixo é apresentado o fluxograma geral do possível funcionamento e o ciclo de um MCAT, elaborado por Piton-Gonçalves e Aluísio.

GRÁFICO 3 - Fluxograma geral de um MCAT.



Retirado de Piton-Gonçalves e Aluísio (2015). Fonte: Piton-Gonçalves (2012).

Na próxima subseção, destacamos as principais vantagens e desvantagens dos CAT, segundo Dunkel (1999), Walczak (2015), Piton-Gonçalves e Aluísio (2015) e Chapelle e Voss (2016).

4.2.2 Vantagens dos CAT

As principais vantagens dos CAT, em relação a outros tipos de testes computadorizados lineares ou não computadorizados, são, especialmente, o tempo reduzido de realização do teste e a entrega imediata de resultado ao examinando, além da maior flexibilidade e adaptabilidade que oferecem. São muitas as vantagens do CAT: i) maior precisão em relação a testes que apresentam um número fixo de itens, uma vez que cada examinando realiza uma versão diferente do teste e os itens são selecionados com base na capacidade e no padrão de resposta individual; ii) melhoria do armazenamento e do processamento de dados, pois as respostas dos examinandos são armazenadas eletronicamente e podem ser acessadas sempre que necessário; iii) a marcação computadorizada elimina a impressão de folhas de resposta e a ambiguidade em relação a rasuras ou respostas marcadas incorretamente; iv) menos itens são necessários para avaliar o nível de proficiência do examinando, reduzindo, consequentemente, os gastos com a administração e aplicação dos testes; v) elimina algumas ameaças de segurança, tornando a fraude tradicional ("a cola") ineficaz; vi) a velocidade das respostas do examinado pode ser usada como informação adicional na avaliação da proficiência ou habilidade estimada; e vii) o examinando é desafiado, e não desmotivado por itens muito acima ou abaixo de seu nível de proficiência. (Dunkel, 1999; Walczak, 2015; Piton-Gonçalves e Aluísio, 2015; Chapelle e Voss, 2016).

No entanto, algumas desvantagens existem e são apresentadas a seguir.

4.2.3 Desvantagens dos CAT

A principal desvantagem ao produzir um CAT está relacionada com a dimensão do banco de itens. É necessário desenvolver itens com vários graus de dificuldades e variados entre si, já que o teste é mais curto, a validade de cada item para medir a habilidade desejada de um examinando deve ser criteriosa e rigorosamente estudada. Itens repetitivos ou falhos podem "afetar alguns examinandos e não outros, pois nem todos eles recebem o mesmo conjunto de itens. Isso pode afetar a imparcialidade do teste" ou a confiabilidade do mesmo. (WAINER, 2000 *apud* WALCZAK, 2015. p.37). Mesmo contando com mecanismos de controle desenvolvidos para evitar o uso excessivo de itens, itens com certas características tendem a ser superexpostos. "Essa superexposição precisa ser monitorada, o que impõe custos

operacionais ao provedor de teste". (Walczak, 2015. p. 37). Assim, para conter esse risco e elaborar um CAT seguro, algumas possibilidades são apresentadas, a seguir.

4.2.4 Produção de um CAT confiável e válido

Ao longo dos últimos vinte anos, muitos pesquisadores têm demonstrado preocupação a respeito do que é necessário na construção e desenvolvimento de testes confiáveis e válidos em meio computadorizado, lineares ou adaptativos. (DUNKEL, 1999; PITON-GONÇALVES e ALUÍSIO, 2015; CHAPELLE e VOSS, 2016). Questões técnicas, dignas de consideração no projeto e desenvolvimento de um teste adaptativo por computador, foram discutidas e os estudiosos reconhecem que a maioria das preocupações são as mesmas ao se elaborar testes convencionais impressos, e até mesmo testes lineares computadorizados. No entanto, os CAT apresentam exigências e medidas peculiares que devem ser tomadas antes, durante e após o seu desenvolvimento, administração ou aplicação. De acordo com pesquisas direcionadas a encontrar soluções e respostas sobre as prioridades na elaboração e produção de um CAT, como o trabalho de Walczak (2015) e Piton-Gonçalves e Aluísio (2015), podemos destacar que esse tipo de teste deve: i) possuir um banco de itens calibrados[32] e, suficientemente, grande para administrar itens válidos e confiáveis; ii) aplicar a pré-testagem de tarefas e conteúdos, bem como estabelecer a dificuldade de teste e regra de parada apropriada ao construto do teste; iii) possuir *hardware* e *software* adequados para administrar os itens de banco, a velocidade, o armazenamento e os recursos necessários à correção e entrega de resultados; iv) desenvolver programas educacionais para garantir que os examinandos compreendam como o CAT funciona; e v) relatar sistemas seguros, úteis e transparentes para os usuários e examinandos.

Nas próximas seções são descritos testes computadorizados e suas principais características, respectivamente: o TOEIC, o PTE-*Academic*, o TOEFL, o Projeto DIALANG e os Testes do ACTFL.

[32] **Calibração** é a estimação dos parâmetros dos itens e ocorre antes da aplicação de um teste formal. Os dados para a calibração são obtidos a partir da aplicação de um pré-teste em uma amostra de examinandos da população em questão. Se o teste for definido para respostas dicotômicas, então, a partir dos acertos e dos erros dos examinados, determinam-se os parâmetros de cada item. (PITON-GONÇALVES & ALUÍSIO, 2015. p. 395).

4.3 O exame TOEIC

As informações desta seção foram selecionadas da página oficial do exame[33].

O *Test of English for International Communication* – TOEIC, é um exame americano, produzido pelo *Educational Testing Service* – ETS e oferecido uma vez por mês em centros aplicadores credenciados por este órgão particular. O *TOEIC* foi criado em 1977 a pedido do Ministério da Indústria e Comércio Exterior do Japão, para facilitar a comunicação de seus servidores e empresas japonesas com empresas ao redor do mundo. Sua primeira aplicação ocorreu em 1979, com os próprios servidores do Ministério japonês para 2700 servidores. O *TOEIC* é utilizado em mais de 150 países e por mais de 14 mil empresas, entre organizações, entidades governamentais, Ministérios de Educação e programas de ensino do idioma inglês em todo o mundo. No Brasil, o exame é oferecido por cursos livres de idiomas e pelos institutos federais, incluindo o Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), em Belo Horizonte.

As questões do exame são baseadas em ambientes de trabalho da vida real (reuniões, viagens, conversas telefônicas, etc.). O TOEIC é estruturado em duas partes, sendo que o examinando pode fazer as duas ou escolher apenas uma delas para testar sua proficiência em LI. Desta forma, os testes TOEIC avaliam as competências linguísticas que são utilizadas na vida quotidiana e no local de trabalho, além de utilizar um vocabulário cotidiano comum em frases e expressões-chave usadas nos negócios.

4.3.1 Formato do Teste

A **Parte I** do exame avalia a leitura e a escuta, em formato papel e caneta, e é realizado em centros aplicadores credenciados, com duração de 2h30min. Esta parte do exame é pontuada entre 10 e 990 pontos.

A **Parte II** do exame avalia a fala e a escrita, em um ambiente *online*. Esta parte é realizada em centros credenciados, com duração total de 90 minutos. Esse formato de teste *online* está disponível em apenas alguns países, por ser um teste relativamente novo. A pontuação desta parte do exame fica entre 0 a 200 pontos.

[33]Disponível em <https://www.ets.org/toeic>; <http://www.examenglish.com/TOEIC/>; e <http://mastertest.com.br/blog/toeic-exame/>. Último acesso em janeiro de 2016.

Considerando que a presente pesquisa concentra-se em verificar, também, as tecnologias de informação e comunicação utilizadas em alguns testes de proficiência linguística, descrevemos, a seguir, apenas a Parte II do exame e suas seções.

4.3.2 Parte II - Teste de Fala e Escrita

As duas seções (Fala e Escrita) são independentes uma da outra e, portanto, uma atribuição diferente de pontuação é dada para cada seção, separadamente. Sendo assim, as descrevemos de forma separada.

4.3.3 O formato do teste de Fala

O teste de Fala inclui 11 (onze) perguntas que medem os aspectos da habilidade de fala do examinando. Para cada pergunta, ele recebe instruções específicas, incluindo o tempo permitido para preparar e falar sua resposta. Com duração de cerca de 20 (vinte) minutos, essa seção do teste segue uma escala de pontuação de 0 a 200. Além disso, já que esta seção é realizada por computador, em ambiente *online*, a voz do examinando é gravada e enviada a um programa de computador de correção, para ser pontuada. A nota final é enviada ao examinando dentro de 3 (três) a 4 (quatro) semanas. Essa seção conta com cinco tarefas, a saber: i) Ler um parágrafo em voz alta (testa a pronúncia); ii) Descrever uma imagem (testa vocabulário / gramática); iii) Responder a perguntas gravadas (testes de fluência / discurso prolongado); vi) Propor uma solução; v) Expressar uma opinião.

Finalmente, as tarefas do teste de Fala do TOEIC estão organizadas para apoiar três alegações sobre o desempenho do examinado: i) O examinando deve gerar uma interação oral com falantes nativos e não-nativos; ii) O examinando deve selecionar o idioma alvo para interações sociais e ocupacionais rotineiras, como dar e receber instruções, pedir e dar informações, solicitar e dar esclarecimentos, além de simular compras, saudações e apresentações; e iii) O examinando deve criar um discurso coeso e sustentar uma conversa apropriada à vida cotidiana típica e ao local de trabalho.

O quadro abaixo mostra o número de questões, as tarefas e a descrição de cada tarefa que fazem parte da seção de Fala da Parte II do exame.

QUADRO 6 – TOEIC – Visão Geral do Teste de Fala - 20 minutos

Número de questões	Tarefa	Descrição da tarefa
1–2	Leia o texto em voz alta	▪ Ler, em voz alta, o texto mostrado na tela. ▪ São 45 segundos para preparar e 45 segundos para ler.
3	Descreva a imagem	▪ Descrever a figura ou imagem na tela como o máximo de detalhes possíveis. ▪ São 30 segundos para preparar a resposta e 45 segundos para gravar a fala sobre a imagem.
4–6	Responda as questões	▪ Responder 3 questões, sem tempo para preparação. ▪ São 15 segundos para responder as questões 4 e 5. ▪ E mais 15 segundos para responder a questão 6.
7–9	Responda as questões usando a informação fornecida.	▪ Responder 3 questões baseadas em uma informação provida. ▪ São 30 segundos para ler a informação antes da pergunta começar. Não é dado tempo extra para preparar a resposta. ▪ São 15 segundos para responder as questões 7 e 8. ▪ E 30 segundos para responder a questão 9.
10	Proponha uma solução	▪ Propor uma solução a um problema apresentado. ▪ São 30 segundos para preparar a resposta. ▪ E 60 segundos para falar. Na resposta deve apresentar o problema e a proposta para lidar com ele.
11	Expresse uma opinião	▪ Dar uma opinião sobre um tópico específico. ▪ São 15 segundos para preparar a resposta. ▪ E, então, 60 segundos para falar.

Retirado e disponível em < https://www.ets.org/toeic/test_takers/speaking_writing/about/content>. Acesso em: jan/2017. Traduzido pela autora.

A seguir, são descritas as principais características da seção do teste de Escrita do exame.

4.3.4 O formato do teste de Escrita

O teste de Escrita também é feito em ambiente *online*, com duração de aproximadamente 60 minutos e enviado para um programa de computador para ser pontuado, de acordo com uma escala de 0 a 200 pontos. As pontuações são enviadas ao examinando dentro de 3 (três) a 4 (quatro) semanas. Essa seção é subdividida em três tarefas, com um total de oito questões, com os seguintes comandos: i) Escreva frases sobre imagens; ii) Escreva uma resposta por e-mail; ii) Escreva um ensaio argumentativo.

Assim, as tarefas do teste de Escrita do TOEIC buscam avaliar: i) O examinando deve produzir frases bem formadas, incluindo frases simples e complexas; ii) O examinando deve produzir um texto de comprimento variado para transmitir informações diretas, perguntas, instruções, narrativas, etc.; e iii) O examinando deve produzir texto de comprimento variado para expressar ideias complexas, opiniões próprias, evidências e explicações.

O quadro 7, a seguir, apresenta uma visão geral do teste de Escrita, que inclui oito perguntas destinadas a medir diferentes aspectos da habilidade de escrita do examinando, com duração de cerca de uma hora. Além disso, há instruções específicas para cada tipo de pergunta, incluindo o tempo permitido para realização da tarefa.

QUADRO 7 – TOEIC – Visão Geral do Teste de Escrita - 60 minutos

Questões	Tarefa	Descrição
1–5	Escreva uma sentença baseada em uma imagem	▪ Escrever uma frase ou sentença baseada na imagem fornecida na tela. ▪ Para cada figura, serão disponibilizadas duas palavras ou frases a ser utilizadas para construir a sentença. ▪ A ordem (ou formas das palavras) pode ser modificada.
6–7	Responda a uma pergunta escrita	▪ Responder a um e-mail da melhor forma possível. ▪ São 10 minutos para ler e responder cada e-mail.
8	Escreva um ensaio de opinião	▪ Escrever um ensaio em resposta a uma pergunta. ▪ É preciso declarar, explicar e apoiar uma opinião sobre o problema proposto. ▪ Tipicamente, um ensaio eficaz contém um mínimo de 300 palavras.

Retirado de <https://www.ets.org/toeic/test_takers/speaking_writing/about/content>. Acesso em: jan/2017. Traduzido pela autora.

A seguir, são apresentadas algumas características da pontuação do TOEIC.

4.3.5 Pontuação do teste

As respostas aos testes de Fala e Escrita do TOEIC são pontuadas com base em critérios de avaliação específicos. As pontuações obedecem a intervalos ou níveis de proficiência que resumem as habilidades e dificuldades específicas. O teste TOEIC Parte Oral e Escrita é realizado no ETS *Online Scoring Network* (OSN). OSN é um sistema de *Internet* seguro que fornece uma plataforma para a exibição das tarefas dos examinandos, com as respostas, os materiais de suporte para os avaliadores e as ferramentas para monitorar a precisão dos avaliadores[34].

O quadro a seguir apresenta as respectivas pontuações dos testes de Fala e de Escrita do TOEIC, a fim de permitir uma melhor compreensão.

QUADRO 8 – Pontuações dos Testes de Fala e Escrita do TOEIC

Pontuações do Teste de Fala	**Pontuações do Teste de Escrita**
As perguntas 1-9 são avaliadas em uma escala de 0-3.	As perguntas 1-5 são avaliadas em uma escala de 0-3.
As perguntas 10-11 são classificadas em uma escala de 0-5.	Perguntas 6-7 são classificadas em uma escala de 0-4. A pergunta oito é avaliada em uma escala de 0-5.
A soma das classificações é convertida para uma pontuação escalonada de 0-200.	A soma das classificações é convertida para uma pontuação escalonada de 0-200.
Oito níveis de proficiência são fornecidos.	Nove níveis de proficiência são fornecidos.

Elaborado pela autora. Informações retiradas e disponíveis em: https://www.ets.org/toeic/test_takers/speaking_writing/about/content>. Acesso em: jan/2017. Traduzido pela autora.

Para finalizar esta seção, o processo de pontuação e correção do exame TOEIC é descrito em seguida.

[34] Para ler a descrição dos níveis de proficiência, bem como os descritores para cada questão, ver o Manual do Examinador da Parte Oral e Escrita do TOEIC (2016. p. 9-23), fornecido na página oficial do Exame.

4.3.6 Visão Geral do Processo de Pontuação

De acordo com o Manual do Examinador da Parte Oral e Escrita do TOEIC (2016. p. 23-28), o ETS usa um rigoroso processo de pontuação para garantir resultados justos e confiáveis. As respostas dos testes são enviadas pelo sistema *online* e pontuadas por vários avaliadores do ETS certificados, que são monitorados de perto durante todo o processo[35]. Esse processo de correção e pontuação é descrito, detalhadamente, ao final da descrição do próximo exame, o TOEFL.

A próxima seção apresenta a descrição do Exame TOEFL.

4.4 O exame TOEFL

O *Test of English as a Foreing Language* - TOEFL é um exame de proficiência americano que apresenta a versão *Internet Based Test* (iBT) desde 2000. Segundo pesquisa feita por Anchieta (2010), em sua dissertação, o TOEFL foi criado em 1964, pelo ETS e é um teste que avalia as habilidades linguístico-comunicativas em LI de falantes não nativos desse idioma, por meio de tarefas voltadas mais especificamente ao ambiente acadêmico. De 1964 até 1998, de acordo com dados históricos, o exame avaliava somente as habilidades de compreensão oral, de gramática, de vocabulário e de leitura. A parte de escrita poderia ser feita separadamente com a pontuação distinta do resultado do teste. Naquela época, o teste escrito era, portanto, complementar e chamado de TWE (*Test of Written English*). Somente em 1999, com uma nova versão, o TOEFL-CBT (*computer based test*), a parte escrita tornou-se obrigatória, mas ainda com a pontuação e o resultado entregues separadamente. Portanto, passou a existir duas versões do exame: a versão em papel e caneta, que foi mantida na maioria dos países, incluindo o Brasil, e a versão eletrônica. Além disso, de acordo com a página oficial do exame, mais de 9.000 instituições de ensino superior, agências e outras instituições em mais de 130 países utilizam os resultados do TOEFL. Os níveis de certificação ou classificação variam de intermediário a avançado. O teste é realizado, principalmente, por estudantes que planejam estudar em uma instituição de Nível Superior fora do seu país de origem, para admissões e conclusões de programas de aprendizado em LI. O teste é também

[35] Para mais informações sobre o sistema de correção e pontuação dos exames, consultar o Manual do Examinador da Parte Oral e Escrita do TOEIC (2016) , fornecido na página oficial do Exame.

realizado por candidatos a bolsa de estudo e certificação, por estudantes de LI que desejam monitorar o próprio progresso e por estudantes e trabalhadores solicitando vistos.

De acordo com Anchieta (2010):

> [...] as versões oficiais do teste TOEFL PBT (*paper based test*) e CBT sempre apresentaram a mesma sequência de seções, sendo elas: 1) a produção escrita de um texto argumentativo sobre um assunto não-técnico; 2) a compreensão oral com três tipos de exercícios - diálogos curtos; conversas curtas; trechos de palestras ou aula; 3) a de questões de estrutura e expressão escrita com dois tipos de exercícios – complementação de orações e reconhecimento de erros; 4) a compreensão de leitura, com vários textos e questões objetivas de múltipla escolha." (ANCHIETA, 2010. p.119).

Somente no fim dos anos noventa, surgiu, então, a versão iBT (*internet-Based Test*), resultado de um projeto denominado TOEFL 2000. Atualmente, ele é aplicado de 30 a 40 vezes por ano, por administradores treinados, em centros de aplicação credenciados em todo o mundo. Desde 2005, um simulado completo do TOEFL-iBT é oferecido gratuitamente e disponibilizado pelo ETS na página oficial do exame, com a parte de avaliação oral paga. Esse exame talvez seja o mais lucrativo para a organização do ETS, que o viabiliza com a validade da certificação de dois anos. Por não ser custeado pelo governo norte-americano e sim por uma instituição privada, existe uma visão mercadológica por trás desse exame, seu principal objetivo não é a difusão da LI, pois esta ainda desempenha o papel de língua hipercentral da ciência e da tecnologia no sistema mundial, mas sim o lucro que pode ser obtido com as suas aplicações ao redor do mundo e os testes simulados oferecidos na página oficial do exame. Além do ETS, muitos cursos livres de idioma, editoras e professores particulares lucram por meio de ofertas de cursos rápidos, presenciais ou virtuais, voltados para a preparação do examinando para realizar o exame. No Brasil, a primeira aplicação desta versão iBT aconteceu em 2006.

Seguimos com o formato e estrutura do TOEFL.

4.4.1 O formato do exame

As seções da versão *online* são estruturadas em quatro partes, a saber, Leitura, Compreensão Auditiva, Expressão Oral e Redação ou Escrita, como mostra o quadro 9, abaixo.

QUADRO 9 – Visão geral do exame TOEFL-iBT

TOEFL-iBT Anchieta (2010)

SEÇÃO DE TESTE	NÚMERO DE QUESTÕES	LIMITE DE TEMPO
Leitura	3 a 5 textos escritos de origem acadêmica não especializada, extraídos de vários periódicos, com até setecentas palavras. 12 a 14 questões sobre cada texto	60 a 100 min
Compreensão Auditiva	4 a 6 palestras ou debates (geralmente de origem universitária ou social); 6 questões sobre cada uma delas; 2 a 3 diálogos com 5 questões sobre cada um deles.	60 a 90 min
Intervalo		
Expressão Oral	6 tarefas: 2 independentes (geralmente opiniões sobre assuntos da vida acadêmica) e 4 integradas: duas de compreensão oral e leitura (resumir a relação entre as ideias apresentadas nos textos); e duas tarefas de compreensão e produção orais, (resumir o que ouviu e apresentar opinião sobre o assunto).	20 min
Redação	1 tarefa integrada (elaboração de uma dissertação (essay)); na segunda tarefa, e 1 tarefa independente, integrando leitura e compreensão oral (expressar opinião sobre os conteúdos apresentados).	20 min e 30 min
Total		4 horas

Elaborado por ANCHIETA (2010) e adaptado pela autora.

As seções de Fala e Escrita do teste TOEFL-iBT incluem outros sotaques de falantes nativos do inglês, como, por exemplo, sotaques do Reino Unido, Nova Zelândia ou Austrália, além dos sotaques da América do Norte.

Em todas as seções são permitidas anotações feitas pelo examinando e há sempre um cronômetro na tela para que o mesmo possa controlar o tempo gasto a cada seção. Além disso, as quatro seções são realizadas no mesmo dia e as tarefas têm como objetivo focar em situações de fala semelhantes àquelas da vida real do meio acadêmico, avaliando a capacidade do examinando de usar e compreender o inglês no nível universitário, além de combinar habilidade de compreensão oral, leitura, expressão oral e escrita para realizar tarefas acadêmicas.

