SUMÁRIO EXECUTIVO ESTENDIDO

INOVAÇÃO EM AVALIAÇÃO PARA MEDIR E DAR SUPORTE A COMPETÊNCIAS COMPLEXAS

Publicação orginal em inglês pela:

OCDE
POLÍTICAS MELHORES PARA UMA VIDA MELHOR

SOBRE ESTA PUBLICAÇÃO

Este documento resume o conteúdo da publicação da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) intitulada *Innovating Assessments to Measure and Support Complex Skills* (Foster e Piacentini, 2023). O *Innovating Assessment* nasceu do esforço colaborativo entre a Secretaria da OCDE e o grupo conhecido como *Research and Innovation Group* (RIG) do *Program for International Student Assessment* (PISA), além de diversos outros especialistas e colaboradores internacionais no campo de medição e desenvolvimento de avaliações educacionais.

Tanto a publicação *Innovating Assessments to Measure and Support Complex Skills* como este sumário só foram possíveis com o apoio do Instituto Unibanco, na figura do seu superintendente, Ricardo Henriques, do gerente de Pesquisa e Inovação João Marcelo Borges e da colaboração da equipe da Coordenação de Articulação e Disseminação de Conhecimento integrada por Djana Contier Fares, Carolina Fernandes, Valquiria A. N. Parlagreco, além da consultoria de Tatiana F. Laganá na revisão crítica.

Natalie Foster e Mario Piacentini editaram o volume original e contribuíram com vários capítulos. Membros do RIG – incluindo Kadriye Ercikan, Xiangen Hu, Cesar A. Amaral Nunes, James Pellegrino, Ido Roll e Kathleen Scalise, e colaboradores convidados, como Miri Barhak-Rabinowitz, Hongwen Guo, Han Hui Por, Errol Kaylor, Cassie Malcom, Argenta Price, John. P. Sabatini, Keith Shubeck e Carl Wieman – contribuíram para o desenvolvimento dos capítulos restantes, prestaram aconselhamento especializado e deram seu *feedback* sobre a publicação de forma geral. Andreas Schleicher, Diretor de Educação e Habilidades da OCDE e Yuri Belfali, Diretor da Divisão de Educação Infantil e Escolar da OCDE, forneceram orientações e comentários adicionais. Esta publicação foi preparada por Mario Piacentini, Natalie Foster e Marc Fuster (OCDE).

Referência à versão original em inglês

Esta tradução não foi criada pela OCDE e não deve ser considerada uma tradução oficial. O texto original foi publicado pela OCDE sob o título FOSTER, N. and PIACENTINI, M. (eds.) (2023). *Innovating Assessments to Measure and Support Complex Skills*. Disponível em: https://doi.org/10.1787/e5f3e341-en. A qualidade da tradução e sua coerência com o texto original da obra são de responsabilidade exclusiva do autor ou autores da tradução. Em caso de discrepância entre o conteúdo original e o traduzido, apenas o texto presente no original será considerado válido.

FICHA TÉCNICA

REALIZAÇÃO

OCDE

Diretor de Educação e Habilidades
Andreas Schleicher

Chefe de Divisão, Primeira Infância e Escolas
Yuri Belfali

PISA Inovação
Mario Piacentini
Natalie Foster
Marc Fuster

Coordenação de Comunicação
Cassandra Morley
Della Shin
Sophie Limoges

Instituto Unibanco

Superintendente Executivo
Ricardo Henriques

Gerência de Pesquisa e Inovação
João Marcelo Borges

Coordenação de Articulação e Disseminação do Conhecimento
Djana Contier Fares
Carolina Fernandes
Valquiria Allis Nantes Parlagreco

Coordenação de Comunicação
Rosane Serro

ELABORAÇÃO DO SUMÁRIO EXECUTIVO ESTENDIDO

Produção de conteúdo (versão em inglês)
Natalie Foster
Mario Piacentini

Tradução
Gabriele Lima

Leitura crítica
Tatiana F. Laganá
Valquiria Allis Nantes Parlagreco

PRODUÇÃO EDITORIAL

Coordenação da Produção Editorial
Fabiana Hiromi

Revisão
Carmen Nascimento
Ofício do texto

Projeto gráfico e diagramação
Fernanda Aoki

Capa
Design da capa com base em imagens de © Shutterstock/treety; © Shutterstock/Merfin
Fernanda Aoki

ÍNDICE

PREFÁCIO

A inovação nas avaliações educacionais é crucial para lidar com os desafios atuais e exige um compromisso incontornável com uma educação de qualidade com equidade. O Sistema Nacional de Avaliação da Educação Básica (Saeb) teve sua primeira aplicação em 1990 e passou por significativos aprimoramentos, responsáveis por alavancar a capacidade de monitoramento e avaliação da aprendizagem no Brasil. A promulgação da Base Nacional Comum Curricular (BNCC), no entanto, demanda alterações mais profundas na prova, para que ela seja capaz de avaliar as competências e as habilidades esperadas ao longo da educação básica, orientando-se por uma visão de desenvolvimento integral do estudante. A reformulação do Saeb também passa pela necessidade de acompanhar os avanços tecnológicos, de formato e de metodologias das avaliações de larga escala.

No cenário das avaliações internacionais, o Programa Internacional de Avaliação de Estudantes (PISA) tem se configurado como um importante referencial de qualidade, com aprimoramento contínuo, contemplando, a partir de 2012, um domínio interdisciplinar em cada ciclo da prova. Competências complexas como Resolução de Problemas (2012), Resolução Colaborativa de Problemas (2015), Competência Global (2018) e Pensamento Criativo (2022) já foram avaliadas ao longo das últimas edições; para a prova de 2025, será avaliada a competência Aprender no Mundo Digital. O que essas avaliações têm em comum é uma visão sobre a importância de avaliar competências relevantes para a vida dos estudantes em uma sociedade com ciclos de inovação cada vez mais curtos (e, portanto, mais velozes), que lida com problemas mais complexos e sistêmicos.

Diante da necessidade de avançarmos no debate sobre inovação em avaliação educacional no Brasil, o Instituto Unibanco firmou em 2019 uma parceria com a Organização para Cooperação e Desenvolvimento Econômico (OCDE) para apoiar a constituição de um grupo de especialistas, o *Research and Innovation Group* (RIG), com objetivo de discutir as tendências e desafios relacionados ao tema no longo prazo. Até aquele ano, especialistas eram convidados de acordo com o domínio de inovação escolhido para aquela edição para elaborar o constructo teórico, acompanhar os testes e estabelecer métricas que possibilitassem comparabilidade. No entanto, esse modelo ensejava muitos desafios para conduzir os aprendizados de uma edição para a outra. A criação desse grupo permanente, que contou com apoio do Instituto e de outras instituições internacionais, possibilitou uma reflexão mais perene, a constituição de linhas de pesquisa e a construção de estratégias para inovações de curto e médio prazo.

Por meio da parceria, o Instituto também pôde acompanhar o desenvolvimento do constructo teórico do domínio de inovação do PISA de 2025 (Aprender no Mundo Digital) e a testagem da *Platform for Innovative Learning Assessments* (PILA) – uma plataforma de avaliação formativa para competências do século 21 – por um grupo de escolas em diferentes países, incluindo escolas da rede estadual do Ceará.

O livro I***nnovating Assessments to Measure and Support Complex Skills*** (Inovar Avaliações para Medir e dar Suporte a Competências Complexas, em tradução livre) é um dos produtos desse apoio à OCDE. A publicação reúne artigos que tratam de aspectos relevantes para o desenvolvimento de avaliações de habilidades complexas, apresentam novas metodologias e abordam questões relevantes tanto para a elaboração das avaliações quanto para os tomadores de decisão nesse processo. Um resumo da obra pode ser conferido no presente Sumário Executivo Estendido.

A ideia não é propor a replicação dos modelos de avaliação apresentados neste material e sim disseminá-los para que possam contribuir para o avanço das discussões acerca da inovação de avaliações no país e inspirar novos caminhos para suprir as demandas do cenário brasileiro.

Boa leitura!

Ricardo Henriques
Superintendente Executivo
Instituto Unibanco

EDITORIAL

Mais de 20 anos depois de seu primeiro ciclo, o PISA, que é um programa internacional de avaliação de estudantes, passou a ser uma força estabelecida e influente na reforma educacional. A ideia transformadora por trás do PISA consiste em avaliar as habilidades dos estudantes utilizando uma métrica internacional, associar a análise aos dados de estudantes, professores, escolas e sistemas para entender as diferenças de desempenho e aproveitar o poder da colaboração internacional para agir sobre os dados.

Desde o início, o PISA diferiu das avaliações tradicionais. Para atingirem um resultado satisfatório no PISA, os estudantes precisam se superar, pensar fora dos limites das disciplinas e aplicar seus conhecimentos com criatividade em situações novas – em vez de, basicamente, só reproduzir o conteúdo adquirido em sala de aula. O mundo moderno não nos recompensa mais pelo que sabemos, mas pelo que podemos fazer com o que sabemos. À medida que a informação se torna cada vez mais acessível e mais tarefas cognitivas rotineiras são digitalizadas e terceirizadas, o foco deve se transformar para que as pessoas se tornem aprendizes de uma vida toda. O conhecimento epistêmico – ou seja, pensar como um cientista ou matemático – e suas formas de trabalhar estão ganhando mais destaque do que o conhecimento de fórmulas, nomes ou lugares específicos.

Essa visão da educação se reflete em muitas estruturas contemporâneas que exigem o desenvolvimento das chamadas habilidades do século XXI – incluindo o *Learning Compass 2030* da OCDE. No entanto, sem que haja mudanças substanciais nos nossos sistemas educacionais, a lacuna entre o que os sistemas ensinam aos nossos jovens e o que nossa sociedade exige provavelmente ficará ainda mais espaçada.

Um componente integral dos sistemas educacionais é a avaliação. A forma como os estudantes são avaliados exerce grande influência no futuro da educação, porque sinaliza as prioridades para o currículo e o ensino. O propósito da avaliação é direcionar o foco a aspectos que são importantes, ou seja: professores e gestores escolares, bem como estudantes, devem prestar atenção ao que é examinado e se adaptarem. Uma questão crucial é como podemos obter uma avaliação assertiva e garantir que ela ajude professores e formuladores de políticas a acompanhar o progresso do que realmente importa na educação.

O problema é que muitos sistemas de avaliação estão mal alinhados com o currículo, com o conhecimento e com as habilidades que os jovens precisam dominar para prosperar. Ao desenvolver as avaliações, muitas vezes trocamos ganhos em validade e relevância por ganhos em eficiência e confiabilidade. No entanto, essas prioridades têm um preço: a avaliação mais confiável e eficiente é aquela que convoca os estudantes a responderem de maneira a não permitir ambiguidades – geralmente em um formato de múltipla escolha. Mas uma avaliação relevante é aquela em que se testa uma ampla gama de conhecimentos e habilidades consideradas relevantes para o sucesso na vida e no trabalho.

Para que a análise seja confiável, são necessários vários formatos de resposta, incluindo os abertos, que provocam explicações mais complexas. Questões que pedem por respostas abertas requerem processos de correção mais sofisticados. Avaliações de qualidade também devem investigar o pensamento e a compreensão dos estudantes, revelando as estratégias que eles adotam para resolver determinado problema, além de fornecer *feedback* produtivo com detalhes necessários para alimentar as decisões de melhoria. As avaliações digitais, ao registrarem também as ações dos estudantes e não apenas suas respostas, oferecem várias oportunidades para o avanço da avaliação nesse sentido.

Além disso, essas avaliações precisam ser justas e garantir medições adequadas em diferentes níveis de detalhamento para que possam atender às necessidades de tomada de decisão em níveis distintos do sistema educacional. Também precisamos trabalhar mais para reduzir a lacuna entre avaliações somativas e formativas. As origens da educação estavam na aprendizagem, na qual os estudantes adquiriam conhecimento de e com pessoas, recebendo *feedback* imediato e pessoal sobre seu progresso. Após alguns séculos, a industrialização da educação separou a aprendizagem da avaliação, encaminhando os estudantes ao acúmulo de anos de aprendizagem para que, mais tarde, reproduzissem o que aprenderam em ambientes frequentemente limitados e com restrições de tempo. Esse hábito tem contribuído para um ensino e aprendizagem muitas vezes superficial e focado no que pode ser medido com mais facilidade. Agora, a digitalização nos oferece a oportunidade de reintegrar aprendizagem e avaliação, combinar seus elementos somativos e formativos e criar sistemas coerentes de avaliação multicamadas que se estendem dos estudantes às salas de aula, das escolas aos níveis regional, nacional e até mesmo internacional. Com uma melhor integração entre avaliação e aprendizagem, os professores não verão mais as avaliações como um desperdício de tempo que poderia ser mais bem aproveitado na aprendizagem, mas, sim, como um instrumento complementar.

É claro que tudo isso também se aplica ao PISA. O PISA é visto como uma medida importante do sucesso dos sistemas escolares mundialmente e, estando nessa posição, precisa liderar a reforma educacional. Desde 2012, e graças à introdução da prova digital, o PISA expandiu sua gama de métricas para incluir um novo domínio interdisciplinar em cada ciclo – incluindo resolução de problemas (2012), resolução colaborativa de problemas (2015), competência global (2018) e, mais recentemente, pensamento criativo (2022).

Em 2020, o PISA deu um passo adiante: apesar de circunstâncias globais mais desafiadoras, os países optaram por investir mais recursos no desenvolvimento de avaliações inovadoras, instaurando um novo programa de Pesquisa, Desenvolvimento e Inovação (PDI) liderado por um grupo sênior, de especialistas internacionais avaliação.

De certa forma, esta publicação resultou da nossa colaboração com diferentes especialistas nos últimos três anos, desde o início do nosso programa de pesquisa. Ela explica por que precisamos inovar nas avaliações, aborda o que deve ser transformado e como podemos nos beneficiar da tecnologia para chegar lá. Também deixa claro que essa mudança não acontecerá da noite para o dia: há muito trabalho a ser feito e será necessária a convergência dos capitais político, financeiro e intelectual para ampliar essas ideias.

O PISA pode se tornar um mecanismo que impulsiona essa mudança, aproveitando o poder da colaboração internacional entre educadores, pesquisadores e formuladores de políticas, e compartilhando os custos – financeiros e políticos – entre os países na busca por práticas inovadoras. Pesquisa e inovação na avaliação em larga escala sempre foram uma parcela essencial do DNA do PISA, e estamos comprometidos em continuar atuando como líderes globais no caminho que está por vir.

Andreas Schleicher
Diretor de Educação e Habilidades
Conselheiro Especial do Secretário-Geral para Políticas Educacionais

O CASO DA PUBLICAÇÃO *INNOVATING ASSESSMENTS TO MEASURE AND SUPPORT COMPLEX SKILLS*

Este documento resume o conteúdo da publicação da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) intitulada *Innovating Assessments to Measure and Support Complex Skills* (FOSTER; PIACENTINI, 2023), produto de uma aprofundada pesquisa colaborativa e plurianual entre especialistas internacionais na área de medição e avaliação educacional e a Secretaria da OCDE.

A razão para se engajar neste trabalho – na verdade, o caso de avaliações inovadoras – é impulsionada por um conjunto de premissas que se conectam entre si. A primeira é que devemos nos preocupar com a avaliação. As avaliações educacionais são sinalizadores importantes que indicam o que os estudantes devem aprender e o que podem fazer. Nesse sentido, elas estão intrinsecamente ligadas a programas de ensino e pedagogias, conduzindo ou retendo mudanças nos objetivos e práticas educacionais. A segunda premissa decorre da primeira: a avaliação educacional deve focar no que importa. O que vale a pena saber, fazer e ser está em constante debate, com uma narrativa global que nos convida a repensar o que é ensinado e aprendido nas escolas, com o propósito de melhor preparar os estudantes como cidadãos e futuros profissionais. A conexão dessas duas premissas consiste na ideia de que qualquer discussão sobre a necessidade de munir os indivíduos das chamadas "competências do século XXI" também deve ser uma discussão sobre avaliação. Dito isso, mudar o foco das avaliações para "o que importa" só será algo valioso na medida em que essas avaliações forem capazes de medir o que afirmam medir. A terceira premissa, portanto, indica que as avaliações devem medir o que importa, e devem fazer isso de forma satisfatória.

A AVALIAÇÃO IMPORTA

Professores, estudantes e formuladores de políticas locais e nacionais muitas vezes seguem indícios sobre os objetivos de ensino e aprendizagem a partir dos tipos de tarefas encontradas nas avaliações locais, nacionais e internacionais. As avaliações sinalizam para vários públicos quais conhecimentos, habilidades e capacidades são importantes e ilustram os tipos de desempenho que queremos que os estudantes sejam capazes de demonstrar. Assim, o que escolhemos avaliar em áreas como Ciências, Matemática, Alfabetização, resolução de problemas, colaboração e pensamento crítico é o que acabará sendo o foco da instrução. Portanto, é fundamental que nossas avaliações representem melhor as formas de conhecimento

e competência, além dos tipos de aprendizagem que queremos enfatizar em nossas salas de aula, para que funcionem positivamente dentro do sistema educacional.

Do ponto de vista do sistema, não faz muito sentido investir no currículo e na reforma da formação de educadores sem investir também na avaliação. O currículo, a pedagogia e a avaliação são aspectos que estão intrinsecamente ligados e devem estar alinhados em sistemas educacionais para que funcionem bem. Transformações no currículo e na pedagogia podem ser motivadas por mudanças no foco da avaliação e pelas lacunas educacionais que elas revelam, por sua vez, norteando a formulação de políticas e reformas. O foco na avaliação traz clareza sobre as expectativas de ensino e aprendizagem em diferentes níveis educacionais, contribuindo para estabelecer um entendimento compartilhado sobre tópicos que importam e como devem ser ensinados. No fim das contas, a principal questão é: o que exatamente importa?

MUDANÇA NOS OBJETIVOS EDUCACIONAIS: UM FOCO NAS COMPETÊNCIAS DO SÉCULO XXI

Há mais de 20 anos, um número crescente de líderes empresariais, organizações educacionais e pesquisadores começaram a exigir novas políticas educacionais centradas no desenvolvimento de habilidades e conhecimentos gerais e transferíveis, muitas vezes chamados de "habilidades do século XXI" (PELLEGRINO; HILTON, 2012; BELLANCA, 2014). Essas exigências são baseadas na ideia de que o sucesso na sociedade contemporânea global e em um mundo de trabalho em evolução exige um conjunto mais amplo de capacidades que vão além da educação tradicional feita de Leitura, Matemática e Ciências.

Basicamente, essa retórica argumenta que a educação deve se concentrar na capacidade de processar (novas) informações e resolver problemas, o que inclui munir os indivíduos de um sólido conhecimento disciplinar, mas também de habilidades analíticas, criativas e de pensamento crítico. Assim, o ensino também deve estar pautado em habilidades mais gerais com relação a si mesmo e ao próximo, como habilidades sociais e emocionais, tolerância e respeito mútuo, além da capacidade de autocontrole e entendimento mais amplo de seus próprios processos de pensamento e aprendizagem.

Certamente, essas capacidades sempre foram importantes. No entanto, em um mundo onde o trabalho era formado por tarefas manuais e rotineiras, e onde as tecnologias de comunicação e informação instantâneas de hoje não passavam de um produto da imaginação, é de se esperar que apenas alguns indivíduos as desenvolvessem. Nas economias baseadas no conhecimento de hoje – caracterizadas por estruturas mais dinâmicas e multiculturais nas quais os cidadãos se comunicam instantaneamente e conseguem se auto-organizar, tanto local quanto globalmente –, é esperado que competências cognitivas e sociocognitivas avançadas sejam a norma.

ENTENDIMENTO DAS COMPETÊNCIAS DO SÉCULO XXI

Começando antes da virada do século, um crescente movimento de pesquisa examinou narrativas globais, produzindo uma variedade de estruturas internacionais que descrevem o conhecimento, as habilidades e os comportamentos que são importantes para o futuro dos jovens. Há uma diversidade de terminologias empregadas de forma intercambiável dentro deste espaço relativamente sobrecarregado: "Habilidades/competências do século XXI", "*soft skills*", "habilidades interdisciplinares", "habilidades transferíveis", entre outras. Essa ambiguidade terminológica se estende às maneiras pelas quais as diferentes estruturas definem competências específicas (por exemplo, educação em TIC *vs.* educação digital *vs.* educação midiática).

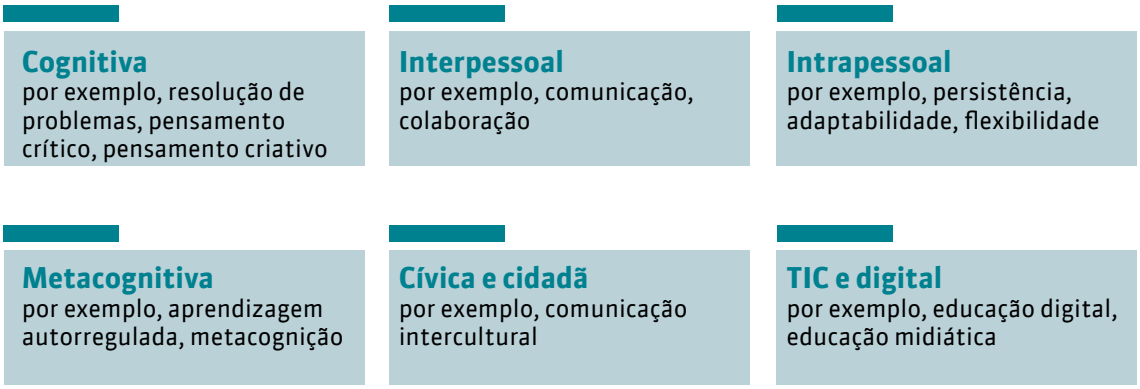
Por uma questão de clareza, a publicação *Innovating Assessments to Measure and Support Complex Skills* usa o termo "competências do século XXI" para se referir à visão ampla da educação estabelecida por essas estruturas e às várias competências que elas descrevem. Embora as estruturas variem, elas tendem a descrever as competências do século XXI como sendo:

- transversais (ou seja, relevantes ou aplicáveis em muitos campos);
- multidimensionais (ou seja, abrangem conhecimentos, habilidades e comportamentos); e
- associadas a habilidades e comportamentos de ordem superior que representam a capacidade de transferir conhecimento, lidar com problemas complexos e se adaptar a situações imprevisíveis (VOOGT; ROBLIN, 2012).

