Avaliação Psicológica
ISSN 2175-3431 (versão on-line)

# Avaliação Psicológica e Avaliação da Aprendizagem em Larga Escala: Diretrizes para Pesquisadores

*Jacob Arie Laros*[1], *Josemberg Moura de Andrade*
Universidade de Brasília – UnB, Brasília-DF, Brasil

**RESUMO**

A avaliação psicológica (AP) e a avaliação educacional (AE) estão entre as contribuições mais importantes das ciências cognitivas e comportamentais para a sociedade atual, pois proporcionam importantes fontes de informações sobre os indivíduos e os grupos. O presente artigo objetivou apresentar diretrizes para os pesquisadores em relação a AP e a avaliação da aprendizagem em larga escala (AALEs). São discutidos os caminhos da AP em um mundo em crise sanitária devido a pandemia da Covid-19. Em termos das AALEs, são discutidos aspectos teóricos, metodológicos e analíticos que devem ser considerados pelos avaliadores e pesquisadores da área. Concluímos que a AP e AALE se relacionam na medida em que ambas cumprem a função social de identificar lacunas que merecem atenção, bem como aspectos funcionais que devem ser mantidos e incentivados. Outra importante característica é a exigência de constante aprimoramento técnico por parte dos avaliadores e pesquisadores.

*Palavras-chave*: testes psicológicos; delineamento; invariância; bifactor; teoria de resposta ao item.

**ABSTRACT – Large-Scale Psychological Assessments and Learning Assessments: Guidelines for Researchers**

Psychological assessment (PA) and educational assessment (EA) are among the most important contributions of cognitive and behavioral sciences to modern society. They provide important information about individuals and groups that are a part of the society. The aim of this article is to present guidelines for researchers regarding PA and large-scale learning assessments (LSLAs). In the context of LSLAs, we discuss theoretical, methodological and analytical aspects that must be considered by evaluators and researchers in the area. We conclude that PA and LSLAs are related to the extent that both fulfill the social function of identifying gaps that deserve attention, as well as functional aspects that must be maintained and encouraged. Another important characteristic is the requirement for constant technical improvement by both evaluators and researchers.

*Keywords*: psychological tests; research design, invariance; bifactor; item response theory.

**RESUMEN – Evaluación Psicológica y Evaluación del Aprendizaje a Gran Escala: Directrices para Investigadores**

La evaluación psicológica (EP) y la evaluación educativa (EE) se encuentran entre las contribuciones más importantes de las ciencias cognitivas y del comportamiento hechas a la sociedad contemporánea, ya que brindan importantes fuentes de informaciones sobre los individuos y grupos que forman parte de ella. Este artículo tiene como objetivo presentar a los investigadores directrices sobre EP y la evaluación del aprendizaje a gran escala (EAGE). Se discuten los caminos de la EP en un mundo en crisis sanitaria por la pandemia del Covid-19. Con respecto a EAGE, se discuten aspectos metodológicos y analíticos que deben ser considerados por evaluadores e investigadores del área. Concluimos que la EP y la EAGE están relacionadas en la medida en que ambas cumplen la función social de identificar brechas que merecen atención, así como aspectos funcionales que deben ser mantenidos y fomentados. Otra característica fundamental es la exigencia de una mejora técnica constante, tanto por parte de los evaluadores, como de los investigadores.

*Palabras clave*: tests psicológicos; delineamiento; invarianza; bifactor; teoría de respuesta al ítem.

A avaliação psicológica (AP) e a avaliação educacional (AE) estão entre as contribuições mais importantes das ciências cognitivas e comportamentais para a sociedade atual, pois proporcionam importantes fontes de informações sobre os indivíduos e os grupos que estes fazem parte. Existe ampla evidência documentando a utilidade de testes psicológicos e educacionais quando são construídos e utilizados de forma adequada. Testes bem construídos e que apresentem parâmetros psicométricos adequados para os propósitos pretendidos possuem o potencial de promover benefícios significativos para todas as partes envolvidas. O uso adequado desses instrumentos nos processos avaliativos resulta em melhores decisões sobre os indivíduos, grupos e sistemas educacionais do que resultaria sem seu uso (American Educational Research Association [AERA] et al., 2014).