De acordo com Anchieta (2010), após análise das mudanças ocorridas entre a versão impressa e a versão *online,* verificou-se que no TOEFL-iBT, "as seções relacionadas às compreensões escrita e oral foram minimamente modificadas, a não ser no que diz respeito à maior ênfase dada a questões de coesão e coerência." (*Ibid*. p. 122). Houve também a inclusão de questões mais subjetivas, "nas quais se pede ao candidato a escolha de possíveis sentenças que melhor resumam trechos específicos de cada texto apresentado." (*Ibid*). Além disso, para que o exame pudesse ser aplicado via *Internet,* a modificação do teste envolveu considerações sobre a proficiência e sobre o desenvolvimento de recursos computacionais. Por exemplo, o teste pode ser aplicado apenas em máquinas devidamente equipadas e em locais específicos para esse fim. Outro fator que a autora destaca é a inclusão da seção de Fala no TOEFL-iBT, designada para a avaliação da habilidade de comunicação oral dos examinandos em ambientes acadêmicos, passando, assim, a abranger as cinco habilidades: fala, escuta, escrita, leitura e gramática.

A seguir, é apresentada a forma de correção e pontuação dos testes do ETS, em que se enquadram o TOEIC e o TOEFL.

4.4.2 O sistema de correção e pontuação dos exames do ETS

Podemos destacar três tipos de corretores automatizados e mais utilizados pelos órgãos ou instituições desenvolvedoras de testes, utilizados para avaliar a escrita e a fala, a saber: o *e-Rater Criterion ou SpeechRater*, o *Phone Pass SET-10* e o PLN.

Os testes do ETS baseiam-se na tecnologia de Processamento de Linguagem Natural (PLN) para avaliar a escrita, em que o critério ou algoritmo analisa a entrada textual, ou seja, "o construto, ou significado, da habilidade de escrita quando definido pelo Critério é derivado das características do ensaio (texto) que o computador é capaz de reconhecer: o conteúdo de vocabulário, os marcadores de discurso e certas categorias sintáticas preestabelecidas." [tradução nossa][36]. (ETS, 2005a *apud* Chapelle e Douglas, 2006. pp. 216-215). A página oficial do ETS apresenta, além da divisão da pontuação de cada parte dos seus testes, o cálculo que é feito para a pontuação e como a qualidade do teste é assegurada, entre outras

[36] Based on NLP technology, Criterion parses textual input, assigning grammatical labels to constituents in the examinee's response and identifying markers of discourse structure and content vocabulary. The score Criterion assigns is derived from the NLP analysis: the lexical analysis signals the extent to which the examinee has addressed the task and the complexity of the response, the discourse analysis rates the organization and development, and the parser's analysis results in a rating of syntactic quality. In other words, the construct, or meaning, of "writing ability" as defined by Criterion is derived from features of the essay that the computer is able to recognize: content vocabulary, discourse markers, and certain syntactic categories.

coisas. Segundo os dados da página, o ETS usa avaliadores e métodos automatizados de pontuação, deixando claro que embora os modelos de pontuação automatizados tenham vantagens, eles não avaliam a eficiência da resposta no idioma e a adequação ao conteúdo, mas tendem a oferecer uma visão mais completa e precisa em aspectos lexicais e gramaticais da capacidade do indivíduo que se submete ao teste. Assim, os avaliadores humanos são necessários para atender a uma variedade mais ampla de recursos, como a qualidade das ideias e do conteúdo, bem como a forma em que são apresentados. Portanto, os exames do ETS utilizam a pontuação automatizada para complementar a pontuação dos avaliadores humanos das tarefas da seção de leitura, escrita e audição.

Para o processamento da correção de resposta oral, os exames do ETS[37] utilizam o mecanismo *SpeechRater*. Esse mecanismo é a aplicação de pontuação de resposta oral mais avançada do momento, direcionada para a obtenção de respostas espontâneas, em que a gama de respostas válidas é aberta e não limitada pelo estímulo de item. Esse sistema é utilizado pelo exame desde 2006. O *SpeechRater* baseia-se numa concepção ampla da construção da proficiência em LI, englobando aspectos da fala (como pronúncia e fluência), a facilidade gramatical e a capacidades de nível superior relacionadas com a coerência tópica e a progressão de ideias do examinando. Assim, ele processa cada resposta com um sistema automatizado de reconhecimento de fala especialmente adaptado para uso da LI como LE ou L2. Com base na saída desse sistema, o PLN e os algoritmos de processamento de fala são usados para calcular um conjunto de características que definem um perfil da fala em um número de dimensões linguísticas, incluindo fluência, pronúncia, uso de vocabulário, complexidade gramatical e prosódia. Um modelo de proficiência de fala é aplicado a esses recursos, a fim de atribuir uma pontuação final para a resposta. Enquanto este modelo é testado em dados previamente observados e pontuados por avaliadores humanos, também é revisado por especialistas em conteúdos, a fim de maximizar sua validade. Além disso, se a resposta for considerada não corrigível, devido à qualidade de áudio ou outros problemas, o mecanismo *SpeechRater* pode deixá-lo de lado para um processamento especial humanizado.

A seguir são apresentadas as principais características do Exame PTE- *Academic.*

[37]Disponívem em : <https://www.ets.org/research/topics/as_nlp/speech/>. Acesso em: jan./2017.

4.5 O exame PTE-ACADEMIC

Todas as informações aqui relatadas sobre o teste, bem como as imagens de quadros, advêm de guias, manuais e tutorial retirados da página oficial do exame[38].

O *PTE Academic* (*The Pearson Test of English Academic*) é um teste britânico, instituído em 2009, e faz parte dos *Pearson Language Tests* do grupo *PLC Pearson,* órgão dedicado a avaliar e validar o uso da LI por falantes não-nativos. Além do *PTE Academic*, fazem parte do grupo outros dois exames: o *PTE General* e o *PTE Young Learners*. Os três exames são realizados em locais específicos, credenciados pela Qualifications and Curriculum Development Agency (QCA) e administrados em associação com o *Edexcel*, o maior órgão examinador do Reino Unido.

Atualmente, o *PTE Academic* pode ser realizado em 200 instituições ao redor do mundo, em datas flexíveis para realização durante todo o ano. Aprovado pelo governo australiano para pedidos de visto e aceito por várias instituições no Reino Unido, Austrália, EUA, Canadá, Nova Zelândia e Irlanda, incluindo as Universidades de Harvard e Yale, o exame testa a língua utilizada em ambientes acadêmicos por meio de material autêntico retirado de palestras universitárias e assuntos sobre o cotidiano no campus, mapas, gráficos, entre outros. O exame é estruturado em tarefas integradas e não-integradas, além de expor o examinando a vários sotaques, inclusive de falantes não-nativos de LI. Na página oficial do exame, existe uma seção que apresenta questões das quatro habilidades avaliadas, exemplificando o teste, ou seja, um simulado, oferecido gratuitamente. O examinando pode baixar esses conteúdos e estudar quando e onde quiser. Para ter acesso a outros materiais e questões é possível pagar pela liberação dos mesmos, caso haja interesse por parte do examinando.

O *PTE Academic* avalia as habilidades de escuta, leitura, fala e escrita via computador em 3 (três) horas de teste, aproximadamente. Para realizar o exame, o examinando deve se dirigir a um centro de aplicação credenciado. O teste é estruturado em três partes principais: i) produção e compreensão de fala e escrita; ii) compreensão auditiva e; iii) compreensão de leitura. Há 20 (vinte) formatos diferentes de questões que vão desde questões de múltipla-escolha até interpretação de textos e redação de pequenos textos escritos. Na seção seguinte é abordado o formato do *PTE Academic.*

[38] Disponível em: <http://pearsonpte.com>. Acesso em: jan/2017.

4.5.1 Formato do exame:

O quadro 10 mostra como o exame é estruturado, suas partes, seu conteúdo e o tempo para realização das tarefas. Cada parte é estruturada em seções ou itens e cada seção é temporizada individualmente.

QUADRO 10 – Visão Geral do Exame – PTE ACADEMIC

PARTE	CONTEÚDO	TEMPO PERMITIDO
Introdução	Introdução pessoal	Não medido
Parte 1	Expressão oral e escrita	77 – 93 minutos
Parte 2	Leitura	32 – 41 minutos
Intervalo opcional		***10 minutos***
Parte 3	Compreensão auditiva	45 – 57 minutos

Retirado do PTE-Academic Tutorial. Fonte: Pearson Education Ltd 2011. Disponível em: <http://pearsonpte.com/wp-content/uploads/2014/10/Tutorial.pdf>. Acesso em: jan/2017. Traduzido pela autora.

A seguir, é apresentada a Parte 1 do exame, em que é avaliado as habilidades de Fala e Escrita.

4.5.2 Parte 1: Fala e Escrita

A primeira parte do exame testa as habilidades de fala e de escrita é estruturada em sete seções, com propósitos e questões diversas, integrando as quatro habilidades, como expostas no quadro abaixo, que apresenta o item, a tarefa, as habilidades avaliadas, o conteúdo e o formato do estímulo para a escrita ou fala, isto é, o tamanho do texto de entrada (*prompt*) e o tempo para resposta de cada item.

O quadro 11 abaixo apresenta os itens que fazem parte da Parte 1 do exame e suas principais características, com o propósito de explicitar a forma como as ferramentas tecnológicas são utilizadas em cada seção do teste.

QUADRO 11 – PTE ACADEMIC – Visão geral da seção de Fala e Escrita

Item	**Tarefa**	**Habilidade**	**Tamanho do *Prompt***	**Tempo para Resposta**
Leia em voz alta	Ler, em voz alta, um texto de até 60 palavras que aparece na tela, sendo a fala gravada. A caixa ou ícone de gravação exibe uma contagem regressiva até que o microfone seja aberto. A gravação é permitida apenas uma vez, procedimento adotado ao longo de todo o teste.	Leitura e Fala	Texto até 60	Varia de acordo com o item, dependendo do tamanho do texto.
Sentença repetida	Repetir a frase que ouve. O áudio começa a ser reproduzido automaticamente. Quando o áudio termina, o microfone é aberto e a fala deve ser imediata. Não é possível reproduzir o áudio duas vezes, e a resposta dada é gravada apenas uma vez.	Escuta e Fala	3 a 9 segundos	15 segundos
Descreva a imagem	Descrever uma imagem. São disponibilizados 25 segundos para analisá-la e preparar a resposta, que deve ser uma frase curta.	Fala	N/A	40 segundos
Recontando a palestra	Recontar o que ouvir, de forma resumida, com as próprias palavras, priorizando os pontos principais. Geralmente, há uma imagem relacionada ao áudio, que é reproduzido automaticamente. Enquanto o áudio está sendo reproduzido. É permitido anotações no bloco de notas apagável fornecido na tela, sendo 10 segundos para preparar a resposta.	Escuta e Fala	Até 90 segundos	40 segundos
Responda uma breve pergunta	Responder a pergunta em uma ou poucas palavras. Há uma imagem relacionada à pergunta.	Escuta e Fala	3 a 9 segundos	10 segundos
Resumo do texto escrito	Escrever um resumo do texto apresentado na tela em uma única frase completa, contendo os pontos principais da passagem de leitura, utilizando não mais de 75 palavras. Há uma contagem de palavras escritas na parte inferior da tela, além de três botões de ajuda: cortar, copiar e colar.	Leitura e Escrita	Texto até 300 palavras	10 minutos

Escreva uma redação	Escrever um texto argumentativo com o mínimo de 200 e o máximo de 300 palavras em resposta a um texto de entrada. Na parte inferior da tela, aparece o número de palavras escritas. Existem também botões de cortar, copiar e colar, que servem como opção de uso durante a construção da resposta.	Escrita	2-3 sentenças	20 minutos

Retirado do PTE-Academic Tutorial. Fonte: Pearson Education Ltd 2011. Disponível em: <http://pearsonpte.com/wp-content/uploads/2014/10/Tutorial.pdf>. Acesso em: jan/2017. Traduzido e adaptado pela autora.

A seguir, é descrita a Parte 2 do exame e suas principais características.

4.5.3 Parte 2: Leitura

A segunda parte do exame testa as habilidades de leitura e reescrita e é estruturado em uma única seção cronometrada, divididas em cinco itens, com propósitos e questões diversas, integrando as habilidades de leitura e escrita, como expostas no quadro abaixo, que apresenta o item, a tarefa, as habilidades avaliadas e o tamanho do *prompt* de cada questão.

O quadro 12 abaixo apresenta os cinco itens que compõem a Parte 2 do exame.

QUADRO 12 – PTE ACADEMIC – Visão geral da seção de Leitura

Item	**Tarefa**	**Habilidade**	**Tamanho do *Prompt***
Múltipla-escolha: selecione uma única resposta	Ler a passagem e responder a pergunta de múltipla-escolha. Existem várias opções de resposta possíveis, mas apenas uma correta. Clicar na resposta correta, usando o botão esquerdo do mouse, para selecionar uma opção. Caso mude de ideia, clicar novamente para desmarcá-la. A opção selecionada é realçada em amarelo.	Leitura	Texto até 110 palavras

Múltipla-escolha: selecione várias respostas	Ler a passagem e responder a pergunta de múltipla-escolha. Há mais de uma resposta correta. Selecionar todas as opções de resposta que julgar corretas na lista de opções possíveis.	Leitura	Texto até 300 palavras
Reordene os parágrafos	Restaurar a ordem original do texto, selecionando as caixas de texto e arrastando-as na tela. Existem duas formas de mover o texto: i) Clicar com o botão esquerdo do mouse em uma caixa para selecioná-lo (ele será esboçado em azul), manter pressionado o botão esquerdo do mouse e arrastar para o local desejado; e ii) Clicar com o botão esquerdo do mouse em uma caixa para selecioná-lo e, em seguida, clicar com o botão esquerdo nos botões de seta para a esquerda e para a direita para movê-lo. No painel direito, há a opção de utilizar os botões de seta para cima e para baixo para reordenar as caixas. Para desmarcar uma caixa, é necessário clicar com o botão esquerdo do mouse em outro lugar na tela.	Leitura	Texto até 150 palavras
Preencha o espaço em branco	Arrastar e soltar as palavras na tela para preencher corretamente as lacunas no texto. Há uma passagem com algumas palavras faltando, as quais podem ser encontradas em uma lista de palavras em uma caixa azul. Há mais palavras do que lacunas, portanto, nem todas as palavras da lista serão usadas.	Leitura	Texto até 80 palavras
Preencha o espaço em branco	Selecionar as palavras mais apropriadas de uma lista *drop-down*[39] para restaurar o texto. Há uma passagem com algumas palavras faltando e, ao lado de cada lacuna, há um botão com uma lista suspensa. Selecionar a opção que julgue melhor para preencher a lacuna e clicar em um botão ao lado da lacuna em branco, a fim de revelar a lista suspensa de opções para essa lacuna.	Leitura e escrita	Texto até 300 palavras

Retirado do PTE-Academic Tutorial. Fonte: Pearson Education Ltd 2011. Disponível em: <http://pearsonpte.com/wp-content/uploads/2014/10/Tutorial.pdf>. Acesso em: jan/2017. Traduzido e adaptado pela autora.

A terceira parte do exame é descrita a seguir.

[39] Um lista *drop-down* ou lista suspensa é um elemento de interface similar a uma lista, que permite que o usuário escolha um valor de uma lista de opções que "cai para baixo". Quando está inativa, esta lista esconde as suas opções, economizando espaço na tela.

4.5.4 Parte 3: Escuta

A terceira parte do exame testa as habilidades de escuta e consiste em perguntas baseadas em clipes de áudio ou vídeo que são reproduzidos automaticamente. O examinando ouve cada áudio ou vídeo apenas uma vez e é autorizado a tomar notas. O volume pode ser ajustado enquanto o áudio ou o vídeo estiver sendo reproduzido, sendo que o examinando pode fazer anotações no bloco de notas apagável fornecido. Esta parte é estruturada em oito seções, com propósitos e questões diversas, integrando as habilidades de leitura e escrita, como expostas no quadro abaixo, que apresenta o item, a tarefa, as habilidades avaliadas e o tamanho do *prompt* de cada questão.

A seguir, no quadro 13, são descritas as características de cada item, separadamente.

QUADRO 13 – PTE ACADEMIC – Visão geral da seção Auditiva

Item	**Tarefa**	**Habilidade**	**Tamanho do *Prompt***
Resumo do texto falado	Escrever um resumo de 50 a 70 palavras sobre o áudio, em 10 minutos. Há uma contagem de palavras na parte inferior da tela para orientação, e existem também botões de corte, cópia e colagem que podem ser utilizados durante a construção do resumo.	Escuta e Escrita	60-90 segundos
Questão de múltipla-escolha: selecione as respostas corretas	Responde a pergunta de múltipla- escolha após escutar a gravação. Há mais de uma resposta correta. É necessário selecionar todas as opções de resposta que julgar corretas na lista de opções possíveis. É possível marcar e desmarcar a resposta escolhida, clicando no botão esquerdo do mouse, sendo que as opções selecionadas são realçadas em amarelo.	Escuta	40-90 segundos
Preencha os espaços em branco	Ouvir e ler a transcrição de uma gravação de áudio e digitar as palavras ausentes.	Escuta e Escrita	30-60 segundos
Selecione o resumo correto	Selecionar o resumo que melhor corresponda à gravação. Há várias opções possíveis de resposta e uma correta. Por ser difícil ler e ouvir ao mesmo tempo, o próprio teste sugere ouvir primeiro, fazer anotações no bloco apagável e, em seguida, ler os resumos.	Escuta e Leitura	30-90 segundos

Questão de múltipla-escolha: escolha uma única resposta	Ouvir a gravação e responder a pergunta de múltipla-escolha. Há várias opções possíveis de resposta, mas apenas uma correta.	Escuta	30-60 segundos
Selecione a palavra ausente	Selecionar a opção mais apropriada, de uma lista apresentada, para concluir ou completar a gravação. A última palavra ou grupo de palavras na gravação são substituídos por um bipe. Há várias opções possíveis de resposta, mas apenas uma correta.	Escuta	20-70 segundos
Escrever um ditado	Ouvir e ler a transcrição da gravação de áudio que contém erros, para, então, selecionar as palavras no texto que diferem do que o orador diz.	Escuta e Leitura	15-50 segundos
Destaque as palavras incorretas	Ouvir uma frase curta e digitar a frase na caixa de resposta na parte inferior da tela, exatamente como ouvir.	Escuta e Escrita	3-5 segundos

Retirado do PTE-Academic Tutorial. Fonte: Pearson Education Ltd 2011. Disponível em: <http://pearsonpte.com/wp-content/uploads/2014/10/Tutorial.pdf>. Acesso em: jan/2017. Traduzido e adaptado pela autora.

Após essa descrição dos itens do teste, é apresentada a pontuação no *PTE Academic*.

4.5.5 Sistema de pontuação do exame

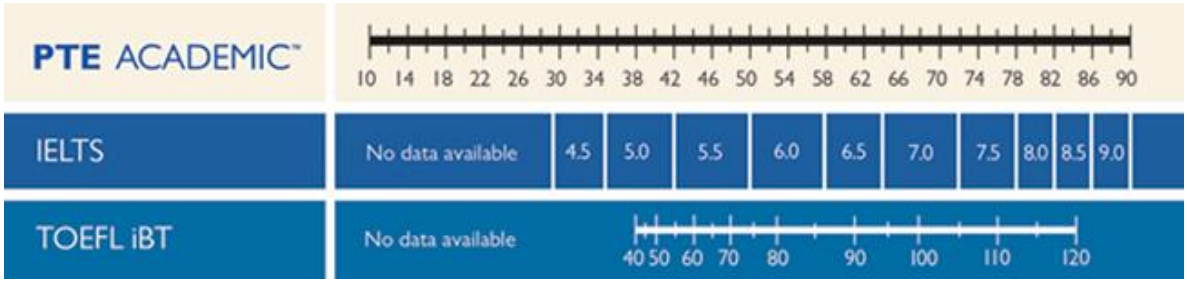
O *PTE Academic* é pontuado de acordo com a Escala Global de Inglês (*Global Scale of English*) e está diretamente relacionado com os níveis do Quadro Europeu Comum de Referência para Línguas (QECR)[40]. (CONSELHO DA EUROPA, 2001). A Escala Global de Inglês é uma escala de pontuação que vai de 10 a 90 pontos. O exame utiliza uma pontuação automatizada baseada em algoritmos complexos anteriormente testados por meio de uma amostra de dados de mais de 10.000 alunos de mais de 120 idiomas. O propósito principal da

[40] O Quadro Europeu Comum de Referência para as Línguas (QECR) foi elaborado pelo Conselho da Europa como forma de uniformizar os níveis dos exames de línguas nas diferentes regiões. É muito utilizado internacionalmente. Existem seis níveis: A1, A2, B1, B2, C1, C2. Veja a descrição dos níveis no Anexo III. Disponível em: <http://www.examenglish.com/CEFR/cefr.php>. Acesso em: jan/2016.

correção totalmente automatizada, sem intervenção humana, é de reduzir o tempo de espera do resultado pelos examinandos, disponíveis em até 5 (cinco) dias úteis após a realização do teste. Conforme informações da página oficial do exame, a medida de erro padrão entre o sistema do exame e um avaliador humano é menor do que entre um avaliador humano para outro[41].

O quadro 14 abaixo mostra a comparação entre as escalas globais de pontuação dos testes *PTE Academic*, TOEFL-iBT e IELTS.

QUADRO 14 - PTE ACADEMIC – Escala Global de pontuação.
Comparação com as escalas do IELTS e TOEFL-iBT.



Retirado de e disponível em <http://pearsonpte.com/teachers/pte-academic-scoring/>. Acesso em: jan/2017.

Em seguida é apresentada a descrição do Projeto DIALANG.

