



Revista on line de Política e Gestão Educacional  
Online Journal of Policy and Educational Management



<sup>1</sup> Universidade de Ordu, Ordu, Turquia. Docente do Departamento de Línguas Estrangeiras, Escola de Línguas Estrangeiras.

<sup>2</sup> Universidade de Hacettepe, Ancara, Turquia. Professora Associada do Departamento de Ensino de Língua Inglesa, Faculdade de Educação.

## TESTES ON-LINE E TECNOLOGIA NA AVALIAÇÃO DE LÍNGUAS

*PRUEBAS EN LÍNEA Y TECNOLOGÍA EN LA EVALUACIÓN DEL LENGUAJE*

*ONLINE TESTS AND TECHNOLOGY IN LANGUAGE ASSESSMENT*

Nurgül BEKDEMİR<sup>1</sup>  
nurgulbekdemir@odu.edu.tr  
İsmail Firat ALTAY<sup>2</sup>  
ifaltay@hacettepe.edu.tr



### Como referenciar este artigo:

Bekdemir, N., & Altay, I. F. (2025). Testes on-line e tecnologia na avaliação de línguas. *Revista on line de Política e Gestão Educacional*, 29(esp.1), e025039. DOI: 10.22633/rpge.v29iesp1.20474

**Submetido em:** 19/05/2025

**Revisões requeridas em:** 13/06/2025

**Aprovado em:** 30/06/2025

**Publicado em:** 31/07/2025

**RESUMO:** A tecnologia é uma parte indispensável do processo de ensino e aprendizagem de línguas, incluindo a avaliação. Este estudo investiga as ferramentas e plataformas tecnológicas na testagem e avaliação on-line de línguas em diversos aspectos. Explora tanto a avaliação formativa quanto a somativa em testes realizados em sala de aula e em larga escala, bem como ferramentas on-line que facilitam a avaliação alternativa. Modelos como a Avaliação Baseada em Computador (CBA) e a Avaliação Adaptativa por Computador (CAT) também são ilustrados. A inteligência artificial, que tem um efeito revolucionário na avaliação linguística, é analisada juntamente com as plataformas atuais. Além disso, questões críticas na avaliação linguística, como problemas de confiabilidade e validade e críticas básicas à avaliação on-line com base em autenticidade, interatividade, impacto e praticidade, são apresentadas com suas vantagens e desvantagens. Consequentemente, o estudo visa ilustrar como a avaliação linguística está sendo transformada pela tecnologia e oferecer insights significativos para designers de testes, educadores e formuladores de políticas para aprimorar as práticas de testagem.

**PALAVRAS-CHAVE:** Avaliação on-line. Avaliação baseada em computador. Avaliação adaptativa por computador. Avaliação alternativa on-line. Inteligência artificial.

**RESUMEN:** La tecnología es una parte indispensable del proceso de enseñanza y aprendizaje de idiomas, incluida la evaluación. Este estudio analiza herramientas y plataformas tecnológicas en la evaluación y pruebas lingüísticas en línea desde diversas perspectivas. Explora tanto las evaluaciones formativas como sumativas en pruebas aplicadas en el aula y a gran escala, así como herramientas en línea que facilitan la evaluación alternativa. Se ilustran modelos como la Evaluación Basada en Computadora (CBA) y las Pruebas Adaptativas por Computadora (CAT). También se examina la inteligencia artificial, que tiene un efecto revolucionario en la evaluación lingüística, junto con las plataformas actuales. Asimismo, se abordan cuestiones críticas como la fiabilidad, validez y críticas sobre la autenticidad, interactividad, impacto y practicidad de la evaluación en línea, presentando ventajas y desventajas. El estudio busca demostrar cómo la tecnología transforma la evaluación lingüística y ofrecer perspectivas relevantes para diseñadores de pruebas, docentes y responsables de políticas educativas.

**PALABRAS CLAVE:** Evaluación en línea. Evaluación basada en computadora. Evaluación adaptativa por computadora. Evaluación alternativa en línea. Inteligencia artificial.

**ABSTRACT:** Technology has an indispensable part of the language learning and teaching process, including assessment. This study investigates technological tools and platforms in on-line language testing and assessment within various aspects. It explores both formative and summative assessments at classroom-based and large-scale tests and also on-line tools having a facilitating effect on alternative assessment. The models, such as Computer-Based Assessment (CBA) and Computer-Adaptive Testing (CAT), are further illustrated. Artificial intelligence, which has a revolutionary effect on language assessment, is scrutinized with the current platforms. Also, critical issues in language assessment, such as reliability and validity issues and basic criticism of on-line language assessment based on authenticity, interactivity, impact, and practicality, are presented with the advantages and disadvantages. Consequently, the study aims to illustrate how language assessment is transformed through technology and offer significant insights for test designers, educators, and policymakers to enhance testing practices.

**KEYWORDS:** On-line Assessment. Computer-based Assessment. Computer Adaptive Testing. On-line Alternative Assessment. Artificial Intelligence.

Artigo submetido ao sistema de similaridade



Editor: Prof. Dr. Sebastião de Souza Lemes

Editor Adjunto Executivo: Prof. Dr. José Anderson Santos Cruz.