No caso do Brasil, quando se fala em AP, especificamente, caminha-se para o aprimoramento dos testes desenvolvidos. Por exemplo, Bandeira et al. (2021) assinalaram que, a partir de uma visão comparativa de alguns

[1] Endereço para correspondência: Campus Universitário Darcy Ribeiro – ICC SUL. Instituto de Psicologia. META – Laboratório de Métodos e Técnicas de Avaliação. Asa Norte, 72910-000, Brasília, DF. Tel.: (61) 3107-6624. *E-mail*: jalaros@gmail.com

países, é possível afirmar que os avanços empreendidos na área da AP no Brasil, principalmente, após a criação do Sistema de Avaliação dos Testes Psicológicos (Satepsi) pelo Conselho Federal de Psicológica (CFP), são perceptíveis. Por meio da definição de critérios psicométricos mínimos mais rigorosos para aprovação de testes psicológicos (Andrade & Valentini, 2018), o Satepsi impactou de forma direta na área da AP ao diferenciar instrumentos que apresentavam embasamento teórico, metodológico e analítico daqueles que, simplesmente, não o apresentavam.

Outro importante avanço que destacamos foi a discussão sobre a justiça e proteção dos direitos humanos na AP, inserida na Resolução CFP nº 09/2018 (CFP, 2018) e, posteriormente, revogada pela Resolução CFP nº 31/2022 (CFP, 2022). Não podemos falar em avaliação, seja ela psicológica ou educacional, sem que os direitos dos indivíduos avaliados estejam totalmente assegurados. A avaliação deve ser entendida como um ato a favor do desenvolvimento humano, na medida em que identifica potencialidades que devem ser incentivadas, bem como afetos, crenças e comportamentos disfuncionais que, ao trazerem sofrimento psicológico para os indivíduos que os possuem, devem ser ressignificados.

Apesar dos avanços na área de AP (Bandeira et al., 2021), novas necessidades de aprimoramento são identificadas. Por exemplo, o contexto da pandemia da Covid-19 com a necessidade de isolamento social e todos os seus impactos em termos de saúde mental (Zanini, Peixoto, Andrade, Campos et al., 2021; Zanini, Peixoto, Andrade & Tramonte, 2021) apontou para lacunas que precisam ser preenchidas na AP. A necessidade de distanciamento social apontou para a necessidade de desenvolvimento de testes informatizados, principalmente a partir da testagem adaptativa computadorizada (*Computerized Adaptive Testing*) (AERA et al., 2014; Magis & Barrada, 2017; Peres, 2019; Wainer, 2015), bem como para a realização de estudos de equivalência de testes lápis e papel e testes de aplicação remota/*on-line* (testes nos quais avaliando/a e avaliador/a não estão no mesmo tempo e espaço) (ITC, 2005), além da construção e obtenção de evidências de validade de testes inéditos para aplicação remota/*on-line*. Tais aspectos requerem atenção tanto para a AP realizada no contexto micro, tal como em clínicas e organizações, quanto para AP realizadas no contexto de larga escala, tal como em concursos públicos. Neste último, a avaliação pode ser realizada simultaneamente em várias cidades com múltiplos locais de aplicação.

Paralelamente, no contexto das avaliações de aprendizagem em larga escala (AALEs), conhecidas em língua inglesa por *large-scale learning assessments* (LSLAs) (Unesco, 2019), o alvo principal da avaliação geralmente não é o indivíduo (nível individual) que responde ao teste, e sim uma unidade maior (nível grupal), tal como uma escola, município, rede escolar, unidade da federação ou mesmo país. Nessas avaliações, é comum os participantes receberem diferentes conjuntos de itens (diferentes blocos que formam um caderno), seguindo um plano de amostragem cuidadosamente balanceado (e.g., Blocos Incompletos Balanceados - BIB; Bekman, 2001). Dessa forma, os resultados de tais avaliações adquirem sentido quando agregados a partir de muitos indivíduos que respondem a diferentes amostras de itens. Tais avaliações podem não fornecer informações adequadas e suficientes para decisões válidas e confiáveis no nível individual, pois cada estudante pode responder a um caderno específico formado por diferentes blocos. Todavia, quando os dados são agregados, os resultados das avaliações educacionais em larga escala podem ser válidos e confiáveis para interpretações sobre o desempenho do nível grupal (AERA et al., 2014).