Além da convergência geral em torno dessas características centrais, as estruturas identificam, organizam e classificam as competências do século XXI de diferentes maneiras. Algumas agrupam as competências com base em suas características conceituais, como competências cognitivas, interpessoais e intrapessoais (PELLEGRINO; HILTON, 2012). Outros agrupam de acordo com seu propósito ou contexto de uso, como "formas de pensar", "formas de viver no mundo", "formas de trabalhar" e "ferramentas para trabalhar" (BINKLEY *et al.*, 2011).

Abstraindo das especificidades de cada estrutura, algumas categorias amplamente distintas de competências emergem com consistência (consulte a Figura 1). Em geral, alguma combinação dessas seis categorias captura a essência de listas de competências identificadas em diferentes estruturas, com competências de pensamento crítico, pensamento criativo, comunicação e outras relacionadas à TIC, além da dimensão cívica e cidadã, que aparecem regularmente. Observe, no entanto, que nem todas as estruturas incluem cada uma das categorias gerais identificadas abaixo e nem sempre atribuem competências específicas às mesmas categorias mais gerais.

Figura 1. Categorias gerais de competências do século XXI



Fonte: Foster (2023)

A identificação de categorias comuns de competências do século XXI disponibiliza informações úteis sobre como os objetivos mais gerais da educação estão mudando. No entanto, essas competências são conceitos complexos e obter evidências e interpretações válidas sobre o que os estudantes pensam e podem fazer ao desenvolvê-las apresenta vários desafios. Avaliar bem as competências do século XXI requer projetos e experiências de avaliação que sejam inovadores, desde a definição de conceitos de avaliação até a elaboração de tarefas avaliativas, além da descoberta dos métodos corretos para interpretar as evidências que emergem dela.

A ANÁLISE DAS COMPETÊNCIAS DO SÉCULO XXI EXIGE UM PROJETO DE AVALIAÇÃO INOVADOR

A primeira questão ao avaliar as competências do século XXI refere-se à definição do que avaliar. Essas competências são complexas; envolvem múltiplos componentes que estão fortemente interligados na prática. Por um lado, envolvê-los implica ativar uma combinação de conhecimentos, habilidades e comportamentos – por exemplo, a capacidade de se comunicar efetivamente envolve o conhecimento da linguagem, certo grau de habilidade escrita, verbal ou digital e certas atitudes em relação à pessoa com quem você se comunica. Esses elementos constitutivos também podem variar em diferentes contextos de prática. Por outro lado, engajar um "tipo" específico de competência na vida real geralmente requer engajamento de outros "tipos" simultaneamente. A resolução bem-sucedida de problemas, por exemplo, envolve aspectos de metacognição e autorregulação e, dependendo do contexto e da tipologia do problema, pode envolver pensamento criativo e colaboração. Esses vínculos complexos dificultam a divisão de conceitos em componentes pontuais e mensuráveis de forma independente, além de isolarem e atribuírem evidências geradas pelos estudantes a uma ou outra competência específica.

Paralelamente, as competências do século XXI são definidas, pelo menos parcialmente, por processos de pensamento e comportamentos que vão além da capacidade de reproduzir o conhecimento do conteúdo. Por exemplo, a capacidade de avaliar criticamente

informações desconhecidas depende da capacidade de entender quais informações adicionais precisam ser pesquisadas e como, de planejar e executar uma estratégia para fazer isso e de persistir na resolução da tarefa e/ou decidir a quem recorrer para obter ajuda ou *feedback*. Esses comportamentos e formas de pensar precisam ser visíveis nas avaliações para que seja feita qualquer afirmação referente à competência do estudante. Para muitas competências do século XXI, isso significa preparar ambientes de avaliação que forneçam aos estudantes ferramentas para fazer e criar e ofereça escolhas e oportunidades para explorar e repetir suas ideias. Para gerar um conjunto mais rico de dados sobre como os estudantes pensam e agem, essas possibilidades exigem a movimentação de tarefas e recursos de avaliação que vão além de itens estáticos e respostas fechadas, comumente usados em avaliações aplicadas em larga escala.

Criar uma próxima geração de avaliações educacionais que respondam a essa visão da educação do século XXI apresenta, portanto, uma sequência de desafios que devem ser superados pelos desenvolvedores de avaliação. Por exemplo, ser capaz de definir os conceitos-alvo da avaliação, identificar as situações relevantes nas quais essas podem ser observadas, replicar seus principais recursos em ambientes de avaliação, converter os registros das ações desses ambientes em evidências e desenvolver modelos adequados para interpretá-las e pontuá-las, fazendo afirmações robustas sobre o desempenho.

Com base nas principais mensagens e nos exemplos mais avançados de prática incluídos na publicação *Innovating Assessments to Measure and Support Complex Skills* da OCDE, as próximas seções elucidam o trajeto para os desenvolvedores de avaliação – incluindo descompactar as principais decisões e as ferramentas emergentes que podem ajudá-los no caminho. No encerramento deste documento há considerações sobre o papel que as autoridades educacionais podem desempenhar, juntamente com outras partes interessadas, integrando uma estrutura mais ampla de colaboração internacional para levar adiante a agenda de "próxima geração de avaliações".

PRÓXIMA GERAÇÃO DE AVALIAÇÕES: DESIGN DE PRINCÍPIOS E EXEMPLOS

Avaliar os resultados educacionais não é tão simples quanto medir a altura ou o peso de alguém. As avaliações não oferecem um canal direto para a mente do aluno, e os atributos a serem medidos são mentais e não visíveis externamente. Assim, a avaliação é uma ferramenta projetada para observar o comportamento dos estudantes e produzir dados que possam ser usados para fazer inferências razoáveis sobre o que os estudantes sabem e são capazes de fazer. Decidir o que avaliar e como fazer essa análise não é tão simples quanto parece. O cenário fica ainda mais difícil quando os alvos da avaliação são conceitos e desempenhos complexos.

A publicação *Innovating Assessments to Measure and Support Complex Skills* traz ideias-chave sobre como projetar a próxima geração de avaliações que mede as competências que os estudantes precisam e apresenta informações acessíveis aos desenvolvedores de avaliação, educadores e formuladores de políticas. Medir *o que importa* exige inovar todas as fases do projeto de avaliação – desde o que avaliamos até como o fazemos. Medir *bem* o que importa implica em um processo de elaboração pautado em princípios e impulsionar tecnologias digitais para gerar evidências relevantes sobre as competências dos estudantes. Também é necessário aplicar métodos analíticos inovadores para dar sentido a essas evidências.

AVALIAÇÃO COMO UM PROCESSO DE RACIOCÍNIO COM BASE EM EVIDÊNCIAS

O processo de fazer inferências sobre o que os estudantes sabem e são capazes de fazer representa uma cadeia de raciocínio baseada em evidências sobre a competência do aluno que caracteriza todas as avaliações, desde questionários em sala de aula e testes padronizados de aproveitamento, até programas de tutoria computadorizados e conversas que os estudantes têm com o professor enquanto resolvem um problema de Matemática ou discutem o significado de um texto. A primeira pergunta no processo de elaboração da avaliação é "evidência sobre o quê?". Os dados não fornecem seu próprio significado, seu valor como evidência pode surgir apenas por meio de alguma estrutura interpretativa. As avaliações educacionais fornecem dados como ensaios escritos, notas nas folhas de respostas, apresentações de projetos ou explicações dos estudantes sobre suas resoluções de problemas, mas esses dados se tornam evidências ape-

nas com relação a conjecturas sobre como os estudantes adquirem conhecimento e habilidade.

Pellegrino, Chudowsky e Glaser (2001) retratam esse processo de raciocínio com base em evidências como uma tríade de elementos interconectados: o Triângulo de Avaliação (consulte a Figura 2). Os vértices do Triângulo representam os três elementos-chave subjacentes a qualquer avaliação: um modelo de cognição e aprendizagem do aluno no domínio da avaliação; um conjunto de suposições e princípios sobre os tipos de observações que fornecem evidências das competências dos estudantes; e um processo de interpretação para dar sentido às evidências considerando o propósito da avaliação e a compreensão do aluno. Esses três elementos podem ser explícitos ou implícitos, mas uma avaliação não pode ser projetada e implementada, ou avaliada, sem a consideração de cada um deles. Os três são representados como vértices de um triângulo porque cada um se conecta e depende dos outros dois. O Triângulo de Avaliação apresenta uma estrutura útil para analisar as bases das avaliações atuais para determinar até que ponto elas cumprem os objetivos que temos em mente, bem como para projetar avaliações futuras e estabelecer sua validade (PELLEGRINO; DIBELLO; GOLDMAN, 2016).

Figura 2. O Triângulo de Avaliação

COGNIÇÃO
Teorias, modelos e dados sobre como os estudantes representam o conhecimento e desenvolvem competências em um domínio de instrução e aprendizagem.

OBSERVAÇÃO
Tarefas ou situações que permitem observar o desempenho dos estudantes.

INTERPRETAÇÃO
Métodos para atribuir sentido às evidências provenientes do desempenho dos estudantes.

Fonte: Pellegrino, Chudowsky e Glaser (2001).

A ponta da *cognição* no Triângulo refere-se à teoria, aos dados e a um conjunto de suposições sobre como os estudantes representam o conhecimento e desenvolvem competência em um domínio intelectual (por exemplo, frações; leis de Newton; termodinâmica). Em qualquer avaliação, é necessária uma teoria de competência no domínio para identificar o conjunto de conhecimentos e habilidades mensuráveis para o contexto de uso pretendido, seja para fazer um julgamento somativo (caracterizando as competências que os estudantes adquiriram em algum momento) ou para fazer julgamentos formativos

(visando orientar a instrução subsequente de modo a maximizar a aprendizagem). Uma premissa central é que a teoria cognitiva deve representar a compreensão cientificamente mais confiável das formas típicas pelas quais os estudantes representam conhecimento e desenvolvem proficiência no domínio em questão.

Toda avaliação também se baseia em um conjunto de suposições e princípios sobre os tipos de tarefas ou situações que levarão os estudantes a dizer, fazer ou criar algo que demonstre conhecimentos e habilidades importantes. As tarefas que os estudantes são orientados a responder em uma avaliação devem ser cuidadosamente projetadas para fornecer as evidências vinculadas ao modelo cognitivo de aprendizagem e para apoiar os tipos de inferências e decisões com base nos resultados da avaliação. O vértice de ***observação*** do Triângulo de Avaliação representa uma descrição ou conjunto de especificações para tarefas de avaliação que visam provocar respostas esclarecedoras por parte dos estudantes. Na avaliação, existe a oportunidade de estruturar qualquer partícula mínima para fazer observações. O desenvolvedor de avaliação pode usar esse recurso para maximizar o valor dos dados coletados, de acordo com suposições subjacentes sobre como os estudantes aprendem.

As avaliações também requerem suposições e modelos para interpretar as evidências coletadas das observações. O vértice de ***interpretação*** do Triângulo abrange todos os métodos e as ferramentas usadas para raciocinar com base em observações falíveis. Ele expressa como as observações derivadas de um conjunto de tarefas de avaliação constituem evidências sobre o conhecimento e as habilidades que estão sendo avaliadas. No contexto da avaliação aplicada em larga escala, o método de interpretação costuma ser um modelo estatístico, que é uma caracterização ou resumo de padrões que se esperaria ver nos dados, conforme os vários níveis de competência do aluno. No contexto da avaliação em sala de aula, a interpretação muitas vezes é feita de forma menos formal pelo professor e, com frequência, tem por base um modelo intuitivo ou qualitativo, em vez de um modelo estatístico formal. Mesmo informalmente, os professores fazem julgamentos coordenados sobre quais aspectos de compreensão e aprendizagem dos estudantes são relevantes, como um aluno executou uma ou mais tarefas e o que o desempenho aponta sobre o conhecimento e a compreensão do aluno.

Um ponto fundamental é que cada um dos três elementos do Triângulo de Avaliação não deve apenas fazer sentido por si só, mas também deve se conectar a cada um dos outros dois elementos de maneira significativa, resultando em uma avaliação eficaz e inferências sólidas. Assim, para que uma avaliação seja válida e eficaz, todos os três vértices do Triângulo devem trabalhar juntos em sincronia. Ao reconhecer que a avaliação é um processo de raciocínio comprovativo, provou-se útil ser sistemático ao enquadrar o processo de projeto da avaliação como um processo de Design Baseado em Evidências (DBE) (MISLEVY; HAERTEL, 2006; MISLEVY; RICONSCENTE, 2006) – consulte a Figura 3 para uma visão geral dos diferentes componentes do modelo de DBE.

Figura 3. Projeto de avaliação como um processo de Design Baseado em Evidências

Fases de definição da estrutura conceitual de uma avaliação

DEFINIÇÃO DE OBJETIVOS E FOCO

DEFINIR O DOMÍNIO DA AVALIAÇÃO

- **Coletar informações sobre o domínio** (análise de domínio), incluindo seus principais componentes e a gama de problemas e situações em que as pessoas fazem uso dos conhecimentos e habilidades-alvo.
- Modelagem de domínio: **Especificar exigências da avaliação** (o que desejamos medir), dados (como vamos medir) e garantias (explicar por que a abordagem de medição é adequada).

OPERACIONALIZAÇÃO DA AVALIAÇÃO (DESIGN BASEADO EM EVIDÊNCIAS (DBE)

DEFINIR COMO É O DESEMPENHO DO ESTUDANTE NO DOMÍNIO (O MODELO DE ESTUDANTE)

- **Definir as variáveis** (conhecimentos, habilidades e comportamentos) que queremos utilizar, as relações entre essas variáveis e se elas são dinâmicas (se algum aprendizado é esperado).
- Fornecer uma visão detalhada do que os estudantes entendem e são capazes de fazer em **diferentes níveis de proficiência**, do nível mais baixo ao mais alto de domínio em cada variável.

DEFINIR AS SITUAÇÕES EM QUE AS EVIDÊNCIAS DE DESEMPENHO PODEM SER ENCONTRADAS (O MODELO DE TAREFA)

- **Especificar as tarefas** em que os examinados podem demonstrar proficiência, como em questões ou tarefas predefinidas (por exemplo, itens de múltipla escolha, tarefas de reordenação ou conclusão) ou em ambientes onde a situação é moldada pelas ações dos examinados (por exemplo, simulações, jogos).
- **Definir os condutores de complexidade** e conhecimento envolvidos e os recursos incorporados na tarefa, incluindo feedback ou suporte para facilitar o aprendizado (se o aprendizado for esperado).

DEFINIR PONTUAÇÕES E INDICADORES DE DESEMPENHO (O MODELO DE EVIDÊNCIA)

- Definir as regras de evidência: **associar uma pontuação ou valor ao que os estudantes fazem** (por exemplo, responder corretamente/incorretamente a perguntas, tomar certas ações/decisões em determinada situação).
- Criar um modelo estatístico que **resuma os dados ao longo das tarefas** no que se refere a princípios atualizados sobre as variáveis do modelo de estudante.

Fonte: Piacentini (2023).

INOVAÇÃO DO VÉRTICE DE *COGNIÇÃO*: DEFININDO CONCEITOS DE AVALIAÇÃO

No desenvolvimento da avaliação, nenhuma questão é tão crítica quanto delinear com nitidez o domínio-alvo e descrever conhecimentos, habilidades, comportamentos e contextos de aplicação constituintes que sustentam o desempenho nesse domínio. De fato, se o domínio estiver mal definido, nem mesmo o cuidado com outras atividades de desenvolvimento de teste ou análise psicométrica complexa (uma vez que os dados tiverem sido coletados) compensará tal inadequação (MISLEVY; RICONSCENTE, 2006). É muito mais provável que uma avaliação atinja seu objetivo quando a natureza do conceito orienta o projeto de tarefas relevantes, assim como o desenvolvimento de critérios de pontuação e rubricas (MESSICK, 1994).

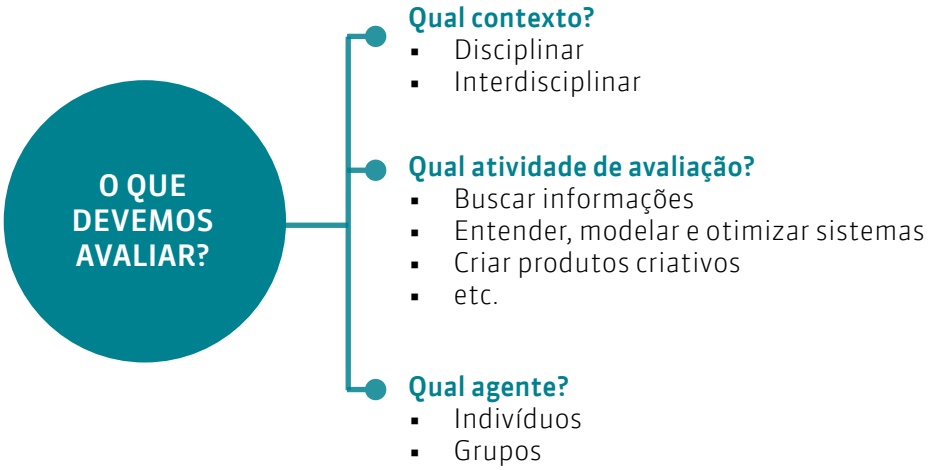
Conforme já discutido, essa atividade crítica se torna mais desafiadora à medida que a complexidade do domínio e do(s) conceito(s)-alvo aumenta. Os tipos de problema ou atividades que envolvem as competências do século XXI requerem uma combinação diferente de conhecimentos, habilidades e atitudes. O contexto de aplicação também é importante para determinar quais desses elementos são mais importantes e como exatamente eles podem ser expressos. Ou seja, desde os estágios iniciais do projeto de avaliação, é fundamental explicitar o que se espera que os estudantes demonstrem por meio do desempenho no teste.

DECISÕES INICIAIS PARA ELABORAÇÃO DAS AVALIAÇÕES COM FOCO NAS COMPETÊNCIAS DO SÉCULO XXI

Quando se trata de decidir o que avaliar, pode não ser o melhor caminho escolher uma estrutura ou lista de competências do século XXI e criar um único instrumento de avaliação para cada competência descrita. Como essas competências são multidimensionais e fortemente interconectadas na prática, uma estratégia mais produtiva pode consistir em desenvolver avaliações de como os estudantes criam conhecimento e solucionam problemas de diferentes complexidades, individualmente ou com a ajuda de outras pessoas, em contextos distintos de aplicação. Atribuir sentido ao que os estudantes fazem em atividades de resolução de problemas abertos e mais longos pode fornecer informações sobre sua capacidade de mobilizar múltiplas competências do século XXI em cenários mais autênticos.

Conforme refletido na Figura 4, três questões inter-relacionadas podem oferecer orientação particularmente útil para determinar o foco da próxima geração de avaliações:

Figura 4. Decisões iniciais de elaboração das avaliações com foco nas competências do século XXI



Fonte: Piacentini e Foster (2023).

- **Em quais tipos de resultados e atividades relacionadas estou interessado para entender a preparação dos estudantes?** Essa decisão está relacionada à definição explícita das atividades de avaliação e das práticas relevantes que queremos que os estudantes demonstrem enquanto participam das atividades.

- **Nas atividades de avaliação, em quais contextos de prática os estudantes podem participar?** Essa decisão refere-se a reconhecer conhecimentos, habilidades e comportamentos que os estudantes precisam ter em determinado tipo de atividade em um dado contexto de prática (ou seja, situar a atividade dentro dos limites de uma disciplina ou torná-la interdisciplinar e especificar o contexto de aplicação).

- **A avaliação será individual ou em grupo?** Essa decisão relaciona-se com a definição explícita de se, quando e com qual finalidade uma avaliação pode proporcionar aos estudantes a possibilidade de interagir com outros agentes, sejam eles reais ou virtuais.

Atividades relevantes para avaliar as competências do século XXI

Resolver problemas complexos envolve uma variedade de habilidades cognitivas, metacognitivas, atitudinais e socioemocionais. No entanto, nem todos os problemas de avaliação podem fornecer um conjunto tão rico de evidências sobre os estudantes. Os modelos tradicionais de resolução de problemas, conhecidos como modelos de fase (BRANSFORD; STEIN, 1984), sugerem que todos os problemas podem ser solucionados quando: (1) o identificamos; (2) criamos soluções alternativas; (3) avaliamos essas soluções; (4) implementamos a solução escolhida; e (5) examinamos a eficácia da solução. Embora sejam úteis, essas descrições de processos gerais podem sugerir erroneamente que a resolução de problemas é uma atividade unifor-

me (JONASSEN, 1991). Na realidade, os problemas variam de muitas formas importantes, incluindo o contexto em que ocorrem, seu nível de estrutura ou abertura e a combinação de habilidades que o solucionador de problemas deve usar para chegar a uma solução.

Para abordar o problema, há diversas metas e atividades que podem ser apresentadas aos estudantes para avaliar as competências do século XXI. Por exemplo, atividades de avaliação que provavelmente fornecerão evidências válidas sobre se as experiências de aprendizagem prepararam os estudantes para o futuro incluem: (1) pesquisar, avaliar e compartilhar informações; (2) compreender, modelar e otimizar sistemas; e (3) desenvolver produtos criativos. Essa não é uma tipologia exaustiva de atividades de avaliação; os tipos de problemas e atividades para os quais os estudantes devem estar preparados continuam a evoluir. Além disso, esses três grupos de atividades não são mutuamente exclusivos e se sobrepõem até certo ponto. No entanto, eles ilustram tipos de problemas que atraem conjuntos distintamente diferentes de competências, conhecimentos, habilidades e comportamentos relacionados. O Box 1 apresenta alguns exemplos de como poderiam ser as próximas gerações de avaliação do primeiro grupo de atividades.

BOX 1.