## INTRODUÇÃO

A tecnologia impactou significativamente praticamente todos os aspectos da vida contemporânea, incluindo o ensino e a aprendizagem de línguas. Atualmente, tanto professores quanto estudantes têm acesso facilitado a ferramentas tecnológicas que potencializam esse processo. Considerando a relevância amplamente reconhecida da avaliação nesse contexto, tornou-se essencial integrar a tecnologia ao universo avaliativo. Conhecida como avaliação mediada por tecnologias da informação e comunicação (TICs) ou avaliação on-line, essa abordagem proporciona oportunidades sem precedentes em áreas como discussões e interações virtuais, sistemas de correção automatizada e acesso a uma ampla gama de recursos, incluindo mídias digitais (Dunn et al., 2005).

É igualmente relevante considerar que os ambientes de aprendizagem devem ser moldados para atender às necessidades e ao perfil de cada indivíduo. A avaliação on-line contempla essas particularidades ao oferecer múltiplos formatos avaliativos que se ajustam ao ritmo, tempo, quantidade de itens, integração de multimídia, finalidade do usuário e diversidade de contextos avaliativos (Chapelle & Douglas, 2006; Dunn et al., 2005). Essa modalidade é amplamente utilizada por docentes, discentes e instituições. Os professores a empregam tanto para fins formativos quanto somativos; já as instituições a utilizam para testes diagnósticos, classificatórios, de proficiência, avaliações intermediárias e exames finais, comumente voltados à aferição somativa. Adicionalmente, os estudantes podem monitorar seu progresso na aprendizagem linguística e se preparar para essas avaliações. Uma das características centrais da avaliação on-line é sua natureza centrada no usuário, permitindo que o processo avaliativo e de aprendizagem ocorra a qualquer momento e em qualquer lugar, respeitando o ritmo individual.

## Revisão de Literatura

Com o avanço tecnológico, os processos de avaliação em línguas passaram por profundas transformações. Esta seção examina as avaliações on-line realizadas em sala de aula, os exames em larga escala, as abordagens avaliativas alternativas em ambientes virtuais, as avaliações baseadas em computador (Computer-Based Assessment – CBA) e o papel da inteligência artificial na testagem linguística.

## Tipos de Avaliação

A avaliação formativa (AF) refere-se às práticas avaliativas voltadas à melhoria do desempenho dos aprendizes (Harmer, 2007), promovendo a continuidade da aprendizagem por meio do feedback docente (Fulcher & Davidson, 2007; Stannard & Skip Basiel, 2013). A AF

fundamenta-se fortemente no retorno qualitativo, cuja finalidade é auxiliar o aluno a reduzir a distância entre seu desempenho atual e o padrão esperado (Lin & Lai, 2013; Taras, 2005). Os profissionais que implementam esse tipo de avaliação devem utilizar instrumentos mensuráveis que possibilitem um feedback construtivo, esperando-se que os estudantes revisem suas produções com base nas orientações recebidas. Esse ciclo retroalimentado promove um impacto significativo no desenvolvimento linguístico (Stannard & Skip Basiel, 2013).

Em contrapartida, a avaliação somativa (AS) ocorre ao final de uma unidade, módulo ou curso e tem como objetivo mensurar os resultados obtidos individual ou coletivamente, bem como o progresso alcançado (Brown, 2000; Harmer, 2007). Alguns estudiosos indicam que métodos de AS aplicados on-line podem substituir as avaliações tradicionais em papel (Navrat & Tvarozek, 2014). Em vez de serem concebidas como abordagens opostas, atualmente recomenda-se a articulação entre AF e AS para se alcançarem melhores resultados educacionais. Enquanto a AF foca na avaliação dos processos, a AS concentra-se nos produtos (Taras, 2005). Ao envolver os estudantes com feedback contínuo e estimular a revisão de suas produções, a AF os prepara para a AS, que verifica a eficácia do processo de aprendizagem.

Um exemplo disso pode ser observado nas práticas de escrita por processo e escrita por produto, amplamente exploradas em contextos de avaliação on-line. Na escrita por processo, os aprendizes utilizam plataformas como *Turnitin Draft Coach*, *Turnitin Feedback Studio* e *ETS Criterion* para elaborar, revisar e editar seus textos com base no retorno docente, buscando alinhar suas produções aos padrões esperados. O *Turnitin Draft Coach*, disponível como complemento para o Microsoft® Word e o Google Docs™, fornece feedback instantâneo aos estudantes, economizando o tempo dos docentes e reduzindo erros gramaticais, problemas de citação e de integridade acadêmica, o que resulta em mais oportunidades de aprendizagem formativa.

Por outro lado, a escrita por produto, típica da AS, exige que os estudantes demonstrem seus conhecimentos em uma única entrega. Ferramentas como o *Turnitin Originality Check* contribuem para garantir a integridade acadêmica, detectando plágio e incentivando a produção autoral. Em vez de manter uma divisão rígida entre AF e AS, os docentes são encorajados a integrar ambas as abordagens, de modo a potencializar os resultados da aprendizagem (Turnitin, s.d.a).

### *Avaliação baseada em computador*

Com a rápida incorporação dos computadores e da internet à vida cotidiana, sua aplicação na avaliação linguística tornou-se indispensável. A CBA, elemento central do ensino a distância (Lin & Lai, 2019), oferece interfaces multimídia diversificadas e cenários baseados em vídeo que simulam contextos reais de uso linguístico. Além disso, disponibiliza análises

automáticas de respostas e permite que os estudantes realizem avaliações a qualquer hora e em qualquer lugar (Burstein et al., 1996; Chapelle & Douglas, 2006).