4.6 O projeto DIALANG

O Projeto DIALANG[42] é um sistema de testes de línguas computadorizados, financiado pelo programa SÓCRATES e por cerca de vinte e cinco instituições europeias, em grande parte universidades. O atual servidor de teste DIALANG começou a ser executado em 2006 e está fisicamente localizado na Universidade de Lancaster, no Reino Unido. Conforme informações da página oficial, em média, há entre 500 a 1000 sessões por dia. Em teoria, os testes DIALANG podem ser usados em qualquer computador que execute um sistema operacional que tenha um navegador *Web*, preferencialmente, o *Chrome*. Seu principal

[41] Para maiores detalhes do guia ou grade de correção das redações do exame, acesse a página: <http://pearsonpte.com/manual-essay-scoring-guide/>. Acesso em: jan/2017.

[42] Maiores informações sobre o teste DIALANG podem ser encontradas na página oficial da Universidade de Lancaster e na Wikipédia: <https://www.lancaster.ac.uk/> e <https://en.wikipedia.org/wiki/DIALANG>. O teste *online* pode ser realizado no endereço: <https://dialangweb.lancaster.ac.uk/>. Acesso em: jan/2017.

objetivo é avaliar o nível de competência linguística de acordo com o QECR. É considerado pelos desenvolvedores como um teste diagnóstico, não voltado para certificação e livre de taxas, podendo ser realizado por quaisquer aprendizes da língua alvo ou professores de línguas com os mais diversos propósitos. Esse teste, segundo sua página oficial, é útil não somente a professores, mas a instituições de ensino de línguas independentes e outras interessadas em promover a proficiência linguística entre seus funcionários, por exemplo. A praticidade do teste é uma das suas principais características. Existe uma nova versão do teste baseada na *Web* (Beta) que pode ser realizada totalmente em formato *online.* Apesar de estar em fase de teste, como informa a página oficial, esta versão encontra-se em funcionamento. Na versão mais antiga o teste pode ser baixado e instalado no computador do usuário para uso posterior.

O DIALANG avalia separadamente as habilidades de leitura, escrita, audição, gramática e vocabulário em catorze línguas diferentes (dinamarquês, holandês, inglês, finlandês, francês, alemão, grego, islandês, irlandês-gaélico, italiano, norueguês, português, espanhol e sueco), sendo as instruções oferecidas nesses idiomas. Os testes podem ser realizados na ordem de preferência do examinando, possuindo cerca de 30 questões para cada teste, a maioria de múltipla-escolha, e sem tempo limite para término, o que possibilita a consulta a materiais didáticos. O DIALANG fornece algumas ferramentas úteis ao usuário, tais como, as instruções do teste, botões de controle, páginas de ajuda, explicações, *feedback* de autoavaliação, resultados dos testes, e conselhos. O DIALANG não é um teste adaptativo. Apenas três níveis de proficiência podem ser medidos: fácil, intermediário e difícil. O interessado deve fazer um teste de nivelamento e uma autoavaliação, antes do início do teste escolhido. Se o teste de nivelamento não for realizado, o examinando receberá automaticamente o teste de nível intermediário.

4.6.1 Algumas desvantagens

Devido às limitações do *design* do teste e a custos para a elaboração do mesmo, o Projeto DIALANG ainda não desenvolveu métodos eficazes para testar a fala e a escrita. O teste utiliza uma abordagem indireta para avaliar as tarefas escritas, por isso, muitas vezes as questões de escrita se assemelham a tarefas de leitura, vocabulário ou gramática. Os usuários têm a oportunidade de completar algumas sentenças, no entanto, eles podem ser limitados a apenas algumas palavras, visto que o banco de itens de palavras destinadas às tarefas é

limitado. Se o examinando fizer um teste no mesmo idioma e mesmo nível de habilidade novamente, ele pode obter o mesmo teste que antes. Além disso, algumas tarefas aparecem em mais de um teste, pois os três testes para uma habilidade se sobrepõem em certa medida. Para um idioma específico, há apenas um teste de nivelamento, que o usuário receberá toda vez que escolher um teste na língua alvo. Da mesma forma, há apenas um conjunto de tarefas de autoavaliação para uma habilidade (leitura, audição ou escrita).

Finalmente, o Projeto DIALANG aceita apenas as respostas corretas mais comuns utilizadas durante o pré-teste. Portanto, uma resposta plausível, mas desconhecida para o sistema do teste, pode ser marcada pelo programa como errada. As desvantagens apontadas comprometem a confiabilidade do teste.

Em seguida, o teste escrito e o teste oral do ACTFL são apresentados.

4.7 Os testes do ACTFL

Segundo a página oficial do teste[43], desde 1992, o *Language Testing International* (LTI) tem sido líder em desenvolvimento de testes de proficiência linguística para mais de 100 idiomas em mais de 40 países. Como licenciado exclusivo do Conselho Americano de Ensino de Línguas Estrangeiras (*American Council on the Teaching of Foreign Language* - ACTFL), o LTI realiza milhares de testes para pequenas empresas, incluindo agências governamentais estaduais, federais e instituições acadêmicas, tais como a *Brigham Young University*, *Ohio State University*, *Defense Language Institute* e o Departamento de Segurança Interna dos Estados Unidos da América. Os testes do ACTFL são cinco: i) ACTFL *Oral Proficiency Interview* (OPI); ii) ACTFL *Writing Proficiency* (WPT): iii) ACTFL *Reading Proficiency Test* (RPT); iv) ACTFL *Listening Proficiency Test* (LTP); e v) ACTFL *Oral Proficiency Interview Computer Test* (OPIc).

Nesta dissertação, são abordados os testes de escrita e os de fala, visto que os de leitura e compreensão oral são similares em formatos e entregues no modelo linear de respostas a questões de múltipla-escolha em meio computadorizado. Iniciemos pela explicitação do teste escrito.

[43] <http://www.languagetesting.com>. Acesso em: jan/2017.

4.7.1 O teste de produção escrita

O ACTFL *Writing Proficiency Test*, ou ACTFL WPT, é um teste *online*, padronizado, e: i) avalia a capacidade global de escrita funcional do examinando, por meio da leitura de *prompts* em inglês e composição de respostas escritas na língua-alvo sem o auxílio de dicionários ou referências gramaticais; e ii) tem como referência os critérios específicos para comparar o desempenho de um indivíduo nas tarefas de comunicação de acordo com cada um dos dez níveis de proficiência, conforme descrito nas Diretrizes de Proficiência do ACTFL 2012 – Teste Escrito.[44]

O ACTFL WPT é uma avaliação estruturada em tarefas escritas que tratam de tópicos práticos, sociais e profissionais encontrados em contextos informais e formais. As tarefas e os *prompts* são articulados em LI e as respostas devem ser na língua-alvo. Cada *prompt* requer uma resposta escrita de um ou mais parágrafos, sendo uma narrativa, outra informativa e a terceira argumentativa. Uma vez que o ACTFL WPT é uma avaliação do desempenho funcional de escrita independente de qualquer currículo específico, é irrelevante quando, onde, por que e em que condições o examinando adquiriu a sua capacidade de escrita na língua. O examinando tem 80 minutos para ler as instruções e completar o teste. Os textos são, então, revisados por um avaliador certificado pelo ACTFL, que atribui a classificação que melhor descreve a proficiência demonstrada pelo indivíduo. O teste é então encaminhado para outro avaliador também certificado pelo ACTFL, para uma segunda avaliação independente.

No que se refere ao teste oral, o ACTFL possui duas versões: uma entrevista síncrona e humanizada, via telefone, e outra computadorizada em que o examinando interage com um *avatar* virtual, ambas descritas a seguir.

4.7.2 O teste oral – a Entrevista de Proficiência Oral do ACTFL via telefone

A Entrevista de Proficiência Oral do ACTFL, ou OPI, é uma conversa telefônica com duração de 20 a 30 minutos entre um entrevistador-avaliador e um examinando. O procedimento de administração é padronizado e avalia a capacidade global de fala, medindo a

[44] Para ver as Diretrizes de Proficiência do ACTFL, visite o endereço eletrônico <www.actflproficiencyguidelines2012.org>. Acesso em: jan/2017.

produção linguística de forma holística. Antes da realização efetiva do teste, o examinando responde uma série de perguntas personalizadas, para que uma amostra de discurso seja elicitada e avaliada de acordo com os níveis de proficiência e os descritores estabelecidos pela ACTFL. Por meio de variados tópicos, o principal objetivo do OPI é avaliar de forma eficiente uma amostra de fala que demonstre o nível mais elevado de desempenho do falante e o nível em que o falante não pode mais sustentar a conversa. O OPI se assemelha a uma conversa, entretanto, o entrevistador respeita um protocolo preestabelecido e estruturado de maneira específica a cada examinando.

Desta forma, o OPI avalia a proficiência linguística em termos da capacidade de uso da língua de forma eficaz e apropriada em situações da vida real. A entrevista é, assim, gravada e avaliada individualmente por dois avaliadores do OPI certificados pela ACTFL, cujas avaliações independentes são comparadas antes que uma classificação oficial seja lançada. Geralmente, as entrevistas podem ser agendadas com antecedência de alguns dias úteis e os resultados são disponibilizados em até dez dias úteis para a entrega dos certificados.

4.7.3 O teste oral – a Entrevista de Proficiência Oral do ACTFL virtual

A Entrevista de Proficiência Oral do ACTFL por computador, ou OPIc, é um teste *online* que simula um contato síncrono com o examinando. Porém, a entrega de perguntas é realizada por meio de um programa que utiliza um *avatar* virtual, assim, o teste pode ser realizado em demanda e em um momento conveniente para o examinando em um computador seguro que atenda às especificações técnicas estabelecidas pelo exame.

Segundo a página oficial do exame, a popularidade e reputação do OPIc resultou em um aumento significativo (de 4.000% em 15 anos) na demanda por esse tipo de teste, por escolas, universidades, empresas, organizações internacionais e agências governamentais.

Antes de iniciar o OPIc, os participantes concluem um levantamento e uma autoavaliação (semelhante ao do Projeto DIALANG), que fornece seis descrições diferentes de quão bem uma pessoa pode falar uma língua, a fim de reconhecer o nível de proficiência linguística inicial do examinando. Desta forma, o examinando seleciona a descrição mais precisa de sua capacidade de uso da língua. Amostras de fala acompanham cada descritor. Ele também recebe uma visão geral completa e uma explicação dos procedimentos do OPIc,

incluindo uma amostra de questões de teste. Após completar essa fase o computador seleciona aleatoriamente as perguntas, que são projetadas individualmente.

As escolhas feitas pelo examinando ao responder o questionário e a autoavaliação asseguram que cada examinando receba um teste adaptativo e único.

A primeira parte de todas as entrevistas é: "Vamos começar a entrevista agora. Diga-me algo sobre você." Esta parte funciona como um 'quebra-gelo' inicial e uma oportunidade para o examinando começar a usar a língua-alvo. A entrevista dura aproximadamente trinta minutos, é interativa e se adapta continuamente à capacidade de fala do examinador. Não há um roteiro ou conjunto prescrito de perguntas, os tópicos discutidos durante a entrevista são baseados nos interesses e experiências do examinando. Estes também são convidados a participar de um *role-play* para demonstrar funções linguísticas que não são facilmente elicitadas por meio de um formato conversacional de entrevista.

Uma vez concluído o teste OPIc, a entrevista é gravada e salva automaticamente em um sistema *online* pelo órgão responsável. A entrevista é então avaliada individualmente por avaliadores que, por sua vez, entram no programa e escutam a amostra, a fim de avaliá-la de acordo com os critérios da escala de classificação.

Na avaliação de uma amostra de fala são considerados os seguintes critérios: i) as funções e as tarefas globais que o examinando é capaz de sustentar; ii) a exatidão ou precisão com que essas tarefas são realizadas e compreendidas; e iii) o tipo de texto oral ou discurso que o indivíduo é capaz de produzir. Um teste ACTFL OPIc é atribuído a uma das seguintes classificações: Superior, Avançado Superior, Avançado Intermediário, Avançado Inferior, Intermediário Superior, Intermediário, Intermediário Inferior, Iniciante Superior, Iniciante Intermediário ou Iniciante Inferior.

...

Finalizamos as descrições de testes em meio computadorizados, apresentando de forma resumida, no quadro 15 abaixo, as oito categorias de análise desses exames, predefinidas em nosso estudo, no intuito de sumarizar as principais características dos testes, por meio de uma visualização objetiva.

QUADRO 15 – Categorias de análise dos exames descritos

CATEGORIAS DE ANÁLISE	***TOEIC***	***TOEFL-iBT***	***PTE-Academic***	***DIALANG***	***ACTFL-WPT/OPI/OPIc***
i) Construto: abstrato x pragmático x comunicativo	Construto abstrato/pragmático	Construto abstrato/pragmático	Construto abstrato	Construto abstrato	Construto comunicativo
ii) Tipo de teste: tradicional x desempenho	Tradicional e desempenho	Tradicional e desempenho	Tradicional	Tradicional	Desempenho
iii) Teste linear x adaptativo	Linear	Linear	Linear	Linear. Nivelamento adaptativo.	Linear e adaptativo
iv) Tarefas integradas x comunicativas x pontos discretos	Integradas e comunicativas	Integradas e comunicativas	Integradas e pontos discretos	Pontos discretos	Integradas e comunicativas
v) Critério de avaliação: analítico x holístico	Analítico e holístico	Analítico e holístico	Analítico	Analítico	Holístico
vi) Nível de Autenticidade e Praticidade	Autenticidade regular. Praticidade alta.	Autenticidade regular. Praticidade alta.	Autenticidade regular. Praticidade alta.	Autenticidade baixa. Praticidade alta.	Autenticidade alta. Praticidade regular.
vii) Nível de certificação /pontuação	Escala de 0 a 200 pts. Oito níveis de certificação.	Escala de 40 a 120 pts. Nível intermediário a avançado.	Escala global de 10 a 90 pts. Níveis de acordo com o QECR.	Níveis: fácil, intermediário e difícil.	Dez níveis: de superior a iniciante inferior.
viii) Artefatos tecnológicos	Uso de multimídias e sistema de correção automatizado: PLN e *SpeechRater*. Programa seguro para armazenamento de dados.	Uso de multimídias e sistema de correção automatizado: PLN e *SpeechRater*. Programa seguro para armazenamento de dados	Uso de multimídias e sistema de correção totalmente automatizado. Programa seguro para armazenamento de dados.	Sistema de correção totalmente automatizado.	Uso de sistema síncrono de comunicação, *webcam*, telefone e um programa com *avatar* virtual.

Elaborado pela autora.

4.8 Considerações parcias

Com base nas análises descritivas sobre as NTIC e as especificidades de alguns testes de proficiência em LE em meios eletrônicos, presentes neste capítulo, percebemos que a tecnologia existente se mostra apta a possibilitar que o Exame Celpe-Bras seja realizado em

um formato eletrônico ou *online*, tanto a Parte Escrita quanto a Parte Oral. Evidentemente, a praticidade de tal formato para o Celpe-Bras seria ampliada, visto que o resultado seria rapidamente disponibilizado.

Como discorrido anteriormente, uma grande quantidade de programas gratuitos e/ou pagos, de fácil manuseio e criação, destinados à elaboração de testes formativos, diagnósticos ou de nivelamento são desenvolvidos e utilizados por universidades conceituadas e instituições voltadas para esse fim, porém, devido ao formato dos itens, questões de segurança e confiabilidade do teste, a área de avaliação de proficiência linguística em larga escala necessita de programas ou sistemas mais desenvolvidos e aprimorados, pois não é confiável para os testes de alta relevância empregar algo pronto e engessado. Consideramos programas ou sistemas mais desenvolvidos aqueles que utilizam uma tecnologia mais segura, blindados contra vazamento de informação ou resultado, como os programas elaborados para testes adaptativos, que avaliam de forma individual, evitando fraudes ou cópias, por exemplo. Ou seja, sistemas que possibilitem o armazenamento seguro das respostas dos examinandos, que possam ser revistas a qualquer momento pelos avaliadores ou responsáveis pelo teste. Programas que ofereçam uma maior variedade de insumos e opção de tarefas, e não somente possibilitem a construção de testes de questões fixas de múltipla-escolha, verdadeiro ou falso, completar lacunas, respostas curtas, com opções de *feedback* limitadas. Por isso, seria interessante que fosse desenvolvido um *software* ou "um sistema próprio para alcançar a flexibilidade em especificar o alcance total dos componentes e as características almejadas" para o Exame e, assim, guiar a correção e a entrega dos resultados desejados. (CHAPELLE e DOUGLAS, 2006. pp. 886). No caso de uma versão virtual do Celpe-Bras, para que isso ocorra, são necessárias outras pesquisas, especialmente, de cunho experimental, além do interesse do órgão responsável e do envolvimento conjunto de especialistas de diversas áreas, tais como, engenheiros de *softwares* ou programadores, especialistas em linguística aplicada à avaliação e *designers*, além de recursos humanos e aparatos tecnológicos, maquinários de qualidade e em funcionamento e recursos financeiros. Assim, de acordo com o propósito do teste, os elaboradores poderão decidir quais as ferramentas ou as especificações ideais para a construção de uma versão computadorizada ou *online*. Essas especificações de teste e os aspectos essenciais a serem considerados na construção ou desenvolvimento de um teste em meio computadorizado são resumidos por Chapelle e Douglas (2006) da seguinte forma:

> Uma avaliação de alta relevância requer um conjunto de especificações de tarefas de teste, um procedimento para selecionar as tarefas específicas a partir de um conjunto de tarefas desenvolvidas para estas especificações, um mecanismo para a entrega individual de perguntas de teste, um meio de entrada individual de respostas, um método para coleta e análise das respostas, e um meio de armazenar os resultados para uma análise subsequente. [...] Os tipos de ferramentas selecionadas, obviamente, dependem da finalidade do teste e de uma série de questões práticas, tais como seu custo, disponibilidade e adequação para fazer o trabalho. [tradução nossa][45]. (Chapelle e Douglas, 2006. pp. 752-753).

No que se refere às mudanças do Celpe-Bras impresso para um formato eletrônico, como vimos no capítulo 2, o Exame depende, totalmente, de seus avaliadores, tanto na Parte Oral, quanto na Parte Escrita. Esses avaliadores são previamente treinados e, apesar de ser uma avaliação subjetiva, os corretores seguem a grade de correção de ambas as partes, durante o processo de avaliação. O resultado, então, é reproduzido em conceitos, de acordo com os cinco níveis de certificação, e uma nota única é estabelecida. Além disso, a aplicação da Parte Oral - a interação face a face - difere de examinando para examinando, já que os temas, ou os EP, e as questões do roteiro da entrevista são todos selecionados por cada dupla de avaliadores (o entrevistador e o observador), de maneira individualizada a cada examinando, podendo ou não haver a repetição de EP anteriormente utilizado. Enquanto que nos testes europeus, como os testes do PERSON e do ACTFL, existem listas de conteúdos preestabelecidos para estudo, fornecido aos examinandos, e um teste específico para cada nível de certificação ou propósito: para fins acadêmicos ou para negócios, por exemplo. Para a elaboração e o desenvolvimento de um formato eletrônico e/ou *online* do Celpe-Bras, principalmente no que concerne à aplicação do teste, essas diferenças apontadas podem ser levadas em consideração, criando-se testes específicos a cada nível ou de acordo com o interesse do examinando, seja acadêmico ou profissional. Além disso, tendo em mente as especificidades de teste, a elaboração de testes voltados para populações específicas possibilita a criação de tarefas que simulem o uso real da língua-alvo de forma autêntica e interativa, utilizando ferramentas tecnológicas disponíveis, como, por exemplo, os dispositivos síncronos de comunicação, *podcastings*, *videocastings*, videoconferência,

[45] A high-stake assessment, in contrast, requires a set of test-task specifications, a procedure for selecting particular tasks from a pool of tasks developed to these specifications, a mechanism for individual delivery of test questions, a means of individual entry of responses, a method for collating and analyzing the responses, and a means of storing results for subsequent analysis. […] The types of tools selected, of course, depend on the purpose of the test and a number of practical issues such as their cost, availability, and adequacy for doing the job.

multimídias audiovisuais, no intuito de gerar a disrupção do modelo tradicional em prol do avanço e da melhoria dos testes de língua.

O Exame Celpe-Bras é um teste de desempenho, que em prol da praticidade e maior alcance, deve buscar uma versão computadoriza/*online*, como o fez o TOEFL. E essa mudança requer estudos e pesquisas pré e pós a aplicação da nova versão de teste. Como abordado neste capítulo, muitos trabalhos entre 1996 e 2016 voltaram-se ao estudo da validade das inferências baseadas na pontuação e às mudanças que podem ser geradas nos resultados do teste a partir da transição de uma versão em papel e caneta para o formato computadorizado. Segundo Chapelle e Douglas (2006), pesquisadores da área descobriram poucas ou leves diferenças entre um teste impresso e um teste computadorizado, o que, na visão dos autores, não interfere diretamente na interpretação dos resultados feitos a partir da comparação entre os dois formatos. Entretanto, esses testes analisados comparativamente, muitas vezes, são testes transpostos do meio impresso para o meio computadorizado, ou seja, seu conteúdo, formato e forma de avaliar permanecem inalterados de um meio para outro. Esse tipo de atitude é chamado de transposição. Ela ocorre também com materiais didáticos impressos que são usados para ensino à distância. Trata-se de um equívoco que não pode se repetir na área dos testes de proficiência *online.*

Ora, se resultados semelhantes são esperados, podemos questionar se realmente vale ou não a pena criar uma nova versão de um teste anteriormente estabelecido em formato impresso. Aí está o perigo de apenas transpor o exame impresso, sem nenhuma modificação, para o computador, pois se o meio de entrega muda, mudanças na concepção do uso da língua ou na forma de interpretar os resultados, por exemplo, consequentemente, serão geradas. Neste caso, a praticidade que defendemos ser uma qualidade essencial em testes deve ser repensada, para que a validade de construto do teste não seja comprometida, pois se a finalidade for apenas a transposição, podemos indagar: i) até que ponto pode ser útil e válido a busca por testes mais eficientes e práticos, porém, sem nenhuma inovação ou novos métodos para avaliar as habilidades e as competências do examinando?; e ii) se o meio de entrega do teste muda, não deveria este apontar a um desempenho diferente e, portanto, ser pensado com um construto diferente, específico dentro do que se pretende medir?