ATIVIDADES RELEVANTES PARA AVALIAÇÕES DE COMPETÊNCIAS DO SÉCULO XXI

Pesquisar, avaliar e compartilhar informações

Nesse grupo de atividades, o principal objetivo de resolução de problemas ou de aprendizagem consiste em pesquisar e usar informações para chegar a uma conclusão com argumentos. A sequência de tarefas em uma avaliação deve estimular os estudantes a identificar suas necessidades de informação, localizar fontes de informação em ambientes on-line ou off-line, extrair, organizar e comparar as informações de cada fonte, conciliar conflitos de informação e tomar decisões sobre quais informações compartilhar e como. Esse conjunto de atividades costuma ser definido como resolução de problemas de informação (BRAND-GRUWEL; WOPEREIS; VERMETTEN, 2005, WOLF; BRUSH; SAYE, 2003). Pesquisas apontam que muitos estudantes não são capazes de resolver esses problemas com êxito (BILAL, 2000; LARGE; BEHESHTI, 2000). Essas atividades se concentram em como os estudantes interagem com vários tipos de mídia e podem ser aplicadas a praticamente qualquer área do conhecimento (ou seja, contexto de prática). Elas enfatizam o pensamento crítico, a síntese, a argumentação, a comunicação responsável e as habilidades de aprendizagem autorreguladas como competências essenciais.

Há vários exemplos de avaliações que se concentram em problemas de informação. Em alguns casos, a avaliação é totalmente integrada a uma experiência de aprendizagem e as evidências são extraídas de maneira "discreta", analisando as sequências de escolhas que os estudantes fazem e o resultado da sua busca de informações. Por exemplo, no ambiente Betty Brain (BISWAS; SEGEDY; BUNCHONGCHIT, 2015), estudantes ensinam uma agente virtual, Betty, sobre um fenômeno científico. Para isso, eles pesquisam recursos por meio de hiperlinks e criam um mapa conceitual que representa sua compreensão emergente do fenômeno. Os estudantes podem pedir a Betty para fazer alguns testes, e ela

responde utilizando as informações representadas no mapa conceitual. Nesse teste, o desempenho de Betty orienta os estudantes sobre elementos incorretos ou ausentes no mapa.

Outros exemplos incorporam ferramentas de busca e gerenciamento de informações em mundos virtuais. O projeto NAEP SAIL Virtual World for Online Inquiry (COIRO et al., 2019) desenvolveu uma plataforma virtual que simula uma microcidade, onde os estudantes são apresentados a um desafio de aprendizagem aberta (por exemplo, descobrir se um artefato histórico deve ser exibido no museu local) e desenvolvem seu conhecimento planejando uma estratégia de consulta com um parceiro virtual, fazendo perguntas a especialistas virtuais, buscando informações em um ambiente web ou em uma biblioteca virtual e adotando diferentes ferramentas digitais para fazer anotações e redigir um relatório. O ambiente inclui recursos de design adaptativo, como dicas, alertas e nivelamento para ajudar os estudantes a regular seus processos de pesquisa e incentivar uma coleta eficiente de informações úteis.

Outros exemplos interessantes estão relacionados às habilidades de verificação de fatos e compartilhamento de informações dos estudantes em ambientes abertos e em rede. Jogos como Fake It To Make It (URBAN; HEWITT; MOORE, 2018), Bad News (ROOZENBEEK; VAN DER LINDEN, 2019) ou Go Viral! (BASOL et al., 2021) ensinam aos jogadores técnicas comuns para promover a desinformação na esperança de que a experiência os prepare para combatê-la. No jogo The Misinformation Game, os participantes podem interagir com postagens de maneiras ecologicamente válidas, escolhendo um comportamento de engajamento (com opções que incluem curtir, não curtir, compartilhar, sinalizar e comentar) e recebem feedback dinâmico (ou seja, alterações em sua própria contagem de seguidores e pontuação de credibilidade simulada) dependendo de como é sua interação com informações confiáveis ou não confiáveis (VAN DER LINDEN; ROOZENBEEK; COMPTON, 2020).

Fonte: Piacentini e Foster (2023).

Contextos de prática ou domínios de aplicação

Embora as competências do século XXI sejam amplamente vistas como transversais ou interdisciplinares, o significado de solucionar problemas, pensar criticamente ou ser criativo em determinado contexto pode ser completamente diferente em uma situação diversa. Essas habilidades não são exercidas nem observadas no vazio, e dificilmente podemos avaliá-las de maneira neutra em termos de domínio. Portanto, ao definir o foco de uma avaliação, o papel e a importância do conhecimento específico do domínio devem ser explícitos desde o início. Em um contexto de avaliação, a capacidade dos estudantes para desempenhar essas competências será sempre observada em determinado contexto ou situação, e o seu conhecimento sobre esse contexto ou situação influenciará o tipo de estratégias que utilizam, bem como o que são capazes de realizar. A tentativa de projetar problemas ou cenários completamente descontextualizados também ameaça a validade: se um aluno não precisa de conhecimento algum para resolver uma tarefa, a avaliação realmente pretende medir os tipos de competências de resolução de problemas complexos nos quais afirma estar interessada?

As próximas gerações de avaliação podem ser contextualizadas em um domínio específico de conhecimento ou abranger várias disciplinas. Aqui, interdisciplinar não se refere ao domínio geral, pois as competências que os estudantes demonstram em tarefas interdisciplinares ainda dependem de um conjunto de conhecimentos bem definido; ou seja, significa apenas que o conhecimento não é limitado a uma única disciplina. As avaliações de resultados de aprendizagem mais amplamente utilizadas são definidas em uma única disciplina (por exemplo, Matemática, Biologia, História) e se concentram na reprodução de conhecimentos adquiridos e procedimentos relevantes para a matéria. Ao pensar em uma avaliação de competências do século XXI no contexto de um domínio disciplinar, novas avaliações poderão trazer um melhor equilíbrio entre o exame de conhecimentos disciplinares e a avaliação da capacidade que estudantes têm de aplicar esses conhecimentos em contextos autênticos e a novos problemas. As avaliações podem convidar os estudantes a participarem de práticas que reflitam como o conhecimento disciplinar é usado para lidar com problemas profissionais e cotidianos. Em História, por exemplo, os estudantes podem ser convidados a investigar colaborativamente e encontrar tendenciosidades no relato histórico de determinado evento. Em Ciências, a avaliação pode pedir para que os estudantes participem da exploração de um fenômeno científico em um laboratório virtual, usando ferramentas relevantes e passando pela sequência de decisões que os cientistas reais tomam em sua prática profissional (consulte o Box 2 para um exemplo mais detalhado).

BOX 2.

AVALIAÇÕES ESPECÍFICAS DE DOMÍNIOS DE HABILIDADES COMPLEXAS

Analisar a tomada de decisão dos estudantes nos campos da Ciência e da Engenharia

A resolução de problemas complexos, particularmente nos campos da Ciência e da Engenharia, é uma competência central do mundo moderno, e muitos padrões científicos mais recentes a tem em seu núcleo. No entanto, as avaliações dos estudantes não costumam capturar os principais processos e as decisões que fazem parte da resolução de problemas na vida real e, portanto, são limitadas no sentido de tirar conclusões significativas sobre as competências dos estudantes.

Resolver os tipos de problema normalmente encontrados em exames escolares e livros didáticos requer reconhecer e seguir um procedimento único e bem estabelecido. Esses problemas podem ser complicados, na medida em que requerem várias etapas, mas envolvem poucas decisões – ou o aluno conhece o procedimento correto ou não. Não é assim que a resolução de problemas complexos funciona. Cientistas e engenheiros especialistas não são qualificados por serem bons em seguir um procedimento ou técnica específica, mas por aplicar seus conhecimentos e habilidades técnicas para resolver problemas quando não há informações completas ou um conjunto definido de etapas para a resolução. Ao contrário dos "problemas de escola", os da vida real têm uma mistura de informações relevantes e irrelevantes, e alguns dos aspectos mais desafiadores que atrapalham a resolução referem-se a questões como Qual informação é necessária?; Quais conceitos são relevantes?; Qual seria um bom plano?; Que conclusões são justificadas pelas evidências?

Wieman e Price (2023) argumentam que os problemas aplicados na escola (e, portanto, na avaliação) precisam se assemelhar mais a problemas autênticos: eles devem fornecer aos estudantes oportunidades para interagir e praticar o tipo de tomada de decisão que os profissionais enfrentam no mundo real, ou seja, aprender a pensar e raciocinar como um cientista ou engenheiro. O problema pode ser autêntico, exigindo que os estudantes tomem decisões em vez de seguir um procedimento prescrito e restrito para exigir o conhecimento esperado dos estudantes em determinado nível. A chave é ter uma boa compreensão das decisões que os profissionais enfrentam (vértice de cognição) e usar esse conhecimento para nortear o projeto de tarefas e métodos de pontuação.

Encontrar o equilíbrio ideal entre autenticidade e praticidade na avaliação envolve a escolha de tarefas e perguntas que restrinjam as soluções dos problemas em um nível apropriado. Ao restringir demais, os recursos importantes e os processos de decisão não serão investigados. Por outro lado, não limitar o suficiente resulta em respostas que podem variar a ponto de ser impossível avaliar e comparar detalhadamente os pontos fortes e fracos dos estudantes.

Fonte: Wieman e Price (2023).

Embora decidir e incorporar problemas autênticos nas avaliações disciplinares represente caminhos importantes para inovar as práticas atuais de avaliação, situar as próximas gerações de avaliação em vários domínios também pode ser uma abordagem valiosa. Uma forma de envolver os estudantes em tarefas interdisciplinares pode ser propor situações de avaliação em que sejam convocados a agir como cidadãos responsáveis, confrontando problemas que envolvam um grupo de colegas, um bairro ou comunidades mais amplas. Avaliações modernas baseadas em simulação podem incorporar muitas dessas situações de aprendizagem experiencial, oferecendo oportunidades para fazer escolhas sociais e desenvolver compreensão empática ao se projetar por meio de um avatar (RAPHAEL *et al.*, 2009). Esses contextos podem ser particularmente adequados para avaliar habilidades socioemocionais, como comunicação, cooperação, regulação emocional e empatia. Um número crescente de jogos de RPG tem sido projetado para avaliar essas habilidades de forma furtiva, como o *Hall of Heroes* (IRAVA *et al.*, 2019). Mesmo assim, um desafio significativo no desenvolvimento de avaliações interdisciplinares é que faltam teorias sólidas sobre o desenvolvimento de conhecimentos e habilidades nesses "domínios". Definir exatamente quais fatores são relevantes ou não para o conceito e o que constitui um "bom desempenho" de uma forma válida em todas as culturas são desafios relacionados à próxima geração de avaliações.

Tarefas individuais *vs.* colaborativas

O trabalho em grupo é cada vez mais utilizado mundialmente como prática pedagógica, apesar dos desafios que os professores enfrentam para estruturar e moderar efetivamente a aprendizagem colaborativa (GILLIES, 2016). Pesquisadores e professores estão cada vez mais conscientes dos efeitos positivos que a colaboração pode ter na capacidade de aprendizagem dos estudantes. Pesquisas mostram

que o trabalho colaborativo promove o desempenho acadêmico e as habilidades de socialização, e esses efeitos positivos se mantêm em todas as idades e disciplinas (BAINES; BLATCHFORD; CHOWNE, 2007; GILLIES; BOYLE, 2010). As práticas de avaliação formativa seguiram essa tendência, pois mais professores ao redor do mundo aplicam o método de rubrica para avaliar a capacidade de seus estudantes trabalharem em grupo. Nas avaliações somativas, o progresso tem sido muito mais instável, embora com exceções notáveis (consulte o Box 3 para conferir dois exemplos em avaliações de larga escala).

BOX 3.

ANÁLISE DE COLABORAÇÃO DO ALUNO EM AVALIAÇÕES DE LARGA ESCALA

Os casos do PISA 2015 e do *Assessment and Teaching of 21st Century Skills* (ATC21S)

Dentro da estrutura da avaliação do PISA para solução colaborativa de problemas, três competências formam o núcleo da dimensão da colaboração: estabelecer e manter um entendimento compartilhado, adotar as devidas medidas para resolver o problema e estabelecer e manter a organização do grupo. O ATC21S identifica dimensões semelhantes de colaboração: participação, tomada de perspectiva e regulação social.

Há uma diferença fundamental entre essas duas experiências: no PISA, os estudantes interagiram com agentes virtuais, enquanto o ATC21S optou pela colaboração entre humanos. A escolha do PISA foi justificada pelo objetivo de padronizar a experiência de avaliação para permitir o uso de métodos de pontuação estabelecidos. A interação entre os estudantes e o agente limitou-se a declarações pré-definidas em formato de múltipla escolha, e toda intervenção possível dos estudantes foi anexada a uma resposta específica dos agentes virtuais ou evento no cenário do problema. Esse ambiente de teste altamente controlado e a falta de formatos de resposta aberta para os estudantes inevitavelmente reduziram a autenticidade da avaliação.

Por outro lado, a abordagem "de humano para humano" do ATC21S tem mais validade de expressão, pois os estudantes podem escolher quando e como interagir com colegas usando um chatbot, isto é, um chat automatizado no qual o programa de computador tenta simular um humano para conversar com as pessoas. No entanto, nesse ambiente mais aberto, é difícil prever o comportamento dos estudantes, criando desafios óbvios para a pontuação. Além disso, o sucesso de um estudante depende do comportamento de outros estudantes, bem como dos estímulos e das reações que eles oferecem. Esse aspecto evidencia o problema de como criar pontuações separadas para os estudantes e seu grupo, e manifesta a preocupação de saber se é justo penalizar um aluno pela falta de habilidade ou motivação de outro.

Essas experiências sugerem que é possível imaginar um futuro não tão distante em que as tarefas colaborativas sejam um componente integral das avaliações. Hu, Shubeck e Sabatini (2023) apresentam exemplos de como o Processamento de Linguagem Natural (PLN) pode ser aproveitado para promover a autenticidade da interação com os agentes virtuais, projetando agentes inteligentes que conseguem "entender" o que os estudantes escrevem ou dizem e responder

devidamente. Da mesma forma, os avanços no PLN têm o potencial de permitir a replicação automatizada de julgamentos de especialistas para grandes conjuntos de dados de conversação, melhorando a qualidade e reduzindo os custos de análise de conversas gravadas e chats por escrito entre estudantes. Independentemente da abordagem, a realização de tarefas colaborativas autênticas requer inovação paralela substancial na medição, pois os modelos analíticos padrão não podem lidar com as muitas interdependências ao longo do tempo e os agentes que surgem em ambientes colaborativos.

Fonte: Piacentini e Foster (2023); Hu, Shubeck e Sabatini (2023).

Estabelecer bases conceituais sólidas

Com maior clareza sobre atividades, contextos e agentes visados para uma nova avaliação, é necessário fazer uma listagem dos conceitos, linguagem e ferramentas que as pessoas usam no domínio-alvo e definir as características de um bom desempenho nesse domínio. Nas avaliações tradicionais de assuntos disciplinares (por exemplo, Matemática), descrições detalhadas do domínio já estão disponíveis para uso no projeto de avaliação. Por exemplo, se o que queremos é avaliar a capacidade de leitura, os desenvolvedores da avaliação podem contar com diversas referências bibliográficas que definem o conhecimento e as habilidades necessárias e que examinaram como as crianças aprendem a ler e melhoram sua proficiência. No entanto, o mesmo entendimento ou conhecimento sobre o progresso da aprendizagem não está disponível para competências complexas, como resolução colaborativa de problemas ou comunicação.

Para gerar essas informações, os desenvolvedores de avaliação podem contar com a contribuição de um grupo de especialistas capazes de construir novas representações do que significa especialização nesses domínios, usando observações empíricas na medida do possível. A Análise Cognitiva da Tarefa (ACT) adota uma variedade de estratégias de entrevista e observação, incluindo rastreamento de processos, para capturar e descrever como os especialistas executam tarefas complexas (CLARK ***et al***. 2008). Por exemplo, uma estratégia definida usada para a ACT é a técnica de incidente crítico, na qual um especialista é chamado a relembrar e a descrever as decisões que tomou durante uma situação autêntica (consulte o capítulo 4 em *Innovating Assessments to Measure and Support Complex Skills* para conferir um exemplo dessa prática). A seguir, as descrições geradas pela ACT são utilizadas para desenvolver experiências de treinamento e avaliações, pois permitem identificar recursos de tarefas que faz sentido incluir e decisões que são mais indicativas de competência.

A definição de um modelo empírico do domínio pode contar com o respaldo de estudos observacionais que abordam como os estudantes trabalham em tarefas que envolvem as habilidades-alvo. Por exemplo, em uma avaliação de habilidades de colaboração, os desenvolvedores podem criar alguns modelos de atividades colaborativas que reflitam sua compreensão inicial de situações relevantes no domínio. Eles podem usar métodos da ACT para identificar os estudantes que são mais ou menos bem-sucedidos em direcionar a colaboração até o resultado esperado e preparar uma listagem do que os estudantes em diferentes níveis de proficiência dizem e fazem (por exemplo, como compartilham informações dentro de um grupo, como negociam a divisão de tarefas, etc.). Estudos observacionais fornecem clareza sobre a sequência de ações que devem ser executadas para atingir uma meta de desempenho e produzir exemplos de produtos reais de trabalho, ou outras evidências tangíveis baseadas em desempenho que podem ser associadas a declarações de proficiência.

Nas fases subsequentes de desenvolvimento, os desenvolvedores de avaliação colaboram com especialistas do domínio para organizar em argumentos de avaliação as informações coletadas na análise de domínio. Isto é, as afirmações que desejam fazer sobre o desempenho do aluno, os dados que servem como evidência para essas afirmações e as garantias ou as razões que explicam por que certos dados devem ser interpretados como evidências suficientes para determinada afirmação (TOULMIN, 2003; MISLEVY; RICONSCENTE, 2006). Conforme exemplificado no Box 4, os argumentos de avaliação podem ser proveitosamente formalizados com o uso de "padrões de projeto", que descrevem o conhecimento, as habilidades e os comportamentos do aluno que são o foco da avaliação. Há também as observações potenciais, os produtos de trabalho e as rubricas que os desenvolvedores de avaliação podem querer usar, além das características de potenciais tarefas de avaliação. Essa estrutura de padrão de projeto ajuda a identificar e a consolidar as bases conceituais de uma avaliação e serve como alicerce para elaborar as especificações técnicas que orientam a operacionalização da avaliação – ou seja, os modelos de aluno, a tarefa e a evidência da estrutura de DBE na Figura 3.

BOX 4.

PADRÕES DE PROJETO NA AVALIAÇÃO: UM EXEMPLO DO PISA

Padrão de projeto para modelagem computacional na avaliação Aprender no Mundo Digital, PISA 2025

Explicação (justificativa)	A modelagem é uma prática central no raciocínio científico, mas os estudantes raramente interagem com ela durante o ensino obrigatório. Os computadores tornam a modelagem mais acessível e significativa para os estudantes, principalmente os principiantes. Observar como os estudantes constroem, refinam e usam modelos computacionais fornece evidências relevantes e interpretáveis sobre como os estudantes são capazes de desenvolver seu próprio conhecimento e compreensão de fenômenos complexos por meio de computadores.
Conhecimentos, habilidades e comportamentos focais	▪ Entender o conceito de variáveis, incluindo variáveis dependentes, independentes, de controle e variáveis moderadoras. ▪ Criar uma representação abstrata de um sistema que pode ser executado por um computador; garantir que o modelo funcione conforme o esperado (por exemplo, observar comportamentos de agentes em uma simulação baseada no modelo). ▪ Identificar tendências, anomalias ou correlações nos dados. ▪ Fazer uma experimentação utilizando a estratégia de controle de variáveis. ▪ Usar um modelo computacional para fazer previsões sobre o comportamento de um sistema.
Conhecimentos, habilidades e comportamentos adicionais	▪ Conhecimento funcional das TICs. ▪ Autoeficácia em TIC. ▪ Conhecimento prévio do fenômeno a ser modelado. ▪ Orientação para perseverança, consciência e maestria.
Observações potenciais e produtos de trabalho	▪ O modelo de aluno representa as informações disponíveis sobre a situação do mundo real. ▪ O aluno consulta recursos de informações relevantes e coleta dados significativos para definir os parâmetros do modelo. ▪ O aluno modifica um modelo incompleto ou defeituoso e justifica suas modificações. ▪ O aluno identifica os pontos fracos do modelo. ▪ O aluno usa seu modelo para traçar previsões corretas (conforme os dados disponíveis).
Recursos característicos de tarefas	▪ Os estudantes recebem informações sobre um fenômeno social ou científico real para modelar ou as ferramentas para obter essas informações. ▪ O aluno pode verificar seu modelo ao comparar seu resultado com dados reais. ▪ Os estudantes podem usar o modelo para traçar previsões.
Recursos variáveis de tarefas	▪ Nível de familiaridade do fenômeno para modelar. ▪ Complexidade das ferramentas de TIC usadas para modelagem. ▪ O aluno aperfeiçoa um modelo básico (fornecido a ele) ou cria o modelo do zero. ▪ O aluno deve encontrar dados relevantes (por exemplo, em um recurso de informação) ou gerar seus próprios dados por meio de experimentação. ▪ Número de variáveis a serem modeladas e estrutura do sistema (simples vs. multinível).
Restrições e desafios	▪ Tempo limitado para aprender a usar a ferramenta de modelagem. ▪ Tempo limitado para aprender conceitos de modelagem desconhecidos (por exemplo, controle de estratégia variável). ▪ Grandes diferenças de conhecimento prévio na população de estudantes-alvo, o que indica dificuldade em desafiar todos os estudantes na mesma tarefa da forma correta.

Fonte: Piacentini (2023).

CONSIDERAR DIFERENÇAS SOCIOCULTURAIS AO DEFINIR CONCEITOS DE AVALIAÇÃO

Inferências comparativas requerem equivalência de medição e comparabilidade de pontuações quando os testes são administrados em vários idiomas ou quando estudantes de grupos culturais distintos são examinados na mesma língua. As questões de validade e comparabilidade intercultural têm particular relevância para avaliações de conceitos complexos em contextos multiculturais e multilíngues, como em avaliações internacionais e em países com populações culturalmente diversas.

A equivalência de conceito é um aspecto importante para definir o foco de uma avaliação. Ela determina o grau em que suas definições se assemelham para as populações visadas pela avaliação. Isto é, espera-se que os indivíduos desenvolvam e progridam nesses conceitos de maneira semelhante e se os conceitos são acessíveis de maneira semelhante para todas as populações. Ela é fundamental para todas as avaliações destinadas a grupos multiculturais e multilíngues, mas assume relevância específica para avaliações de conceitos complexos e multidimensionais em larga escala ERCIKAN; OLIVERI, 2016).