A CBA também é conhecida como Teste de Língua Assistido por Computador (Chapelle & Douglas, 2006; Suvorov & Hegelheimer, 2013) e Avaliação de Língua Assistida por Computador (Chapelle & Voss, 2008). Desde a década de 1980, a CBA passou por transformações e aperfeiçoamentos significativos, como a introdução de testes adaptativos, novos tipos de itens, avaliações integradas de múltiplas habilidades e sistemas automatizados de correção (Suvorov & Hegelheimer, 2013, p. 1).

O desenvolvimento de um teste baseado em computador envolve diversas etapas: (1) planejamento do teste, (2) construção do teste, (3) pré-teste dos itens, (4) aplicação dos itens, (5) gerenciamento dos itens, (6) correção dos itens, (7) análise e interpretação dos resultados e (8) divulgação das pontuações (Burstein et al., 1996, p. 241–242). Especificamente: o planejamento do teste define a estrutura geral e determina se todas as quatro habilidades serão avaliadas para aferir a competência linguística global; a construção do teste envolve o uso de tecnologia para apresentar os textos e registrar diferentes tipos de resposta, escrita ou oral; o pré-teste refere-se à aplicação piloto antes da implementação definitiva; a aplicação dos itens garante a verificação da identidade dos participantes e a coleta segura das respostas; o gerenciamento dos itens concentra-se na armazenagem segura, assegurando a integridade do teste; a correção dos itens pode ser automatizada ou realizada manualmente; a análise e interpretação permitem que os participantes compreendam seus desempenhos e pontuações; por fim, a divulgação dos resultados é feita por meio de tecnologia, com envio rápido das notas, atendendo à expectativa dos candidatos. Em termos gerais, a CBA consiste na aplicação de testes por meio de computadores e pode ser utilizada tanto para avaliações formativas quanto somativas.

A CBA pode ser estruturada de três formas distintas: (1) linear, (2) semi-adaptativa e (3) adaptativa (Suvorov & Hegelheimer, 2013). Os testes lineares apresentam duas variações: na primeira, todos os itens são exibidos sequencialmente, permitindo ao participante visualizar toda a prova, semelhante ao formato tradicional em papel; na segunda, o participante não pode retornar às questões anteriores.

Os testes semi-adaptativos e adaptativos selecionam, respectivamente, conjuntos de itens (*testlets*) ou itens individualizados com base nas respostas anteriores. Esses modelos são considerados eficazes por ajustarem o nível de dificuldade conforme o desempenho do candidato: ao acertar uma questão, o sistema apresenta um item mais difícil; em caso de erro, um item mais fácil é selecionado. Por isso, os CATs são reconhecidos por sua precisão na mensuração da proficiência linguística e por fornecerem relatórios confiáveis a profissionais e instituições.

Além disso, os itens são extraídos de bancos de dados específicos (Chapelle & Douglas, 2006). Destaca-se ainda o papel fundamental da Teoria da Resposta ao Item (Item Response

Theory – IRT) no desenho dos CATs. A IRT parte de dois pressupostos: (1) mensuração politômica e (2) mensuração dicotômica (Jiao et al., 2012). Quando os itens são corrigidos de forma binária (certo/errado), aplica-se a mensuração dicotômica. Já a mensuração politômica é empregada quando os itens contribuem para um único construto e são avaliados com base em escalas. Por exemplo, Jamieson (2005) apresenta a habilidade de leitura avaliada dicotomicamente, com seis itens pontuados de 0 a 1, conforme a correção. Por outro lado, a escrita de ensaios é avaliada por meio de mensuração politômica, com notas variando de 0 a 5 por categoria, refletindo múltiplos níveis de desempenho.

### *Testes em larga escala: testes baseados na internet e no computador*

A eficácia dos exames de proficiência linguística em larga escala depende da consistência em aspectos como administração, correção, divulgação e interpretação dos resultados, especialmente em testes de nivelamento (Kunnan, 2008, p. 135). Esta seção apresenta exemplos de testes de línguas realizados por meio de plataformas on-line ou por computador.

O IELTS, disponível nas modalidades *IELTS Academic* e *IELTS General Training*, refere-se ao *International English Language Testing System*, administrado conjuntamente por Cambridge ESOL, British Council e IDP Education Australia (Kunnan, 2008, p. 140). Embora as duas versões compartilhem diversas características, o IELTS pode ser realizado tanto em papel quanto por meio digital, sendo o foco aqui a versão computadorizada.

O exame é amplamente aceito por empregadores, autoridades de imigração e entidades profissionais em diversos países (IELTS, s.d.a). Avalia quatro habilidades: compreensão auditiva, produção oral, leitura e escrita. As seções de escuta, leitura e escrita são realizadas no computador, enquanto a parte oral ocorre presencialmente com um examinador. A produção escrita é avaliada com base nos critérios de cumprimento da tarefa, coerência e coesão, repertório lexical e correção gramatical; a parte oral é avaliada quanto à fluência e coerência, vocabulário, gramática e pronúncia. Os resultados costumam estar disponíveis em até sete dias (IELTS, s.d.b).