O interesse pelas AALEs tem aumentado substancialmente nas últimas décadas (Hernández-Torrano & Courtney, 2021; Unesco, 2019). Cada vez mais, novas habilidades estão sendo incluídas nas AALES no decorrer do tempo. Além de matemática e leitura, competências socioemocionais (Carias et al., no prelo; Primi, 2018), interesses profissionais (Ambiel et al., 2018), lócus de controle (Hwang, 2019), compreensão de conceitos e questões relacionadas à educação cívica e à cidadania estão sendo avaliadas (Unesco, 2019). Também, o número de países que participam de avaliações internacionais aumentou consideravelmente. Por exemplo, o número de países participantes no *Trends in International Mathematics and Science Study* (TIMSS) passou de 42 países em 1995 para 64 países em 2019 e, no caso do *Programme for International Student Assessment* (PISA), o número de participantes aumentou de 42 países em 2000-2001 para 79 países em 2018, juntamente com sete países adicionais envolvidos no PISA para o desenvolvimento (*PISA for Development*) (Raudonyte, 2019). Este último – *PISA for Development* – inclui países de renda média e baixa (Addey, 2016).

Conforme citam Hernández-Torrano e Courtney (2021), a pesquisa moderna sobre avaliação internacional em larga escala é um campo relativamente interdisciplinar e que se desenvolveu alicerçado em percursos históricos diferenciados. Os pesquisadores da área já abordaram uma ampla variedade de tópicos que incluem desde qualidade e equidade da educação, globalização, política educacional e mensuração, até autoeficácia do aluno, motivação e relacionamentos interpessoais. Todavia, ainda existem aspectos que, idealmente, devem continuar a evoluir nos próximos anos. Por exemplo, a escassez de pesquisas em países de renda média e baixa pode estar limitando o desenvolvimento da área. Estudos futuros devem examinar se a incorporação progressiva de territórios como a América do Sul, Sudeste Asiático, Médio Oriente e Norte da África (região MENA – *Middle East and North Africa*), nos programas de avaliações internacionais, impulsionará o desenvolvimento da referida área.

Não menos importante, aspectos metodológicos e analíticos assumem grande importância na operacionalização das AALEs. Pesquisas com grandes coletas de

dados, comumente, utilizam amostragem probabilística para obter amostras representativas. Tais conjuntos de dados são ferramentas valiosas para pesquisadores em todas as áreas da ciência. No entanto, muitos pesquisadores não possuem formação e preparação para utilização adequada desses recursos. A modelagem equivocada de dados complexos pode distorcer substancialmente os resultados da análise ocasionando erros de inferência e estimativas incorretas dos parâmetros dos dados (Osborne, 2011).

Diante do exposto, o presente artigo objetivou apresentar diretrizes para pesquisadores no que diz respeito a avaliação psicológica e avaliação educacional em larga escala. São discutidos os caminhos da AP em um mundo em crise sanitária devido a pandemia da Covid-19. Em termos das AALEs são discutidos aspectos teóricos, metodológicos e analíticos que devem ser considerados pelos avaliadores e pesquisadores da área. Entre tais aspectos são discutidos: (a) a importância da transparência dos cálculos dos escores das avaliações educacionais, (b) a necessidade de utilização de pesos amostrais para amostras complexas, (c) a necessidade de avaliação do pressuposto da unidimensionalidade, (d) a utilização de métodos adequados para análise de dados com estrutura hierárquica, (e) a importância dos modelos *bifactor* para instrumentos com fatores de segunda ordem e, por fim, (f) o uso adequado de métodos para tratamento de dados ausentes (*missing values*).