Conceitos como criatividade, inteligência, pensamento crítico e colaboração não são ensinados da mesma forma nas escolas, sendo conceituados e definidos de forma distinta em diferentes culturas. Por exemplo, o modo como a criatividade se desenvolve e como os comportamentos criativos se manifestam é algo que difere entre os grupos culturais (LUBART, 1990; NIU; STERNBERG, 2001). Outros pesquisadores também argumentaram que a noção de inteligência é fundamentada em contextos culturais e, como tal, os conceitos têm diferentes definições nesses contextos (STERNBERG, 2013).

Considerando que habilidades complexas estão inseridas em contextos sociais e são caracteristicamente moldadas por normas e expectativas culturais, é previsto que suas manifestações e o valor atribuído aos resultados dos estudantes variem entre as culturas. Devido a essas diferenças entre os grupos culturais, há uma necessidade de avaliar claramente quais aspectos de um conceito podem ser avaliados de forma significativa em um contexto comparativo e, portanto, incluídos na avaliação, mesmo que isso possa resultar no estreitamento do conceito. A Avaliação de Pensamento Criativo do PISA 2022 (OECD, 2022) exemplifica como a avaliação de um conceito complexo em grupos linguísticos e culturais pode, no entanto, concentrar-se em certos aspectos do conceito que otimizam a comparabilidade (consulte o Box 5).

BOX 5.

CONSIDERAÇÃO DE DIFERENÇAS SOCIOCULTURAIS NA DEFINIÇÃO DO CONCEITO EM AVALIAÇÕES DE LARGA ESCALA

Garantindo a equivalência de conceito na Avaliação do Pensamento Criativo, PISA 2022

A Avaliação do Pensamento Criativo, do PISA 2022, enfatiza que os itens de avaliação devem se pautar em conhecimentos e experiências comuns à maioria dos estudantes mundialmente, e para os quais os estudantes podem produzir trabalho criativo de forma significativa e realista dentro das restrições de um ambiente do PISA.

Para que isso acontecesse, os desenvolvedores dessa avaliação consideraram cinco questões em particular:

- Centraram a avaliação no conceito mais restrito do pensamento criativo (em vez de no conceito mais amplo da criatividade), definido como a "competência para se engajar produtivamente na criação, avaliação e melhoria de ideias". Esse foco mais estreito enfatizou os processos cognitivos relacionados à criação de ideias, enquanto a criatividade também abrange traços de personalidade e requer julgamentos subjetivos sobre o valor criativo das respostas dos estudantes.

- Definiram o pensamento criativo, o modo como ele é ativado (ou seja, indicadores de oportunidades para desenvolvê-lo) e como ele acontece no contexto de uma sala de aula frequentada por estudantes de 15 anos, com foco nos aspectos do conceito com maior probabilidade de serem desenvolvidos em contextos escolares (e não fora da escola).

- Identificaram domínios de avaliação interculturalmente relevantes com os quais os jovens de 15 anos poderiam interagir e praticar o pensamento criativo (por exemplo, escrever histórias curtas, criar produtos visuais, debater ideias sobre problemas sociais e científicos comuns).

- Na pontuação, focaram tanto na originalidade (definida como raridade estatística) quanto na diversidade das ideias (definida como pertencente a diferentes categorias de ideias), em vez de focarem no seu valor criativo (considerado mais sujeito a diferenças socioculturais).

- Engajaram-se na verificação intercultural significativa das rubricas de codificação usadas por avaliadores humanos para analisar as respostas, incluindo o refinamento dessas rubricas por meio da análise de respostas de estudantes em diversos países.

Fonte: OECD (2022).

INOVAÇÃO DO VÉRTICE DE *OBSERVAÇÃO*: INCLUIR TAREFAS DE AVALIAÇÃO MAIS VARIADAS E INTERATIVAS

Considerando a avaliação como um processo de raciocínio com base em evidências, as tarefas de avaliação devem extrair evidências relevantes dos estudantes, e elas precisam estar claramente conectadas ao conceito. Em outras palavras, tarefas ou situações de avaliação devem permitir a observação dos tipos de desempenho que esperamos que os estudantes dominem. Para conceitos como conhecimento matemático, a ligação entre os indicadores do teste e o conceito é bastante direta: uma resposta correta a uma pergunta específica demonstra conhecimento do tópico. No entanto, essa lógica pode não ser suficiente para captar a complexidade das competências do século XXI.

Um argumento central da publicação *Innovating Assessments to Measure and Support Complex Skills* aponta que as avaliações têm maior probabilidade de gerar evidências válidas sobre o que os estudantes sabem e podem fazer se forem confrontados com situações autênticas. O que motiva o apelo à inovação é o fato de que as avaliações existentes muitas vezes impedem que isso aconteça, parcialmente porque a capacidade técnica que essa abordagem em escala demanda tem sido lenta. Avaliações educacionais, particularmente testes padronizados em larga escala, foram projetadas dentro de um conjunto de restrições – custos de impressão e transporte, segurança, ambiente e tempo de teste e custo de pontuação – enquanto precisam atender a padrões psicométricos de confiabilidade, validade, comparabilidade e imparcialidade. As principais características do design, a administração, a pontuação e o relatório de testes "tradicionais", como itens de múltipla escolha, tomaram forma devido a essas restrições (OECD, 2013), e sua capacidade de capturar aspectos mais complexos e multifacetados do desempenho permaneceu, consequentemente, limitada.

Mesmo assim, muitas das restrições no design e na administração de testes não fazem mais sentido, foram transformadas ou podem ser flexibilizadas – em grande parte devido aos avanços tecnológicos e em análises de dados. Em particular, a caixa de ferramentas digitais disponível agora para desenvolvedores de teste expande consideravelmente as oportunidades e os recursos do projeto de avaliação, com potencial para tornar as experiências de teste menos artificiais e mais válidas ao aproximar ou simular situações ou contextos em que os conceitos-alvo são aplicados na vida real.

REPENSAR O DESIGN DA TAREFA

Piacentini, Foster e Nunes (2023) fornecem um conjunto de inovações para tarefas e itens, (Capítulo 2 em *Innovating Assessments to Measure and Support Complex Skills*) o que inclui (1) permitir tarefas aprofundadas de desempenho com uma abordagem "chão baixo, teto alto" (ou seja, acessível a todos os estudantes, sem deixar de desafiar estudantes em níveis iniciante ou mais proficiente); (2) contabilizar o conhecimento do domínio de forma explícita; e (3) oferecer oportunidades de falha produtiva e aprendizagem no teste, apresentando *feedback* e suporte instrucional durante a avaliação. A ideia

por trás desses princípios não é eliminar formas mais tradicionais de experiências de avaliação e formatos de resposta, pois elas ainda podem fornecer informações relevantes para usos interpretativos (por exemplo, identificação de lacunas de conhecimento). Em vez disso, o argumento consiste em complementar essas formas estabelecidas de avaliação com um conjunto diferente de experiências avaliativas que incorporam esses recursos inovadores.

Princípio de projeto nº 1: Permitir tarefas aprofundadas de desempenho com uma abordagem "chão baixo, teto alto"

Na avaliação, particularmente nas somativas em larga escala, as considerações de eficiência resultaram na prioridade por tarefas de avaliação curtas e pontuais sobre atividades de desempenho mais longas. Em geral, o uso excessivo de itens curtos fornece dados mais confiáveis sobre se os estudantes dominam um conhecimento específico e conseguem executar determinado conjunto de procedimentos, pois as informações são acumuladas em um número maior de observações. A medição também é menos complexa: a evidência é acumulada pela aplicação de modelos psicométricos estabelecidos a itens completamente independentes. No entanto, se o objetivo da avaliação passa a ser identificar se os estudantes conseguem construir novos conhecimentos em ambientes com uma diversidade de escolhas, então esses estudantes devem ser submetidos a tarefas de avaliação e a ambientes adequados a esse propósito.

Para isso, é importante considerar como uma avaliação pode proporcionar aos estudantes um desafio que seja proposital e que conceda tempo suficiente para que eles demonstrem suas competências. A inclusão de unidades que se aprofundam aos poucos, em que várias são sequenciadas como etapas para atingir um objetivo principal, pode proporcionar aos estudantes uma experiência de avaliação mais autêntica e motivadora. Incentivar uma mudança na mentalidade do examinado, de "tenho que acertar o máximo de questões do teste", para "tenho um desafio para enfrentar e superar", pode, em última análise, fornecer evidências mais válidas do que os estudantes são capazes de fazer fora das restrições de um contexto de teste, que é estressante e focado no prazo em que precisa ser entregue.

É mais desafiador projetar tarefas aprofundadas que são baseadas em desempenho, principalmente porque é preciso estabelecer um enredo coerente que mantenha os estudantes envolvidos e para abordar possíveis problemas de dependência – por exemplo, ao fornecer dicas para movimentar estudantes com dificuldades de uma atividade para outra. Ao mesmo tempo, as avaliações devem permitir que todos os estudantes demonstrem sua capacidade de aprender e progredir, independentemente de seu nível inicial de conhecimento ou habilidade, projetando tarefas com a abordagem "chão baixo, teto alto", o que significa que são acessíveis a todos os estudantes, sem deixar de desafiar aqueles em níveis iniciante ou mais proficiente (consulte o Box 6 para ver um exemplo da *Platform for Innovative Learning Assessments,* PILA, da OCDE).

Uma maneira de projetar problemas "chão baixo, teto alto" é solicitar que os estudantes produzam algo original: pode ser uma história, um jogo, um design para um novo produto, um relatório de investigação sobre alguma notícia, um discurso etc. Tarefas

de desempenho mais aberto geram uma ampla gama de respostas qualitativamente distintas, e até mesmo os estudantes mais avançados têm incentivos para usar recursos que podem ajudá-los a produzir uma solução mais rica, completa e exclusiva. O projeto de uma avaliação com abordagem "chão baixo, teto alto" também pode ser usado no contexto de tarefas de resolução de problemas mais padronizados, deixando claro para os estudantes que há metas intermediárias a serem atingidas e que é esperado que eles progridam ao máximo em direção a uma solução sofisticada.

Os projetos adaptativos também podem abordar a complexidade de medir a aprendizagem em ação entre populações heterogêneas de estudantes. Uma maneira relativamente simples de fazer isso envolve a criação de cenários em que os estudantes têm um objetivo complexo a atingir e progridem conforme concluem uma sequência de tarefas que, gradualmente, ficam mais difíceis (semelhante ao esquema de aumento de nível nos videogames). Estudantes mais proficientes concluirão rapidamente o conjunto inicial de tarefas simples para, então, chegar aos problemas mais desafiadores. Os estudantes menos preparados ainda serão capazes de interagir com as tarefas mais simples, mesmo que não concluam a sequência completa. Dentro desses projetos, ambos os grupos de estudantes trabalham no limite de suas habilidades, com benefícios óbvios em termos de qualidade de medição e engajamento no teste. Com as tecnologias atuais, esse projeto pode ser aprimorado ainda mais com a introdução de vários caminhos adaptativos dentro de um cenário: com base na qualidade de seu trabalho, os estudantes são direcionados imediatamente para subtarefas mais fáceis ou mais difíceis.

BOX 6.

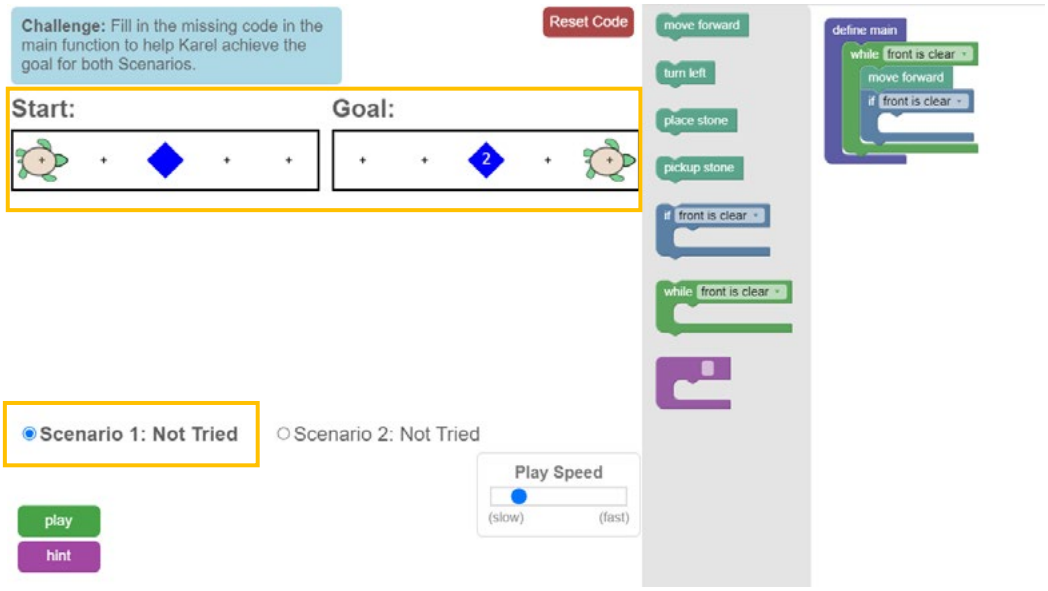
TAREFAS DE AVALIAÇÃO COM ABORDAGEM "CHÃO BAIXO, TETO ALTO"

Atender a estudantes com habilidades diferentes na avaliação do PILA para resolução de problemas computacionais

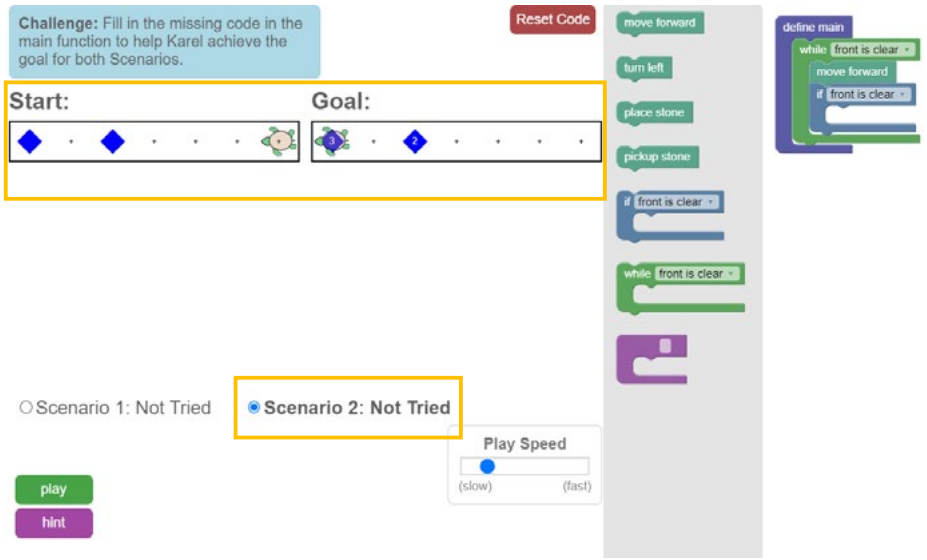
A Platform for Innovative Learning Assessments (PILA) é um laboratório de pesquisa coordenado pela OCDE. Na PILA, as avaliações são projetadas como experiências de aprendizagem e fornecem feedback em tempo real sobre o progresso dos estudantes. Elas podem, portanto, ser usadas também no contexto da instrução em sala de aula. O objetivo geral da PILA é fazer com que desenvolvedores de avaliação, programadores, especialistas em medição e educadores trabalhem juntos para explorar novas maneiras de reduzir a lacuna entre aprendizagem e avaliação.

Um aplicativo desenvolvido na PILA se concentra na resolução de problemas computacionais. Os estudantes usam uma interface de programação visual baseada em blocos para instruir um robô tartaruga ("Karel") a executar determinadas ações. A avaliação é baseada na abordagem "chão baixo, teto alto": a intuitividade da linguagem visual e as ferramentas instrucionais incorporadas (por exemplo, tutorial interativo, exemplos trabalhados) permitem que os estudantes sem experiência em programação participem com sucesso de tarefas algorítmicas simples. No entanto, o mesmo ambiente também pode ser usado para criar problemas que podem desafiar até mesmo programadores experientes.

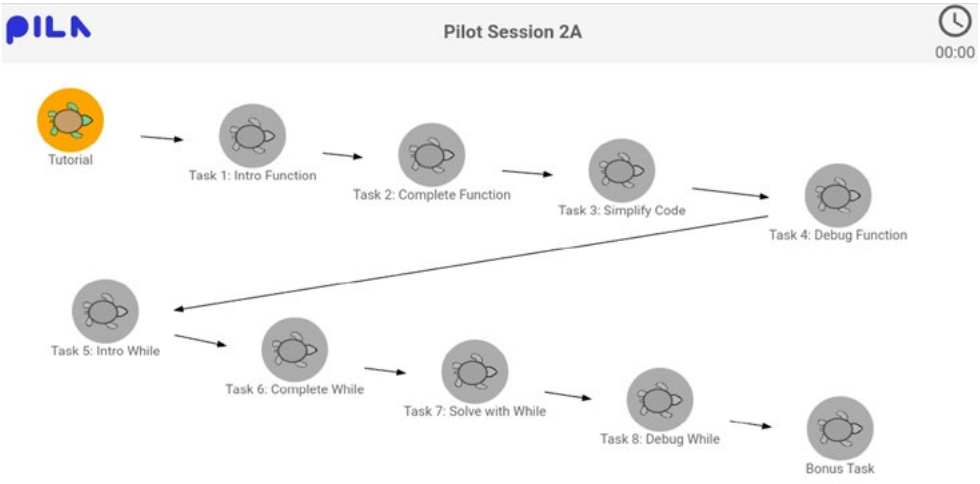
As imagens abaixo mostram um exemplo de problema: os estudantes são convidados a desenvolver um único programa no qual o Karel possa atingir o objetivo em dois cenários diferentes. Para resolver o problema, os estudantes podem alternar entre as duas situações para observar visualmente as diferenças no ambiente e até que ponto o programa resolve o problema em ambos os cenários. Nesse tipo de tarefa, mesmo os estudantes com uma sólida experiência em programação costumam desenvolver e executar várias iterações de seu programa antes de encontrar uma solução. Os modelos de pontuação consideram soluções parciais (por exemplo, a capacidade de um aluno resolver o problema em determinada realidade) e os painéis de relatórios incluem indicadores de desempenho mais complexos (por exemplo, o número de iterações testadas pelos estudantes).



Tarefa: os estudantes precisam programar Karel para avançar e distribuir uma pedra ao longo do caminho para corresponder ao estado de objetivo do cenário 1 (imagem acima). O mesmo código também deve resolver o cenário 2 (imagem a seguir).



Cada experiência de avaliação da PILA também é estruturada como uma progressão de tarefas cada vez mais complexas que têm um objetivo de aprendizagem comum (por exemplo, usar funções com eficiência). Os desenvolvedores de avaliação e os professores têm a opção de bloquear os estudantes em uma tarefa específica até que consigam resolvê-la (ou seja, um mecanismo de aumento de nível) ou possam controlar como se movem pela sequência de tarefas. É esperado que apenas estudantes altamente qualificados concluam toda a sequência de tarefas, o que é comunicado claramente aos estudantes no início para reduzir as experiências de frustração. Futuramente, a PILA planeja incluir caminhos adaptativos (ou seja, sequências de problemas que se adaptam em tempo real ao desempenho do aluno), a fim de alinhar ainda mais a experiência com os conhecimentos e as habilidades prévias dos estudantes.



Exemplo de Experiência de Avaliação (“Mapa”) na sessão com Karel.

Fonte: Piacentini, Foster e Nunes (2023), capítulo 2 em Innovating Assessments to Measure and Support Complex Skills.

Princípio de projeto nº 2: contabilizar o conhecimento do domínio de forma clara

Conforme discutido anteriormente, ao projetar avaliações de competências do século XXI, é importante identificar claramente o conhecimento que os estudantes precisam ter para participar de forma significativa das atividades de teste e avaliar até que ponto as diferenças no conhecimento prévio influenciam a evidência que se pode extrair das habilidades-alvo. No contexto de avaliações somativas em larga escala, pode ser enganoso fazer afirmações gerais como "os estudantes do país A são melhores em resolver problemas do que os estudantes do país B". De fato, com base em uma única avaliação somativa pode-se apenas afirmar que os estudantes do país A são melhores do que os estudantes do país B na resolução de problemas dentro das situações apresentadas no teste (muito provavelmente, um número limitado de situações contextualizadas em um ou poucos domínios do conhecimento).

Medir o conhecimento relevante que os estudantes têm quando realizam uma tarefa de desempenho (por exemplo, por meio de uma pequena bateria de itens no início do teste) deve se tornar parte integrante dos processos de elaboração e planejamento da próxima geração de avaliação. Essas informações também podem ajudar a interpretar os comportamentos e as escolhas dos estudantes em avaliações com tarefas de desempenho complexas. As avaliações também podem minimizar a variabilidade no conhecimento prévio relevante, fornecendo aos estudantes tutoriais, exemplos e problemas passo a passo que podem ajudá-los a interagir com determinada tarefa. Essas abordagens podem ser úteis para contabilizar a compreensão do domínio, como também o conhecimento de recursos ou ferramentas incorporadas no ambiente de avaliação (ou seja, ajudar os estudantes a navegar por ele).

Princípio de projeto nº 3: oferecer oportunidades de falha produtiva e aprendizagem no teste, por meio de feedback e mecanismos de apoio

Nos testes tradicionais, o objetivo é avaliar os conhecimentos adquiridos pelos estudantes antes da tarefa. Normalmente, nenhum *feedback* é passado para eles, as tarefas provavelmente diferem muito uma da outra (para evitar que a resposta seja dada no decorrer do exame) e os tipos de resposta são limitados principalmente às categóricas, ou seja, respostas corretas ou incorretas. Esses instrumentos são insuficientes quando os objetivos da avaliação se expandem da aplicação de conhecimentos estáticos (*resultados* de aprendizagem) para a aquisição dinâmica e o desenvolvimento de novos conhecimentos (*processos* de aprendizagem) diante de tarefas complexas.