O TOEFL Junior® Comprehensive é voltado para estudantes a partir de 11 anos de idade. Trata-se de um exame computadorizado que avalia habilidades de leitura, escuta, fala e escrita. As pontuações variam de 200 a 300 em cada seção, totalizando de 600 a 900 no geral. A correção das redações é realizada pelo mecanismo *e-rater*®, enquanto a performance oral é avaliada pelo sistema *SpeechRaterSM* (Evanini et al., 2015). O *SpeechRater* analisa fluência, pronúncia, prosódia, gramática, precisão e conteúdo; já o *e-rater* examina a proficiência escrita com base em dez variáveis, incluindo gramática, vocabulário e aspectos mecânicos do texto (Evanini et al., 2015).



O Oxford Test of English (OTE) é um exame geral de proficiência em língua inglesa que utiliza tecnologia de teste adaptativo por computador nas seções de compreensão auditiva (*listening*) e leitura (*reading*), e aplica randomização de tarefas nas seções de escrita (*writing*) e fala (*speaking*). O teste avalia as quatro habilidades linguísticas. Nas seções de *listening* e *reading*, os tipos de questões incluem múltipla escolha, preenchimento de lacunas e associação de informações. Na seção de escrita, os participantes devem redigir um e-mail e um ensaio ou uma resenha. Já a seção oral exige que discutam temas cotidianos e apresentem discursos breves sobre assuntos propostos (OTE, s.d.a). As pontuações variam de 51 a 140, e os resultados — aceitos por universidades e instituições de ensino — são divulgados em até 14 dias (OTE, s.d.b).

O TOEIC, sigla para Test of English for International Communication, destina-se à avaliação do nível geral de proficiência em inglês de profissionais em contextos corporativos. O exame é administrado pela Educational Testing Service (ETS) e possui três versões: *Listening & Reading*, *Speaking & Writing* e TOEIC Bridge (ETS, s.d.a). As duas primeiras são aplicadas em formato computadorizado, sendo o foco desta seção.

O exame *TOEIC Listening and Reading*, com duração de duas horas e meia, é composto por sete partes: a seção de compreensão auditiva apresenta diálogos curtos relacionados ao ambiente de trabalho, enquanto a seção de leitura inclui textos voltados a contextos empresariais (ETS, s.d.b). Na seção de fala, os candidatos devem responder às instruções e expressar opiniões sobre temas diversos. A seção de escrita é composta por três tarefas: redação de frases, de um e-mail e de um ensaio opinativo (ETS, s.d.c). A pontuação máxima possível é de 990 pontos.

O PTE Academic é um exame computadorizado destinado a estudantes ou imigrantes, aplicado em centros de testes distribuídos em diversos países (Pearson, s.d.a). O teste é dividido em três seções: *Speaking & Writing*, *Reading* e *Listening*, com tempo total de até três horas. A primeira seção inclui atividades como apresentação pessoal, leitura em voz alta, recontagem de palestras e produção de resumos e ensaios. A seção de leitura apresenta questões de múltipla escolha, reorganização de parágrafos e preenchimento de lacunas. A seção de escuta envolve tarefas como resumir textos orais, responder a questões objetivas, identificar palavras ausentes e escrever a partir da audição (Pearson, s.d.b).

A correção é realizada por um sistema automatizado de pontuação do PTE Academic, que classifica as respostas como corretas, incorretas ou parcialmente corretas. A pontuação é calculada por habilidade e também de forma global, com variação entre 10 e 90 pontos (Pearson, s.d.c).

O e-YDS (*Electronic Foreign Language Exam*) é um exame de proficiência aplicado em diferentes idiomas na Turquia. Seu formato é semelhante ao do exame tradicional em papel (YDS), com questões de múltipla escolha que avaliam gramática, vocabulário, compreensão de

leitura e habilidades de tradução (OSYM, 2019). A principal diferença é que o e-YDS é realizado por computador, e os resultados são divulgados imediatamente após o término do exame.

O TOEFL (*Test of English as a Foreign Language*), também administrado pela ETS, é uma avaliação amplamente reconhecida da proficiência em inglês. Sua versão *internet-based* (TOEFL iBT) indica que o teste é realizado on-line. Os resultados são aceitos em processos de admissão universitária e por instituições em mais de 150 países (ETS, s.d.d).

O TOEFL iBT avalia as quatro habilidades linguísticas: leitura, escuta, fala e escrita. Além das tarefas tradicionais de leitura e escuta, o teste inclui atividades independentes e integradas de produção oral e escrita. As tarefas independentes exigem que os candidatos expressem suas ideias de forma autônoma, por meio da fala ou da escrita. Já as tarefas integradas avaliam a capacidade de combinar competências — por exemplo: escutar e falar; ler e escutar, e em seguida falar; ou ainda ler, escutar e escrever. A pontuação total máxima é de 120 pontos.

No processo de correção do TOEFL iBT, as seções de leitura e escuta são avaliadas automaticamente, por se basearem em questões objetivas. Por outro lado, a produção oral é analisada por três a seis avaliadores humanos, enquanto a produção escrita é avaliada por uma combinação entre corretores humanos e o sistema automatizado da ETS, o que garante a confiabilidade dos resultados (ETS, s.d.e).

### **Avaliação alternativa on-line**

A avaliação alternativa, fortemente associada ao uso autêntico da linguagem em contextos reais, é referida na literatura por diferentes denominações: “avaliação autêntica”, “avaliação de desempenho”, “avaliação direta” e “avaliação descritiva” (Hamayan, 1995, p. 213). Essa abordagem contempla tanto o processo de aprendizagem quanto os seus resultados, permitindo procedimentos interativos de avaliação. Além disso, a avaliação alternativa promove maior autonomia, pois os aprendizes assumem mais responsabilidade pelo próprio desenvolvimento linguístico (Tsagari, 2004).