## Novos Caminhos para a Avaliação Psicológica no Brasil

A AP pode ser entendida como um processo planejado de investigação dos fenômenos psicológicos. Tal processo é composto por métodos, técnicas e instrumentos que objetivam prover informações para a tomada de decisão consciente, no âmbito individual, grupal ou institucional (CFP, 2018). Ainda, considerando o caráter dinâmico e integrativo da AP (Andrade & Sales, 2017), esse processo deve servir como instrumento para atuar sobre os indivíduos e grupos, modificando possíveis visões cristalizadas que venham trazer prejuízos para estes, sendo necessário levar em consideração os diferentes momentos do processo (Amorim-Gaudêncio et al., 2013).

No processo de AP, compete a(o) psicóloga(o) planejar e realizar o processo avaliativo com base em aspectos técnicos e teóricos. A definição do número de sessões para a sua realização e de quais instrumentos/técnicas de avaliação devem ser utilizados, será baseada em alguns critérios, entre eles podemos citar: (a) contexto no qual a avaliação psicológica se insere; (b) propósitos da AP; (c) construtos psicológicos a serem investigados; (d) adequação das características dos instrumentos/técnicas aos indivíduos avaliados; e (e) condições técnicas, metodológicas e operacionais do instrumento de avaliação (Reppold et al., 2019). Nesse contexto, inovações metodológicas, como alteração do formato de aplicação do teste, alterações no cômputo dos escores etc., não são bem-vindas, a não ser que seus efeitos sejam previamente pesquisados e submetidos para avaliação do Satepsi.

Os(as) profissionais que atuam com testagem e avaliação em larga escala, também, devem garantir acomodações padronizadas para todos(as) os(as) sujeitos avaliados(as). A padronização das condições de testagem deve ser realmente garantida a fim de evitar tratamento injusto e discriminação. Variações não podem ser realizadas nem mesmo com uma suposta intenção de manter a comparabilidade dos escores. Se inovações metodológicas são realizadas, essa comparabilidade pode ser comprometida e o teste pode não medir os mesmos construtos para todos os sujeitos que participam da avaliação (AERA et al., 2014).

Ainda no que se refere à padronização da aplicação dos instrumentos psicológicos e educacionais, a excepcionalidade do contexto da pandemia da Covid-19 trouxe novos desafios e demandas que precisam ser atendidas. Com as medidas de isolamento social para contenção da pandemia (Zanini, Peixoto, Andrade, Campos et al., 2021; Zanini, Peixoto, Andrade & Tramonte, 2021), a necessidade de instrumentos psicológicos para aplicação remota via *internet* ficou mais evidenciada. Importante destacar que a simples transposição do formato de aplicação do teste de lápis e papel para o formato remoto/*on-line* não deve ser realizado. No caso de adaptação de um instrumento já disponibilizado no mercado para o formato remoto/*on-line*, uma pesquisa psicométrica prévia – sem fins avaliativos dos sujeitos – deve ser realizada a fim de avaliar a equivalência dos escores (Lakens et al., 2018) dos dois formatos de aplicação (lápis e papel versus remoto/on-line).

Os testes de equivalência estão se tornando cada vez mais populares em muitas áreas das ciências humanas e devem ser aplicados sempre que o objetivo do estudo não seja mostrar diferenças, mas, sim, concluir pela similaridade (Meyners, 2012). Quando a invariância de um instrumento não é assegurada, isso significa dizer que os grupos (e.g., homens *versus* mulheres, trabalhadores de organizações públicas *versus* trabalhadores de organizações privadas) interpretaram e/ou responderam diferentemente os itens ou estímulos do teste. Nesse caso, tais grupos não podem ser comparados (Schoot et al., 2012).

Assim, identificada a equivalência dos dois formatos de aplicação, as mesmas tabelas normativas do teste formato lápis e papel podem ser utilizadas para o formato remoto/*on-line*. Por outro lado, se a equivalência não for comprovada, novos estudos psicométricos de evidências de validade, estimativas de fidedignidade, análise dos itens e interpretação dos escores devem ser realizados para o instrumento cujo formato de aplicação é remoto/*on-line*. No caso do desenvolvimento de um novo teste no formato de aplicação remoto/*on-line*, os mesmos estudos psicométricos de um teste no formato de aplicação lápis e papel devem ser realizados. Por fim, quando

o instrumento é utilizado no contexto da AP, o estudo de equivalência também deve ser apresentado ao Satepsi/CFP para avaliação da sua adequação.