Um método promissor para abordar as deficiências atuais envolve o uso de "atividades de invenção" na avaliação. Ele convida os estudantes a resolverem problemas que aparentemente não estão relacionados ao material de aula e que envolvem conceitos ou procedimentos ainda não ensinados. Os estudantes precisam inventar suas próprias soluções originais para esses novos problemas e, nesse processo, tendem a cometer erros e não conseguem gerar soluções canônicas. No entanto, as atividades de invenção ajudam os estudantes a entender profundamente os conceitos, a abandonar velhas interpretações e procedimentos que não funcionam e a procurar novos

padrões e interpretações – e, no contexto de uma avaliação, podem fornecer evidências sobre se os estudantes conseguem aplicar seu conhecimento em contextos desconhecidos com flexibilidade, como fazem os especialistas adaptativos. É certo que a exploração totalmente aberta e não guiada pode não fornecer a evidência mais útil do que os principiantes são capazes de fazer; no entanto, as atividades de aprendizagem devem ser cuidadosamente projetadas para apoiar os estudantes no desenvolvimento de sua compreensão à medida que inventam e interagem com problemas que têm aspectos desconhecidos.

As próximas gerações de avaliação devem considerar a inclusão de orientação e suporte durante o processo de resolução na forma de aconselhamento, *feedback* ou alertas. Esse suporte pode desempenhar uma variedade de funções: (1) atrair o interesse do estudante quando ele parece desinteressado; (2) aumentar sua compreensão dos requisitos da tarefa quando demonstra estar confuso; (3) reduzir os graus de autonomia ou o número de atos constituintes necessários para chegar a uma solução; (4) manter a direção; (5) marcar aspectos críticos, incluindo discrepâncias entre o que o estudante produziu e o que ele reconheceria como correto; (6) demonstrar ou modelar soluções, por exemplo, reproduzir e concluir uma solução parcial que o estudante tentou desenvolver; e (7) provocar articulação e reflexão (GUZDIAL, 2001).

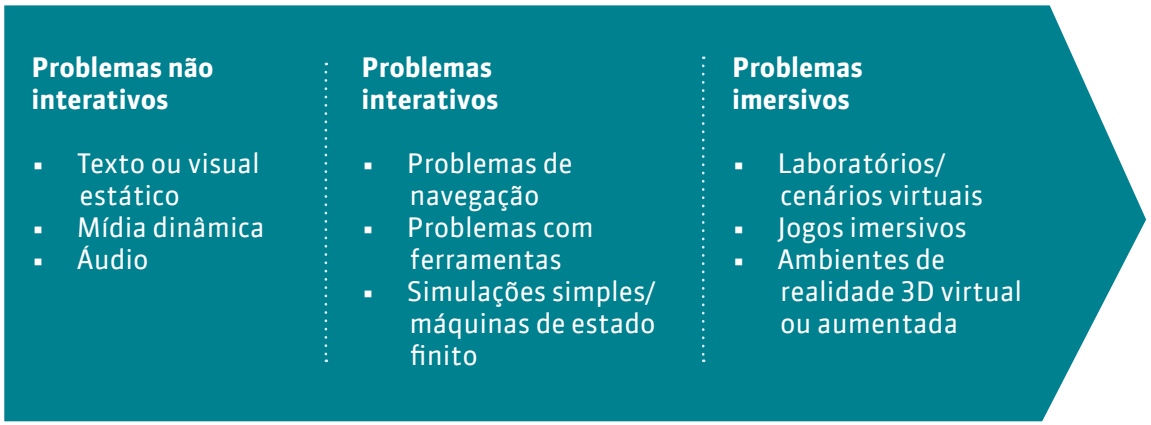
APROVEITAR TECNOLOGIAS MODERNAS PARA INOVAR O PROJETO DA AVALIAÇÃO

Os desenvolvimentos tecnológicos em curso viabilizam cada vez mais as inovações de projeto acima mencionadas, expandindo as ferramentas disponíveis para o preparo da avaliação. Conforme discutido por Sabatini e colegas (*Innovating Assessments to Measure and Support Complex Skills*, Capítulo 7), as tecnologias modernas expandem o alcance do que é possível quando se trata de projetar formatos de tarefas, recursos de teste e fontes de evidência.

Formato de tarefa: das situações de avaliação estáticas às interativas e dinâmicas

Muitas avaliações são caracterizadas por problemas não interativos, o que normalmente inclui texto escrito estático ou estímulos visuais (por exemplo, fotos, desenhos, tabelas, mapas ou gráficos) e, em alguns casos, estímulos mais dinâmicos, como áudio, animações, vídeo e outros conteúdos multimídia. Em problemas não interativos, o material de estímulo costuma fornecer aos estudantes todas as informações necessárias para a resolução da tarefa, as respostas normalmente assumem a forma de itens escritos ou fechados com pouca ou nenhuma interatividade possível do examinador e o ambiente de teste não evolui conforme o participante interage com ele.

Figura 5. *Continuum* de formato de tarefa
Problemas de avaliação não interativos, interativos e imersivos



Fonte: Sabatini *et al.* (2023).

Em contraste, ao criar cenários de resolução de problemas que caracterizam tipos mais complexos de desempenho, os problemas interativos permitem que os estudantes se envolvam ativamente nos processos de fazer e produzir. Esses tipos de formato de tarefa são mais abertos e responsivos a ações e comportamentos dos participantes do teste. Eles costumam ser formados por várias etapas, envolvem o uso de aplicativos de computador, ferramentas ou mecanismos de pesquisa que refletem melhor os contextos contemporâneos da prática e, normalmente, requerem navegação dentro e entre as telas.

Auxiliadas pela tecnologia, as avaliações também podem incorporar problemas verdadeiramente imersivos. Ou seja, laboratórios simulados, jogos imersivos ou modelagem 3D e ambientes de realidade virtual. Os problemas imersivos permitem que os examinados naveguem por uma versão bidimensional ou tridimensional de um mundo virtual – imaginário ou real – em uma tela ou por meio de fones de ouvido de realidade virtual. Os problemas imersivos costumam empregar elementos baseados em jogos para potencializar a motivação, bem como estruturar ou controlar a experiência do aluno (PELLAS et. al., 2018). Os exemplos incluem simulações usadas com mais frequência para treinamento profissional, como aviação virtual ou simulações de intervenção médica, embora esses tipos de tarefa estejam se tornando cada vez mais viáveis para projetar e implementar em escala.

É importante ressaltar que uma maior interatividade e imersão nas tarefas de avaliação precisa ser equilibrada com o conceito e as considerações práticas. Tarefas que evoluem à medida que os examinados interagem podem resultar em experiências de tarefas menos uniformes e, portanto, em cobertura desigual dos conceitos-alvo, criando desafios ao fazer inferências entre as populações de estudantes. Tarefas autênticas e interativas também podem levar mais tempo para serem concluídas do que as estáticas mais simples. A elaboração de tarefas para problemas interativos e imersivos envolve necessariamente a otimização do equilíbrio entre a autenticidade e as restrições da tarefa: em projetos imersivos, é fundamental que as tarefas do mundo virtual sejam sensíveis a variações de desempenho entre indivíduos (por exemplo, principiantes e especialistas

do mundo real), que realmente reflitam o conhecimento e as habilidades de interesse (ou seja, que tenham validade de conceito) e que não distraiam negativamente os estudantes da tarefa em questão.

Recursos de teste: apresentação da adaptabilidade de teste e elementos de aprendizagem

A tecnologia digital também pode servir para inovar os recursos de teste, que se referem a recursos ou características que podem ser sobrepostas a qualquer um dos formatos de tarefas mencionados acima. Dois tipos de recurso são particularmente considerados aqui para as próximas gerações de avaliação: adaptabilidade e recursos de aprendizagem.

Primeiro, a validação digital permitiu o Teste Adaptativo Computadorizado (TAC). Uma das inovações mais pesquisadas na elaboração de teste (WAINER et al., 2000) são as regras de decisão ou como os algoritmos selecionam, com base em um banco de testes, itens individuais para os usuários. Embora diferentes usuários possam receber itens distintos ou módulos maiores na mesma avaliação, suas pontuações são distribuídas em uma escala comum e permanecem comparáveis. Em geral, a adaptabilidade do teste aumenta a eficiência, a precisão e a imparcialidade no projeto, além da administração e da interpretação da avaliação, embora diferentes projetos do TAC tenham diferentes pontos positivos e negativos (consulte o Box 7).

BOX 7.

TESTE ADAPTATIVO COMPUTADORIZADO: POSSIBILIDADES E DESAFIOS

Pontos positivos e negativos de diferentes projetos do Teste Adaptativo Computadorizado (TAC)

Vários projetos do TAC foram pesquisados e implementados em avaliação em larga escala. Em projetos adaptativos mais simples, os itens de teste são agrupados em módulos que diferem em dificuldade e um algoritmo de computador direciona os estudantes para um módulo ou outro, dependendo do desempenho. Os testes podem incluir vários estágios que, por sua vez, incluem diversos módulos (dependendo do módulo e da duração do teste). Diferentes algoritmos podem ser usados para tomar decisões de ramificação entre os estágios do teste. Nesses projetos, a adaptabilidade ocorre no nível do estágio. Outros projetos empregam adaptabilidade instantânea, que acontece no nível do item (ou seja, cada item é adaptado ao aluno com base em seu desempenho nos itens anteriores). Uma vantagem dos Múltiplos Estágios de Teste Adaptativo (META) em relação ao teste adaptativo no nível do item é que ele permite que os módulos incluam formatos de tarefas maiores e mais complexos, que tenham sua própria lógica interna para itens contidos na tarefa. Por outro lado, abordagens dinâmicas (em que os formulários de teste não são definidos antecipadamente pelos desenvolvedores de teste, mas durante o tempo do exame pelo computador) são eficientes na entrega de itens para o conjunto de restrições fornecido. Elas também podem fornecer uma estimativa mais assertiva

da capacidade por unidade de tempo de teste. O ponto negativo de basear as decisões do próximo item apenas no desempenho é que pode resultar na redução da cobertura do conceito e em uma trajetória arbitrária (em vez de coesa ou temática) por meio do domínio de conteúdo. Avanços recentes no CAT podem ajudar a resolver esse problema com a integração de modelos de medição híbridos, embora esses projetos sejam muito menos maduros do que suas contrapartes bem pesquisadas.

Um projeto diferente do TAC adapta tarefas com base em escolhas ou ações anteriores do usuário – como os videogames fazem com ações e comportamentos dos jogadores. Essa abordagem tem a vantagem de refletir melhor as contingências em ambientes reais de resolução de problemas e, se projetada para conceder ao usuário algum grau de escolha ou controle, pode aumentar o engajamento. No entanto, permitir a adaptabilidade baseada por completo na escolha do usuário pode introduzir variações irrelevantes para o conceito quando a escolha não integra claramente a estrutura de avaliação. Mesmo nos casos em que a escolha é explicitamente avaliada, podem surgir problemas semelhantes aos dos modelos adaptativos instantâneos sem mecanismos de restrição suficientes. Tarefas adaptáveis internamente também exigem algoritmos complexos a serem entregues. Ainda não surgiram técnicas para desenvolver tais projetos com agilidade e eficiência, o que encarece o desenvolvimento e os testes desse tipo de adaptabilidade, dificultando a pontuação para efeitos de uma avaliação padronizada. No entanto, avaliações inovadoras podem integrar esse tipo de adaptabilidade multinível mais complexa, adotando algumas das soluções técnicas já usadas em videogames que são projetadas para manter a atenção do jogador, por exemplo, com alternância entre estados de aprendizagem e domínio.

Fonte: Sabatini *et al.* (2023).

Em segundo lugar, as tecnologias digitais facilitam a inclusão de recursos de aprendizagem nas avaliações. Quando o foco avaliativo consiste em identificar em que medida os estudantes sabem ou podem fazer algo em determinado momento, não há necessidade de incorporar recursos de aprendizagem na tarefa. No entanto, avaliações inovadoras podem fazer afirmações sobre como os estudantes lidam com situações-problema autênticas e como adaptam suas estratégias de resolução de problemas à medida que expandem sua compreensão do problema, fazendo uso de uma diversidade de recursos. O Box 8 descreve três tipos de recurso que se tornam possíveis com a integração de recursos de aprendizagem em avaliações digitais.

BOX 8.

DESIGN INOVADOR DE TAREFAS COM RECURSOS DE APRENDIZAGEM

Três tipos de recursos em avaliações ricas em tecnologia

Os recursos de aprendizagem oferecem múltiplas maneiras para decretar comportamentos orientados para objetivos. Roll e Barhak-Rabinowitz agrupam essas possibilidades em três famílias: Experimentação, feedback explícito e busca de informações.

A experimentação permite que os estudantes interroguem e representem suas ideias, executando-as de modo a produzir respostas do ambiente. Por exemplo, ambientes de codificação permitem que os estudantes codifiquem, compilem, executem e observem os resultados (por outro lado, tarefas de codificação em que os estudantes inserem o código, mas não podem executá-lo, não são consideradas recursos de aprendizagem conforme essa definição). Outro exemplo são as simulações científicas interativas em que os estudantes podem manipular elementos e observar o resultado de sua exploração (WIEMAN; ADAMS; PERKINS, 2008). O principal benefício dos recursos de experimentação vem de suas respostas às ações do aluno, muitas vezes conhecidas como feedback situacional (NATHAN, 1998; ROLL et al., 2014). Por exemplo, uma simulação interativa de eletricidade ajustará a intensidade da luz mostrada com base na voltagem que os estudantes definirem (JONG et al., 2018; ROLL et al., 2018). O feedback situacional é implícito e se origina na própria situação da tarefa, consistente com sua lógica interna. Ou seja, os estudantes não estão sendo sinalizados ou avaliados por um modelo onisciente externo. Em vez disso, eles têm a oportunidade de extrair, observar e interpretar as informações relevantes da resposta do ambiente (NATHAN, 1998). A observação de como os estudantes respondem ao feedback situacional pode ser usada para avaliar seus comportamentos de monitoramento e os ajustes correspondentes que eles aplicam em suas estratégias cognitivas.

As pregnâncias de feedback explícito fornecem aos estudantes uma avaliação de suas ações, o que pode incluir uma variedade de entradas, desde sinalização de erro até explicações sobre a natureza do erro ou sugestões para trabalhos futuros (DEEVA et al., 2021). O feedback pode ser acionado sob demanda (por exemplo, com o uso de um botão "teste") ou automaticamente (por exemplo, seguindo um número definido de tentativas malsucedidas). Ao contrário do feedback situacional, que é incorporado à narrativa do desafio, o feedback explícito é externo. Ele assume um agente ou ambiente "onisciente" capaz de comparar a entrada do aluno com o estado desejado. Seu uso sob demanda oferece uma medida direta das estratégias metacognitivas dos estudantes, como monitoramento, ou quais objetivos secundários eles buscam (WINSTONE et al., 2016). Assim como acontece com o feedback situacional, os estudantes que optam por ajustar devidamente suas estratégias cognitivas após o feedback explícito demonstram uso produtivo de estratégias metacognitivas (KINNEBREW; SEGEDY; BISWAS, 2017).

As funcionalidades de busca de informações apoiam os estudantes fornecendo comunicação adicional sobre a tarefa em questão. Os recursos

informativos incluem dicas (ALEVEN et al., 2016), (SEO et al., 2021), exemplos trabalhados (GANAIEM; ROLL, 2022; GLOGGER-FREY et al., 2015) bancos de dados pesquisáveis etc. As fontes de informação podem ser fixas (como na maioria dos tutoriais) ou adaptáveis – como em dicas sobre a etapa específica do problema (VANLEHN et al., 2007). Ao usar fontes de informação, os estudantes fazem escolhas sobre quando usá-las (por exemplo, quando solicitar dicas), como usá-las (por exemplo, vídeos de navegação) e como aplicar as informações ao desafio em questão. Estudantes eficazes e estratégicos buscam informações no momento certo para preencher suas próprias lacunas de conhecimento (SEO et al., 2021; WOOD, 2001). Assim, as interações com os recursos de informação podem fornecer insights significativos sobre os processos de busca de ajuda e monitoramento dos estudantes (ROLL et al., 2014).

Para qualquer um dos itens acima, é necessário observar que permitir a escolha durante o fornecimento de suportes de aprendizagem integra um conceito adicional na avaliação. A escolha, portanto, precisa ser explicitamente refletida na definição do domínio e incorporada às inferências sobre o desempenho do examinado.

Fonte: Roll e Barhak-Rabinowitz (2023).

As decisões sobre o tipo exato e a natureza do suporte fornecido aos estudantes devem ser guiadas pelos objetivos da avaliação, conforme especificado na estrutura avaliativa (o vértice de cognição). Por exemplo, quando o uso de *feedback* é considerado relevante, mecanismos inteligentes encaixam *feedback* úteis para os estudantes nas tarefas. Em outras palavras, se todos os estudantes receberem o mesmo *feedback,* mas se a informação dada não for útil para alguns deles, é possível que não consigam demonstrar a habilidade almejada. Da mesma forma, quando a escolha do examinado é relevante para o conceito, talvez um mecanismo sob demanda seja adequado; no entanto, permitir a escolha também pode impedir oportunidades de observar tais comportamentos, por isso pode ser desejável também criar mecanismos de *feedback* acionados por ação ou evento.

Um dos principais desafios da introdução de recursos de aprendizagem na avaliação está relacionado à decisão de quais modelos de pontuação aplicar. Esses sistemas de suporte podem alterar potencialmente o estado de conhecimento do usuário à medida que o teste avança, influenciando seu desempenho em itens de teste futuros. O que resta é a extensa pesquisa sobre *feedback*, suportes e recursos, como dispositivos de aprendizagem, a serem conduzidos em modelagem psicométrica para o projeto da avaliação. Por exemplo, quando os examinados recebem mais de uma chance de responder (por exemplo, após receber *feedback*), os modelos de pontuação podem atribuir um peso maior a quem acertar mais na primeira vez do que nas tentativas subsequentes. Por outro lado, pode ser que chegar a uma resposta correta, mesmo com suporte, justifique o crédito total. A interação próxima com uma equipe de psicometria durante o processo de desenvolvimento da avaliação é fundamental para entender os tipos de inferências que podem ser feitas e como integrar esses recursos no modelo estatístico.

NOVAS FONTES DE EVIDÊNCIA: DADOS DE PRODUTO E PROCESSO DE RESPOSTA

Os testes digitais expandem a gama de possíveis fontes de evidências nas avaliações. A paleta de potenciais evidências vai muito além das respostas tradicionais de múltipla escolha ou construídas (escritas) que dominaram a elaboração de avaliações tradicionais, sobretudo os exames em larga escala. Uma distinção conceitual central nesse sentido é entre produtos de resposta e processos de resposta, e os diferentes tipos de evidência que essas fontes distintas de dados geram (consulte a Tabela 1).

TABELA 1.
FONTES DE EVIDÊNCIA

Dados do produto de resposta e dados do processo de resposta

DADOS DO PRODUTO	DADOS DO PROCESSO
Várias respostas selecionadas (por exemplo, múltipla escolha, verdadeiro/falso, arrastar e soltar, ponto de acesso etc.)	Dados de tempo (por exemplo, tempo na tarefa, tempo para a primeira ação, tempo inativo)
Resposta escrita	Estados intermediários da solução (ou seja, antes de enviar a solução final)
Resposta falada	Registros de ação (por exemplo, uso de recursos, pressionar de teclas, cliques do mouse, eventos)
Resposta de desempenho (por exemplo, obtenção de nível em um jogo, estado de simulação, produto)	Medidas fisiológicas (por exemplo, dados de rastreamento ocular)

Fonte: Sabatini *et al.* (2023).

Os produtos de resposta referem-se às respostas finais dos estudantes em uma tarefa de avaliação ou em determinado item; os dados do produto de resposta, portanto, costumam se referir a dados resultantes de respostas selecionadas (por exemplo, em um item de múltipla escolha), respostas escritas que sejam curtas ou extensas ou o produto final em uma demonstração de desempenho real ou simulada. Os processos de resposta referem-se aos processos de pensamento, estratégias e abordagens dos examinados quando leem, interpretam e formulam soluções para tarefas de avaliação (ERCIKAN; PELLEGRINO, 2017). Os processos de resposta vão além do domínio cognitivo, incluindo emoções, motivações e comportamentos (HUBLEY; ZUMBO, 2017). Os dados que capturam possíveis evidências desses processos podem, portanto, ser entendidos como dados de processo (resposta), que costumam incluir informações que representam ações ou sequências de ações, dados de rastreamento ocular, de tempo e os que ultrapassam o formato de resposta específico, como chats e diálogos internos com agentes virtuais ou humanos.

A forma mais simples de dados do produto é gerada por meio de formatos de resposta selecionados, como itens de múltipla escolha ou verdadeiro/falso que apresentam respostas predefinidas aos estudantes. Esses formatos de resposta são mais fáceis e econômicos de pontuar do que outros, mas os participantes do teste podem adivinhar a resposta correta e, de maneira mais geral, esses formatos não são capazes de fornecer evidências diretas de habilidades de produção.

Outras formas de dados do produto (respostas construídas) podem fornecer essa evidência, como respostas escritas (variando de frases curtas e pontuais a ensaios extensos), respostas faladas ou por meio da criação de um produto ou representação (por exemplo, participar de um projeto de construção realista em um exame de arquitetura ou realizar uma operação em um simulador médico). Ao exigir que os estudantes participem de uma atividade de produção, as respostas construídas são menos suscetíveis a recompensar indevidamente os estudantes por comportamentos de adivinhação, além de serem mais adequadas para gerar evidências de aprendizagem bem-sucedida e resolução de problemas. Contudo, também exigem maior investimento por parte dos examinados e os dados gerados podem ser mais complexos para pontuar de maneira confiável e comparável. Por exemplo, modelos de pontuação típicos podem assumir a forma de rubricas ou diretrizes, mas podem, por outro lado, restringir o projeto de tarefas autênticas ao exigir os tipos de resposta para as quais avaliadores capacitados podem obter julgamentos confiáveis de qualidade.

Avanços em tecnologia e análise de dados (por exemplo, processamento de linguagem natural, *software* de reconhecimento de fala) estão convergindo para remover algumas dessas barreiras. Por exemplo, ferramentas de análise sintática podem ser usadas para avaliar a estrutura das respostas dos estudantes, e algoritmos de *machine learning* podem ser treinados para identificar semelhança semântica entre as respostas dos estudantes e o gabarito (HU; SHUBECK; SABATINI, 2023).