Queremos esclarecer que consideramos a ampliação da praticidade a maior justificativa para a construção de um teste computadorizado ou *online*, por este gerar menos custos no longo prazo, entre outros benefícios. No entanto, entendemos que todos os seis critérios de teste, explicitados no capítulo 3 deste estudo, devem ser considerados na

elaboração de um teste em meio computadorizado, de maneira equilibrada. Assim, as NTIC devem ser utilizadas para favorecer ou facilitar a construção de tarefas que possibilitem uma avaliação apurada ou ofereça informações precisas a respeito do desempenho do examinando, de acordo com as habilidades ou competências que o teste se propõe a medir, em comparação com um teste impresso. Em outras palavras, é preciso um entendimento conciso tanto da natureza das tarefas de testes, quanto da correlação entre a filosofia do teste e o método de pontuação. As tecnologias, portanto, devem ser utilizadas a favor da concepção de língua em uso e não apenas como mais um meio ou forma de avaliar. Tal atitude abre espaço para a inovação na área. Logo, mudanças precisam ser incentivadas e apoiadas por avanços conceituais na compreensão da LE e no uso e aprendizado da mesma.

...

Neste capítulo, apresentamos as NTIC utilizadas na elaboração de testes de proficiência linguística, lineares e adaptativos, e descrevemos cinco exames, suas principais características e uso das tecnologias para a avaliação do desempenho dos examinandos. Percebemos que os testes adaptativos, apesar de serem uma tendência e uma promessa de maior praticidade na aplicação e correção, nem sempre são considerados de alta relevância ou de larga escala, mas com maior frequência para diagnósticos e nivelamento. Os testes que possuem versões impressas e computadorizadas como o TOEFL e o TOEIC buscam em suas versões a praticidade e a precisão na avaliação de seus usuários, com propósitos predefinidos. Com exceção do Projeto DIALANG, que é um teste diagnóstico gratuito, todos os testes descritos são oferecidos por instituições privadas e, portanto, visam ao lucro, por isso, quanto mais rápida a aplicação e a entrega de resultados, melhor para a organização responsável. Podemos, ainda, elencar alguns pontos observados nos testes aqui descritos. Por exemplo, as questões fixas de múltipla-escolha, *cloze*, completar lacunas, pequenos resumos, repetição de sentenças, respostas curtas e o uso de multimídias audiovisuais são recursos muito utilizados em testes computadorizados. A parte oral e a escrita seguem sendo partes diversificadas e complexas para medição do desempenho do examinando, sendo, muitas vezes, aplicadas separadamente da parte de leitura e audição e escrita, com questões ou tarefas integradas, porém, nem sempre autênticas. No ACTFL OPI e OPIc, a Parte Oral é avaliada por meio de entrevista via telefone e por um *avatar* virtual, respectivamente, por meio de tarefas comunicativas, simulando ações da vida real, algo parecido com o que é feito na interação

face a face do Celpe-Bras, porém com menos interatividade. Outro fato observado é o predomínio da correção totalmente humanizada ou em conjunto com a correção automatizada, com exceção do DIALANG e do *PTE Academic* que utilizam apenas a correção automatizada. Além disso, por questão de segurança e confiabilidade, os testes de alta relevância são, geralmente, aplicados em centros credenciados.

No próximo capítulo, são apresentadas a análise do estudo e as propostas para a elaboração de uma nova versão computadorizada e/ou *online* do Exame Celpe-Bras, baseado em aspectos discutidos nos capítulos anteriores.

CAPÍTULO 5

ANÁLISE E PROPOSTAS: CELPE-BRAS *ONLINE*

Versions of the test specification constitute the history of test evolution, and act as a record of decisions taken over time to further define and improve the measurement constructs.

Glenn Fulcher

Neste capítulo, com o objetivo de sistematizar os capítulos anteriores, retomamos inicialmente a pergunta de pesquisa para, em seguida, apresentar a análise do estudo juntamente com as propostas para um formato *online* ou computadorizado do Exame Celpe-Bras. São apontadas algumas possibilidades julgadas viáveis, de acordo com este estudo, para a elaboração desse formato, bem como algumas vantagens e questões problemáticas a serem consideradas pelos responsáveis e pelos desenvolvedores do Exame. A intenção não é definir um padrão fixo e/ou pronto, entendido como a versão adequada do Exame, mas sim apresentar as possibilidades frente à tecnologia oferecida aos testes de LE. As NTIC trazem avanços e alternativas variadas para a área de avaliação em LE, resta decidir quais delas seriam úteis de acordo com o propósito e a filosofia do exame, também entendida como construto de validade que o define.

Ressaltamos que as proposições aqui apresentadas referem-se, especialmente, à aplicação *online*/computadorizada do Exame. Apesar de abordamos todo o processo do Exame, pois as etapas de desenvolvimento de um teste são indissociáveis, este estudo não tem o objetivo de propor soluções para a correção em meio virtual. Este tema necessitaria um estudo específico e minucioso, desdobrando-se em outra pesquisa. Assim, apontaremos algumas direções e possibilidades para a correção, sem, no entanto, aprofundar nesta questão.

5.1 Retomada da pergunta de pesquisa

Considerando que o Celpe-Bras possui apenas a versão impressa, nos desafiamos a abordar um tema pouco discutido na área de avaliação de proficiência linguística no Brasil: as NTIC em testes de alta relevância. Desta maneira, procuramos investigar neste estudo:

- ✓ Como utilizar as NTIC existentes na área de avaliação de proficiência linguística, para a elaboração e aplicação de uma versão eletrônica/*online* do Celpe-Bras?

A pesquisa aponta para várias direções em busca de respostas a essa questão. Desta forma, na próxima seção, apresentamos as possibilidades e propostas para a utilização das NTIC no Exame Celpe-Bras de acordo com nossa pesquisa.

5.2 O Celpe-Bras em versão eletrônica/*online*: algumas propostas

O principal objetivo do Exame Celpe-Bras é medir o domínio do uso da LP pelo examinando, por meio de tarefas comunicativas de uso real da língua. Assim, a utilidade do teste está em seu propósito específico, e pretende estar baseado nos seis critérios fundamentais que devem guiar o desenvolvimento de qualquer exame de alta relevância, explicitados no capítulo 3 desta dissertação, a saber: *o construto, a validade, a confiabilidade, a praticidade, a autenticidade* e *o efeito retroativo.* Em um meio computadorizado, esses critérios tendem a permanecer como fundamento de qualquer teste de larga escala, não diferente com o Exame Celpe-Bras, que tem sido considerado ousado e inovador, se comparado aos testes lineares, computadorizados ou não, como os descritos neste trabalho. Além disso, outras questões como a *segurança*, o *espaço*, o *tempo de duração e o tamanho do teste*, os *custos* e os *aparatos tecnológicos* devem ser, da mesma forma, criteriosamente definidos em testes eletrônicos. Desta forma, a partir do estudo realizado nos capítulos anteriores, observamos e julgamos essencial que o processo de elaboração, desenvolvimento, administração e correção do Celpe-Bras em meio eletrônico ou *online*, primeiramente, seja norteado pela '*Lista de verificação para avaliar a utilidade de testes*' e pela '*Lista para descrever as tarefas de uso da língua alvo*', ambas de Bachman e Palmer (1996. pp. 150-155 e pp. 108, respectivamente), ver ANEXO IV e V. A primeira compreende uma série de questionamentos a serem respondidos pelos elaboradores de testes a respeito dos seis critérios ou

qualidades de utilidades de teste, previamente discutidos no capítulo 3 deste estudo. A segunda compreende uma lista que engloba as qualidades das tarefas que compõem o teste, como as características: i) da configuração ou local do teste; ii) do *input* de teste (formato, canal, língua que será administrado e tipo de questão); iii) da resposta esperada; e iv) relação entre o *input* e a resposta. Além dessas duas listas, propomos a verificação da '*Lista de ameaças à validade e às respostas*' e a '*Lista dos aspectos positivos e negativos do CALT, relacionados à tecnologia*', de acordo com os mesmos seis critérios ou qualidades acima mencionados, ambas de Chapelle e Douglas (2006. pp. 741 e pp.1007-1008, respectivamente), ver ANEXOS VI e VII[46]. Consideramos primordial que os desenvolvedores ou responsáveis pelo Exame Celpe-Bras estejam atentos a essas especificidades antes, durante e depois da testagem final.

Em seguida, frente a tantas possibilidades e desafios, estabelecemos, no quadro abaixo, quatro questões essenciais que esperamos servirem como norte aos desenvolvedores e elaboradores do Exame em uma nova versão eletrônica/*online*. Essas questões são baseadas nos questionamentos de Corbel (1993, pp. 53 *apud* CHAPELLE, 2006. pp. 182-185) e adaptadas ao nosso estudo a respeito das NTIC e do Exame Celpe-Bras.

QUADRO 16 - Questionamentos para elaboração e aplicação do Exame Celpe-Bras em meio computadorizado.

QUESTIONAMENTOS PARA ELABORAÇÃO E APLICAÇÃO DO EXAME CELPE-BRAS *ONLINE*/ELETRÔNICO	
I.	**Quais são as mudanças específicas implicadas pelo uso da tecnologia em um Exame Celpe-Bras *online*, na administração, na aplicação e na correção e entrega de resultados?**
II.	**O computador pode fornecer os meios para apresentar técnicas de teste inovadoras, mantendo a confiabilidade e validade do Exame?**
III.	**O construto e a validade do teste permaneceriam o mesmo?**
IV.	**O conceito de testes baseados em tarefas comunicativas, seguido pelo Celpe-Bras, pode ser operacionalizado pelo computador? Quais tarefas mudariam ou seriam adaptadas ou criadas?**

Adaptado pela autora a partir dos questionamentos de Corbel (1993, pp.53 apud CHAPELLE e DOUGLAS, 2006. pp. 182-185).

[46] As quatro listas aqui mencionadas foram mantidas na língua original (LI). Optamos por colocar essas listas em anexo devido à extensão das mesmas, sem, no entanto, diminuir sua importância em nosso estudo.

Diante das questões acima elaboradas, apresentamos abaixo nossas proposições de respostas, de acordo com o estudo realizado até aqui.

I. **Quais são as mudanças específicas implicadas pelo uso da tecnologia em um Exame Celpe-Bras *online*, na administração, na aplicação e na correção e entrega de resultados?**

Após a descrição das NTIC e os testes de proficiência que as utilizam, apresentados no capítulo anterior, é possível destacar as principais mudanças entre um teste impresso e um computadorizado. Tais mudanças devem ser consideradas pelos desenvolvedores de testes eletrônicos e se relacionam, principalmente, ao método de teste e suas características, que fazem parte do processo de elaboração e operacionalização de um teste de LE. Para entender melhor o que seria **método de teste**, utilizamos os conceitos de McNamara (2000) e Chapelle e Douglas (2006) ao verificarem os aspectos de mudanças geradas.

Segundo McNamara (2000. p. 23-33), a elaboração de um teste é um processo cíclico, pois a qualquer momento podem surgir novas situações tais como, mudanças sociais, culturais e/ou políticas e educacionais, que podem levar à necessidade de revisão e reavaliação de todo o processo. Esse processo cíclico conta com quatro estágios ou fases principais, a saber: o *design* ou fase de concepção do teste, a construção do teste, a experimentação ou pré-testagem do teste e a administração do teste. Na fase de *design* do teste, os desenvolvedores de testes devem ter em mente a finalidade ou propósito do teste, além do conteúdo e o método do teste. A visão de língua e língua em uso contida no teste determinará todo o processo de concepção e, consequentemente, a interpretação da pontuação final. Logo, torna-se essencial reconhecer as principais diferenças de um teste eletrônico, pois a tecnologia, por tornar possível diferentes formas de tarefas, pode afetar as interpretações de desempenho do examinando no teste de LE.

Assim, definir o conteúdo implica na visão do construto do teste, que pode ser uma visão operacional/prática da língua, avaliando o desempenho do examinando de forma integrada e comunicativa, ou um construto abstrato, seguindo a teoria do conhecimento linguístico (gramatical, léxical, sintático, etc) e/ou das habilidades de leitura, escrita, fala, escuta, avaliadas separadamente, por exemplo. Definir o método de teste corresponde ao formato de resposta e relaciona-se com a autenticidade do conteúdo e das tarefas de teste. Assim, se o teste tiver uma visão pragmática da língua, os desenvolvedores de teste devem, primeiramente, identificar as tarefas comunicativas, por meio da análise criteriosa de material

autêntico, a fim de determinar que tipo de texto deve ser utilizado e como os examinandos vão interagir com ele, além de estabelecer, previamente, o tamanho e o formato de resposta esperada, as instruções e os critérios para avaliação do desempenho. Além desses, outros aspectos também devem ser considerados durante todo o processo de construção do teste, tais como: i) as restrições e as questões práticas, os recursos financeiros, humanos e aparatos tecnológicos necessários e disponíveis para a elaboração e operacionalização do teste; e ii) a segurança de conteúdo do teste e armazenamento das respostas dadas pelo examinando. Todas estas questões, definidas por McNamara como **especificações de teste**, são essenciais para o processo de construção de testes, seja em meio computadorizado ou impresso.

Desta forma, Chapelle e Douglas (2006. p. 284-479) afirmam que as principais mudanças de um meio para outro estão no método de teste que, por sua vez, se divide em cinco aspectos: i) as circunstâncias físicas e temporais; ii) as instruções; iii) a entrada de resposta e a resposta esperada; iv) a interação entre a entrada e a resposta esperada; e v) os aspectos de avaliação. Os autores exemplificam essas especificações de teste e estabelecem uma comparação entre as vantagens e as limitações dessas características do método de testes em meio computadorizado. Essa comparação pode ser encontrada no Anexo VIII deste estudo e, com base nessa comparação estabelecemos no quadro abaixo, sucintamente, as principais mudanças no método de teste entre um teste computadorizado/*online* e um teste impresso.

QUADRO 17 – Principais mudanças entre um teste impresso e um teste computadorizado/*online*.

MÉTODO DE TESTE	**TESTE IMPRESSO**	**TESTE COMPUTADORIZADO/*ONLINE***
i. Circunstâncias temporais e físicas	Teste realizado em postos aplicadores fixos, em data e horário previamente marcados.	Possibilidade de realizar os testes *online* em quaisquer lugares, ou computador, e em qualquer período ou data, sem a necessidade de presença de instrutores humanos. Rapidez na aplicação.
ii. Instruções de resposta	Tarefas e instruções impressas em papel e apresentadas por aplicadores treinados.	Uniformidade e consistência nas instruções e na forma de apresentação das tarefas na tela do computador, sem a necessidade da presença de um instrutor humano.
iii. Texto de entrada e Resposta esperada	Recursos limitados para a elaboração de tarefas, que podem ser contextualizadas ou descontextualizadas, a depender da escolha de autenticidade de conteúdo do teste e do formato das questões.	Possibilidade de uso de multimídias diversas, vídeos, áudios, hipertextos, contextualizados e autênticos, em tarefas e tipos de respostas variadas.

iv. Interação entre a entrada e a resposta esperada	Formato de tarefas fixo e linear e maior tempo necessário para *feedback*.	Possibilidade de uso de testes lineares e/ou adaptativos que permitem adequar a resposta do examinando ao seu nível, e rapidez de *feedback*.
v. Aspectos avaliativos: correção e pontuação	Correção humanizada e necessidade de mais recursos materiais e humanos e maior tempo para a entrega de resultados.	Correção e pontuação automatizada de respostas complexas, por NLP, *SpeechRater*, por exemplo, e rapidez na correção e entrega de resultados. Segurança de armazenamento de dados. Necessidade de adequação dos critérios de correção e definição do construto do teste.

Elaborado pela autora e adaptado do trabalho de Chapelle e Douglas (2006. p. 327).

Seguimos para a próxima questão.

II. **O computador pode fornecer os meios para apresentar técnicas de teste inovadoras, mantendo a confiabilidade e validade do Exame?**

Apesar de ser uma questão complexa e que necessita de pesquisas experimentais, o que percebemos, diante do estudo dos testes estrangeiros que utilizam as NTIC, é que as ferramentas tecnológicas disponíveis aos testes podem servir como um meio para inovação do Celpe-Bras, mantendo-se, ao mesmo tempo, a confiabilidade e validade do Exame. Entendemos que a inovação não precisa estar presente em todo o processo de desenvolvimento do teste. Ela pode surgir devagar, em detalhes e aspectos mínimos, seja na aplicação, correção ou administração do teste. A inovação a qual nos referimos consiste na utilização das NTIC em prol do desenvolvimento de testes que avaliem o desempenho do examinando de forma integrada, interativa e pragmática, considerando os aspectos socioculturais e os sociocomunicativos envolvidos nesse processo, em um determinado contexto. Um teste elaborado pensando-se no impacto que causará no sistema de avaliação, de ensino e de aprendizagem, que utilize ferramentas adequadas ao construto definido e, finalmente, traga praticidade na administração, na correção, na pontuação e na entrega de resultados. Um teste que, de acordo com seu propósito, inove na elaboração e *design* de tarefas comunicativas que consigam extrair do examinando o conhecimento de uso da língua-alvo, observado por meio de seu desempenho final. Enfim, consideramos inovador aquele teste de desempenho integrativo, pragmático e comunicativo aliado à tecnologia. Entretanto, tal inovação é onerosa, difícil e requer interesse, tempo e recursos materiais e humanos para

alcançá-la. Todavia, a tecnologia, se utilizada de forma adequada, pode ser uma grande aliada nessa árdua tarefa.

Como vimos no capítulo 3, comparar testes computadorizados com testes impressos foi tema de muitos trabalhos na área de avaliação, visto que diferentes contextos de uso da língua requerem diferentes métodos, inclusive de correção e pontuação. No entanto, muitos desses trabalhos buscavam demonstrar, por meio da comparação, a maior eficiência dos testes computadorizados frente aos de caneta e papel e, portanto, não enfatizavam a importância da busca pela inovação dos mesmos. Muitas vezes, utilizar o meio eletrônico ou *online* para aplicação de testes não é o suficiente para rotular o teste de **inovador** e **tradicional**. Por exemplo, um teste computadorizado linear pode ser considerado tradicional se apresentar apenas questões que avaliem os aspectos de conhecimentos da língua (gramatical, lexical, etc.) ou as habilidades cognitivas (leitura, compreensão auditiva), de forma separada, fora de um contexto de uso, o que muitas vezes se faz por meio de questões de múltipla-escolha. Por outro lado, se um teste computadorizado utiliza tarefas integradas e comunicativas para medir o desempenho do examinando de forma contextualizada, por meio de diferentes tipos de questões e recursos multimídias, por exemplo, pode ser considerado um teste de desempenho não tradicional, quiçá, inovador. O que falta na área são pesquisas que procurem enxergar nas novas tecnologias de informação e comunicação a oportunidade de utilizar ferramentas adequadas, que não somente tornem o teste mais eficiente, mas também inovador, seja na elaboração, *design*, administração ou correção. Chapelle e Douglas entendem o uso da tecnologia como um método diversificado, porém, ainda não revolucionário na área de avaliação:

> [...] a realidade do CALT [*Computer-Assisted Language Test*] hoje não é o que se poderia chamar de revolução na avaliação de língua. Certas características dos métodos do CALT são substancialmente diferentes das dos testes envolvendo outros meios de entrega e resposta, mas a tecnologia não tem reconfigurado radicalmente o papel da avaliação no ensino e na aprendizagem. Até agora, vimos o CALT como uma evolução na avaliação, ampliando o que fazemos em testes, ao invés de uma revolução, mudando a avaliação em relação à educação e à pesquisa de línguas. (*Ibid*, 2006. pp. 1161). [tradução nossa][47].

[47] […] the reality of CALT today is not what one could call a revolution in language assessment. Certain characteristics of CALT methods are substantively different from those of tests involving other means of delivery and response, but technology has not radically reconfigured the role of assessment in teaching and learning. Thus far we have seen CALT as an evolution in assessment, expanding what we do in testing, rather than a revolution, changing what assessment is in relation to language education and research.

Digamos que é preciso mais que a eficiência e a praticidade, é preciso inovar. Essa inovação não se refere a aparatos tecnológicos apenas, pois como vimos, esses se encontram em constante evolução e as ferramentas atuais mostram-se úteis na elaboração de testes eletrônicos. No entanto, é preciso pensar no desempenho do examinando que se pretende alcançar, demonstrado de forma prática, por meio de simulações, de tarefas de uso real da língua, estabelecidas dentro de um contexto situacional ao qual o examinando é exposto no momento do teste. Estamos falando do que Chapelle e Douglas e outros especialistas chamam de construto de contexto, em que o avaliador levará em conta não só o fator cognitivo ou competência linguística, mas também a competência tecnológica, estratégica e o fator contextual no desempenho final do examinando. Entender, a partir de múltiplas perspectivas, como o uso da tecnologia pode estar a serviço do construto do teste e escolher quais as tecnologias mais apropriadas para cada situação de teste, eis o desafio para a nova versão do Exame Celpe-Bras.

Prosseguimos, assim, para a próxima questão.

III. **O construto e a validade do teste permaneceriam o mesmo? Quais são as possibilidades de uso da NTIC caso o construto seja mudado?**

Antes de responder essas questões é importante lembrar que as NTIC apresentam relevantes mudanças para os professores de LE, os usuários de testes de proficiência, os elaboradores ou desenvolvedores e, também, para os pesquisadores da área. Utilizar a tecnologia em testes de proficiência linguística influencia diretamente o desempenho do examinando e difere, singularmente, de testes impressos, por áudio e vídeo, por telefone ou entrevistas face a face. Segundo Chapelle e Douglas (2006), as principais áreas para entendimento e estudo da criação de uma versão eletrônica ou *online* de um teste, geralmente, referem-se à definição do(s) construto(s) do exame e à exploração das questões de validação dos testes.