Nesse cenário, faz-se necessário, também, destacar que o formato de aplicação informatizada (via computador) não é equivalente ao formato de aplicação remoto/*on-line* (CFP, 2019). No primeiro, a aplicação é mediada pelo computador, ou seja, os itens ou estímulos do teste foram transpostos para a tela do computador. No segundo caso – aplicação remota/*on-line* – psicólogo(a) e avaliando(a) estão a distância, ou seja, não estão no mesmo espaço. No caso da aplicação remota/*on-line*, na qual psicólogo(a) e avaliando(a) não estão no mesmo ambiente, faz-se necessário atentar para variáveis que possam comprometer a qualidade técnica do processo de AP. Certamente, isso é um grande desafio no caso das aplicações remotas/*on-line*. Tais especificações devem conter informações sobre limites de tempo, procedimentos de acomodação dos(as) avaliados(as), instruções padronizadas, procedimentos de monitoramento e garantia da segurança do teste, descrição do *hardware* (*web* câmera, tamanho mínimo e resolução de monitor requeridos), informações sobre conectividade, como velocidade mínima da *internet*, entre outros aspectos (AERA et al., 2014).

Importante também destacar que uma série de técnicas pode ser considerada para fins de avaliação da equivalência entre os testes de aplicação lápis e papel e de aplicação remota/*on-line*, sendo necessário justificar o emprego de cada uma delas. Entre tais possibilidades, destaca-se: análise fatorial confirmatória multigrupo (Damásio, 2013), funcionamento diferencial dos itens (*differential item functioning* – DIF) (Martinková et al., 2017; Russel & Kaplan, 2021), Factor Mixture Models (Lubke & Muthén, 2005) e *multiple-indicators multiple-causes* (MIMIC) (Kim et al., 2012).

Por fim, quando se fala em futuro da avaliação e testagem informatizada, não se pode deixar de mencionar a testagem adaptativa computadorizada (*computerized adaptive testing* – CAT). O desenvolvimento de uma CAT requer conhecimentos psicométricos avançados (Magis & Barrada, 2017; Wainer, 2015). A CAT é um tipo de teste informatizado que apresenta itens adequados ao nível de habilidade do participante. Ao longo da aplicação, o algoritmo da CAT seleciona itens para cada respondente, de acordo com a resposta dada (acerto ou erro) ao item anterior, sendo, assim, mais eficaz para situar o indivíduo no *continuum* do traço latente em comparação com os testes tradicionais. Comumente, a CAT é mais curta e rápida, tornando-se mais atrativa para o(a) respondente, além de eliminar possíveis diferenças entre aplicadores. Considerada como uma das técnicas mais modernas para o desenvolvimento de testes psicológicos da atualidade, a utilização da CAT é crescente em todo o mundo, embora no Brasil sua aplicação ainda seja incipiente (Peres, 2019; Sales et al., 2018; Santos, 2015; Veldkamp & Sluijter, 2019). Após considerações sobre a AP, são discutidos, a seguir, alguns aspectos metodológicos e analíticos que devem ser considerados pelos avaliadores e pesquisadores na área das AALEs.

## A Importância da Transparência dos Cálculos dos Escores das AALE

Quando se fala em AALE, a transparência da pontuação ou nota é um dos principais requisitos para a aceitação de uma avaliação por todas as partes interessadas (*stakeholders*), como alunos, professores e pais. Isso é provavelmente a razão pelo qual a utilização dos escores brutos totais baseados no número de itens corretos ainda é predominantemente preferido na educação (Glas, 2019). Os escores baseados em número de itens corretos apresentam as vantagens de serem fáceis de calcular e de serem entendidos pelo público-alvo. A lógica desse tipo de pontuação é evidente: quanto maior o número de acertos, maior a proficiência do(a) aluno(a).