O surgimento de dados do processo de resposta

Além dos produtos de resposta, um avanço notável nas avaliações apoiadas pela tecnologia é a capacidade de gerar evidências a partir dos processos de resposta. As interações dos alunos com ambientes de avaliação digital podem ser registradas para fornecer dados sobre como eles participam de determinados processos, o que pode ser crítico para entender as operações que realizam quando resolvem uma tarefa e por quê. Os dados do processo de resposta oferecem a oportunidade de revelar essas ações, incluindo onde e como os alunos gastam seu tempo e quais escolhas fazem em ambientes interativos e imersivos, o que pode ser útil para traçar inferências sobre o pensamento do aluno (ERCIKAN; PELLEGRINO, 2017).

Os dados do processo podem variar drasticamente (por exemplo, comportamento on-line, gestos e expressões faciais, interação verbal, movimento dos olhos) e cada fonte desses dados pode contribuir para a compreensão de como os usuários participam das tarefas de avaliação. Nesse sentido, os dados do processo podem constituir evidência de desempenho se métodos de interpretação adequados forem empregados para traçar inferências válidas. No entanto, também podem se tornar uma ferramenta altamente valiosa nos

esforços de validação avaliativa, ajudando os desenvolvedores de avaliação a entender a participação de diferentes estudantes em determinado ambiente de avaliação (ERCIKAN; GUO; POR, 2023).

INOVAÇÃO DO VÉRTICE DE *INTERPRETAÇÃO*: DAR SENTIDO ÀS OBSERVAÇÕES DA AVALIAÇÃO

As seções anteriores destacaram um crescente consenso sobre a necessidade de concentrar a avaliação no que importa; que, para medir competências mais complexas, as avaliações devem apresentar aos estudantes problemas de avaliação abertos e interativos situados em contextos autênticos; e que as avaliações que contam com o apoio da tecnologia podem expandir os tipos de evidências nas quais se pode confiar para fazer afirmações de medição, incluindo fontes de dados que podem elucidar o modo como os estudantes pensam, agem e aprendem, desde que se tenha acesso às devidas ferramentas de interpretação – isto é, garantias robustas. É aqui que reside o terceiro argumento que evoca uma avaliação inovadora: embora definir conceitos de avaliação de competências complexas e captar novas formas de evidência sejam relativamente "fáceis", com o apoio de especialistas no domínio da tecnologia digital, fazer interpretações justificáveis do que a evidência significa é bem mais complicado.

O desafio decorre do fato de que o vértice de interpretação do Triângulo de Avaliação refere-se, na verdade, a dois aspectos: o afastamento de fragmentos de evidências e o acúmulo dessas evidências para fazer uma inferência sobre os Conhecimentos, Habilidades ou Comportamentos (CHC) dos estudantes. Ambos devem ser justificáveis, devendo mostrar exatidão e precisão das métricas envolvidas e descartar hipóteses alternativas, além de verificar se a avaliação é justa e equitativa para subpopulações. Antes de documentar, tanto a eliminação quanto a agregação de evidências devem ser transparentes, justificadas e garantidas.

UMA ABORDAGEM PROJETADA COM BASE EM PRINCÍPIOS PARA ATRIBUIR SENTIDO A DADOS COMPLEXOS: AS REGRAS DE EVIDÊNCIA E OS MODELOS ESTATÍSTICOS NA AVALIAÇÃO

Conforme resumido anteriormente na Figura 3, dois componentes são necessários no processo de construção de garantias ou interpretações justificáveis em avaliações de larga escala: as regras de evidência e o modelo estatístico. Eles especificam como atribuir valores a variáveis observáveis e resumir os dados em indicadores ou escalas.

Criar regras de evidência

As regras de evidência associam uma pontuação a ações e comportamentos do aluno. A formulação dessas regras é bastante direta em avaliações tradicionais e não interativas, sobretudo quando itens de múltipla escolha são usados: se o estudante selecionar a resposta correta, receberá crédito. Tarefas de desempenho mais complexas exigem que os desenvolvedores de avaliação descrevam as características dos produtos de trabalho ou outras evidências tangíveis

que os especialistas do domínio associariam a CHC no domínio de interesse. Em avaliações baseadas em simulação ou jogo, as regras de evidência costumam depender da interpretação de ações e comportamentos que são registrados como dados de processo (consulte o Box 9 para conferir um exemplo).

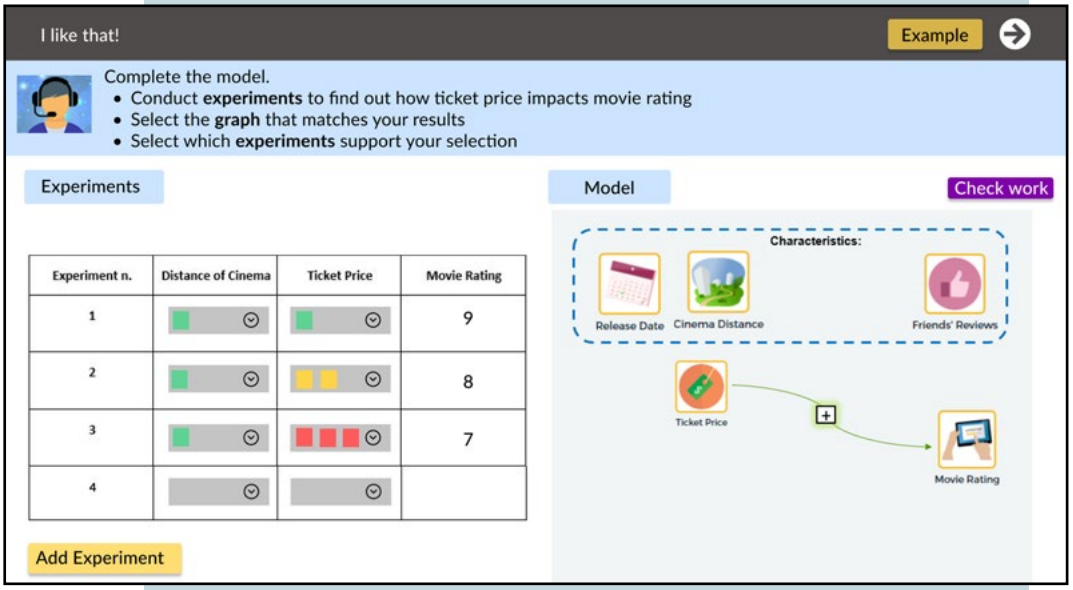
No entanto, a interpretação desses dados é suscetível a erros, pois as ações em ambientes digitais abertos e interativos podem ser frequentemente interpretadas de maneiras variadas. Por exemplo, observar que um usuário interage com todos os recursos (ou seja, o estudante explora as possibilidades com confiança) ou, inversamente, como baixo nível de interação (isto é, ele não participa significativamente da tarefa). Portanto, definir regras de evidência nesses ambientes requer: (1) reconstruir o universo de ações possíveis que o estudante pode realizar e classificá-las em grupos significativos; (2) definir até que ponto as ações dependem do estado do ambiente (e, portanto, das ações anteriores); e (3) usar essas informações para identificar sequências de ações contextualizadas que demonstrem o domínio de CHC visados e que possam ser transformadas em indicadores descritivos ou pontuações.

BOX 9.

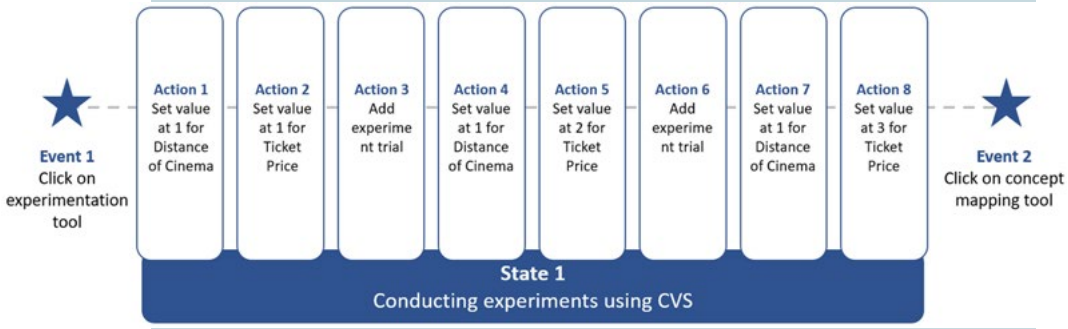
USO DE DADOS DO PROCESSO COMO FONTES DE EVIDÊNCIA

O caso da unidade Eu gosto disso na avaliação Aprendendo no Mundo Digital (AMD) PISA 2025

Em uma tarefa protótipo para a avaliação AMD do PISA, que foi projetada para obter evidências sobre a capacidade dos estudantes de "conduzir experimentos e analisar dados" (imagem abaixo), os estudantes devem usar uma ferramenta de experimentação para conduzir testes nos quais aplicam a estratégia de controle de variáveis (a ECV, ou seja, alterar os valores da variável independente mantendo todas as outras constantes).



Interface da unidade "I like that!"



Sequência de ações para implementar o controle da estratégia variável

Para atribuir uma pontuação ao trabalho de um aluno, as sequências de ações capturadas nos dados do registro são comparadas a uma solução especializada (imagem acima). Regras de pontuação parcial podem ser desenvolvidas para reconhecer os estudantes cujos dados do processo revelam que entenderam a lógica dos experimentos controlados, mas que cometeram erros de procedimento na execução da estratégia (por exemplo, testar apenas alguns valores da variável independente). Como outras tarefas semelhantes de melhoria da tecnologia, é importante considerar a ameaça de variação irrelevante de construção ao definir regras de evidência. Um exemplo de variação irrelevante do conceito nessa tarefa de protótipo pode ser a incapacidade do aluno de conduzir a EVC (ou qualquer experimento) por não conseguir usar os menus suspensos da ferramenta de experimentação.

Fonte: Organisation for Economic Cooperation and Development (OECD, no prelo).

No processo de definição de regras de evidência para avaliações complexas, normalmente os desenvolvedores precisam revisar sua elaboração de tarefas para adicionar recursos que capturem ações direcionadas ou para deixar o ambiente mais buscando reduzir a gama de possíveis ações e interpretações. Um ciclo iterativo de análises empíricas e discussões com especialistas no assunto é, portanto, crucial para identificar evidências em ambientes interativos. Esse processo costuma combinar hipóteses antecipadamente sobre as relações entre observáveis e CHC com análise exploratória e mineração de dados.

Mislevy et al. (2012) descrevem essa interação entre teoria e descoberta para uma atividade de avaliação que abrange a configuração de uma rede de computadores. Os pesquisadores realizaram análises confirmatórias em um conjunto de regras de pontuação definidas por especialistas, que consideravam as características dos produtos de trabalho enviados pelos estudantes (por exemplo, determinada seção da rede é considerada "correta" se os dados forem transferidos de um computador para outro). Eles complementaram essa evidência de produtos de trabalho com a aplicação de métodos de mineração de dados a entra-

das de arquivos de registro com data e hora. Essa análise identificou certos recursos, incluindo o número de comandos usados para configurar a rede, o tempo total gasto e o número de vezes que os estudantes alternaram entre os dispositivos de rede, como possíveis evidências adicionais que poderiam ser combinadas em uma medida de eficiência.

Seleção de um modelo estatístico adequado

O segundo componente do vértice de interpretação é o modelo estatístico que resume os dados entre tarefas ou situações de avaliação, em termos de crenças atualizadas sobre as variáveis do modelo do aluno. No modelo estatístico, o objetivo é expressar, em termos probabilísticos, a relação entre as variáveis observadas (respostas, produtos do trabalho, sequências de ações) e CHC dos estudantes. As especificações de modelagem descritas na estrutura de avaliação fornecem uma base para decisões operacionais durante a construção do teste, como definir quantas tarefas são necessárias para tirar conclusões justificáveis com base nas pontuações.

Os modelos de medição mais simples somam as respostas corretas para tirar conclusões sobre a proficiência, enquanto os modelos de medição mais complexos adotam estruturas de variáveis latentes, como a teoria de resposta ao item (AYALA, 2009; RECKASE, 2009), modelos de classificação de diagnóstico (RUPP; TEMPLIN; HENSON, 2010) e redes bayesianas (LEVY; MISLEVY, 2004; CONATI, 2002).

Avaliações inovadoras que simulam aprendizagem aberta e resolução de problemas podem gerar evidências sobre as capacidades dos estudantes que sejam de alto valor, mas mais difíceis de acumular em modelos de medição existentes. Como a estrutura e a natureza dos dados coletados em tarefas ricas em tecnologia podem variar amplamente entre os examinados, e como os itens do teste podem de fato se tornar interdependentes em tarefas abertas e aprofundadas, torna-se difícil ou inadequado aplicar os mesmos métodos psicométricos utilizados em avaliações mais tradicionais (QUELL-MALZ *et al.*, 2012). Essa evidência só pode ser totalmente explorada com o uso de novas técnicas psicométricas computacionais. Um importante desafio para as avaliações inovadoras é refinar e aproveitar o potencial dos métodos computacionais para lidar com os dados mais ricos de ambientes abertos e interativos, preservando as forças inferenciais dos métodos psicométricos estabelecidos.

UM CONTO DE DOIS MUNDOS: ABORDAGENS DE MACHINE LEARNING E DESIGN BASEADO EM EVIDÊNCIAS

Estudiosos em Análise de Aprendizagem (AA) e Mineração de Dados Educacionais (MDE) obtiveram um tremendo progresso na aplicação de técnicas de *machine learning* (ML) para extrair *insights* úteis dos fluxos de dados gerados em ambientes abertos de aprendizagem digital. Os objetivos desta pesquisa são descrever como os estudantes aprendem ou encontrar maneiras de adaptar e personalizar o conteúdo para estudantes individualmente. Esses novos métodos e os rápidos avanços na tecnologia de computação que os sustentam deram as ferramentas para identificar padrões no pensamento dos estudantes, mesmo em larga escala. Os desenvolvedores de avaliação agora precisam aproveitar esses novos algoritmos computacionais,

orientados por dados, para estabelecer novos modelos analíticos que os ajudem a fazer afirmações de medição, preservando um bom alinhamento com os conceitos fundamentais da psicometria.

Conseguir isso não é tão fácil quanto parece, afinal, os dois campos da psicometria e da análise de aprendizagem seguiram trajetórias de pesquisa bastante distintas. Mais de seis décadas de pesquisa em psicometria e tecnologia de medição estabeleceram procedimentos bem aceitos para questões importantes referentes à avaliação somativa, que incluem calibração e estimativa da pontuação geral, informações de confiabilidade e precisão, criação de formulários de teste, vinculação e equacionamento, administrações adaptativas, avaliação de suposições, verificação do ajuste do modelo de dados, funcionamento diferencial e invariância. Modelos de *machine learning* "caixas preta" como redes neurais de aprendizagem profunda não podem depender de tais procedimentos, com confiabilidade mais baixa quando se trata de fazer afirmações sobre as habilidades dos estudantes – sobretudo quando essas afirmações têm grandes riscos. Inferências robustas desses modelos de ML são possíveis sem a capacidade de calibrar e estimar pontuações gerais, gerar informações de confiabilidade e precisão, conduzir análises de subgrupos e interagir com vinculação e equacionamento?

À primeira vista, o projeto de avaliação centrado em evidências e a mineração de dados educacionais parecem estar em conflito: o primeiro refere-se a uma abordagem pautada em princípios para projetar situações de tarefas que evocam tipos particulares de evidências a serem pontuadas e acumuladas, enquanto o segundo método se concentra em identificar padrões significativos nos dados disponíveis. No entanto, é possível usar métodos de ML dentro de um processo de projeto avaliativo pautado em princípios, em que os modelos de ML geram informações adicionais sobre os examinados. Esses dados podem ser vinculados a regras de evidência e "agregados" a outras evidências (por exemplo, respostas a itens de múltipla escolha), com o objetivo de fazer afirmações mais refinadas e robustas.

A ideia simples, mas poderosa, por trás dessa abordagem é que os métodos estatísticos que possuem propriedades de medição bem estabelecidas, como a Teoria de Resposta ao Item (TRI), podem ser ampliados com técnicas de análise de aprendizagem para explorar por completo a riqueza dos dados disponíveis em tarefas ricas em tecnologia. A evidência agregada resultante do processo pode ser avaliada com procedimentos de diagnóstico padrão e, portanto, considerada mais "confiável" pelos usuários da avaliação. Um exemplo desse método usando um modelo Bayesiano multidimensional TRI (SCALISE, 2017) é apresentado na Caixa 10. O modelo mTRI-Bayes emprega pequenas redes bayesianas para ajudar a gerar pontuações com base em padrões de ações e, em seguida, utiliza um modelo Bayesiano mutidimensional TRI para acumular pontuações e produzir inferências.

Essas oportunidades de fortalecimento em diferentes disciplinas são evidentes no contexto de tarefas tecnológicas autênticas, como simulações ou jogos sérios. Essas atividades incorporam pequenas experiências que geram padrões de dados muitas vezes considerados significativos em termos de afirmações de avaliação. Por exemplo, um avatar controlado pelo aluno pode acabar em uma sala com duas portas, de modo que o estudante deve decidir qual abrir e o que fazer

na próxima sala. Essas escolhas podem ser vinculadas a um modelo de características e habilidades dos estudantes e, portanto, podem ser usadas como evidência para atualizar as crenças sobre o domínio dessas características e habilidades por parte dos estudantes. O caminho a seguir é desenvolver uma estrutura de medição que englobe as perspectivas de ambas as disciplinas e que apoie o projeto e a análise de avaliações tradicionais e inovadoras (MISLEVY *et al.*, 2012).

BOX 10.

APLICAÇÃO DE MODELOS HÍBRIDOS À TAREFA

Tem um novo sapo na cidade

O Novo Sapo VPA é um ambiente virtual imersivo que se parece com um videogame. Cada participante se conecta a um avatar que pode se movimentar pelo ambiente virtual. Os objetivos do relatório da avaliação eram multidimensionais e envolviam exploração e investigação científica (conforme refletido nos padrões científicos da época).

Uma tela de exemplo do O Novo Sapo VPA

Em o Novo Sapo, os examinados foram convidados a explorar o problema de um sapo que tem seis pernas. Eles poderiam escolher examinar diferentes sapos para investigar a questão, e essa escolha em si não seria nem certa nem errada (portanto, o problema não era um "item" típico com uma resposta pré-definida). No entanto, os padrões sobre o tipo e o número de sapos examinados (por exemplo, aqueles localizados em diferentes fazendas, juntamente com amostras de água dos locais) foram considerados informações importantes, e esses padrões podem ser representados em uma rede de Bayes que é pequena, mas informativa.

A acumulação da rede de Bayes adicionou informações consideráveis ao modelo da TRI, mostrou um ajuste aceitável dos padrões da tarefa naturalística e resultou em uma redução do erro padrão de medição (SCALISE; CLARKE-MIDURA, 2018). De

fato, as pontuações geradas pelas duas sub-redes de Bayes provaram estar entre os três "itens" mais informativos da tarefa, em termos de ajuste do modelo no estudo, apesar de terem sido projetadas com base em dados originalmente descartados. Esse fato não surpreende, considerando que a pontuação consistiu em um padrão sobre observações salientes, ao passo em que o outro item mais informativo foi considerado de resposta construída, avaliado por humanos, e que teve um custo mais elevado. Em geral, uma pequena inferência foi viável na tarefa sem tempo adicional de teste ou recursos de pontuação, e os pontos fortes dos estudantes com baixo desempenho na condução de perguntas foram mais evidentes.

Fonte: Scalise, Malcom e Kaylor (2023).

A JUSTIFICATIVA PARA TAREFAS MAIS COMPLEXAS E FORMAS PRÁTICAS DE USÁ-LAS EM RELATÓRIOS

Projetar os tipos de tarefas autênticas modeladas com base em ambientes reais de aprendizagem e resolução de problemas descritos ao longo da publicação *Innovating Assessments to Measure and Support Complex Skills* é parte central de instituir o argumento de validade para as próximas gerações de avaliação que visam às competências do século XXI (como resolução colaborativa de problemas, que são essencialmente definidas por processos).

A inclusão de tarefas mais complexas e autênticas na avaliação também desempenha um papel significativo. Professores, estudantes e formuladores de políticas locais e nacionais tomam decisões referentes aos objetivos de instrução e aprendizagem a partir dos tipos de tarefa encontradas nas avaliações locais, nacionais e internacionais. O que é avaliado muitas vezes acaba sendo o foco da instrução; portanto, é fundamental que as avaliações representem as formas de conhecimento e competência e os tipos de experiências de aprendizagem que queremos ver mais nas salas de aula. Quando é esperado que os estudantes alcancem as proficiências complexas e multidimensionais necessárias para a realidade presente e futura, eles devem conseguir demonstrar sua proficiência. É provável que, incorporando a gestão e a relevância nas avaliações, também seja promovida a interação dos estudantes e, portanto, as chances de observar o que eles podem fazer com a melhor versão de sua capacidade.

Muitos envolvidos no processo de avaliação ainda se sentem à beira do precipício com relação ao efeito do projeto e à inclusão de tarefas mais autênticas em suas práticas de trabalho. Custos, versão, compatibilidade com plataformas de entrega de avaliação e outras considerações práticas existem, sobretudo no contexto de avaliações em larga escala que devem ser replicáveis e comparáveis. Normalmente, essas restrições desencorajam a criação de tarefas complexas para alguns protótipos. Mesmo quando são feitos investimentos na elaboração dessas tarefas, a dificuldade em aplicar abordagens de medição padrão com dados mais complexos, conforme descrito acima, muitas vezes resulta em atalhos que reduzem o valor da integração de tarefas autênticas e abertas em primeiro lugar. Por exemplo, os dados do processo podem ser coletados pela plataforma de entrega,

mas não usados no modelo de evidência, sendo que apenas a resposta final será codificada como correta ou incorreta e fornecerá informações sobre a proficiência do aluno.

Para que a comunidade de medição encontre pontos de entrada práticos para incluir esse tipo mais complexo de evidência, talvez seja necessário que as avaliações educacionais (ao menos por enquanto) incluam uma mistura de itens e tipos de tarefas mais novos e mais antigos, além de investigar como as evidências produzidas por diferentes tipos de formatos de tarefas e experiências são trianguladas. Essa triangulação pode ajudar a desenvolver uma compreensão compartilhada do valor de tarefas inovadoras para inferências e, ao mesmo tempo, torná-las mais justificáveis e "confiáveis" na visão de várias partes interessadas.