Em relação ao construto, concordamos que a maneira como o teste é definido trará consequências em termos de pontuação final. (BACHMAN e PALMER, 1996. pp. 117). Portanto, os elaboradores devem estabelecer, claramente, como e quais tarefas e conteúdos serão utilizados na construção do teste, além de decidir se as competências estratégicas do examinando serão previstas no construto final do exame. Ou seja, é necessário definir em detalhes o construto, ao invés de simplesmente determinar, por exemplo, que apenas a habilidade de leitura e interpretação de texto será avaliada em um teste de leitura. Ora, torna-

se necessário pensar nos aspectos da leitura que serão medidos: estrutura gramatical, lexical, coesão textual, adequação ao gênero proposto, além de questões de *design* como a estipulação ou não do tempo de leitura, definir quais comandos o examinando terá acesso durante o teste, e assim por diante. Outro exemplo: consideremos um teste de audição/escuta. O simples fato de decidir se o examinando ouvirá duas ou apenas uma vez o áudio ou se esse áudio vai ser processado em vídeo, com imagens, ou permitirá anotações enquanto ouve e/ou apresentará um botão de ajuda para esclarecimento das instruções do teste são detalhes que dizem respeito à operacionalização do construto de teste, durante a elaboração do mesmo. Assim, é preciso questionar se tais aspectos comprometerão ou não a validade do construto em prol do aumento de autenticidade e de praticidade do teste.

Desta forma, sendo o Celpe-Bras um exame diferenciado em relação ao seu construto baseado em tarefas comunicativas, ao se criar um formato computadorizado ou *online*, defendemos que o ideal é que se mantenha o construto atual do teste e que seja criado um *software* próprio, exclusivo, elaborado de acordo com os principais fundamentos do teste, em busca de uma avaliação mais justa e inovadora, tanto na aplicação quanto na correção. Consideramos que o Exame já é por si só ousado e inovador, pois busca avaliar o desempenho do examinando de forma holística e analítica, por meio de tarefas integradas que simulam o cotidiano da vida dos brasileiros. O formato do teste não apresenta questões consideradas tradicionais, como múltipla-escolha, *cloze*, respostas curtas, mas utiliza conteúdos contextualizados retirados de fontes autênticas. Assim, entendemos que a visão subjacente ao Celpe-Bras é, em sua maior parte, pragmática, avaliando tanto os aspectos pragmáticos, quanto os socioculturais e os discursivos, e não apenas o nível de entendimento do examinando de sintaxe e de semântica da LP, separadamente. Portanto, para que haja a manutenção do construto do Exame, é necessário um programa que assegure condições para que as tarefas permaneçam com seu caráter interativo e comunicativo, possibilitando a junção de multimídias, áudio, vídeo, textos com linguagem verbal e não-verbal, ferramentas de ajuda e instrução ao examinando, botões de repetição (caso seja permitido), além dos cuidados com as questões de segurança e armazenamento das respostas, assim como a facilidade na correção e entrega de resultados.

A validação de um teste, por sua vez, é totalmente dependente do que o teste pretende medir, qual a sua proposta, a que tipo de grupo é destinado e como será administrado. Desta forma, pensar de maneira minuciosa na validação de um teste torna-se algo imprescindível durante a elaboração de qualquer exame de larga escala. Em outras palavras, é preciso que os

desenvolvedores de testes e os pesquisadores da área entendam melhor o que muda em meio eletrônico, quais os aspectos ou habilidades linguísticas podem ser medidas por meio do computador e o que se espera do examinando ao realizar um teste computadorizado. Essas mudanças referem-se, principalmente, ao método de teste, às circunstâncias físicas e temporais, às condições de teste, às instruções de teste, ao texto de entrada de resposta e resposta esperada, à interação entre entrada e resposta e às características de avaliação do desempenho e pontuação, conforme apresentadas na questão I do quadro 17, anteriormente, neste capítulo.

Em contrapartida, caso haja mudança no construto do teste, há a possibilidade de se desenvolver um formato CAT do Exame Celpe-Bras, não linear, o que pode ser interessante, caso seja considerada a busca pela inovação discutida anteriormente. O formato CAT parece adequado ao Exame se considerarmos que este modelo pode fornecer dados relevantes para avaliar as subabilidades, tais como: o tempo gasto em cada item ou tarefa, que pode ser registrado e utilizado como um indicador da automaticidade ou fluência dos examinandos. Assim, o tempo gasto nesse modelo de teste não é desperdiçado com itens muito fáceis ou muito difíceis. Já no caso de testes lineares, se os itens ou tarefas forem mal elaborados, de acordo com o nível de proficiência a ser medido, o teste corre o risco de não fornecer informações de medição ou resultados confiáveis. (CHAPELLE e VOSS, 2016. p.118).

Porém, baseado no construto presente na versão impressa do Exame, este formato adaptativo já utilizado em testes internacionais deve ser repensado e/ou reestruturado de acordo com o construto do Celpe-Bras. Apesar da flexibilidade de tempo e espaço, a praticidade na administração, a entrega rápida de resultado e do fato de a grande maioria das pessoas que utilizam testes estarem mais confortáveis ao utilizar um computador, a escolha por um CAT pode diminuir a validade de conteúdo do exame, por utilizar poucas questões para medir o desempenho do examinando de maneira individualizada. Além disso, a praticidade na aplicação e na correção pode não compensar o alto custo da construção desse formato, além do tempo necessário para a pré-testagem das questões que comporão o banco de itens que, por sua vez, deve ser grande e diversificado o suficiente para não gerar repetição constante, contendo tarefas ou questões de todos os níveis a serem medidos. Parece, então, que o ideal, no caso do Exame Celpe-Bras seria a construção de um MCAT, retratado no capítulo anterior. Porém, este formato ainda encontra-se em desenvolvimento e testagem.

Finalmente, seguimos para o último questionamento.

IV. O conceito de testes baseados em tarefas comunicativas, seguido pelo Celpe-Bras, pode ser operacionalizado pelo computador? Quais tarefas mudariam ou seriam adaptadas ou criadas?

Entendemos que o conceito de tarefas comunicativas do Exame pode sim ser operacionalizado pelo computador em uma nova versão *online*, entretanto, para que isso aconteça, alguns cuidados são necessários. Como discutido anteriormente, o Exame Celpe-Bras foi criado com uma natureza comunicativa e possui diferentes tipos de tarefas que procuram se assemelhar a situações da vida real. Tanto a Parte Escrita quanto a Parte Oral procuram avaliar a habilidade do examinando em usar a LP na resolução de problemas cotidianos. Essa natureza comunicativa não deverá ser alterada no formato eletrônico. Embora a utilização de questões de múltipla-escolha torne a correção do Exame mais prática, questões como as utilizadas no DIALANG, por exemplo, ou itens curtos demais podem limitar a interatividade das tarefas e comprometer não apenas a autenticidade de conteúdo, mas também a validade e a confiabilidade do construto. Portanto, seria interessante manter a utilização de diferentes tipos de tarefas, para que o examinando seja capaz de mostrar suas diversas habilidades linguísticas e o seu desempenho ser avaliado de forma mais justa. Prognosticar o potencial de uso da língua-alvo em diversas situações comunicativas requer diversidade não apenas do formato do teste, mas também da forma como as tarefas são apresentadas e da diversidade de conteúdo e itens do teste.

Para responder de forma sistemática quais tarefas mudariam, seriam adaptadas ou criadas, e explicitar as propostas apresentadas, a Parte Escrita e a Parte Oral e suas respectivas correções são apresentadas em momentos separados nas próximas subseções.

5.2.1 A Parte Escrita do Celpe-Bras *online*

A Parte Escrita do Exame Celpe-Bras na versão impressa mede a produção escrita e a compreensão oral e escrita do examinando de forma integrada, e é estruturada em quatro tarefas como destacado no capítulo 2 deste trabalho. Ora, se apenas colocarmos essas quatro tarefas, de forma linear, em um programa de computador destinado à avaliação de proficiência de língua, diferenciando somente o meio de aplicação, haverá alteração apenas na praticidade de administração do teste. Apesar de parecer o caminho mais prático e seguro, fazer uso da tecnologia visando apenas à eficiência e à evolução do mesmo, não o tornará uma avaliação melhor, nem mesmo provocará inovação na área. Tendo em mente que o Exame é

voltado a qualquer pessoa, maior de 16 anos, falante não-nativo de LP, e que o propósito do exame é avaliar o desempenho do examinando por meio de tarefas comunicativas e habilidades integradas, é importante se certificar que os itens, as tarefas ou as questões estejam adequados para a administração de um teste em meio computadorizado. Para isso, as seguintes reflexões podem ser de interesse dos elaboradores de itens: O que será exatamente medido ou avaliado? E como será avaliado? Qual conteúdo abordar? O número de tarefas continuaria o mesmo? Se sim, o que justificaria? Quais tipos de tarefas seriam selecionados?

Entendemos que o formato CAT para um possível formato *online* do Celpe-Bras surge como uma opção para, de forma prática e rápida, adequar o Exame de acordo com o nível de cada examinando, individualmente. Neste formato podem ser criadas questões iniciais - mesmo que de múltipla-escolha, completar lacunas, identificar ideias gerais e específicas de textos escritos e/ou apresentados em áudio e vídeo - e, então, apresentar tarefas de cunho dissertativo dentro dos gêneros escolhidos para o Exame. Neste caso, o banco de itens do Exame Celpe-Bras deve conter uma vasta coleção de textos e gêneros e questões de diversos níveis para cada tarefa, por exemplo, bem como áudios e vídeos variados.

De acordo com Walczak (2015), um banco de itens ideal e grande o suficiente para não gerar repetição constante de tarefas, em um CAT, deve conter para cada 10 mil examinandos um total de 400 tarefas ou questões dentre os níveis a serem avaliados. Atualmente, o Exame Celpe-Bras, conta com mais de 5 mil inscritos em cada edição semestral em todo o mundo, crescendo gradualmente como mostramos no capitulo 2. Assim, para cada 5 mil examinandos, o banco de itens deveria conter 200 tarefas ou questões. Desde a primeira edição, em 1998, o Exame conta com aproximadamente 30 aplicações, o que, se multiplicarmos por quatro, teremos um número próximo de 120 tarefas já realizadas da Parte Escrita. Esse número não é baixo se considerarmos que, ao serem elaboradas questões iniciais ou diferenciadas para o teste, poderíamos elaborar, a partir do áudio, vídeo e texto, questões referentes à compreensão oral e escrita dos mesmos, de acordo com os níveis de proficiência certificados pelo Exame. É claro que o custo gerado na elaboração desse teste seria inicialmente alto, devido aos vários profissionais envolvidos na realização do teste e, posteriormente, para manutenção de um banco de itens diversificado a cada edição. Desta forma, "uma regra para projetos de tecnologia é que eles normalmente levam mais tempo e custam mais do que o esperado. Portanto, o melhor conselho é começar pequeno e planejar o

mais cuidadosamente possível". [tradução nossa][48]. (Chapelle e Douglas, 2006. pp. 1049-1050). No entanto, esse custo pode ser compensado, em parte, pela economia de recursos humanos na aplicação e correção do Exame e de material impresso. As instituições e/ou órgãos responsáveis por testes de larga escala, como o Celpe-Bras, devem pensar o Exame no longo prazo, já que com o passar do tempo, a tendência é que as NTIC se tornem mais modernas e acessíveis, economicamente falando, seguindo o curso natural das novidades tecnológicas. Consideremos o fato que os primeiros computadores portáteis eram caríssimos e inacessíveis à maioria das pessoas, porém, já são produzidos e comercializados de forma a atingir incontestáveis consumidores, ao menos na maioria das sociedades, como é o caso do Brasil. No entanto, ao que tudo indica, há pouca tecnologia de ponta sendo utilizada na elaboração e no aperfeiçoamento de testes de proficiência linguística ou destinado ao setor educacional. A área de avaliação não se desenvolve no mesmo ritmo das NTIC existentes em outras áreas ou presentes no mercado, como aplicativos para jogos, programas síncronos de comunicação, entre outros. Além disso, muitos programas são utilizados primeiramente ao ensino de língua e posteriormente adaptados à avaliação, o que não consideramos ideal. Outro problema é o fato de que as novas tecnologias, em muitos casos, foram incorporadas nos testes sem que houvesse uma mudança no modelo tradicional, transferindo os exames impressos para os computadores. E isso não gera inovação ou aprimoramento na forma de avaliação de proficiência em LE.

Finalmente, concordamos que o fato de os computadores oferecerem uma diversidade de insumos multimodais, como texto escrito, som, imagens fotográficas, *gifs* e vídeos, permite que os textos e vídeos selecionados para a Parte Escrita sejam ainda mais explorados em uma nova versão. As tarefas poderão ser expostas aos examinandos no computador, bastando um *clic* para abrir um vídeo ou imagem, texto, *hiperlink*, podendo oferecer botões de ajuda e instruções de execução da tarefa, como acontece nos exames descritos nesse trabalho. Cabe aos elaboradores decidirem se os vídeos e o áudio serão ouvidos uma ou duas vezes, como ocorre na versão impressa, e se o cronômetro de contagem do tempo dispendido pelo examinando vai parar, caso ele consulte as ferramentas de ajuda e instruções, como acontece nos exames do ETS, por exemplo.

A seguir são apresentadas algumas considerações e propostas para a correção e pontuação da Parte Escrita de uma versão *online* do Celpe-Bras.

---

[48] A rule of thumb for technology projects is that they typically take longer and cost more than expected; therefore, the best advice is to start small and plan as carefully as possible.

5.2.2 Correção e pontuação em meio computadorizado – a Parte Escrita

Não é intuito desta pesquisa se aprofundar na questão dos critérios de correção e na pontuação que o Exame deverá seguir, no entanto, podemos traçar algumas sugestões de acordo com o estudo feito até aqui.

Há muito tempo o computador é usado para correções de testes em larga escala, na verdade, ele começou a ser incorporado, inicialmente, na correção dos testes, como vimos no capítulo 3, podemos citar alguns exames que possuem uma correção parcialmente automatizada pelo computador, como o TOEFL e o TOEIC, e outros totalmente automatizados na correção, como é o caso do PTE *Academic* e o Projeto DIALANG. No entanto, mesmo que administrados por computadores, alguns testes ainda são corrigidos por humanos, e por mais de um avaliador, como, por exemplo, o teste escrito do ACTFL.

A escolha sobre como será a correção, muitas vezes, baseia-se em critérios de praticidade e custo, além de os desenvolvedores de testes defenderem a ideia, por estudos realizados, que o tipo de correção que apresentam é o meio mais confiável dentro do construto da validade de cada um deles. Entretanto, o maior diferencial de se utilizar o computador para a correção e entrega de resultados é a precisão e a velocidade com que essa tarefa é executada, se comparado à correção humana.

No caso do Exame Celpe-Bras, como as tarefas designadas à Parte Escrita, muitas vezes, são de cunho dissertativo, e nunca de múltipla-escolha, é preciso pensar de forma criteriosa sobre como será a correção, caso haja alguma mudança nas tarefas do Exame. Todavia, desde a edição 2015/01, a correção da Parte Escrita é realizada no espaço *online*, muito semelhante ao que é feito no teste de escrita do ACTFL, apesar das diferenças de entrega e formato das tarefas entre eles. Na realidade, um programa elaborado pela instituição responsável pelo Exame disponibiliza as avaliações dos examinandos no sistema, por meio de digitalização, e "pares cegos" de avaliadores humanos, pré-selecionados e treinados, corrigem individualmente e de maneira não presencial os textos dos examinandos segundo os critérios e a grade de correção do Exame. Sendo assim, o computador serve apenas como um meio prático e econômico para a realização da correção, que continua sendo totalmente humanizada. Logo, não podemos dizer que essa correção seja automatizada, mas sim virtual, pois não há um sistema automatizado de correção, apenas um ambiente de entrega de dados *online*.

Portanto, a combinação do julgamento de um avaliador de conteúdo e significado com a pontuação automática de recursos de língua asseguram pontuações consistentes e de qualidade, e pode ser um caminho para a correção do Exame Celpe-Bras *online*, caso criem-se tarefas distintas das atuais ou este seja entregue em formato CAT, por exemplo. Apenas nessas condições, o Exame poderia, então, utilizar um sistema *e-Rater* ou PLN, a fim de obter uma correção mais precisa e confiável ao avaliar alguns critérios da grade que ainda causam dubiedade entre os avaliadores, como a diferença entre 'raras' e 'poucas' inadequações lexicais e gramaticais ou 'raras' e 'poucas' interferências de outras línguas, por exemplo. Para isso, destacamos a necessidade da criação de um algoritmo para o programa de correção que seja sensível tanto a fatores linguísticos quanto aos conteúdos de tópicos e tipos de tarefas do exame, juntamente com a revisão humanizada de um ou mais avaliadores. Todavia, mantendo-se o construto, o que parece ser a tendência, este tipo de correção seria inviável. Portanto, com a manutenção do construto, defendemos a correção humanizada.

A seguir apresentamos nossas considerações sobre a Parte Oral do Exame.

5.2.3 A Parte Oral do Celpe-Bras *online*

Como vimos no capítulo 2, a Parte Oral do Exame Celpe-Bras procura medir a capacidade do indivíduo de compreensão e produção oral em LP. Essa parte consiste em uma entrevista de 20 minutos, na qual são utilizados três Elementos Provocadores (EP) de diversos temas e orientada por um roteiro de questões. Assim, na construção de um formato *online* do teste, esta parte pode sofrer mudanças em alguns aspectos que devem ser discutidos cuidadosamente, tais como as circunstâncias econômicas, físicas e temporais. Por exemplo, o uso de computadores ao invés de papéis e gravadores na realização de um teste de proficiência oral em LE gera por um lado uma economia e, por outro, um possível gasto com o maquinário e o artefato tecnológico necessário, além da construção e manutenção de páginas na *Internet*, programas de instruções e recursos humanos, como a contratação de técnicos da área da ciência da computação que trabalhem, juntamente com os linguistas, na elaboração de um sistema adequado ao novo formato do Exame em meio virtual.

No que tange à nova versão *online* do Celpe-Bras, consideramos que o uso de um dispositivo síncrono aprimorado ou a elaboração de um aplicativo de comunicação instantânea, criado especialmente para o Celpe-Bras, sejam as opções mais próximas da Parte Oral do Exame, em uma interação quase real de comunicação, e possibilitaria, ao mesmo

tempo, a redução para somente um examinador no momento do teste, o qual continuaria avaliando o examinando de forma holística em tempo real. Sendo assim, o teste poderia ser gravado e (re) visto diversas vezes pelo mesmo examinador e/ou por outros, caso necessário, antes de se atribuir a nota final ao examinando, como acontece no ACTFL OPIc. Isso ajudaria também a diminuir a pressão e a ansiedade do examinando que passaria a ter somente um examinador no momento do teste. Essa maior facilidade de armazenamento seria útil também para o treinamento de examinadores, que passariam a visualizar e não somente ouvir os áudios das entrevistas em cursos de capacitação.

Talvez o maior desafio em se escolher um aplicativo de comunicação instantânea, seria pensar nos entrevistadores, visto ser um teste que ocorre em diversos países com um grande número de inscritos. Além disso, há a questão de fuso horário, caso os entrevistadores se encontrem no Brasil e o examinando não. Uma possibilidade seria seguir o que o ACTFL-OPIc faz em seus testes, ou seja, a criação de um *avatar* virtual, atendendo a crescente demanda de forma unificada e onipresente. Para isso, o formato do teste passaria a contar, inicialmente, com perguntas variadas de cunho pessoal para 'quebrar o gelo' e, posteriormente, questões de diferentes níveis sobre assuntos diversos do cotidiano social, o que não eliminaria, necessariamente, o uso dos EP. No entanto, esse modelo do OPIc é assíncrono, seguindo a proposta de interação face a face virtual e automatizada. O que significa que o Exame perderia em interatividade e autenticidade das tarefas. Além desses aspectos, o custo na elaboração desse programa seria alto, mas com diminuição de aplicadores, contando apenas com uma correção humanizada e individualizada, sendo gravada e armazenada em um sistema seguro, para possíveis revisões ou consultas.

Por isso, defendemos que as duas versões sejam oferecidas, mesmo que não simultaneamente. O Exame pode começar com uma aplicação *online* por ano, fora a aplicação presencial, que seria administrada em outra época.

Com respeito ao tema ou conteúdo da entrevista, se este seria o mesmo a todos os examinandos ou não, torna-se interessante imaginarmos a criação de um sistema síncrono de comunicação visual, multimodal, integrado e adaptativo. As tecnologias atuais e seus avanços na área de avaliação permitem que algo desse porte seja elaborado, em que o programa, com a interferência humana, selecione de forma aleatória os EP ao examinando em sequência, de acordo com o desempenho e nível de proficiência demonstrado e entendido pelo entrevistador nos minutos iniciais de 'quebra-gelo', que seriam mantidos. Assim, o entrevistador e o

examinando ficariam mais confortáveis em prosseguir no ritmo adequado ao teste, de maneira individualizada.

Como discutido na Parte Escrita, a Parte Oral possui ainda mais possibilidades de construir um banco de itens grande e variado para um CAT, por exemplo. Com cerca de 30 aplicações em edições anteriores, com 20 EP por teste, a Parte Oral começaria com um banco de 600 EP. Entendendo que continuariam sendo três elementos por examinando, podemos dizer que o número ideal de 200 tarefas para cada 5 mil inscritos estaria alcançado se essa aplicação fosse hoje. O desafio, no entanto, seria montar questões de níveis variados sobre cada EP ou dividi-los entre os quatro níveis de certificação oferecidos, além da criação de um sistema no qual o entrevistador pudesse contar com tarefas diversificadas dentre os níveis propostos. O ACTFL-OPIc, como vimos, mostra que essa possibilidade existe, caso seja de interesse das instituições ou órgãos responsáveis pelo Exame. Neste caso, o banco de dados deve ser ampliado a cada edição.