Por exemplo, a autoavaliação, a avaliação por pares, portfólios e diários são utilizados para avaliar o desenvolvimento linguístico dos aprendizes (Brown, 2000; Dikli, 2003). Ademais, a tecnologia potencializou a avaliação alternativa. Brown (2000) destaca que a autoavaliação e a avaliação por pares aprimoram a capacidade dos estudantes de monitorar o próprio desempenho e o dos colegas, favorecendo a revisão e o aprimoramento.

No âmbito da avaliação por pares, plataformas como Turnitin Feedback Studio e ETS Criterion oferecem oportunidades valiosas para que os aprendizes comentem as produções dos colegas por meio de ferramentas on-line guiadas (Turnitin, s.d.b). Para a autoavaliação, Turnitin Draft Coach e ETS Criterion são plataformas eficazes, onde os estudantes podem aprimorar suas habilidades de escrita acadêmica, identificando áreas que necessitam correção e



revisando seus textos até alcançarem uma versão ideal. O sistema proporciona feedback instantâneo, permitindo foco no desenvolvimento contínuo sem atrasos.

Em substituição aos portfólios tradicionais, recomenda-se o uso de e-portfólios, como o European Language Portfolio (ELP), LinguaFolio On-line e ePEARL. Esses instrumentos auxiliam os alunos a registrar tarefas, acompanhar seu progresso e gerenciar o processo de aprendizagem, além de estimular a autoavaliação, elemento fundamental na trajetória educacional.

Os diários, frequentemente associados às práticas reflexivas, exercem impacto significativo tanto sobre os aprendizes quanto sobre os professores. A tecnologia apoia a manutenção desses registros por meio de blogs e redes sociais, possibilitando o compartilhamento de vídeos e gravações de áudio. Ferramentas como Blogger, WordPress e Edmodo Blog contribuem para o aprimoramento das habilidades de escrita e oralidade.

De modo geral, em comparação aos métodos tradicionais de avaliação, a avaliação alternativa on-line enriquece o processo de aprendizagem por meio da integração multimídia e promove o crescimento acadêmico.

### *Inteligência Artificial (IA) na Avaliação de Línguas*

A IA está revolucionando a avaliação linguística. Com a incorporação de técnicas de processamento de linguagem natural (PLN) e algoritmos de aprendizado de máquina (ML), as habilidades linguísticas podem ser avaliadas de forma mais rápida e personalizada. Na última década, sistemas automatizados de correção baseados em IA — também chamados de sistemas automatizados de pontuação ou *automarkers* — têm sido amplamente adotados para avaliar as habilidades de fala e escrita, à medida que algoritmos passaram a ser treinados para analisar ensaios e respostas orais (Xu et al., 2025).

A IA oferece diversos benefícios, como correção rápida, aplicação de testes sob demanda sem necessidade de avaliador humano, consistência na pontuação, testes adaptativos, feedback imediato e integração fluida da avaliação ao processo de aprendizagem (Xi, 2021). Contudo, persistem desafios, entre eles preocupações quanto à autenticidade e validade dos modelos de correção, riscos de desonestidade acadêmica e a necessidade de alfabetização em IA.

A autenticidade permanece como limitação, dado que sistemas baseados em IA ainda demandam aprimoramento para lidar eficazmente com avaliações orais interativas. Problemas de validade surgem na análise de aspectos linguísticos complexos, como qualidade léxico-gramatical, organização do discurso, pragmática e fatores interacionais, por exemplo, a alternância de turnos. Ademais, o uso inadequado de ferramentas de IA, como o ChatGPT, pode favorecer práticas desonestas por meio da produção de conteúdos semelhantes aos humanos (Karani & Mwancha, 2025). Portanto, o desenvolvimento da alfabetização em IA é fundamental para que os usuários dos testes evitem usos indevidos e garantam a aplicação apropriada da tecnologia na avaliação linguística (Xu et al., 2025).

Diversas plataformas já incorporaram avaliação linguística baseada em IA, especialmente para as habilidades de escrita e fala. O Duolingo English Test (DET), por exemplo, é um exame adaptativo e baseado em IA que avalia a proficiência linguística em curto espaço de tempo. As habilidades produtivas, como escrita e fala, são avaliadas por sistemas automatizados de pontuação.

A ETS desenvolveu as ferramentas SpeechRater® e e-rater®, amplamente utilizadas em exames de larga escala, como o TOEFL, para avaliação confiável da fala e da escrita. Essas ferramentas são adotadas mundialmente por aprendizes. O Write & Improve by Cambridge, apoia o desenvolvimento da escrita, fornecendo feedback instantâneo gerado por IA. Para a habilidade oral, a Cambridge também está cuidando do Speak & Improve, projeto alinhado ao Quadro Europeu Comum de Referência para Línguas (QECR), voltado à prática da fala. Ferramentas como projeto alinhado ao Quadro Europeu Comum de Referência para Línguas (QECR), voltado à prática da fala. Ferramentas como TalkPal AI e SpeakNow AI complementam o aprendizado com práticas de fala e pronúncia, sendo o SpeakNow AI também capaz de realizar avaliações baseadas nos níveis do QECR.

Esses exemplos demonstram como as ferramentas baseadas em IA tornam a avaliação linguística mais acessível, flexível e personalizada, apoiando os aprendizes na conquista de maior proficiência.