Outro caminho promissor consiste em usar diferentes métodos para tipos distintos de afirmações. Por exemplo, modelos de medição estabelecidos podem ser usados para criar uma escala que descreva, de forma confiável e comparável, quais problemas os estudantes são capazes de solucionar. Os métodos de análise de aprendizagem podem ser usados para diagnosticar, de forma mais descritiva, estratégias e processos dos estudantes nas tarefas para atingir um resultado. Isso pode ser feito por meio de uma análise de agrupamento que descreve diferentes "tipos" de solucionadores de problemas, por exemplo. As descrições do trabalho dos estudantes em cada grupo diferente podem ser muito úteis para professores e estudantes, além de fornecer ilustrações tangíveis de como as competências do século XXI são usadas em contextos relevantes ao ensino.

INNOVATING ASSESSMENTS TO MEASURE AND SUPPORT COMPLEX SKILLS: PRÓXIMOS PASSOS

A publicação Innovating Assessments to Measure and Support Complex Skills revela o progresso obtido na conceituação e na operacionalização de aspectos críticos das "próximas gerações de avaliação". O material fornece uma visão dos pontos que precisam estar no foco dessa nova avaliação, como ela será e funcionará. Dessa forma, há o início do mapa do trajeto a ser percorrido para atingir esse objetivo e alguns destinos ao longo do caminho. O mapa inclui os conceitos de interesse, as inovações e as práticas necessárias para progredir, bem como muitos dos obstáculos conceituais e técnicos a serem superados para concretizar a visão de uma avaliação inovadora.

INVESTIR NAS PRÓXIMAS GERAÇÕES DE AVALIAÇÃO

Uma jornada do tipo imaginada pela Innovating Assessments to Measure and Support Complex Skills não pode ser concretizada nem terá sucesso sem um investimento de múltiplas formas de capital. Na discussão que se segue, três formas particulares são consideradas, juntamente com uma explicação de sua relevância. Elas incluem capital intelectual, capital fiscal e capital político. Apesar de cada um deles ser necessário, são insuficientes quando considerados isoladamente. No entanto, quando reunidos, eles fornecem o capital necessário para avançar a teoria e a prática da avaliação educacional e maximizar seu benefício social no século XXI.

CAPITAL INTELECTUAL

Ao considerar a inovação da avaliação, nenhuma disciplina ou área de especialização será suficiente para realizar o que precisa ser feito. Até o momento, os avanços revelam que o desenvolvimento de avaliações da próxima geração é, inerentemente, um empreendimento multidisciplinar. Diferentes comunidades de especialistas precisam trabalhar juntas, e de forma colaborativa, com o objetivo de encontrar soluções para muitos dos desafios conceituais e técnicos já observados e para aqueles que ainda serão descobertos. É fundamental alistar pessoas criativas de várias origens e perspectivas para o empreendimento do projeto e o uso da avaliação, facilitando a colaboração entre esse grupo de pessoas. Sinergias precisam ser promovidas entre desenvolvedores de avaliação e de tecnologia,

cientistas de aprendizagem, especialistas de domínio, especialistas em medição, cientistas de dados, profissionais da educação e formuladores de políticas.

Considerando que a aprendizagem está inserida em contextos sociais e é caracteristicamente moldada por normas e expectativas culturais, é previsto que o desempenho varie entre as culturas. Projetar avaliações válidas, particularmente aquelas que examinam habilidades complexas para as quais não há progressões de aprendizagem estabelecidas, é algo que requer equipes multidisciplinares e especialização no ramo. Portanto, é necessário considerar o complexo contexto sociocultural ao decidir o que e como avaliar e de que maneira os resultados da avaliação serão interpretados e usados. A Avaliação do Pensamento Criativo, PISA 2022 (OECD, 2022), exemplifica a importância de considerar as ameaças à comparabilidade entre idiomas e grupos culturais ao projetar a avaliação de uma construção complexa.

Além das questões de elaboração e validação decorrentes do contexto e da cultura, a comunidade de desenvolvimento de avaliação em larga escala precisará lidar com questões complexas. Isso inclui planejar tarefas que possam simular contextos autênticos e obter comportamentos e evidências relevantes, como interpretar e acumular as inúmeras fontes de dados geradas pelas avaliações aprimoradas pela tecnologia e comparar estudantes de forma significativa em ambientes de teste cada vez mais dinâmicos e abertos. Para abordar essas e outras questões, um número considerável de pesquisas deve se concentrar na modelagem e na validação de desempenhos complexos habilitados por tecnologia que geram conjuntos de dados multifacetados. Esse aspecto inclui dependências de modelagem e dados ausentes não aleatórios em tarefas de avaliação abertas e aprofundadas.

Estudos emergentes mostraram que as técnicas de *machine learning* e inteligência artificial podem ajudar os pesquisadores a entender e a modelar melhor os processos de aprendizagem (KLEINMAN *et al.*, 2022), além de ajudar os especialistas em conteúdo a registrar de maneira eficiente e em escala os processos completos de resolução de problemas dos estudantes (GUO *et al.*, 2022). Trabalhos desse tipo são necessários para complementar as evidências derivadas de estudos laboratoriais cognitivos em pequena escala e promover o avanço da ciência da aprendizagem.

Em um nível pragmático, Schwartz e Arena (2013) argumentam que precisamos "democratizar" o projeto de avaliação, da mesma forma que o design de videogames se tornou mais acessível com a proliferação de comunidades on-line. As plataformas colaborativas, como o sistema PILA na OECD (2023), fornecem aos desenvolvedores modelos de tarefas que eles podem iterar, além de incorporarem instrumentos de coleta de dados que simplificam o trabalho dos pesquisadores de validar e medir. Tais ambientes e bancadas de teste podem facilitar consideravelmente o envolvimento em alguns dos trabalhos intelectuais multidisciplinares mencionados acima.

Resumidamente, há vários desafios intelectuais e pragmáticos na fusão da ciência da aprendizagem, ciência de dados e ciência da medição para entender como as fontes de evidência que se pode extrair

de tarefas complexas podem ser mais bem analisadas e interpretadas utilizando modelos e métodos de inteligência artificial, *machine learning*, estatísticas e psicometria. O envolvimento colaborativo de cientistas de aprendizagem, cientistas de dados, especialistas em medição, designers de avaliação, especialistas em tecnologia e profissionais da educação nessas questões é capaz de gerar uma nova disciplina de Engenharia de Avaliação de Aprendizagem.

CAPITAL FISCAL

O desenvolvimento de avaliações para aplicação e uso em qualquer nível razoável de escala é um empreendimento demorado e de alto custo. A maior parte dos fundos substanciais atualmente gastos em programas de avaliação, nos níveis nacional e internacional, é direcionado ao projeto e à execução de avaliações em larga escala focadas em domínios disciplinares tradicionais, como Matemática, Alfabetização e Ciências (por exemplo, o programa NAEP, nos Estados Unidos, e o programa PISA, da OCDE). Grande parte dessas avaliações se enquadra nos parâmetros convencionais para desenvolvimento, entrega, captura de dados, pontuação e relatórios de tarefas. Esse tem sido o cenário há algum tempo, apesar de muitos programas de avaliação em larga escala terem migrado para um modelo de apresentação, captura de dados e relatórios de tarefas pautado na tecnologia. Capitalizar muitas das possibilidades tecnológicas, conforme descrito anteriormente, não tem sido uma característica distinta desses programas de avaliação.

Desenvolver e validar tarefas e ambientes ricos em tecnologia são atividades de custo muito mais elevado do que a atualização das avaliações atuais com a criação de itens tradicionais que utilizam projetos e especificações de tarefas padrão, apresentados por meio da tecnologia e não por escrito. Esses novos instrumentos requerem pesquisa e desenvolvimento consideráveis em relação a projeto, implementação, análise de dados, pontuação, relatórios e validação de tarefas. Conforme observado acima, esse escopo de trabalho precisa ser executado por grupos interdisciplinares representando especialistas de domínio, desenvolvedores de problemas, psicometristas, designers de interface do usuário e programadores. O financiamento sustentado para o tipo de pesquisa e desenvolvimento necessário é um elemento-chave no avanço da próxima geração de avaliação.

Um obstáculo significativo para alcançar a avaliação do conhecimento e as habilidades do século XXI é a escassez de exemplos de instrumentos de avaliação que tenham um conceito cognitivo complexo, principalmente exemplos que foram criados com base em princípios de planejamento sistemático, como o Design Baseado em Evidências, e posteriormente validados em campo. Os casos em que o trabalho avançou a ponto de oferecer argumentos de validade, incluindo evidências de viabilidade para implementação em escala, raramente foram além dos laboratórios de pesquisa e desenvolvimento (P&D) onde foram prototipados. Esse fato vale até mesmo para casos que atingiram um alto nível de visibilidade na comunidade técnica de pesquisa e desenvolvimento de avaliação. Lamentavelmente, o presente trabalho não conseguiu alterar a forma como a avaliação é conceituada e executada em escala.

É igualmente necessário investir para levar os esforços de avaliações inovadoras existentes à plena maturidade, ampliando sua implementação quando houver evidências de que podem enfrentar o desafio de medir os conceitos que importam. É provável que soluções de avaliação inovadoras atuais e futuras fiquem estagnadas no laboratório de P&D, a menos que seja fornecido o financiamento necessário para encaminhá-las do laboratório para o espaço de implementação em larga escala, no qual sua eficácia e utilidade podem ser devidamente avaliadas. Só então haverá a possibilidade de utilizá-las para substituir as formas de funcionamento atuais.

CAPITAL POLÍTICO

Conforme praticado hoje, a avaliação educacional é um empreendimento altamente arraigado, sobretudo quando falamos do uso de avaliações padronizadas em larga escala para monitoramento educacional e decisões políticas. A padronização inclui o que é avaliado, como é avaliado, como os dados são coletados e analisados e como os resultados são interpretados e relatados. Esse processo é produto de muitos anos operando dentro de uma perspectiva particular sobre o que se quer e se precisa saber sobre o conhecimento, as habilidades e as capacidades dos indivíduos. Atrelado a esse processo temos uma tecnologia altamente refinada de desenvolvimento e administração de testes, atrelada a uma epistemologia de interpretação sobre o mundo mental enraizada em uma metáfora de medição derivada do mundo físico.

É difícil fazer grandes mudanças nos sistemas existentes quando há programas operacionais bem estabelecidos enraizados na prática e na política. Mudanças do tipo considerado necessário requerem forte inclinação e visão política para encorajar as pessoas a enxergar além do que é possível agora, ou mesmo no futuro próximo. Sem vontade política, será impossível gerar capital fiscal suficiente para reunir o capital intelectual necessário na busca por desenvolvimento e implementação da próxima geração de avaliação e alcance de uma mudança significativa na avaliação educacional.

O capital político necessário não se limita aos formuladores de políticas estaduais e federais. Ele abrange vários segmentos das comunidades de desenvolvimento de avaliação educacional, de medição e psicométrica e a de prática educacional. Cada uma delas tem suposições e práticas arraigadas quando se trata de avaliação. Assim, cada comunidade precisa adquirir uma visão de transformação capaz de produzir resultados em desacordo com os aspectos do procedimento operacional padrão atual. Por exemplo, se o conhecimento e as habilidades de um aluno deixarem de ser vistos como pontuais e independentes, avaliá-lo pode exigir o exame de todo o comportamento/processo interativo em ambientes de aprendizagem adaptáveis que imitam cenários do mundo real. Independentemente de onde o processo possa levar, essas comunidades devem trabalhar juntas para gerar a quantidade de vontade política e capital necessários para organizar, apoiar e sustentar esse processo.

AVALIAÇÕES INTERNACIONAIS EM LARGA ESCALA: POSSIBILIDADES DE INOVAÇÃO EM ESCALA

Conforme as questões destacadas anteriormente, é visível que muitos pontos são necessários para o avanço da agenda de inovação em avaliação. Um dos maiores desafios para que essa transformação aconteça é que é necessário visualizá-la em larga escala para verificar o que seria possível. Ampliar ideias promissoras é fundamental para testar o grau de flexibilidade ou fragilidade que essas ideias e abordagens podem ter, além do que é necessário para colocá-las em prática em escala. Felizmente, há exemplos de esforços nesse sentido, que apontam o que é possível e onde permanecem os desafios.

Normalmente, as avaliações internacionais servem como ferramentas para monitorar o desempenho nos padrões disciplinares contemporâneos. Como tal, esses programas fazem declarações sobre o que é valorizado globalmente e apresentam informações sobre a proficiência do aluno em escala. Além disso, ilustram um exemplo operacional do agrupamento de capital (intelectual, fiscal e político) necessário para levar adiante uma agenda de avaliação inovadora e em larga escala. Por exemplo, além de seus programas regulares de avaliação em Matemática, Leitura e Ciências, o Programa PISA da OCDE começou a incluir uma avaliação "inovadora" em cada um de seus ciclos de avaliação. Por meio dessa iniciativa, a OCDE sinalizou as formas importantes de conhecimento e habilidades do século XXI que devem ser avaliadas como parte do monitoramento de metas e objetivos educacionais mais amplos. Será considerado brevemente um exemplo recente desse programa para ilustrar parte do que foi aprendido com as tentativas de colocar em prática ideias inovadoras sobre a avaliação da aprendizagem.

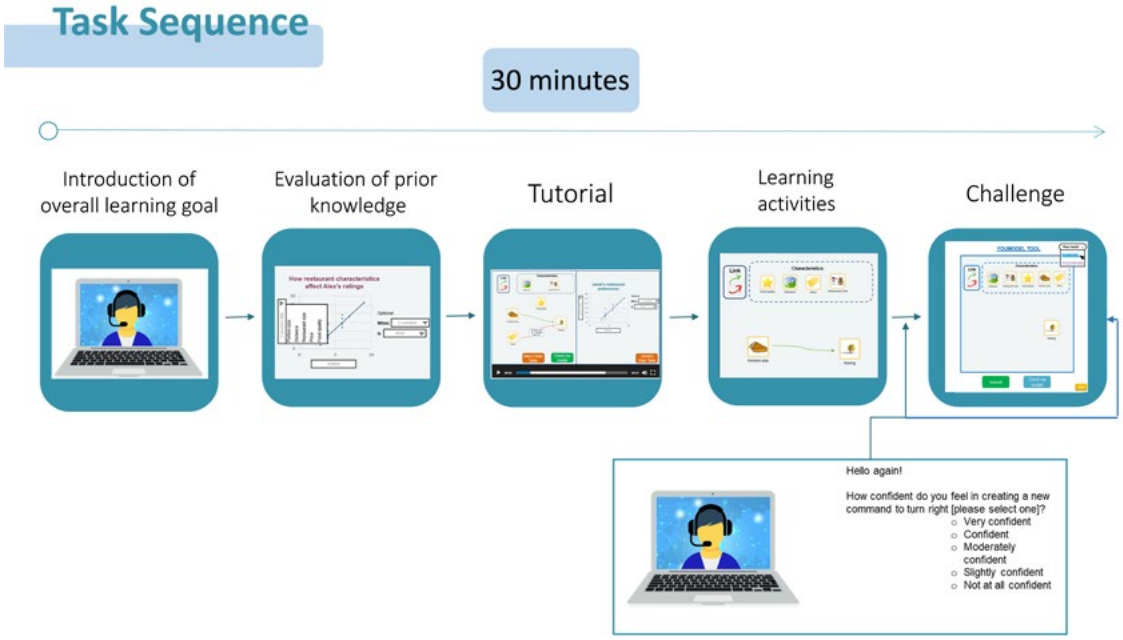
APRENDENDO NO MUNDO DIGITAL PISA 2025

Em seu ciclo de 2025, o PISA incluirá uma avaliação da aprendizagem no mundo digital. Quando o Conselho de Administração do PISA embarcou nesse novo desenvolvimento, em 2020, havia expectativas claras sobre o valor agregado que deveria trazer: os países estavam interessados em dados comparáveis sobre a prontidão dos estudantes para aprender e solucionar problemas com ferramentas digitais. Mesmo antes da pandemia global da COVID-19, as partes interessadas já sabiam que as tecnologias digitais geram grande impacto na educação, apesar de não haver informações suficientes sobre se os estudantes têm as habilidades necessárias para aprender com essas novas ferramentas, e se as escolas estão equipadas para apoiar essas novas formas de aprendizagem.

Essa demanda política orientou várias decisões de projeto. Conforme já foi abordado, uma avaliação de habilidades de aprendizagem tem requisitos diferentes de uma avaliação de conhecimento. Para distinguir aprendizes mais eficazes de aprendizes menos eficazes, a avaliação teve que fornecer oportunidades para os estudantes se envolverem em algum tipo de atividade de construção de conhecimento. Em outras palavras, os desenvolvedores de avaliação tiveram que estruturá-la como uma experiência de aprendizagem, na qual

seria possível avaliar como o conhecimento dos estudantes mudou ao longo da avaliação. Consequentemente, a estrutura das unidades de avaliação divergiu do formato tradicional do PISA, com uma série de estímulos e questões independentes, para um novo formato estruturado como uma série de aulas conectadas (Figura 6).

Figura 6. Sequência de tarefas na avaliação Aprendendo no Mundo Digital PISA 2025



Fonte: OECD (no prelo).

Um tutor virtual orienta os alunos durante o teste, explicando como eles podem resolver problemas relativamente complexos com o uso de ferramentas digitais, que incluem codificação baseada em blocos, simulações, coleta de dados e interfaces de modelagem. Em cada unidade há um tutorial interativo com vídeos para ajudar os alunos a entender como usar essas ferramentas e atenuar as diferenças na familiaridade dos estudantes com ferramentas digitais ou ambientes de aprendizagem específicos. A seguir, eles resolvem uma série de tarefas que progridem do nível mais básico ao avançado, apresentando conceitos e práticas que eles precisam aprender na unidade e que devem aplicar na tarefa final (e mais complexa) do "desafio".

Parte do conceito da avaliação está relacionada à capacidade dos estudantes em participar de uma aprendizagem autorregulada, exigindo, portanto, o desenvolvimento de medidas como monitoramento e adaptação ao *feedback* e avaliação de conhecimento e desempenho. A fim de gerar parâmetros para esses processos de aprendizagem autorregulada, uma série de recursos foram incorporados ao ambiente de avaliação. Ao longo do teste, os estudantes podem receber *feedback*, verificando se alcançaram os resultados esperados ao solicitar que o tutor examine o trabalho. Eles podem optar por conferir as soluções para as tarefas de treinamento depois de enviarem suas respostas e, para cada tarefa, podem acessar dicas e exemplos resolvidos que são úteis para ajudá-los a solucionar o problema. No final de cada tarefa desafiadora, os estudantes são convidados a avaliar seu desempenho

e a relatar o grau de esforço que o trabalho na unidade exigiu, além de compartilhar o que sentiram no decorrer da tarefa. A avaliação, portanto, integra a ideia de que é possível medir melhor os conceitos sociocognitivos complexos proporcionando escolha aos estudantes. Ou seja, não se mede apenas o grau de acerto dos problemas como também o desempenho dos estudantes enquanto aprendem a solucioná-los.

Essas inovações representam respostas a necessidades comprovativas bem definidas. A avaliação foi projetada para fornecer respostas a três questões interconectadas: que tipos de problema no domínio do projeto e modelagem computacional os estudantes conseguem resolver? Até que ponto eles são capazes de aprender novos conceitos nesse domínio ao resolver sequências de tarefas conectadas e estruturadas? E até que ponto essa aprendizagem é sustentada por comportamentos produtivos, como decisões de usar recursos de aprendizagem quando necessário ou monitorar o progresso em direção a seus objetivos de aprendizagem? Essas perguntas definiram o modelo de cognição da avaliação, orientaram o projeto das tarefas necessárias para obter as devidas observações. Além disso, conduzem os planos de análise para interpretar os dados de forma consistente com os propósitos do relatório da avaliação, considerando também a natureza complexa dos dados.

A expectativa é produzir relatórios multidimensionais do desempenho dos estudantes nesse exame, incluindo medidas de (1) desempenho geral dos estudantes nas tarefas (representadas em uma escala, como em outras avaliações do PISA); (2) ganhos de aprendizagem, ou seja, em que grau o conhecimento dos estudantes sobre determinados conceitos e sua capacidade de concluir operações específicas aumenta após o treinamento; e (3) capacidade de autorregular sua aprendizagem e gerenciar seus estados afetivos. Essas diferentes medidas serão trianguladas na análise, por exemplo, explicando parte da variação dos ganhos de aprendizagem com os indicadores de comportamentos de aprendizagem autorregulada. O objetivo é fornecer aos formuladores de políticas informações acionáveis que não se limitem a uma pontuação e posição em um *ranking* internacional, mas que incluam descrições mais sutis do que os estudantes são capazes de fazer, revelando quais aspectos de seu desempenho merecem mais atenção.

CONSIDERAÇÕES FINAIS: VOLTANDO AOS TRÊS TIPOS DE CAPITAL

O desenvolvimento da avaliação de Aprendizagem no Mundo PISA 2025 só foi possível devido à convergência dos diferentes tipos de capital descritos acima. O apoio político de uma agenda de pesquisa e desenvolvimento por parte dos países participantes do PISA tem sido forte. A avaliação inovadora incluída em cada ciclo do PISA é, agora, vista como um espaço seguro para testar inovações importantes no projeto de tarefas e modelos analíticos que podem ser transferidos para os domínios de tendência de Leitura, Matemática e Ciências, ou que podem servir de inspiração para o desenvolvimento de avaliações nacionais, uma vez que seu valor seja comprovado.

Reconhecendo a necessidade de múltiplas iterações na concepção de tarefas e de extensos processos de validação para escolhas analíticas e de projeto por meio de laboratórios cognitivos e estudos-piloto, o Conselho de Administração do PISA forneceu os apoios financeiro e político necessários para iniciar o desenvolvimento do teste cinco anos antes da coleta dos dados principais. Recursos adicionais foram disponibilizados por fundações de pesquisa que reconheceram o valor das avaliações inovadoras.