Outras questões que merecem aprofundamento na discussão são: a entrevista seria em dias e horários diferentes de acordo com o número de examinandos inscritos? Neste caso, fariam apenas a Parte Escrita em um posto aplicador e voltariam outro dia para fazer a Parte Oral individualmente? Ou a entrevista seria feita em quaisquer computadores e locais escolhidos pelo examinando? E se este não tiver acesso a uma boa conexão de *Internet*, ou no momento da entrevista a *rede* travar ou cair? Sobre essas questões, embasados no que foi apresentado a respeito de ameaças existentes em testes computadorizados, consideramos que, por questões de segurança e autenticidade, o ideal seja a realização do Exame em um posto aplicador credenciado, em salas adequadamente equipadas, seguindo a grade de horário por número de examinandos inscritos e espaço disponível.

Outra possibilidade para o Exame Celpe-Bras *online*, considerando a permanência de seu construto, seria uma parte da entrevista oral ser substituída por questões de repetição de pronúncia, *role-play* virtual, além das respostas ao roteiro de perguntas dos EP. No entanto, entendemos que a entrevista face a face ainda é a maneira mais completa e confiável de avaliação, dentro da proposta do Celpe-Bras impresso. Além disso, fatores emocionais e de letramento digital também poderiam fazer parte do construto, visto que o vídeo ficaria armazenado para futuras reavaliações.

Finalmente, descartamos a tecnologia PLN por entrevista telefônica, como acontece no ACTFL OPI, não apenas pelo custo, mas, principalmente, porque este modelo privaria o

examinando de estímulos ou de ser exposto a conteúdos multimodais, não simulando, assim, tarefas comunicativas autênticas, contextualizadas e interativas.

Essas e muitas outras questões, principalmente de cunho orçamentário e de manutenção do teste, devem ser consideradas, visto que são desconhecidos testes de proficiência oral que sejam realizados por meio de dispositivos de comunicação síncrona oferecidos pela *Internet*.

A seguir são discutidas algumas propostas para a correção e pontuação da Parte Oral de uma versão *online* do Celpe-Bras.

5.2.4 Correção e pontuação em meio computadorizado – a Parte Oral

Segundo Chapelle e Douglas (2006), embora a pontuação automatizada e a realização de testes assistidos por computador sejam realidades, não há, ainda, "evidências para a avaliação de desempenho tornada prática por meio do uso amplo de simulação, geração autêntica de itens ou mudanças significativas nos propósitos de testes por meio da tecnologia". (pp. 1123). [tradução nossa][49].

Assim, a correção da Parte Oral é um aspecto que merece atenção. Sobre a questão da correção em testes de proficiência oral, Salomão (2010) afirma que;

> No caso específico dos testes de proficiência oral, a tecnologia necessária para correção é mais complexa: a tecnologia de processamento natural da fala, que ainda é nova, cara e limitada. Além disso, um teste de habilidade de produção oral ou escrita requer respostas linguisticamente construídas de possibilidades infinitas, e os dispositivos de correção automática ainda não estão bem desenvolvidos, podendo haver erros e comprometimento da validade do teste. Pode haver também um efeito retroativo negativo, fazendo com que os candidatos fiquem mais preocupados em aprender como o *e-Rater* corrige o teste do que em estudar o conteúdo. (SALOMÃO, 2010. p. 337).

A tecnologia de PLN a qual a autora se refere não seria, portanto, a mais adequada para ser utilizada na Parte Oral do Celpe-Bras. Não somente por ser cara e limitada, mas

[49] Although automated scoring and computer-assisted test delivery are realities, we were unable to show evidence for performance assessment made practical through widespread use of simulation, authentic item generation, or significant changes in testing purposes through technology. Such revolutionary changes need to be prompted and supported by conceptual advances in our understanding of language, language use, and language learning.

também por não se adequar ao construto do teste, visto que seu objetivo é criar uma interação face a face com o examinando para melhor avaliar suas habilidades linguísticas de maneira holística e analítica. Além disso, a autenticidade do teste e de seu conteúdo poderia ser comprometida.

Sendo assim, se o Exame Celpe-Bras continuar sendo uma entrevista, porém, virtual e síncrona, os sistemas PLN *e-Rater* ou *SpeechRater*, apresentados no quarto capítulo, não ajudariam no sentido de que esses mecanismos não são ainda capazes de realizar uma correção automatizada de um grande insumo de resposta, apenas frases e respostas curtas, como as tarefas dos exames do ETS descritos no capítulo anterior. Nossa proposta é que o entrevistador interaja por um sistema síncrono de comunicação virtual com o examinando e avalie de forma holística o seu desempenho final, logo após a interação. Um avaliador externo seria uma alternativa de substituição do observador, deixando o entrevistador e o examinando mais a vontade no decorrer da entrevista. Neste caso, obter uma amostra da oralidade do examinando por meio da gravação e armazenamento da entrevista poderia favorecer uma correção entre dois avaliadores, individualmente, realizada após o teste, como acontece no teste oral virtual da LTI, o ACTFL OPIc.

Por fim, ressaltamos que a interpretação do desempenho do examinando em novos tipos de tarefas precisa ser examinada qualitativa e quantitativamente, para que sejam utilizadas de maneira apropriada tanto na Parte Oral quanto na Parte Escrita, caso o Exame opte por um teste adaptativo ou por mudanças em prol da inovação e praticidade, mesmo que mínimas.

5.3 Outros aspectos para o Celpe-Bras *online*/eletrônico

Além das propostatas apresentadas neste capítulo, outros aspectos podem ser observados a partir de nosso estudo, servindo, também, como sugestões para o Celpe-Bras *online*/eletrônico. Primeiramente, o Celpe-Bras pode seguir o exemplo do TOEFL, TOEIC e PTE *Academic* nos aspectos temporais e espaciais, embora apresentem propósitos diferentes entre si, com abrangências geográficas e públicos diferentes. Em um formato CBT linear, o exame deve manter o que já é feito na versão original em relação à estipulação do tempo de realização de suas tarefas, bem como a aplicação somente em locais autorizados, com examinadores devidamente treinados, atendendo, assim, as várias localidades do Brasil e do mundo. Esta atitude proporciona um melhor atendimento aos examinandos, que, por sua vez,

devem buscar uma preparação para o teste, se julgam ainda não a possuir, principalmente no que diz respeito à ansiedade e ao letramento digital, de informática e de mídia necessários para o bom desempenho no teste. Por exemplo, o examinando do Celpe-Bras *online* precisa estar familiarizado com o uso do teclado e leitura de textos na tela do computador, saber utilizar os botões de ajuda e instrução de cada tarefa, identificar o espaço correto de resposta, localizar o temporizador, além de reconhecer o formato de cada questão e opções de voltar ou seguir em frente, caso seja disponibilizado pelo teste. Anotações durante o teste escrito continuariam sendo permitidas e testes de edições anteriores poderiam ser disponibilizados na página oficial do INEP, destinada à divulgação do exame, para o melhor preparo dos examinandos e a familiaridade com as tarefas e o formato do teste. Além disso, o cuidado com o aparato tecnológico, computadores devidamente equipados, com uma conexão segura e eficiente, devem ser uma preocupação durante a elaboração, para que o teste seja aplicado em formato eletrônico sem que possíveis falhas técnicas ocorram. É interessante que instruções quanto a possíveis problemas técnicos sejam disponibilizadas aos examinandos previamente, como, por exemplo, o que fazer caso o programa trave ou o sinal de acesso à Rede de *Internet* caia. O examinando precisa saber se deve esperar para continuar o teste da questão em que parou, sem prejuízo da contagem de tempo, ou se outro teste será disponibilizado desde o começo para ele, se for um teste adaptativo (o que geraria um transtorno). Todos esses fatores ajudam a aumentar a confiabilidade dos dados do teste, assim como a praticidade e aplicabilidade.

Outro aspecto importante e digno de consideração, destacado por Chapelle e Douglas (2006), é a questão da validade diferencial, aquela que remete às habilidades externas dos examinandos de diferentes regiões ao lidar com o computador. No caso do Celpe-Bras, o exame é aplicado em 65 postos credenciados no exterior e, portanto, considerar que uma versão *online* atinja todos esses postos parece irreal, inicialmente. As condições tecnológicas e de aparatos tecnológicos, maquinários, redes de *internet* compatíveis e com velocidade e tamanho em *megabytes* adequados em algumas regiões internacionais e, até mesmo, nacionais, podem variar distintamente. Por isso, sustentamos a ideia de que uma versão não substitui a outra, ou seja, ambas podem coexistir e, quando possível, o examinando poderia escolher a que melhor lhe convém, a impressa ou a computadorizada.

No que diz respeito à autenticidade das tarefas e dos conteúdos do teste, concordamos com Kyle, Crossley e McNamara (2015) ao afirmarem que a presença de material autêntico nas tarefas afeta as características linguísticas das respostas dos examinandos. Assim, o

domínio de uso da língua-alvo em uma tarefa, isto é, uma situação no campus *versus* um curso acadêmico, por exemplo, pode afetar a linguagem produzida para a resposta de determinada tarefa. Quando os métodos de teste divergem significativamente do uso da língua-alvo, do mundo real, ou interferem na medição das habilidades ou competências linguísticas desejadas, o efeito do método é um fator negativo. Sendo assim, o Celep-Bras em um formato *online* deve manter tarefas autênticas do cotidiano brasileiro, com conteúdos retirados de fontes autênticas, não somente na Parte Escrita, mas na Parte Oral também, disponibilizando essas fontes de acordo com o formato da tarefa definido em meio computadorizado.

Finalmente, em relação ao efeito retroativo, entendemos que uma versão *online* do Celpe-Bras tem o potencial de provocar um impacto no ensino, na medida em que os professores que preparam os examinandos ao Exame podem ser levados a realizar tarefas interativas em meio eletrônico, com os alunos, de acordo com as possibilidades, e, assim, prepará-los melhor para o Exame. Como vimos, programas como o Teletandem ajudam não somente no aprendizado da língua-alvo, mas também na avaliação e *feedback* do examinando, como em um ciclo contínuo. Ao que tudo indica a intimidade em manusear computadores e seus aplicativos e programas ou ler textos e livros digitais são uma constante na formação de alunos que pretendem realizar exames de proficiência em língua estrangeira. Geralmente, as pessoas que precisam e/ou procuram os exames em larga escala têm objetivos de entrar em uma universidade, trabalhar em um país que fala o idioma pretendido, revalidar seus diplomas ou simplesmente medir seu conhecimento em uma língua alvo. No caso da maioria dos inscritos no Exame Celpe-Bras, esse vasto perfil parece ser a regra, portanto, a maioria dos examinandos deve apresentar letramento digital necessário para a realização de um exame *online*, o que não exclui a necessidade de se oferecer tutoriais e manuais de instrução do mesmo.

A seguir são apresentadas algumas considerações parciais do estudo e uma síntese das principais proposições aqui discorridas.

5.4 Considerações parciais e resumo das proposições

Neste capítulo, apresentamos algumas tecnologias que o Exame Celpe-Bras pode fazer uso, empiricamente falando. No entanto, esclarecemos que nossas propostas destacam o que percebemos ser possível utilizar, caso o construto seja mantido e caso sejam realizadas

mudanças ou adaptações no construto do teste. Tentamos, assim, relatar todas as possibilidades, tendo em vista as ferramentas tecnológicas em uso nos testes de proficiência de LE de larga escala observados até aqui. Isso não quer dizer que tais mudanças são esperadas ou serão geradas, apenas são explicitações do que consideramos exequível e digno de pesquisas experimentais. Considerando que o construto do teste seja mantido, como sugerimos, algumas possibilidades apontadas são consideradas menos viáveis, porém, o intuito de nossa pesquisa consiste em apresentá-las com suas vantagens e limitações, como algo possível de ser realizado e digno de futuras pesquisas e experimentos. É o caso da correção da Parte Escrita e da Oral do Celpe-Bras, ambas foram explicitadas de forma geral, não sendo o foco deste estudo e, portanto, necessitam de pesquisas específicas em meio computadorizado. Destacamos que as máquinas não fazem juízo de valor, assim, elementos culturais e situacionais não podem ser avaliados por um sistema eletrônico, o que, no caso do Celpe-Bras, torna a avaliação humanizada uma opção viável, frente à avaliação automatizada.

Além disso, diante de nossas propostas para uma aplicação *online*/eletrônica do Celpe-Bras, entendemos que o novo sistema deve ser flexível no que diz respeito às tecnologias utilizadas, pois, estas estão evoluindo rapidamente e ferramentas tecnológicas ainda não existentes podem se tornar tendência na área em poucos anos. Assim, o Exame *online* ou eletrônico deve deixar uma abertura às NTIC para que estas sejam primeiramente testadas e gradualmente incorporadas. Se um programa de teste for muito rígido em seu formato, as novas tecnologias, que ainda serão conhecidas, podem provocar grandes diferenças no que será possível fazer daqui alguns anos na área de avaliação de proficiência de língua e, assim, correrá o risco de tornar o teste obsoleto ou pouco inovador em algum aspecto, seja na aplicação ou na correção de teste. Assim, o uso das NTIC deve sempre ser pensado como possibilidade de melhorar a interpretação do desempenho do examinando e, também, a praticidade de teste.

Desta forma, estabelecemos abaixo, de maneira resumida, as principais propostas deste estudo para a elaboração e aplicação de uma versão computadorizada ou *online* do Exame Celpe-Bras.

QUADRO 18 – Resumo de propostas para o Celpe-Bras *online*/eletrônico

ASPECTOS GERAIS
✓ Manutenção do construto: Teste de desempenho baseado em tarefas comunicativas; ✓ Aplicação somente em locais autorizados, com a presença de examinadores treinados; ✓ Estipulação do tempo de realização de tarefas; ✓ Computadores devidamente equipados e conexão segura; ✓ Disponibilização de testes de edições anteriores na página oficial do INEP; ✓ Disponibilização de instruções de teste e uso de recursos em tela, instruções quanto a possíveis problemas técnicos e botões de ajuda; ✓ Seleção de tarefas autênticas do cotidiano brasileiro, conteúdos de fontes autênticas; ✓ Elaboração de tarefas e tipos de respostas variadas. ✓ Cuidados com a formatação de tarefas e conteúdos selecionados para o teste, ambos variados; ✓ Uso de multimídias diversas - vídeos, áudios, hipertextos - contextualizados e autênticos.
PARTE ESCRITA
✓ Teste linar computadorizado; ✓ Permissão para anotações durante o teste; ✓ Permanência das quatro macro-tarefas; ✓ Possibilidade de ouvir os vídeos e o áudio duas vezes; ✓ Presença em tela de cronômetro de contagem de tempo dispendido pelo examinando e a possibilidade de pausa para consultar as ferramentas de ajuda e instruções; ✓ Possibilidade de acréscimo de questões iniciais - múltipla-escolha, completar lacunas, identificar ideias gerais e específicas de textos escritos, de áudio e/ou vídeo – além das tarefas de cunho dissertativo. ✓ Correção humanizada.
PARTE ORAL
✓ Teste semi-adaptativo computadorizado; ✓ Utilização de um dispositivo síncrono de comunicação elaborado especialmente para o teste e que possibilite a gravação e o armazenamento seguro do vídeo; ✓ Entrevista ou interação virtual em tempo real; ✓ Redução para um entrevistador-avaliador em tempo real de teste; ✓ Possibilidade de seleção de forma aleatória dos EP de níveis diferenciados; ✓ Permanência dos minutos iniciais de ‘quebra-gelo’; ✓ Banco de dados (EP) ampliado a cada edição; ✓ Avaliação holística e correção humanizada.

Finalmente, julgamos as questões aqui dicorridas úteis e contextualizadas dentro da área de avaliação de línguas e esperamos que sirvam como norteadoras de uma possível implementação do formato eletrônico e/ou *online* do Celpe-Bras. Se a praticidade é a justificativa e o alvo, que os outros critérios mencionados sejam igualmente considerados e, minuciosamente, entendidos e definidos, em especial, o construto e a validade do Exame.

Seguiremos, assim, para a última parte do trabalho, apresentando as discussões e considerações finais.

DISCUSSÃO E CONSIDERAÇÕES FINAIS

Computers aren't the thing.
They're the thing that get you to the thing!

Joe MacMillan - HACF

Esta pesquisa aponta para o desenvolvimento plausível de uma aplicação *online* e/ou eletrônica do Celpe- Bras, já que o Exame deve adaptar-se às novas tecnologias com toda a flexibilidade que isto implicará nas mudanças e inovações das próprias tecnologias em um futuro próximo. Além disso, a aplicação de uma nova versão do Celpe-Bras em ambiente virtual, utilizando sistemas computadorizados e ferramentas tecnológicas apropriadas ao seu construto, possibilita a expansão do Exame não somente em âmbito nacional, mas, também, em âmbito internacional. Como discutido nesta pesquisa, há uma crescente procura do Exame e um aumento significativo do número de inscritos a cada ano, desde a primeira aplicação em 1998. Esse crescimento deve-se a fatores econômicos e políticos que fazem com que muitos estrangeiros queiram utilizar o português para se ter acesso ao Brasil (DINIZ, 2010. p. 92), além do aumento de imigrantes dentro do país por motivos diversos, inclusive, humanitário. Esses fatores favorecem, consequentemente, o desenvolvimento do ensino e da aprendizagem de PLE dentro e fora do território brasileiro, o que leva à necessidade de se ter um instrumento prático e justo para avaliação do desempenho desses aprendizes, de acordo com os seus interesses.

Desta forma, segundo Diniz (2010), a LP é "uma língua que está em toda parte, se inserindo no 'mundo globalizado'." (*Ibid*, p. 80). Nas palavras de Faraco (2016), o Brasil, em sua condição de país com o maior número de falantes da LP no mundo, está cada vez mais assumindo seu papel de protagonista na promoção da LP. De acordo com o autor, isso se deve aos atrativos e às condições naturais que o país possui:

> [...] o Brasil por si só, tem forças de atração inquestionáveis. É o maior país da América do Sul; é o segundo país mais populoso da América; é a terceira economia da América; a sétima ou sexta economia do mundo; é a fonte de vastas riquezas minerais; está entre os maiores produtores de alimentos do mundo; tem uma presença política internacional que se caracteriza, há pelo menos um século, por um posicionamento de equilíbrio e equidistância dos mais poderosos, pela não intervenção e pela conciliação; e, por fim, tem uma cultura singular em sua diversidade e dinâmica. (FARACO, 2016. p. 346).

Tais atrativos, destacados por Faraco, proporcionam a oportunidade de criação de políticas e de iniciativas para a internacionalização da LP. Como afirma o autor, "o Estado brasileiro não está, obviamente, ausente dessa promoção sistemática." (*Ibid*, p. 346). O país conta com a Rede Brasil Cultural, os leitorados, os Institutos binacionais e a criação, em 2010, da Universidade Federal da Integração Latino-Americana (UNILA) e da Universidade da Integração Internacional da Lusofonia Afro-Brasileira (UNILAB), que, como destaca Faraco, têm um papel importante na difusão e promoção da LP. Junto a essas iniciativas está o Exame Celpe-Bras que coopera com as políticas oficiais do Brasil. Desta forma, "talvez o desafio do momento seja realizar uma avaliação de todas essas iniciativas, buscando meio de aprimorá-las e ampliá-las, intensificando seu alcance." (*Ibid*, p. 347). E é exatamente neste ínterim que nossa pesquisa se insere. Pois, aumentar a praticidade do Exame, por meio de uma aplicação em formato *online* e/ou computadorizado, colabora para que a tão desejada ampliação da promoção e difusão da LP possa ser efetivamente concretizada em termos de avaliação de proficiência na língua-alvo. Obviamente, o alcance do Exame será maior, além de exercer a função de pautar o ensino da LP a estrangeiros em território nacional e internacional, função essa que o Exame já exerce e que também será ampliada. Neste sentido, as NTIC utilizadas para a otimização do Celpe-Bras tornam-se essenciais ao propósito de expansão da LP.

Retomemos agora a epígrafe inicial deste estudo, em que ecoam as palavras de Tim McNamara (2000. p. 79): "O teste de língua é uma área em crise, mascarado pelo impressionante aparecimento do avanço tecnológico". Essa crise a qual o autor se refere pode ser entendida como uma conotação positiva, considerando a busca constante da área de avaliação em acompanhar as teorias de língua de cada época, em prol do aperfeiçoamento e da justeza da avaliação do examinando em testes de LE. E é nessa crise que as novas tecnologias podem ser úteis para ajudar a extrair um desempenho mais próximo da realidade, mesmo em meio ao comportamento artificial do examinando em situação de teste. O avanço tecnológico aplicado ao serviço da compreensão da validade e do construto propostos de teste, certamente, auxiliará a área de testes de proficiência a alcançar a eficiência e inovação desejadas pelos desenvolvedores de teste.

Sendo assim, esta pesquisa procurou investigar o uso das NTIC como ferramenta para integrar e aprimorar a realização de tarefas destinadas à avaliação de proficiência em Português como Língua Estrangeira, no Exame Celpe-Bras. Suportada pelo fato de que o Exame ainda não possui uma versão eletrônica/*online* e que o teste tende a aprimorar a

praticidade na aplicação e na correção, procuramos, por meio de um estudo qualitativo, identificar as NTIC utilizadas na realização de tarefas avaliativas em testes de proficiência linguística, discutir as qualidades ou critérios de teste e as características de alguns exames de proficiência em meio eletrônico e propor alguns cuidados e/ou sugestões para uma versão *online* do Exame Celpe-Bras.