### *Crítica básica ao uso da tecnologia na avaliação de línguas*

Esta seção concentra-se na avaliação da CBA, incluindo a avaliação on-line e o CAT. São discutidos cinco critérios: (1) confiabilidade e validade, (2) autenticidade, (3) interatividade, (4) impacto e (5) praticidade.

No que tange à confiabilidade e validade, diversos estudos apresentam perspectivas divergentes sobre a validade e confiabilidade da CBA (Dunkel, 1991; Huff & Sireci, 2001; Öz, 2018; Parhizgar, 2012; Roever, 2001; Shraim, 2018). A confiabilidade do teste depende de fatores como instruções claras, familiaridade com o formato da CBA, condições ambientais durante a aplicação e características individuais dos candidatos (Dunkel, 1991). A correção das respostas desempenha papel crítico na confiabilidade; como o fator humano pode ser eliminado na CBA, a parcialidade na correção é reduzida (Dunkel, 1991; Öz, 2018; Shraim, 2018), tornando a confiabilidade do avaliador menos preocupante no formato computacional. Contudo, a validade de ferramentas como e-rater e SpeechRater tem sido questionada (Parhizgar, 2012), e, atualmente, essas ferramentas não são geralmente usadas como única forma de avaliação das habilidades de escrita e fala.

Outro ponto relevante é que os computadores possibilitam a “atribuição de pontuação parcial, o que permite uma mensuração mais precisa e uma amplitude maior de escores”

(Chapelle & Douglas, 2006, tradução nossa). Além disso, Öz (2018) observou que os candidatos podem ouvir os áudios individualmente, com o uso de fones de ouvido, o que melhora a qualidade sonora e reduz distrações externas. Altay e Altay (2017) também não identificaram impacto significativo nos escores de aprendizes ao lerem textos on-line, corroborando a confiabilidade do uso da tecnologia na avaliação.

Por outro lado, questões técnicas como “computadores lentos, má conexão de rede ou falta de energia” podem levar os candidatos a perceberem a CBA como pouco confiável (Shraim, 2018). Ademais, alguns estudos argumentam que a CBA está limitada a determinados tipos de itens, como múltipla escolha, verdadeiro/falso, associação e respostas curtas (Cook & Jenkins, 2010; Hodgson & Pang, 2012; Roevers, 2001). Outros, entretanto, defendem que CBAs bem elaboradas podem avaliar habilidades cognitivas de ordem superior (Huff & Sireci, 2001; Jordan, 2013; Simkin & Kuechler, 2005; Williamson, 2018).

Itens dissertativos podem minimizar essas limitações, embora o CAT possa enfrentar dificuldades para selecionar tarefas que mensurem eficazmente habilidades específicas (Chapelle & Douglas, 2006). Itens inovadores, como componentes de áudio/vídeo e tarefas simuladas de desempenho, tornam as CBAs mais autênticas que os testes tradicionais em papel, potencialmente aumentando a validade de construto (Huff & Sireci, 2001; Öz, 2018). Não obstante, fatores como letramento digital, restrições de tempo, ansiedade relacionada a formatos CAT, custos elevados, velocidade de digitação e falhas técnicas podem impactar negativamente a validade dos testes (Huff & Sireci, 2001; Roevers, 2001).

Na avaliação linguística, autenticidade refere-se à capacidade do candidato de utilizar a língua de forma eficaz em contextos reais. Dado que a internet disponibiliza vasto material autêntico, torna-se relativamente simples integrar tarefas reais às avaliações linguísticas. Elementos multimídia, como áudio, imagens, gráficos e vídeos, tornam os CBTs mais realistas e envolventes (Chapelle & Douglas, 2006; Shraim, 2018). Por exemplo, o candidato pode assistir a um vídeo de uma conversa e responder a perguntas correlatas. A vantagem do uso de vídeo em tarefas de compreensão auditiva é a visualização do cenário, conferindo maior autenticidade e refletindo experiências reais (Huff & Sireci, 2001).

A interatividade é igualmente fundamental na avaliação linguística, pois permite que o candidato demonstre “estratégias linguísticas e comunicativas, conhecimentos sobre temas e variáveis afetivas de forma integrada” (Chapelle & Douglas, 2006, p. 95, tradução nossa). Os CBTs promovem a interação por meio de tipos de tarefa como interpretação de gráficos e redação de ensaios fundamentados neles, exigindo do aprendiz habilidades de resolução de problemas e expressão escrita. O fator-chave é “a interação entre o candidato e a tarefa” (Bachman & Palmer, 1996, p. 25-26, tradução nossa).

Os testes também geram impactos em níveis micro e macro. No nível micro, influenciam diretamente os indivíduos que os realizam. No nível macro, afetam os sistemas educacionais

e a sociedade como um todo (Bachman & Palmer, 1996). Chapelle e Douglas (2006) destacam tanto os impactos positivos quanto os negativos da CBA. Entre os aspectos positivos, o uso crescente da CBA pode levar as instituições a disponibilizarem mais computadores para a aplicação de testes de língua. Por outro lado, a falta de experiência prévia com esse tipo de avaliação pode gerar ansiedade nos participantes, e os altos custos envolvidos podem impedir que algumas pessoas realizem o teste, mesmo quando ele for necessário.