A opção de utilizar as estimativas de proficiência (*teta* – θ) modeladas a partir da Teoria de Resposta ao Item (TRI) (Andrade et al., 2010; Andrade et al., 2021; Baker & Kim, 2004), por exemplo, é mais difícil de ser explicada para o público leigo. No entanto, esforços de tradução dos conceitos para esse público devem ser realizados em função das vantagens (e.g., os parâmetros dos itens são independentes dos sujeitos, estimativas de erro para cada nível de proficiência etc.) que tais métodos proporcionam. Nessa direção, Glas (2019), por exemplo, desenvolveu um método para combinar a pontuação baseada no número de itens corretos com a pontuação usada nos modelos logísticos de (2PLM) e de três parâmetros (3PLM) da TRI. Wiberg et al. (2019), por sua vez, desenvolveram os chamados escores ótimos baseados na TRI não paramétrica como uma alternativa aos escores totais e os escores usados na TRI paramétrica.

## A Necessidade de Utilização de Pesos Amostrais para Amostras Complexas

Grande parte das AALEs utiliza amostragem complexa em vez de amostragem randômica simples (Osborne, 2011). Nos estudos do tipo *survey* verificou-se, muitas vezes, dependência entre observações no nível micro, devido ao uso frequente de amostragem de grupos por áreas geográficas ou por outros tipos de *clusters* (Hox, 2010). Respondentes de uma mesma área geográfica tendem a ter mais semelhanças entre eles do que respondentes de áreas geográficas diferentes. A homogeneidade entre as observações conduz a estimativas incorretas dos erros padrão das variáveis. Esse tipo de erro é conhecido como o "efeito de delineamento" ou "*design effect*" em inglês (Kish, 1987).

Um procedimento de correção usualmente aplicado consiste em calcular os erros-padrão por métodos de análise tradicionais, estimar a correlação intraclasse entre os respondentes dentro dos *clusters* e, por fim, empregar

uma fórmula de correção para os erros-padrão. Os efeitos de delineamento da amostragem devem ser contabilizados ou o pesquisador corre o risco não apenas de estimar mal os efeitos, mas de cometer erros do Tipo I (Hox 2010; Snijders & Bosker, 2012).

Segundo Osborne (2011), há dois tipos de amostras que precisam de correções na análise de dados, são elas: 1. estudos que empregam técnicas avançadas de amostragem (ou amostras que contém dados ausentes que precisam ser contabilizados) e que não fazem uso de pesos amostrais, e 2. amostras que violam os pressupostos de independência das observações, potencialmente levando a uma estimativa errônea significativa dos níveis de significância em testes estatísticos inferenciais. Essas amostras não ponderadas podem estimar erroneamente, não apenas as suposições dos parâmetros, mas também os erros-padrão. Isso ocorre porque a amostra não ponderada não é representativa da população como um todo e contém muitas excentricidades, como sub-representação de populações.

Na maioria dos pacotes estatísticos modernos, os pesos amostrais podem ser facilmente aplicados para um conjunto de dados. O problema é que a aplicação de pesos aumenta drasticamente o tamanho da amostra para aproximadamente o tamanho da população. A solução é usar pesos amostrais normatizados, de modo que a amostra ponderada tenha o mesmo número de participantes que a original não ponderada. O uso adequado de pesos amostrais resulta em uma melhor estimativa de parâmetros e dos erros-padrão. Embora tal procedimento exija um esforço extra para modelar adequadamente as amostras complexas, é um passo necessário para termos confiança nos resultados decorrentes das análises (Osborne, 2011).

## A Necessidade de Avaliação do Pressuposto da Unidimensionalidade

Estudos mostram que a verificação do pressuposto da unidimensionalidade é de suma importância sempre que os modelos da TRI unidimensional são utilizados, a fim de que a propriedade desejável da invariância dos parâmetros dos itens (dificuldade, discriminação e acerto ao acaso) possa se manifestar (Conde & Laros, 2007; Zopluoglu & Davenport Junior., 2017). Entre as consequências negativas da violação do pressuposto da unidimensionalidade, está o comprometimento da validade do instrumento (Kirisci et al., 2001).