Devido à escassez de estudos sobre o tema de avaliação em meio computadorizado, principalmente em âmbito nacional, buscamos trabalhos que investigaram os testes eletrônicos e as qualidades e/ou critérios de testes em larga escala para servir como conceitos e concepções teóricas da pesquisa.

Assim, após contextualizar nosso estudo, descrevemos o Exame Celpe-Bras em seu formato impresso e verificamos ser este um teste desenvolvido sob a análise criteriosa do uso da LP pelo indivíduo em um contexto específico, visando à interação deste com o seu interlocutor e, consequentemente, à comunicação bem sucedida. Além disso, o Exame não especifica o conteúdo das provas aos examinandos sendo que a escolha dos tópicos que compõem o teste está sob a responsabilidade dos elaboradores e dos profissionais da área. Tais características estão presentes no Exame desde a sua primeira aplicação, o que caracteriza a manutenção de sua filosofia ou de seu construto.

Sendo assim, dando sequência a nosso estudo, após a análise descritiva das NTIC e seus usos em testes computadorizados, verificamos que os computadores abrem uma vasta possibilidade de introduzir novos insumos e tipos de testes, simplificando ou agilizando a coleta e a análise de dados. (WALCZAC, 2015. p. 39). As possibilidades são variadas. Desta forma, retomamos a questão de pesquisa que norteia esta investigação e de forma sucinta traçamos alguns aspectos que podem servir como indícios de respostas ao questionamento: *Como o Celpe-Bras pode fazer uso das NTIC em uma possível versão online?*

Primeiramente, entendemos que cabe aos elaboradores ou órgãos responsáveis pelo Exame Celpe-Bras definir se o construto do teste e a sua validade serão alterados ou adaptados em uma versão *online*, para, então, definir como as NTIC serão utilizadas no novo formato e se o teste será linear ou adaptativo, por exemplo. Como discutido no capítulo anterior, as ferramentas tecnológicas ideais e necessárias para o desenvolvimento de teste dependem do propósito final da avaliação e de questões práticas, tais como, os recursos financeiros, humanos, físicos e de tempo disponíveis. Além de considerações fundamentais como, a escolha da autenticidade de conteúdo das tarefas e a busca pela maior praticidade do teste. Essa busca, no entanto, torna-se controversa, entendendo que há um espaço de tensão

entre a autenticidade e a praticidade, ao se pensar na elaboração de uma versão *online* do Exame, pois, quanto maior a autenticidade de um teste, mais oneroso e complexo ele será, e mais subjetiva a correção se torna. Assim, no caso do Celpe-Bras, é preciso encontrar a melhor maneira de equilibrar as tarefas de teste oferecidas em meio computadorizado, sem comprometer seu conteúdo e ao mesmo tempo encontrar meios práticos de aplicabilidade, execução e pontuação dessas tarefas.

Conforme aponta nosso estudo, na mudança de meio de entrega ou aplicação existe a possibilidade de revisão da validade do construto e, se o construto é minimamente alterado, alteram-se também as formas, as características, os tipos de tarefas e o modo de correção, por exemplo. Todas estas questões devem ser consideradas no desenvolvimento ou na criação de uma nova versão eletrônica do Exame.

Concordamos, portanto, com Fulcher ao afirmar que:

> Uma versão de um teste é uma evolução da especificação de teste. A especificação pode ser alterada para introduzir um novo tipo de item que melhor reproduza o domínio, ou altere a natureza da pontuação, a fim de melhorar a relação entre a pontuação e o desempenho. Tal evolução muda todas as formas subsequentes do teste. Uma versão pode, portanto, ser vista como uma evolução diacrônica, enquanto as formas são sincronizadas a cada versão. [...] Versões da especificação de teste constituem a história da evolução do teste, e atuam como um registro de decisões tomadas ao longo do tempo para definir e melhorar a medição do construto. [tradução nossa][50]. (FULCHER, 2015. p. 3).

No entanto, vamos além da visão de versão de teste de Fulcher para destacar que, mais que uma evolução, uma nova versão deve trazer consigo a oportunidade de uma revolução, a criação de algo novo que venha aprimorar o teste, em qualquer direção, seja na aplicação e/ou na correção. Neste sentido, as ferramentas tecnológicas ou as NTIC funcionam como um meio para se alcançar tal desafio, e não como um fim.

Em seguida, elencamos algumas possibilidades que consideramos viáveis na elaboração do Celpe-Bras em meio computadorizado, tendo em vista a Parte Escrita, a Parte Oral e as suas respectivas correções. Assim, verificamos que, com o foco nos CBT, uma

[50] A version of a test is an evolution of the test specification. The specification may be changed to introduce a novel item type that better replicates the domain, or changes the nature of the scoring to improve the relationship between the score and the performance. Such an evolution changes all subsequent forms of the test. A version may therefore be seen as diachronic evolution, while forms are synchronic to each version. [...] Versions of the test specification constitute the history of the test evolution, and acts as a record of decisions taken over time to further define and improve the measurement of constructs.

dessas possibilidades seria a criação de um CAT da Parte Escrita, por exemplo. Lembramos que este modelo possui suas vantagens, como sugere Piton-Gonçalves e Aluísio (2015. p. 392): "[...] o Teste Adaptativo surge como uma metodologia que possibilita o aumento da precisão da estimativa das habilidades medidas, administrando um número menor de itens em testes educacionais", além da rapidez na entrega de resultados. Entretanto, a busca por eficiência não deve servir como a única direção a ser seguida no desenvolvimento de um teste. Segundo Chalhoub-Deville (2001. p. 97), "os avanços na tecnologia devem encorajar os desenvolvedores de testes a ir além do pensamento que há muito domina o teste impresso e inspirar o uso de aplicações disruptivas," e, assim, por em prática maneiras inovadoras de testagem. Em outras palavras, os CAT que se valem apenas da novidade da adaptação ao nível do item, podem restringir novas possibilidades. A tecnologia deve ser vista, assim, como um recurso para melhorar os métodos e os usos crescentes dos testes. (CHAPELLE e VOSS, 2016. pp. 5-7). Desta forma, consideramos que, mesmo em um formato adaptativo, o Exame possa se valer de tarefas comunicativas integradas, utilizando-se de recursos tecnológicos para o seu aprimoramento.

No que se refere à Parte Oral do Exame, reconhecemos que outra área onde há espaço para o progresso é o reconhecimento da fala. Segundo Chapelle e Voss (2016), medir a proficiência oral usando a tecnologia de computador é uma área de interesse crescente de pesquisa. Porém, as ferramentas para marcar automaticamente o desempenho oral, por meio do reconhecimento da fala, ainda não são suficientemente sofisticadas para incorporar todos os níveis de proficiência do examinando, pois esta é complexa e varia substancialmente. O reconhecimento automático de voz até agora tem sido limitado a respostas curtas e respostas restritas. Este campo tem avançado gradativamente com excertos mais longos de fala, mas muito estudo ainda é necessário para um reconhecimento de fala mais preciso e confiável na avaliação em LE. (*Ibid*. p. 125).

Fato é que "quando um teste possui itens dissertativos, há o aumento do custo e do treinamento de corretores de provas, [...]. Como alternativa, os testes objetivos buscam cada vez mais avaliações justas, precisas e rápidas, de acordo com métricas e critérios educacionais preestabelecidos e consolidados." (PITON- GONÇALVES e ALUÍSIO, 2015. p. 391). Essa busca também existe no caso do Exame Celpe-Bras, apesar de ser moldado em tarefas comunicativas. Entretanto, conforme apontou nosso estudo, existe a possibilidade de inovação mesmo utilizando-se formas subjetivas de testagem, como a entrevista na avaliação de desempenho oral em meio computadorizado, por exemplo, e a elaboração de um sistema

síncrono de comunicação, futuramente. Mesmo entendendo que o que temos até agora são especulações e problematizações, visto existirem poucos trabalhos que possam avaliar a real eficácia deste modelo síncrono de avaliação, ainda não testado na área, consideramos ser este um dos caminhos a ser trilhado. O desafio da elaboração de um CAT na Parte Oral do Celpe-Bras também é grande, devido ao seu construto comunicativo. E como afirmam Chapelle e Douglas (2006. p. 427-428) "o desafio de desenvolver testes de língua computadorizados adaptativos que reflitam a necessidade de avaliar a habilidade comunicativa de uso da língua ocupará os desenvolvedores de testes de língua e pesquisadores nas próximas décadas".[51] [tradução nossa].

Finalmente, concordamos que, no campo de testes de proficiência em ambientes virtuais, as NTIC podem ser de grande ajuda no que se refere à praticidade de testes de proficiência em larga escala, como é o caso do Exame Celpe-Bras, que vem crescendo gradativamente em procura no mundo inteiro. Porém, faz-se necessário a criação de critérios e métodos que possam assegurar a validade e a confiabilidade de testes em meio *online*, bem como a autenticidade, as circunstâncias físicas, temporais e de segurança dos testes e o critério de correção e procedimento de pontuação. Somente quando decisões forem tomadas a respeito dos aspectos da língua-alvo a ser medidos, dos critérios pelos quais o desempenho será avaliado e do procedimento pelo qual o desempenho será marcado poder-se-á, então, lançar mão de um estudo criterioso de um Exame Celpe-Bras *online*. Enquanto isso, mais estudos e pesquisas são necessários para que esse formato se concretize como uma versão válida, confiável e com um construto sólido, bem definido, evitando o risco de ser mal utilizado ou mal interpretado.

Esperamos que este estudo desperte reflexões na área acerca dos critérios que devem compor uma versão *online* do Exame Celpe-Bras e que as sugestões aqui delineadas possam ser úteis para o aprimoramento e a otimização do teste em meio eletrônico. Esperamos, também, que as ferramentas tecnológicas e testes aqui descritos sirvam como um norte para os interessados na área e suscite diferentes pesquisas sobre assuntos não abordados neste estudo.

Temos, ainda, a expectativa que este estudo sirva para futuras pesquisas na área da avaliação de larga escala e o uso das NTIC em testes de proficiência linguística. Reconhecemos os limites deste trabalho que aborda apenas aspectos bibliográficos, de

[51] The challenge of developing computer-adaptive language tests that reflect the need to assess communicative language ability will occupy language test developers and researchers for decades to come.

natureza prática, porém, não voltada ao experimento. Portanto, consideramos que as propostas apontadas aqui devem ser exploradas em futuras pesquisas de cunho experimental, seja na aplicação ou correção do Exame. Futuros estudos podem desenvolver, em meio computadorizado, tarefas da Parte Escrita ou da Parte Oral para uma testagem, por exemplo, do uso de um sistema síncrono de comunicação para a verificação do comportamento do examinando e do entrevistador-avaliador em uma interação virtual. Devido ao tempo e espaço limitados desta pesquisa, apesar da vontade da pesquisadora, tais experimentos não foram desenvolvidos.

Por fim, entendemos que o campo é vasto, as possibilidades são variadas. Destarte, após a conclusão desta dissertação, temos o intuito de dar continuidade aos estudos na área de avaliação e contribuir, de forma significativa, para o aprimoramento do Exame Celpe-Bras, tanto na Parte Oral quanto na Parte Escrita, por meio de pesquisas experimentais.

REFERÊNCIAS

ACERVO CELPE-BRAS. **Elementos provocadores da Parte Oral 2004/2**. [online]. Disponível em:<http://www.ufrgs.br/acervocelpebras/arquivos/elementos-provocadores-da-parteoral/2015_1>. Acesso em: jan/ 2017.

BRASIL. Instituto Nacional de estudos e Pesquisas Educacionais Anísio Teixeira – INEP. ***Certificado de proficiência em língua portuguesa para estrangeiros*: Manual do Examinando.** Brasília: 2011. [online]. Disponível em http://download.inep.gov.br/download/celpebras/2011/manual_examinando_2011_1.pdf Acesso em: 28 dez. 2015.

BRASIL. **Instituto Nacional de estudos e Pesquisas Educacionais Anísio Teixeira – INEP. Certificado de proficiência em língua portuguesa para estrangeiros: Postos Aplicadores**. Brasília: 2016. [online]. Disponível em: < http://download.inep.gov.br/outras_acoes/celpe_bras/postos_aplicadores/2016/postos_aplicadores_CelpeBras.pdf>. Acesso em: 10 jan. 2016.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Certificado de proficiência em língua portuguesa para estrangeiros: manual do examinando.** [online]. Brasília: MEC/INEP, 2010a. Disponível em: <http://www.inep.gov.br/ download/celpebras/2010/manual_do_examinando_2-9-2010.pdf> Acesso em: julho/ 2016.

BRASIL. Ministério da Educação. Secretaria de Ensino Superior. **Manual do aplicador do certificado de proficiência em língua portuguesa para estrangeiros: estrutura do exame**. Brasília: MEC, 2015. [online]. Disponível em <http://portal.inep.gov.br/celpebras-estrutura_exame> Acesso em: dez. 2016.

BRASIL. Ministério da Educação. Secretaria de Ensino Superior. ***Certificado de proficiência em língua portuguesa para estrangeiros*: roteiro de interação face a face: Parte individual – elementos provocadores**. Brasília: MEC, 2015.

BRASIL. Ministério da Educação. Secretaria de Ensino Superior. **Certificado de proficiência em língua portuguesa para estrangeiros: caderno de questões da parte coletiva**. Brasília: MEC, 2015.

AGOSSA, M. O. B. **O exame Celpe-Bras como instrumento de divulgação da cultura brasileira: percepções de candidatos**. 2017. Dissertação de mestrado. Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2017.

ANCHIETA. P. P. **Análise de testes de proficiência em língua inglesa: subsídios à elaboração de um exame para professores de inglês no Brasil.** 219 págs. Dissertação de Mestrado na Universidade Estadual Paulista Júlio de Mesquita Filho. São José do Rio Preto, São Paulo, 2010.

BACHMAN, L. F.; PALMER, A. S. **Language testing in practice: Designing and developing useful language tests.** Oxford: Oxford University Press, 1996.

BARNWELL, D. P. **A history of Foreign Language Testing in the United States: from its beginning to the present.** Arizona: Bilingual Press, 1996.

BENNETT, R. E. (2001). **How the Internet Will Help Large-Scale Assessment Reinvent Itself.** Education Policy Analysis Archives. 9 (5), 2001. ISSN 1068-2341. [online].

BROWN, H. D. **Language Assessment: principles and Classroom Practices.** 2nd Edition. Pearson Longman: 2010.

BROWN, A.; IWASHITA, N. & MCNAMARA, T. **An examination of rater orientations and test-taker performance on English-for-academic-purposes: speaking tasks**. ETS: Monograph Series, 2005.

CHALHOUB-DEVILLE, M. (2001). **Language testing and technology: Past and future.** [online]. Language Learning & Technology, *5*(2), 95–98. Retrieved from http://llt.msu.edu/vol5num2/deville/

CHAPELLE, C.A.; DOUGLAS, D. **Assessing language through computer technology.** Cambridge, UK: Cambridge University Press, 2006. Kindle Edition, pp. 1543.

CHAPELLE, C. A.; ENRIGHT, M. K. & JAMIESON, J. M. **Building a validity argument for the test of English as a foreign language**. New York: Routledge, 2008.

CHAPELLE, C. A. & VOSS, E. **20 years of technology and language assessment in language learning & technology.** [online]. Language Learning & Technology http://llt.msu.edu/issues/june2016/chapellevoss.pdf. 2016, Volume 20, Number 2 pp. 116–128. ISSN 1094-3501 116.

CONSOLO, D. A. & FURTOSO, V. B. **Assessing oral proficiency in computer-assisted foreign language learning: A study in the context of teletandem interactions.** [online]. http://dx.doi.org/10.1590/0102-445022183819328533. D.E.L.T.A, 2015. pp. 665-689.

CONSOLO, D. A. & SILVA, V. L. T. **Em defesa de uma formação linguística de qualidade para professores de línguas estrangeiras: o exame EPPLE**. In: Revista Horizontes de Linguística Aplicada, ano 13, n. 1, 2014. pp. 63-87.

CONSOLO, D. A. **A construção de um instrumento de avaliação da proficiência oral do professor de língua estrangeira.** Trabalhos em Linguística Aplicada, Campinas, v.43, n.2, p. 265-286, jul./dez. 2004.

CORBEL, C. (1993). **Computer-enhanced language assessment.** In G. Brindley (Series ed.) Research Report Series 2. National Centre for English Language Teaching and Research, Marquarie University, Sydney, Australia.

COURA-SOBRINHO, J. **O sistema de avaliação Celpe-Bras**. In: HORA, D. (Org.) Língua(s) e Povos: Universidade e Diversidade**.** João Pessoa: Ideia, 2006, p. 127-132.

DELL´ISOLA, R. L. P. (Org.) **O exame de proficiência Celpe-Bras em foco.** Campinas: Pontes, 2014.

DINIZ, L. R. A. **Política linguística do Estado brasileiro para a divulgação do português em países de língua oficial espanhola.** Trabalhos em Linguística Aplicada (online), v.51, n. 2, Dez, 2012, p.435-458.

DINIZ, L. R. A. **Mercado de línguas – A Instrumentalização brasileira do português como língua estrangeira.** Campinas: Editora RG, 2010.

DORNYEI, Z. **Qualitative data collection**. In: Research methods in Applied Linguistics: quantitative, qualitative and mixed methodologies. Oxford: OUP, 2007b. Ch. 6, p. 96-100.

DUNKEL, P. A. (1999). **Considerations in developing or using second/foreign language proficiency computer-adaptive tests.** [online]. Language Learning & Technology, *2*(2), 77–93. http://llt.msu.edu/vol2num2/article4/

ETS. Educational Testing Service. **Reliability and Comparability of TOEFL iBT Scores.** 2015a. [online]. TOEFL iBT Research Series 1, Volume 3. pp. 1-16. https://www.ets.org/toefl.

ETS. Educational Testing Service. **Validity Evidence Supporting the Interpretation and Use of TOEFL iBT Scores.** 2015b. [online]. TOEFL iBT Research Series 1, Volume 4. pp. 1-16. https://www.ets.org/toefl.

ETS. Educational Testing Service. **TOEIC Bridge Examinee Handbook: speaking and writing.** 2015c. [online]. pp. 1-36. https://www.ets.org/toeic.

FARACO, C. A. **História sociopolítica da língua portuguesa.** 1ª. ed. São Paulo: Parábola Editorial, 2016.

FRANÇA, J. L.; VASCONCELOS, A. C. A. G. **Manual de Normatização de Publicações Técnico-científicas.** 8ª. Edição. Belo Horizonte: Ed. UFMG, 2009.

FERREIRA, L. M. L. **Habilidades de leitura na proposta de interação face a face do exame Celpe-Bras.** 158 págs. Dissertação de Mestrado em Estudos Linguísticos - Faculdade de Letras, Universidade Federal de Minas Gerais. Belo Horizonte, 2012.

FURTOSO, V. B. **Desempenho oral em português para falantes de outras línguas: da avaliação à aprendizagem de línguas estrangeiras em contexto online.** 285 págs. Dissertação de Mestrado na Universidade Estadual Paulista Júlio de Mesquita Filho. São José do Rio Preto, São Paulo, 2011.

FURTOSO, V. B., GOMES, M. J. & CONSOLO, D. A. (2011). **Os serviços de *podcasting* na otimização da aprendizagem e da avaliação de língua estrangeira em contexto online.** In: Paulo Dias e Antonio José Osório (orgs.) Actas da VII Conferência Internacional de Tecnologias de Informação e Comunicação na Educação – Challenges 2011, Braga: Centro de Competência da Universidade do Minho, pp. 767-781. ISBN 978-972-98456-9-7.

FULCHER, G. **Re-examining language testing: a philosophical and social inquiry.** New York: Routledge, 2015.

FULCHER, G. **Practical language testing**. London: Routledge, 2010.

FULCHER, G. **Test use and political philosophy**. Annual Review of Applied Linguistics (2009) 29, 3–20. Printed in the USA. [online]. Cambridge University Press. 0267-1905/09 DOI:10.1017/S0267190509090023.

GODWIN-JONES, R. (2001). **Emerging technologies: Language testing tools and technologies.** Language Learning & Technology, *5*(2), 8–12. Retrieved from http://llt.msu.edu/vol5num2/emerging/ ISSN 1094-3501. [online]. Acesso em: jan/20017.

GONÇALVES, L. **O Ensino de Português como Segunda Língua nos EUA: Desafios Antigos e Recursos Inovadores.** In: LUNA, J. M. F. de (Org). Ensino de Português nos Estados Unidos: história, desenvolvimento, perspectivas. Jundiaí: Paco Editorial, 2012. p. 43-56.

KYLE, K.; CROSSLEY, S. A. & McNAMARA, D. S. **Construct validity in TOEFL- iBT speaking tasks: Insights from natural language processing.** Language Testing. Sage: 2015. pp. 1–21. DOI: 10.1177/0265532215587391.

LARSEN-FREEMAN, D.; LONG, M. H. **An introduction to second language acquisition research.** New York: Addison Wesley Longman, 1991.

LUNA, J. M. F. de**. Por uma Historiografia da Formação de Professores de Português como Língua Estrangeira nos Estados Unidos**. In: LUNA, J. M. F. de (Org). Ensino de Português nos Estados Unidos: história, desenvolvimento, perspectivas. Jundiaí: Paco Editorial: 2012. p. 21-41.

McNAMARA, T. **Language Testing**. Oxford: Oxford University Press, 2000.

MESSICK, S. Validity. In R. L. Linn (ed.): **Educational Measurement.** (3rdedn.) Macmillan, 1989.

NORRIS, J. M. **Task-based Teaching and Testing.** In C. J. Doughty & M. H. Long (eds.), The Handbook of Language Teaching (pp.578-594).Oxford, UK: Wiley-Blackwell. 2009.

OLIVEIRA, G. M. **Política linguística e internacionalização: a língua portuguesa no mundo globalizado do século XXI**. Trabalhos em Linguística Aplicada [*online*]. v. 52, n.2, 2013, p. 409-433. [online]. Disponível em:<http://www.scielo.br/pdf/tla/v52n2/a10v52n2.pdf> Acesso em: novembro 2015.