Por fim, a praticidade constitui um fator essencial, a par dos demais critérios. Esse conceito abrange aspectos como “limitações financeiras, restrições de tempo, administração, pontuação e interpretação” (Brown, 2000, p. 386, tradução nossa). O desenvolvimento de CBA, pode demandar investimentos elevados, devido à necessidade de expertise técnica (Chapelle & Douglas, 2006). No entanto, essa modalidade oferece vantagens, como a análise automatizada de respostas, que reduz o tempo necessário para correção, além da aplicação de testes via web (WBT), o que possibilita que os(as) aprendizes realizem as avaliações a qualquer momento e em qualquer lugar — aumentando, assim, a viabilidade prática do processo avaliativo.

## **CONSIDERAÇÕES FINAIS**

Em síntese, a tecnologia tornou-se completamente integrada à avaliação linguística. Os termos mais utilizados nesse campo — CBA, testes mediados por TIC/WBT, avaliação on-line, CAT e avaliação linguística baseada em inteligência artificial — têm impulsionado uma transformação significativa no modo como se realizam os testes de proficiência linguística, ao oferecer benefícios claros, como maior agilidade, consistência e adaptabilidade nas avaliações.

No que tange aos critérios de fidedignidade, validade, autenticidade, interatividade, impacto e praticidade, há vantagens e desvantagens. A fidedignidade pode ser ampliada por meio da pontuação automatizada, que elimina a interferência de avaliadores humanos e de fatores externos nas CBAs e nas avaliações on-line. Apesar de críticas quanto à validade limitada dessas modalidades, essa limitação pode ser superada com a inclusão de questões que avaliem habilidades cognitivas de ordem superior.

O uso de recursos multimodais torna as avaliações mais autênticas do que os testes em papel (PBT), permitindo aferir não apenas competências linguísticas, mas também o conhecimento temático, além de verificar se os(as) candidatos(as) são capazes de transmitir significados de forma eficaz no idioma-alvo. Assim como os PBT, os testes baseados em computador e os exames on-line exercem impacto relevante tanto sobre os(as) participantes quanto sobre a sociedade.

Quanto à praticidade, embora o desenvolvimento inicial demande investimentos elevados, os processos de aplicação e correção tornam-se mais eficientes e funcionais em comparação com os PBT. Ademais, muitas das desvantagens observadas podem ser progressivamente mitigadas diante do ritmo acelerado dos avanços tecnológicos.

## REFERÊNCIAS

- Altay, İ. F., & Altay, A. (2017). The impact of on-line reading tasks and reading strategies on EFL learners' reading test scores. *Journal of Language and Linguistic Studies*, 13(2), 136-152.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.
- Brown, H. D. (2000). *Principles of language learning and teaching* (Vol. 4). Longman.
- Burstein, J., Frase, L. T., Ginther, A., & Grant, L. (1996). Technologies for language assessment. *Annual Review of Applied Linguistics*, 16, 240–260. <https://doi.org/10.1017/S0267190500001537>
- Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Ernst Klett Sprachen.
- Chapelle, C. A., & Voss, E. (2008). Utilizing technology in language assessment. In N. H. Hornberger (Ed.), *Encyclopedia of language and education*, 2(7), 123–134.
- Cook, J., & Jenkins, V. (2010). *Getting started with e-assessment*. University of Bath. <http://opus.bath.ac.uk/17712/>
- Council of Europe. (s.d.). *European Language Portfolio (ELP)*. Retrieved March 10, 2020, from <https://www.coe.int/en/web/portfolio>
- Dunn, L., Morgan, C., O'Reilly, M., & Parry, S. (2005). *The student assessment handbook: New directions in traditional and on-line assessment*. Routledge.
- Educational Testing Service. (s.d.-a). *The TOEIC® program*. Retrieved April 15, 2020, from <https://www.ets.org/toeic/>
- Educational Testing Service. (s.d.-b). *Content and format of the TOEIC® Listening and Reading Test*. Retrieved April 15, 2020, from <https://www.ets.org/toeic/test-takers/listening-reading/about/content-format/>
- Educational Testing Service. (s.d.-c). *Content and format of the TOEIC® Speaking and Writing Tests*. Retrieved April 15, 2020, from <https://www.ets.org/toeic/test-takers/speaking-writing/about/content-format/>
- Educational Testing Service. (s.d.-d). *How ETS scores the TOEFL iBT® test*. Retrieved April 15, 2020, from [https://www.ets.org/s/toefl/flash/33910\\_toefl-scoring-video\\_transcript.html](https://www.ets.org/s/toefl/flash/33910_toefl-scoring-video_transcript.html)