O desenvolvimento da avaliação também foi conduzido por um grupo de especialistas com diferentes formações disciplinares: especialistas no assunto trabalharam lado a lado com psicometristas, estudiosos em análise de aprendizagem e especialistas em design de UI/UX. Essa fertilização cruzada foi importante para abrir espaço para os novos métodos de identificação de evidências em ambientes de aprendizagem digital, mantendo em mente o objetivo central de atingir métricas comparáveis que gerem interpretações válidas de diferenças de desempenho entre países e grupos.

Esse novo teste do PISA representa apenas uma incursão inicial no empreendimento de avaliações inovadoras. Conforme argumentado na publicação *Innovating Assessments to Measure and Support Complex Skills*, precisamos de novas avaliações disciplinares e interdisciplinares para fornecer uma descrição exaustiva da qualidade das experiências educacionais em todos os países. Vários desafios também permanecem, sobretudo no vértice de interpretação do Triângulo de Avaliação. Fóruns internacionais, como PISA ou IEA, têm um papel a desempenhar na coordenação de demandas políticas e na facilitação de um consenso sobre quais peças do quebra-cabeça precisamos montar, e quais devem ser as prioridades para o curto prazo e além. Há mais do que ampla evidência de que a avaliação inovadora de competências educacional e socialmente significativas é desejável e possível. As evidências também sugerem que a cooperação e a colaboração em escala global podem ser a melhor e a única maneira de alcançar esses avanços.

REFERÊNCIAS

ALEVEN, V. et al. Help helps, but only so much: research on help seeking with intelligent tutoring systems. **International Journal of Artificial Intelligence in Education**, v. 26, n. 1, p. 205-223, jan. 2016. Disponível em: https://doi.org/10.1007/s40593-015-0089-1. Acesso em: 4 abr. 2023.

AYALA, R. de. **The theory and practice of Item Response Theory**. Nova York: Guilford Press, 2009.

BAINES, E.; BLATCHFORD, P.; CHOWNE , A. Improving the effectiveness of collaborative group work in primary schools: effects on science attainment. **British Educational Research Journal**, v. 33, n. 5, p. 663-680, out. 2007. Disponível em: https://doi.org/10.1080/01411920701582231. Acesso em: 4 abr. 2023.

BASOL, M. et al. Towards psychological herd immunity: cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. **Big Data & Society**, v. 8, n. 1, maio 2021. Disponível em: https://journals.sagepub.com/doi/10.1177/20539517211013868. Acesso em: 4 abr. 2023.

BELLANCA, J. **Deeper learning**: beyond 21st century skills. Bloomington: Solution Tree Press, 2014.

BILAL, D. Children's use of the Yahooligans! web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks. **Journal of the American Society for Information Science**, v. 51, n. 7, p. 646-665, 2000. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(2000)51:7%3C646::AID-ASI7%3E3.0.CO;2-A. Acesso em: 4 abr. 2023.

BINKLEY, M. et al. Defining twenty-first century skills. In: GRIFFIN, P.; McGAW, B.; CARE, E. (Org.). **Assessment and teaching of 21st century skills**. Dordrecht: Springer, 2011. Ebook. Disponível em: https://doi.org/10.1007/978-94-007-2324-5_2. Acesso em: 4 abr. 2023.

BISWAS, G.; SEGEDY, J.; BUNCHONGCHIT, K. From design to implementation to practice a learning by teaching system: Betty's Brain. I**nternational Journal of Artificial Intelligence in Education**, v. 26, n. 1, p. 350-364, 2015. Disponível em: https://doi.org/10.1007/s40593-015-0057-9. Acesso em: 4 abr. 2023.

BRAND-GRUWEL, S.; WOPEREIS, I.; VERMETTEN, Y. Information problem solving by experts and novices: analysis of a complex cognitive skill. **Computers in Human Behavior**, v. 21, n. 3, p. 487-508, 2005. Disponível em: https://doi.org/10.1016/j.chb.2004.10.005. Acesso em: 4 abr. 2023.

BRANSFORD, J.; STEIN, B. **The ideal problem solver**: a guide for improving thinking, learning, and creativity. Nova York: Freeman, 1984.

CLARK, R. et al. Cognitive task analysis. In: SPECTOR, J. et al. (Org.). **Handbook of research on educational communications and technology**. Nova York: Macmillan/Gale, 2008. p. 541-551. .

COIRO, J. et al. Students engaging in multiple-source inquiry tasks: capturing dimensions of collaborative online inquiry and social deliberation. **Literacy Research: Theory, Method, and Practice**, v. 68, n. 1, p. 271-292, 2019. Disponível em: https://doi.org/10.1177/2381336919870285. Acesso em: 4 abr. 2023.

CONATI, C. Probabilistic assessment of user's emotions in educational games. **Applied Artificial Intelligence**, v. 16, n. 7-8, p. 555-575, 2002. Disponível em: https://doi.org/10.1080/08839510290030390. Acesso em: 4 abr. 2023.

DEEVA, G. et al. A review of automated feedback systems for learners: classification framework, challenges and opportunities. **Computers & Education**, v. 162, 2021. https://doi.org/10.1016/j.compedu.2020.104094.

ERCIKAN, K.; GUO, H.; POR, H. Uses of process data in advancing the practice and science of technology-rich assessments. In: FOSTER, N.; PIACENTINI, M. (eds.) Innovating Assessments to Measure and Support Complex Skills. Paris: OECD Publishing, 2023. Disponível em: <https://doi.org/10.1787/e5f3e341-en>.

ERCIKAN, K.; OLIVERI, M. In search of validity evidence in support of the interpretation and use of assessments of complex constructs: discussion of research on assessing 21st century skills. **Applied Measurement in Education**, v. 29, n. 4, p. 310-318, 2016. Disponível em: https://doi.org/10.1080/08957347.2016.1209210. Acesso em: 4 abr. 2023.

ERCIKAN, K.; PELLEGRINO, J. **Validation of score meaning for the next generation of assessments**. New York: Routledge, 2017. Disponível em: https://doi.org/10.4324/9781315708591. Acesso em: 4 abr. 2023.

FOSTER, N. 21st century competencies: Challenges in education and assessment. In: FOSTER, N.; PIACENTINI, M. (eds.) **Innovating Assessments to Measure and Support Complex Skills.** Paris: OECD Publishing, 2023. Disponível em: <https://doi.org/10.1787/e5f3e341-en>.

FOSTER, N.; PIACENTINI, M. (eds.). **Innovating Assessments to Measure and Support Complex Skills.** Paris: OECD publishing, 2023. Disponível em: https://doi.org/10.1787/e5f3e341-en.

GANAIEM, E.; ROLL, I. **The effect of different sequences of examples and problems on learning experimental design**. Proceedings of the International Conference of the Learning Sciences, Hiroshima, 2022, p. 727 - 732.

GILLIES, R. Cooperative learning: review of research and practice. **Australian Journal of Teacher Education**, v. 41, n. 3, p. 39-54, 2016. Disponível em: https://doi.org/10.14221/ajte.2016v41n3.3. Acesso em: 4 abr. 2023.

GILLIES, R.; BOYLE, M. Teachers' reflections on cooperative learning: Issues of implementation. **Teaching and Teacher Education**, v. 26, n. 4, p. 933-940, 2010. Disponível em: https://doi.org/10.1016/j.tate.2009.10.034. Acesso em: 4 abr. 2023.

GLOGGER-FREY, I. et al. Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. **Learning and Instruction**, v. 39, p. 72-87, 2015. Disponível em: https://doi.org/10.1016/j.learninstruc.2015.05.001. Acesso em: 4 abr. 2023.

GUO, H.; et al. Understanding students' test performance and engagement. In: INTERNATIONAL MEETING OF PSYCHOMETRIC SOCIETY, 2022, Bologna. **IMPS Proceedings**. Bologna: IMPS, 2022.

GUZDIAL, M.; RICK, J.; KEHOE, C. Beyond adoption to invention: teacher-created collaborative activities in higher education. **Journal of the Learning Sciences**, v. 10, n. 3, p. 265-279, 2001. Disponível em: https://doi.org/10.1207/s15327809jls1003_2. Acesso em: 4 abr. 2023.

HU, X.; SHUBECK, K.; SABATINI, J. Artificial Intelligence-enabled adaptive assessments with Intelligent Tutors. In: FOSTER, N.; PIACENTINI, M. (eds.) **Innovating Assessments to Measure and Support Complex Skills**. Paris: OECD Publishing, 2023. Disponível em: <https://doi.org/10.1787/e5f3e341-en>.

HUBLEY, A.; ZUMBO, B. Response processes in the context of validity: setting the stage. In: HUBLEY, A.; ZUMBO, B. (Org.). **Understanding and investigating response processes in validation research**. S.l.: Springer International Publishing, 2017. p. 1-12. Disponível em: https://doi.org/10.1007/978-3-319-56129-5_1. Acesso em: 4 abr. 2023.

IRAVA, V. et al. Game-based socio-emotional skills assessment: a comparison across three cultures. **Journal of Educational Technology Systems**, v. 48, n. 1, p. 51-71, 2019. Disponível em: https://doi.org/10.1177/0047239519854042. Acesso em: 4 abr. 2023.

JONASSEN, D. What are cognitive tools? In: **Cognitive tools for learning**. Berlim: Springer, 1992. p. 1-6. https://doi.org/10.1007/978-3-642-77222-1_1.

JONG, T. de et al. Simulations, games, and modeling tools for learning. In: **International Handbook of the Learning Sciences**. Nova York: Routledge, 2018. p.256-266 . https://doi.org/10.4324/9781315617572-25.

KLEINMAN, E. et al. Analyzing students' problem-solving sequences. **Journal of Learning Analytics**, v. 9, n. 2, p. 1-23, 2022. Disponível em: https://doi.org/10.18608/jla.2022.7465. Acesso em: 4 abr. 2023.

KINNWBEW, J.; SEGEDY, J.; BISWAS, G. Integrating model-driven and data-driven techniques for analyzing learning behaviors in open-ended learning environments. **IEEE Transactions on Learning Technologies**, v. 10, n. 2, p. 140-153, 2017. Disponível em: https://doi.org/10.1109/tlt.2015.2513387. Acesso em 13 abr. 2023.

LARGE, A.; BEHESHTI, J. The web as a classroom resource: reactions from the users. **Journal of the American Society for Information Science**, v. 51, n. 12, p. 1069-1080, 2000. Disponível em: https://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1017>3.0.CO;2-W. Acesso em: 4 abr. 2023.

LEVY, R.; MISLEVY, R. Specifying and refining a measurement model for a computer-based interactive assessment. **International Journal of Testing**, v. 4, n. 4, p. 333-369, 2004. Disponível em: https://doi.org/10.1207/s15327574ijt0404_3. Acesso em: 4 abr. 2023.

LUBART, T. Creativity and cross-cultural variation. **International Journal of Psychology**, v. 25, n. 1, p. 39-59, 1990. Disponivel em: https://doi.org/10.1080/00207599008246813. Acesso em: 4 abr. 2023.

MESSICK, S. The interplay of evidence and consequences in the validation of performance assessments. **Educational Researcher**, v. 23, n. 2, p. 13, 1994. Disponível em: https://doi.org/10.2307/1176219. Acesso em: 4 abr. 2023.

MISLEVY, R. et al. Design and discovery in educational assessment: evidence-centered design, psychometrics, and educational data mining. **Journal of Educational Data Mining**, v. 4, n. 1, p. 11-48, 2012. Disponível em: https://doi.org/10.5281/zenodo.3554641. Acesso em: 4 abr. 2023.

MISLEVY, R.; HAERTEL, G. Implications of evidence-centered design for educational testing. **Educational Measurement: Issues and Practice**, v. 25, n. 4, p. 6-20, 2007. Disponível em: https://doi.org/10.1111/j.1745-3992.2006.00075.x. Acesso em: 4 abr. 2023.

MISLEVY, R.; RICONSCENTE, M. Evidence-centered assessment design: layers, concepts, and terminology. In: DOWNING, S.; HALADYNA, T. (Org.). **Handbook of test development.** Mahwah: Erlbaum, 2006. p. 61 - 90.

NATHAN, M. Knowledge and situational feedback in a learning environment for algebra story problem solving. **Interactive Learning Environments**, v. 5, n. 1, p. 135-159, 1998. Disponível em: https://doi.org/10.1080/1049482980050110. Acesso em: 4 abr. 2023.

NIU, W.; STERNBERG, R. Cultural influences on artistic creativity and its evaluation. **International Journal of Psychology**, v. 36, n. 4, p. 225-241, 2001. Disponível em: https://doi.org/10.1080/00207590143000036. Acesso em: 4 abr. 2023.

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT. **PISA 2025 Learning in the Digital World assessment framework**. Paris: OECD Publishing, 2023. No prelo.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **Synergies for better learning**: an international perspective on evaluation and assessment. Paris: OECD Publishing, 2013. Disponível em: https://www.oecd.org/education/school/synergies-for-better-learning.htm. Acesso em: 4 abr. 2023.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. **Thinking outside the box:** the PISA 2022 creative thinking assessment. Paris: OECD Publishing, 2022. Disponível em: https://issuu.com/oecd.publishing/docs/thinking-outside-the-box. Acesso em: 4 mar. 2023.

PELLAS, N. et al. Augmenting the learning experience in primary and secondary school education: a systematic review of recent trends in augmented reality game-based learning. **Virtual Reality**, v. 23, n. 4, p. 329-346, 2018. Disponível em: https://doi.org/10.1007/s10055-018-0347-2. Acesso em: 4 abr. 2023.

PELLEGRINO, J.; CHUDOWSKY, N.; GLASER, R. **Knowing what students know**: the science and design of educational assessment. Washigton, DC: National Academies Press, 2001.

PELLEGRINO, J.; DIBELLO, L.; GOLDMAN, S. A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. **Educational Psychologist**, v. 51, n. 1, p. 59-81, 2016. Disponível em: https://doi.org/10.1080/00461520.2016.1145550. Acesso em: 4 abr. 2023.

PELLEGRINO, J.; HILTON, M. **Education for life and work:** developing transferable knowledge and skills in the 21st century. Washigton, DC: National Academies Press, 2012. Disponível em: https://doi.org/10.17226/13398. Acesso em: 4 abr. 2023.

PIACENTINI, M. Defining the conceptual assessment framework for complex competencies. In: FOSTER, N.; PIACENTINI, M. (eds.) **Innovating Assessments to Measure and Support Complex Skills**. Paris: OECD Publishing, 2023. Disponível em: <https://doi.org/10.1787/e5f3e341-en>.

PIACENTINI, M.; FOSTER, N. Framing the focus of new assessments of 21st century competencies. In: FOSTER, N.; PIACENTINI, M. (eds.) **Innovating Assessments to Measure and Support Complex Skills**. Paris: OECD Publishing, 2023. Disponível em: <https://doi.org/10.1787/e5f3e341-en>.

PIACENTINI, M.; FOSTER, N.; NUNES, C. Next-generation assessments of 21st century competencies: Insights from the learning sciences. In: FOSTER, N.; PIACENTINI, M. (eds.) **Innovating Assessments to Measure and Support Complex Skills**. Paris: OECD Publishing, 2023. Disponível em: <https://doi.org/10.1787/e5f3e341-en>.

QUELLMALZ, E. et al. 21st century dynamic assessment. In: MAYRATH, M. et al. (Org.). **Technology-based assessments for 21st century skills**. Charlotte: Information Age Publishing, 2012. p. 55-89. Disponível em: http://www.simscientists.org/downloads/Chapter_2012_Quellmalz.pdf. Acesso em: 4 abr. 2023.

RAPHAEL, C. et al. Games for civic learning: a conceptual framework and agenda for research and design. **Games and Culture**, v. 5, n. 2, p. 199-235, 2009. Disponível em: https://doi.org/10.1177/1555412009354728. Acesso em: 4 abr. 2023.

RECKASE, M. **Multidimensional Item Response Theory**. New York: Springer, 2009. Disponível em: https://doi.org/10.1007/978-0-387-89976-3. Acesso em: 4 abr. 2023.

ROLL, I. et al. Tutoring self- and co-regulation with intelligent tutoring systems to help students acquire better learning skills. *In*: SOTTILARE, R. et al. (Org.), **Design recommendations for intelligent tutoring systems**. Orlando: US Army Research Laboratory, 2014. v. 2: Instructional Management, p. 169 - 182 .

ROLL, I. et al. Understanding the impact of guiding inquiry: the relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning. **Instructional Science**, v. 46, n. 1, p. 77-104, 2018. Disponível em: https://doi.org/10.1007/s11251-017-9437-x. Acesso em: 4 abr. 2023.

ROLL, I.; BARHAK-RABINOWITZ, M. Measuring self-regulated learning using feedback and resources. In: FOSTER, N.; PIACENTINI, M. (eds.) **Innovating Assessments to Measure and Support Complex Skills**. Paris: OECD Publishing, 2023. Disponível em: <https://doi.org/10.1787/e5f3e341-en>.

ROOZENBEEK, J.; VAN DER LINDEN, S. The fake news game: actively inoculating against the risk of misinformation. **Journal of Risk Research**, v. 22, n. 5, p. 570-580, 2018. Disponível em: https://doi.org/10.1080/13669877.2018.1443491. Acesso em: 4 abr. 2023.

RUPP, A.; TEMPLIN, J.; HENSON, R. **Diagnostic measurement**: theory, methods, and applications. Nova York: Guilford Press, 2010.

SABATINI, J.; et al. Designing innovative tasks and test environments. In: FOSTER, N.; PIACENTINI, M. (eds.) **Innovating Assessments to Measure and Support Complex Skills**. Paris: OECD Publishing, 2023. Disponível em: <https://doi.org/10.1787/e5f3e341-en>.

SCALISE, K. Hybrid measurement models for technology-enhanced assessments through mIRT-bayes. **International Journal of Statistics and Probability**, v. 6, n. 3, p. 168, 2017. Disponível em: https://doi.org/10.5539/ijsp.v6n3p168. Acesso em: 4 abr. 2023.

SCALISE, K.; CLARKE-MIDURA, J. The many faces of scientific inquiry: Effectively measuring what students do and not only what they say. **Journal of Research in Science Teaching**, v. 55, n. 10, p. 1469-1496, 2018. Disponível em: https://doi.org/10.1002/tea.21464. Acesso em: 4 abr. 2023.

SCALISE, K.; MALCOM, C.; KAYLOR, E. Analysing and integrating new sources of data reliably in innovative assessments. In: FOSTER, N.; PIACENTINI, M. (eds.) **Innovating Assessments to Measure and Support Complex Skills**. Paris: OECD Publishing, 2023. Disponível em: <https://doi.org/10.1787/e5f3e341-en>.

SCHWARTZ, D.; ARENA, D. **Measuring what matters most:** choice-based assessments for the digital age. Cambridge: The MIT Press, 2013.

SEO, K. et al. Active learning with online video: the impact of learning context on engagement. **Computers & Education**, v. 165, 2021. Disponível em: https://doi.org/10.1016/j.compedu.2021.104132. Acesso em: 4 abr. 2023.

STERNBERG, R. Intelligence. *In*: FREEDHEIM, D.; Weiner, I. (Org.). **Handbook of psychology**: history of psychology. Hoboken: John Wiley & Sons, 2013, p. 155-176.

TOULMIN, S. **The uses of argument.** Cambridge: Cambridge University Press, 2003. Disponível em: https://doi.org/10.1017/cbo9780511840005. Acesso em: 4 abr. 2023.

URBAN, A.; HEWITT, C.; MOORE, J. Fake it to make it, media literacy, and persuasive design: using the functional triad as a tool for investigating persuasive elements in a fake news simulator. **Proceedings of the Association for Information Science and Technology**, v. 55, n. 1, p. 915-916, 2018. Disponível em: https://doi.org/10.1002/pra2.2018.14505501174. Acesso em: 4 abr. 2023.

VAN DER LINDEN, S.; ROOZENBEEK, J.; COMPTON, J. Inoculating against fake news about COVID-19. **Frontiers in Psychology**, v. 11, p. 1-7, 2020. Disponível em: https://doi.org/10.3389/fpsyg.2020.566790. Acesso em: 4 abr. 2023.

VANLEHN, K. et al. **What's in a step?** Toward general, abstract representations of tutoring system log data. In: INTERNATIONAL CONFERENCE ON USER MODELING, 11., 2007, Corfu. Proceedings [...]. S.l.: Springer, 2007. p. 455-459.

VOOGT, J.; ROBLIN, N. A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. **Journal of Curriculum Studies**, v. 44, n. 3, p. 299-321, 2012. Disponível em: https://doi.org/10.1080/00220272.2012.668938. Acesso em: 4 abr. 2023.

WAINER, H. *et al*. **Computerized adaptive testing.** Nova York: Routledge, 2000. Disponível em: https://doi.org/10.4324/9781410605931. Acesso em: 4 abr. 2023.

WIEMAN, C.; ADAMS, W.; PERKINS, K. PhET: simulations that enhance learning. **Science**, v. 322, n. 5902, p. 682-683, 2008. Disponível em: https://doi.org/10.1126/science.1161948. Acesso em: 4 abr. 2023.

WIEMAN, C.; PRICE, A. Assessing complex problem-solving skills through the lens of decision making. In: FOSTER, N.; PIACENTINI, M. (eds.) **Innovating Assessments to Measure and Support Complex Skills**. Paris: OECD Publishing, 2023. Disponível em: <https://doi.org/10.1787/e5f3e341-en>.

WINSTONE, N. et al. Supporting learners' agentic engagement with feedback: a systematic review and a taxonomy of recipience processes. **Educational Psychologist**, v. 52, n. 1, p. 17-37, 2016. Disponível em: https://doi.org/10.1080/00461520.2016.1207538. Acesso em: 4 abr. 2023.

WOLF, S.; BRUSH, T.; SAYE, J. Using an information problem-solving model as a metacognitive scaffold for multimedia-supported information-based problems. **Journal of Research on Technology in Education**, v. 35, n. 3, p. 321-341, 2003. Disponível em: https://doi.org/10.1080/15391523.2003.10782389. Acesso em: 4 abr. 2023.

WOOD, D. Scaffolding, contingent tutoring and computer-supported learning. **International Journal of Artificial Intelligence in Education**, v. 12, n. 3, p. 280-293, 2001.

Publicação orginal em inglês pela:

OCDE
POLÍTICAS MELHORES
PARA UMA VIDA MELHOR

Apoio:

INSTITUTO
UNIBANCO