PEARSON (2011). **Pearson Test of English Academic: Automated Scoring.** pp. 1-5. Disponível em: http://pearsonpte.com/. [online]. Acesso em: jan/2017.

PEARSON (2012). PAE, H. K. Research Note: **Construct validity of the Pearson Test of English Academic: A multitraitmultimethod approach.** [online]. pp. 1-8. Disponível em: http://pearsonpte.com/. Acesso em: jan/2017.

PITON-GONÇALVES, J. **Desafios e perspectivas da implementação computacional de testes adaptativos multidimensionais para avaliações educacionais.** 2012. 153 f. Tese (Doutorado)-Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2012. [online].

PITON-GONÇALVES, J.; ALUÍSIO, S. M. **An architecture for multidimensional computer adaptive test with educational purposes.** In: Brazilian Symposium on Multimedia and the Web (WebMedia '12), 18., 2012, New York, NY. Proceedings. New York: ACM, 2012. p. 17-24. [online].

PITON-GONÇALVES, J.; ALUÍSIO, S. M. **Teste Adaptativo Computadorizado Multidimensional com propósitos educacionais: princípios e métodos.** Ensaio: aval. pol. públ. Educ., Rio de Janeiro, v. 23, n. 87, p. 389-414, abr./jun. 2015. [online].

POWERS, D. E. **Validity: what does it mean for the TOEIC tests?** ETS: Research Report, 2010.

POWERS, D. E.; KIM, H-J & WENG, V. Z. **The redesigned TOEIC listening and reading: relations to test taker perceptions of proficiency in English.** ETS : Compendium Study, 2010.

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico.** 2.ed. Novo Hamburgo/RS: Universidade FEEVALE 2013. p. 54-57.

RACILAN, M. **Leitura comunicativa? A abordagem comunicativa nos livros didáticos de leitura instrumental em língua inglesa.** 200 págs. Dissertação de Mestrado em Estudos Linguísticos- Faculdade de Letras, Universidade Federal de Minas Gerais. Belo Horizonte, 2005.

SALOMÃO, A. C. B**.. Fatores a serem levados em consideração para o Desenvolvimento de testes de proficiência oral em contexto virtual.** Trab. Ling. Aplic., Campinas, v. 49, n.2, Jul./Dez. 2010, p. 323-341.

SELIGER, H. W.; SHOHAMY, E. **Second language research methods**. Oxford: Oxford University Press, 1989.

TOSTES, S. C. **Estudos Linguísticos, Inovação e Tecnologia – Artigos Critérios de avaliação em testes de proficiência à distância.** In: Revista Linguística do Programa de Pós-Graduação em Linguística da UFRJ. [online]. CEP 5 (2), 2009: DOI: http://dx.doi.org/10.17074/linguistica.v5i2.530. pp. 1-14

WAINER, H. (Ed.). (2000). **Computerized adaptive testing: A primer** (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

WALCZAK, A. **Computer-adaptive testing.** Cambridge English: Research Notes Issue 59 (2015). pp. 35-39. ISSN 1756-509X. [online].

ANEXOS

ANEXO I

GRADE DE AVALIAÇÃO HOLÍSTICA CELPE-BRAS – ENTREVISTADOR

GRADE DE AVALIAÇÃO DA INTERAÇÃO FACE A FACE	
Nota	**Descrição do desempenho do examinando**
5	Quando o examinando demonstra autonomia e desenvoltura, contribuindo bastante para o desenvolvimento da interação. Sua produção apresenta fluência e variedade ampla de vocabulário e de estruturas, com raras inadequações. Sua pronúncia é adequada e demonstra compreensão do fluxo natural da fala.
4	Quando o examinando demonstra autonomia e desenvoltura, contribuindo para o desenvolvimento da interação. Sua produção apresenta fluência e variedade ampla de vocabulário e de estruturas, com inadequações ocasionais na comunicação. Sua pronúncia pode apresentar algumas inadequações. Demonstra compreensão do fluxo natural da fala.
3	Quando o examinando contribui para o desenvolvimento da interação. Sua produção apresenta fluência, mas também algumas inadequações de vocabulário, estruturas e/ou pronúncia. Demonstra compreensão do fluxo natural da fala.
2	Quando o examinando contribui para o desenvolvimento da interação. Apresenta poucas hesitações, com algumas interrupções no fluxo da conversa. Sua produção apresenta inadequações de vocabulário, estruturas e/ou pronúncia. Pode demonstrar alguns problemas de compreensão do fluxo da fala.
1	Quando o examinando contribui pouco para o desenvolvimento da interação. Sua produção apresenta muitas pausas e hesitações, ocasionando interrupções no fluxo da conversa ou apresenta alternância no fluxo de fala entre língua portuguesa e outra língua. Apresenta muitas limitações e/ou inadequações de vocabulário, estruturas e/ou pronúncia. Demonstra problemas de compreensão do fluxo natural da fala.
0	Quando o examinando raramente contribui para o desenvolvimento da interação. Sua produção apresenta pausas e hesitações muito frequentes, que interrompem o fluxo da conversa, ou apresenta fluxo de fala em outra língua. Apresenta muitas limitações e/ou inadequações de vocabulário, estruturas e/ou pronúncia, que comprometem a comunicação. Demonstra problemas de compreensão de fala simplificada e pausada.

ANEXO II

GRADE DE AVALIAÇÃO ANALÍTICA CELPE-BRAS – OBSERVADOR

	5	4	3	2	1	0
COMPREENSÃO	Compreensão do **fluxo natural da fala.** Rara necessidade de repetição e/ou reestruturação ocasionada por palavras menos frequentes e/ou por aceleração da fala.	Compreensão do **fluxo natural da fala.** Alguma necessidade de repetição e/ou reestruturação ocasionada por palavras menos frequentes e/ou por aceleração da fala.	Alguns problemas na **compreensão do fluxo** natural da fala. Necessidade de repetição e/ou reestruturação ocasionada por palavras de uso frequente, em ritmo normal da fala.	Alguns problemas na **compreensão do fluxo** natural da fala. Necessidade frequente de repetição e/ou reestruturação ocasionada por palavras de uso frequente, em ritmo normal da fala.	Muitos problemas na compreensão do **fluxo natural da fala.** Necessidade muito frequente de repetição e/ou reestruturação ocasionada por palavras básicas, em ritmo normal da fala.	Problemas sérios na compreensão do **fluxo natural da fala.** Necessidade constante de repetição e/ou reestruturação, mesmo em situação **de fala simplificada e** muito pausada.
COMPETÊNCIA INTERACIONAL	Apresenta muita desenvoltura e autonomia, contribuindo muito para o desenvolvimento da conversa. Quando necessário, faz uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Apresenta desenvoltura e autonomia. Não se limita a respostas breves, contribuindo para o desenvolvimento da conversa. Quando necessário, faz uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Não se limita a respostas breves, contribuindo para o desenvolvimento da conversa. Quando necessário, faz uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Pode se limitar a respostas breves, mas contribui para o desenvolvimento da conversa. Mesmo quando necessário, faz pouco uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Limita-se a respostas breves, contribuindo pouco para o desenvolvimento da conversa. Mesmo quando necessário, faz pouco uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.	Limita-se a respostas breves, raramente contribuindo para o desenvolvimento da **conversa, que fica** totalmente dependente do avaliador. Mesmo quando necessário,não faz uso de estratégias (reformulações, paráfrases, correções) para resolver problemas lexicais, gramaticais e/ou fonológicos.
FLUÊNCIA	Pausas e hesitações para organização do pensamento e, eventualmente, para resolver algum problema de construção linguística, sem **interrupções no fluxo** da conversa.	Pausas e hesitações para organização do pensamento e, eventualmente, para resolver algum problema de construção linguística, com poucas inter**rupções no fluxo da** conversa.	Pausas e hesitações para organização do pensamento e, algumas vezes, para resolver algum problema de construção linguística, com algumas **interrupções no fluxo** da conversa.	Pausas e hesitaçõe para organização do pensamento e para resolver algum problema de construção linguística, com interrupções no **fluxo da conversa.**	Pausas e hesitações frequentes exigem um grande esforço do interlocutor, ou **alternância no fluxo** da fala entre língua portuguesa e outra língua.	Pausas e hesitações muito frequentes in**terrompem o fluxo da conversa, ou fluxo de** fala em outra língua.
ADEQUAÇÃO LEXICAL	Vocabulário amplo e adequado para a discussão de tópicos do cotidiano e para a expressão de ideias e opiniões sobre assuntos variados. Raras interferências de outras línguas.	Vocabulário amplo e adequado para a discussão de tópicos do cotidiano e para a expressão de ideias e opiniões sobre assuntos variados. Poucas interferências de outras línguas.	Vocabulário adequado para a discussão de tópicos do cotidiano e para a expressão de ideias e opiniões sobre assuntos variados. Algumas interferências de outras línguas, com ocasional comprometimento da interação.	Vocabulário adequado para a discussão de tópicos do cotidiano com algumas limitações que podem interferir no desenvolvimento de ideias. Algumas interferências da língua materna, ocasionando algum comprometimento da interação.	Vocabulário inadequado e/ou limitado para a discussão de tópicos do cotidiano e para expressar ideias e opiniões sobre assuntos variados. Muitas interferências de outras línguas, ocasionando frequente comprometimento da interação.	Vocabulário muito inadequado e/ ou limitado para a discussão de tópicos do cotidiano e para expressar ideias e opiniões sobre assuntos variados. Muitas interferências de outras línguas, comprometendo a interação.
ADEQUAÇÃO GRAMATICAL	uso de variedade ampla de estruturas. Raras inadequações na utilização de estruturas.	uso de variedade ampla de estruturas. Poucas inadequações na utilização de estruturas complexas e raras inadequações no uso de estruturas básicas.	uso de variedade de estruturas. Algumas inadequações na utilização de estruturas complexas e poucas inadequações no uso de estruturas básicas.	uso da variedade limitada de estruturas. Inadequações mais frequentes tanto na utilização de estruturas complexas quanto nas básicas.	uso de variedade limitada de estruturas. Muitas inadequações na utilização de estruturas básicas e complexas.	uso de variedade bastante limitada de estruturas. Muitas inadequações na utilização de estruturas básicas e complexas, comprometendo a interação.
PRONÚNCIA*	Pronúncia (sons, ritmo e entonação) adequada.	Pronúncia (sons, ritmo e entonação) com algumas inadequações e/ou interferências de outras línguas.	Pronúncia (sons, ritmo e entonação) com inadequações e/ou interferências de outras línguas.	Pronúncia (sons, ritmo e entonação) com inadequações e/ou interferências frequentes de outras línguas.	Pronúncia (sons, ritmo e entonação) inadequada e/ou interferências acentuadas de outras línguas.	Pronúncia (sons, ritmo e entonação) inadequada e/ou interferências muito acentuadas de outras línguas.

ANEXO III

LEVELS OF THE *COMMOM EUROPEAN FRAMEWORK OF REFERENCE* – CEFR

LEVEL	DESCRIPTION
C2 Mastery	The capacity to deal with material which is academic or cognitively demanding, and to use language to good effect at a level of performance which may in certain respects be more advanced than that of an average native speaker. Example: *CAN scan texts for relevant information, and grasp main topic of text, reading almost as quickly as a native speaker.* All practice tests at this level
C1 Effective Operational Proficiency	The ability to communicate with the emphasis on how well it is done, in terms of appropriacy, sensitivity and the capacity to deal with unfamiliar topics. Example: *CAN deal with hostile questioning confidently. CAN get and hold onto his/her turn to speak.* All practice tests at this level
B2 Vantage	The capacity to achieve most goals and express oneself on a range of topics. Example: *CAN show visitors around and give a detailed description of a place.* All practice tests at this level
B1 Threshold	The ability to express oneself in a limited way in familiar situations and to deal in a general way with nonroutine information. Example: *CAN ask to open an account at a bank, provided that the procedure is straightforward.* All practice tests at this level
A2 Waystage	An ability to deal with simple, straightforward information and begin to express oneself in familiar contexts. Example: *CAN take part in a routine conversation on simple predictable topics.* All exams and practice tests at this level
A1 Breakthrough	A basic ability to communicate and exchange information in a simple way. Example: *CAN ask simple questions about a menu and understand simple answers.*

ANEXO IV

LISTA DE VERIFICAÇÃO PARA AVALIAR A UTILIDADE DE TESTES

(A checklist for evaluating usefulness, by Bachman & Palmer, 1996)

Questions for logical evaluation of usefulness	Extent to which quality is satisfied	Explanation of how quality is satisfied
Reliability		
1 To what extent do characteristics of the test setting vary from one administration of the test to another?		
2 To what extent do characteristics of the test rubric vary in an unmotivated way from one part of the test to another, or on different forms of the test?		
3 To what extent do characteristics of the test input vary in an unmotivated way from one part of the test to another, from one task to another, and on different forms of the test?		
4 To what extent do characteristics of the expected response vary in an unmotivated way from one part of the test to another, or on different forms of the test?		
5 To what extent do characteristics of the relationship between input and response vary in an unmotivated way from one part of the test to another, or on different forms of the test?		
Construct validity		
Clarity and appropriateness of the construct definition, and the appropriateness of the task characteristics with respect to the construct definition		
6 Is the language ability construct for this test clearly and unambiguously defined?		
7 Is the language ability construct for the test relevant to the purpose of the test?		
8 To what extent does the test task reflect the construct definition?		
9 To what extent do the scoring procedures reflect the construct definition?		
10 Will the scores obtained from the test help us to make the desired interpretations about test takers' language ability?		
Possible sources of bias in the task characteristics		
11 What characteristics of the test *setting* are likely to cause different test takers to perform differently?		
12 What characteristics of the test *rubric* are likely to cause different test takers to perform differently?		
13 What characteristics of the test *input* are likely to cause different test takers to perform differently?		
14 What characteristics of the expected *response* are likely to cause different test takers to perform differently?		
15 What characteristics of the relationship between *input* and *response* are likely to cause different test takers to perform differently?		
Authenticity		
16 To what extent does the description of tasks in the TLU domain include information about the setting, input, expected response, and relationship between input and response?		

17 To what extent do the characteristics of the test task correspond to those of TLU tasks?		
Interactiveness		
Involvement of the test takers' topical knowledge		
18 To what extent does the task presuppose the appropriate area or level of topical knowledge, and to what extent can we expect the test takers to have this area or level of topical knowledge?		
Suitability of test tasks to the personal characteristics of the test takers		
19 To what extent are the personal characteristics of the test takers included in the design statement?		
20 To what extent are the characteristics of the test tasks suitable for test takers with the specified personal characteristics?		
Involvement of the test takers' language knowledge		
21 Does the processing required in the test task involve a very narrow range or a wide range of areas of language knowledge?		
Involvement of language functions in the test tasks		
22 What language functions, other than the simple demonstration of language ability, are involved in processing the input and formulating a response?		
Involvement of the test takers' metacognitive strategies		
23 To what extent are the test tasks interdependent?		
24 How much opportunity for strategy involvement is provided?		
Involvement of the test takers' affective schemata in responding to the test tasks		
25 Is this test task likely to evoke an affective response that would make it relatively easy or difficult for the test takers to perform at their best?		
Impact		
Impact on individuals		
Impact on test takers		
26 To what extent might the experience of taking the test or the feedback received affect characteristics of test takers that pertain to language use (such as topical knowledge, perception of the target language use situation, areas of language knowledge, and use of strategies)?		
27 What provisions are there for involving test takers directly, or for collecting and utilizing feedback from test takers, in the design and development of the test?		
28 How relevant, complete, and meaningful is the feedback that is provided to test takers?		
29 Are decision procedures and criteria applied uniformly to all groups of test takers?		

30 How relevant and appropriate are the test scores to the decisions to be made?		
31 Are test takers fully informed about the procedures and criteria that will be used in making decisions?		
32 Are these procedures and criteria actually followed in making the decisions?		
Impact on teachers		
33 How consistent are the areas of language ability to be measured with those that are included in teaching materials?		
34 How consistent are the characteristics of the test and test tasks with the characteristics of teaching and learning activities?		
35 How consistent is the purpose of the test with the values and goals of teachers and of the instructional program?		
Impact on society and education systems		
36 Are the interpretations we make of the test scores consistent with the values and goals of society and the education system?		
37 To what extent do the values and goals of the test developer coincide or conflict with those of society and the education system?		
38 What are the potential consequences, both positive and negative, for society and the education system, of using the test in this particular way?		
39 What is the most desirable positive consequence, or the best thing that could happen, as a result of using the test in this particular way, and how likely is this to happen?		
40 What is the least desirable negative consequence, or the worst thing that could happen, as a result of using the test in this particular way, and how likely is this to happen?		
Practicality		
41 What type and relative amounts of resources are required for: (a) the design stage, (b) the operationalization stage, and (c) the administration stage?		
42 What resources will be available for carrying out (a) (b), and (c) above?		

ANEXO V

LISTA PARA DESCREVER AS TAREFAS DE USO DA LÍNGUA ALVO

(TLU task checklist, by Bachman & Palmer, 1996)

	TLU task
Characteristics of the setting: Physical setting	
Participants	
Time of task	
Characteristics of input: Format	
Channel	
Form	
Language	
Type	
Characteristics of expected response: Format	
Channel	
Form	
Language	
Type	
Relationship between input and response	
Reactivity	
Scope	
Directness	

ANEXO VI

LISTA DE AMEAÇAS À VALIDADE E ÀS RESPOSTAS

(Summary of the potential validity threats and responses, by Chapelle & Douglas, 2006)

Potential threat to validity	Responses
Different test performance	Test users need to consider whether or not this is a real threat, and if so research is needed to compare performance on would-be parallel forms.
New task types	The interpretation of performance on new task types needs to be examined qualitatively and quantitatively so they can be used appropriately.
Limitations due to adaptive item selection	A variety of types of adaptivity needs to be explored.
Inaccurate automatic response scoring	Coordinated research and development efforts need to be directed at development of multifaceted, relevant criteria for evaluating machine scoring on a variety of constructed response items.
Compromised security	Security needs to be considered in view of the level of security required for a particular test purpose.
Negative consequences	The potential positive and negative consequences of CALT need to be understood and planned for, and then consequences need to be documented.

ANEXO VII

LISTA DOS ASPECTOS POSITIVOS E NEGATIVOS DO CALT

(Technology-related positive and negative aspects of CALT, by Chapelle & Douglas, 2006)

Quality	Positive attributes	Negative attributes
Reliability	Partial-credit scoring implemented by computer should provide more precise measurement, larger variance in scores, and high coefficient-alpha reliabilities. Computer-assisted scoring rubrics perform consistently from one occasion to another. CAT algorithms test continuously until a score with the desired reliability has been obtained.	
Construct validity	Constructs of academic reading, listening, and online composing can be reflected in computer-assisted test tasks. Open-ended responses are less likely than multiple-choice to be affected by systematic test-taking strategies.	The types of items that can be developed in computer-assisted formats are different from those that can be developed with other media. Performance on a computer-delivered test may fail to reflect the same ability as that which would be measured by other forms of assessment. Selection of items to be included on an adaptive test by an algorithm may not result in an appropriate sample of test content. Computer-assisted response scoring may fail to assign credit to the qualities of a response that are relevant to the construct that the test is intended to measure.
Authenticity	Computer-assisted test tasks simulate some tasks in the target language use domain.	Computer-assisted test tasks are dissimilar to some tasks in the target language use domain.
Interactiveness	Multimedia input may offer opportunities for enhanced interactiveness.	Short items sometimes used on computer-adaptive tests may limit interactiveness.

Quality	Positive attributes	Negative attributes
Impact	Anticipation of CALT should prompt computer work in L2 classes, which may help L2 learners gain important skills.	Tests may cause anxiety for examinees who do not have extensive experience using technology.
	Language programs may be prompted to make computers available to learners and teachers.	Tests may be so expensive that some examinees may not be able to take them.
		The research, development, and delivery infrastructure for CALT may funnel testing resources to technology in lieu of other, at least equally important, research areas.
Practicality	Computational analysis of responses makes open-ended questions a possibility for an operational testing program.	It may be too expensive to prepare the partial-credit scoring for items on a regular basis.
	Internet-delivered tests add flexibility of time and place for test delivery.	Internet-delivered tests can raise questions about test security in high-stakes testing.

ANEXO VIII

LISTA DAS CARACTERÍSTICAS DO METÓDO DE TESTE E AS VANTAGENS E LIMITAÇÕES DO CALT

(Test method characteristics and CALT advantages and limitations, by Chapelle & Douglas, 2006)

Test method characteristics	CALT advantages	CALT limitations
Physical and temporal circumstances Location, time, personnel	CALTs can be taken at many convenient locations, at convenient times, and largely without human intervention.	Security is an issue in high-stakes tests; equipment not standardized nor universally available; IT expertise required for establishment, maintenance.
Rubric/Instructions Procedures for responding	Test tasks are presented in a consistent manner for all test takers and instructions and input are presented automatically and uniformly, making for enhanced fairness.	Different levels of instructions, voluntary help screens, different languages of instructions can detract from uniformity.
Input and expected response Features of the context: *setting, participants, tone* Format: *visual/audio/ video*	Multimedia capabilities allow for a variety of input and response types, enhancing contextualization and authenticity.	Input and response types are limited by available technology.
Interaction between the Input and Response Reactivity: *reciprocal*	Computers can adapt input in response to test takers' responses and actions, allowing for computer-adaptive tests and rapid feedback.	Interactiveness is more controlled than certain other formats; computer's ability to sample fairly may be limited; CATs are expensive to develop.
Characteristics of assessment Construct definition Criteria for correctness Scoring procedures	Natural language processing (NLP) technology allows for automated scoring of complex responses, affecting the construct definition, scoring criteria, and procedures.	NLP technology is new, expensive, and limited, thus creating potential problems for construct definition and validity.