- Educational Testing Service. (s.d.-e). *About the TOEFL iBT® test*. Retrieved April 15, 2020, from <https://www.ets.org/toefl/test-takers/ibt/about>
- Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). Automated scoring for the TOEFL Junior® comprehensive writing and speaking test. *ETS Research Report Series*, 2015(1), 1–11. <https://files.eric.ed.gov/fulltext/EJ1109726.pdf>
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. Routledge.
- Harmer, J. (2007). *The practice of English language teaching* (4th ed.). Pearson Education Limited.
- Hodgson, P., & Pang, M. Y. C. (2012). Effective formative e-assessment of student learning: A study on a statistics course. *Assessment & Evaluation in Higher Education*, 37(2), 215–225. <https://doi.org/10.1080/02602938.2010.523818>
- International English Language Testing System (IELTS). (s.d.-a). *Computer-delivered IELTS*. Retrieved April 15, 2020, from <https://www.ielts.org/about-the-test/computer-delivered-ielts>
- International English Language Testing System (IELTS). (s.d.-b). *Test format*. Retrieved April 15, 2020, from <https://www.ielts.org/about-the-test/test-format>
- Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, 25, 228–242. <https://doi.org/10.1017/S0267190505000127>
- Jiao, H., Liu, J., Haynie, K., Woo, A., & Gorham, J. (2012). Comparison between dichotomous and polytomous scoring of innovative items in a large-scale computerized adaptive test. *Educational and Psychological Measurement*, 72(3), 493–509. <https://doi.org/10.1177/0013164411422903>
- Jordan, S. E. (2013). E-assessment: Past, present and future. *New Directions for Adult and Continuing Education*, 2013(140), 87–106. <https://doi.org/10.29311/ndtps.v0i9.504>
- Karani, A. O., & Mwancha, C. N. (2025). The role of ChatGPT in STEM education: Assessing its prospects, challenges, ethical implications, and influence on student creativity and innovation competencies. *International Journal of Education, Technology and Science*, 5(2), 116–138. <https://doi.org/10.5281/zenodo.15489162>
- Lin, J. W., & Lai, Y. C. (2019). User acceptance model of computer-based assessment: Moderating effect and intention-behavior effect. *Australasian Journal of Educational Technology*, 35(1), 1–15. <https://doi.org/10.14742/ajet.4684>
- Navrat, P., & Tvarozek, J. (2014). On-line programming exercises for summative assessment in university courses. In *Proceedings of the 15th International Conference on Computer Systems and Technologies* (pp. 341–348).

- Ogrenci Secme ve Yerlestirme Merkezi. (2019). *2019 Yabancı Dil Bilgisi Seviye Tespit Sınavı (YDS) kılavuzu*. Retrieved April 15, 2020, from <https://dokuman.osym.gov.tr/pdfdokuman/2019/YDS-1/kilavuz30012019.pdf>
- Oxford University Press. (s.d.-a). *Oxford Test of English*. Retrieved April 15, 2020, from [https://elt.oup.com/feature/global/oxford\\_test\\_of\\_english/?cc=tr&selLanguage=en](https://elt.oup.com/feature/global/oxford_test_of_english/?cc=tr&selLanguage=en)
- Oxford University Press. (s.d.-b). *Who recognizes the Oxford Test of English?* Retrieved April 15, 2020, from [https://elt.oup.com/feature/global/oxford\\_test\\_of\\_english/recognition?cc=tr&selLanguage=en](https://elt.oup.com/feature/global/oxford_test_of_english/recognition?cc=tr&selLanguage=en)
- Pearson. (s.d.-a). *PTE Academic*. Retrieved April 15, 2020, from <https://pearsonpte.com/the-test/>
- Pearson. (s.d.-b). *Test format*. Retrieved April 15, 2020, from <https://pearsonpte.com/the-test/format/>
- Pearson. (s.d.-c). *How is the test scored?* Retrieved April 15, 2020, from <https://pearsonpte.com/the-test/about-our-scores/how-is-the-test-scored/>
- Stannard, R., & Skip Basiel, A. (2013). A practice-based exploration of technology-enhanced assessment for English language teaching. In Motteram, G. (Ed.), *Innovations in learning technologies for English language teaching* (pp. 145–174). British Council.
- Simkin, M. G., & Kuechler, W. L. (2005). Multiple choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, 3(1), 73–97. <https://doi.org/10.1111/j.1540-4609.2005.00053.x>
- Suvorov, R., & Hegelheimer, V. (2013). Computer-assisted language testing. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 2, pp. 594–613). Wiley-Blackwell.
- Taras, M. (2005). Assessment—summative and formative—some theoretical reflections. *British Journal of Educational Studies*, 53(4), 466–478. <https://doi.org/10.1111/j.1467-8527.2005.00307.x>
- Turnitin Feedback Studio. (s.d.-a). *What can you do with Feedback Studio?* Retrieved March 10, 2020, from <https://www.turnitin.com/products/feedback-studio>
- Turnitin Feedback Studio. (s.d.-b). *How to use PeerMark*. Retrieved March 10, 2020, from <https://www.turnitin.com/videos/how-to-use-peermark>
- Williamson, H. (2018). On-line exams: The need for best practices and overcoming challenges. *The Journal of Public and Professional Sociology*, 10(1), 1–7. <https://doi.org/10.62915/2154-8935.1128>

- Xi, X. (2021). Validity and the automated scoring of performance tests. In Fulcher, G. & Harding, L. (Eds.). *The Routledge handbook of language testing* (2nd ed., pp. 513–529). Routledge.
- Xu, J., Galaczi, E., Schmidt, E., & Somers, A. (2025). Automarking in language assessment: Key considerations for best practice. *Language Assessment Quarterly*.

*CRediT Author Statement*

---

**Reconhecimentos:** Não se aplica.

**Financiamento:** Esta pesquisa não recebeu apoio financeiro.

**Conflitos de interesse:** Não há conflito de interesse.

**Aprovação ética:** A pesquisa respeitou os princípios éticos durante sua realização.

**Disponibilidade de dados e material:** Os dados e materiais utilizados neste trabalho não estão disponíveis para acesso público.

**Contribuições dos autores:** 50% para cada autor. Ambos os autores contribuíram de forma igualitária para a elaboração do trabalho.

---

**Processamento e editoração: Editora Ibero-Americana de Educação**

Revisão, formatação, normalização e tradução