Uma das principais limitações da Teoria Clássica dos Testes (TCT) é que as características dos(as) examinandos(as) e as características dos testes não podem ser consideradas independentes, pois uma só pode ser interpretada no contexto da outra. A TRI, entretanto, assume a propriedade de invariância dos parâmetros, considerada como a sua maior distinção da TCT. Esse princípio afirma que, quando um conjunto total de itens tem um bom ajuste a um modelo da TRI, os parâmetros psicométricos dos itens não dependem da habilidade dos examinandos e tal habilidade pode ser estimada independentemente da dificuldade do teste utilizado (Baker & Kim, 2004). Em pesquisa realizada por Conde e Laros (2007), foi verificada relação negativa entre a propriedade de invariância da TRI e o grau de falta da unidimensionalidade de uma medida. Ainda, os autores verificaram que a exclusão dos itens que apresentam carga fatorial negativa ou muito fraca no fator dominante teve um efeito positivo na propriedade de invariância da TRI.

## Utilização de Métodos Adequados para Análise de Dados com Estrutura Hierárquica

Nas avaliações dos escores brutos de testes educacionais, como, por exemplo, o Sistema de Avaliação da Educação Básica (SAEB) e o Exame Nacional do Ensino Médio (ENEM), o contexto escolar precisa ser sempre levado em consideração (Alves & Soares, 2013). O contexto escolar também chamado de contexto de aprendizagem ou em inglês "*learning environment*" (Mullis et al., 2021) é de suma importância, uma vez que a literatura nacional (Andrade & Laros, 2007; Laros et al., 2010; 2012; Vinha et al., 2016) e a literatura internacional (Agasisti & Cordero, 2017) apresentam uma grande quantidade de estudos que evidenciam a associação entre o contexto de aprendizagem e o desempenho escolar dos alunos. Nesse sentido, Mullis et al. (2021) afirmaram que alunos(as) com ambientes de aprendizagem mais favoráveis apresentam consistentemente maiores chances de melhores desempenhos em matemática e ciências do que aqueles que não o têm. Entre as variáveis do contexto de aprendizagem, o nível socioeconômico (NSE) do(a) aluno(a) é a variável que tem o maior impacto no desempenho escolar.

Nesses estudos provenientes da educação, da psicologia e de áreas afins, a análise de regressão múltipla constitui uma das técnicas de análise de dados frequentemente utilizadas em pesquisas que adotam métodos quantitativos para predição e explicação de fenômenos. O problema central com o uso dessa técnica é que, em muitas ocasiões, um dos pressupostos centrais - a independência das observações - é violada (Puente-Palacios & Laros, 2009).

Os dados coletados nas ciências sociais e humanas são frequentemente de indivíduos agrupados em *clusters*, também conhecidos como conglomerados (e.g., estudantes em escolas, trabalhadores em empresas etc.). Considerando tais agrupamentos, é provável que esses indivíduos compartilhem atributos similares em decorrência do contexto que lhes é comum. Como consequência da dependência das observações ou atributos mensurados, ocorre a subestimação dos erros-padrão dos coeficientes da regressão. Considerando essa problemática, a análise multinível é uma alternativa para a regressão múltipla, que leva em consideração a similaridade dos grupos. Essa técnica é um tipo de análise de regressão que

contempla simultaneamente múltiplos níveis de agregação, tornando corretos os erros-padrão, os intervalos de confiança e os testes de hipóteses (Ferrão, 2003).

O modelo de regressão multinível incorpora a estrutura hierárquica dos dados, tratando o intercepto e os coeficientes de inclinação como variáveis aleatórias. Considerar os coeficientes como variáveis aleatórias significa que cada unidade do segundo nível pode ter seu próprio valor. Assim, a opção por esses modelos é justificada na medida em que permite corrigir erros provenientes da utilização de metodologia inadequada e, também, por permitir utilização mais eficiente dos dados (Snijders & Bosker, 2012).

Estudos que objetivam identificar variáveis preditoras do desempenho acadêmicos ou causas da repetência escolar comumente têm utilizado análise de regressão multinível (Andrade & Laros, 2007; Laros et al., 2010; 2012; Vinha et al., 2016). Por exemplo, Ferrão, Costa e Matos (2017), ao utilizarem dados do PISA 2012 e a técnica de análise de regressão logística multinível, verificaram que maior nível socioeconômico do(a) aluno(a) está associado a menor chance de repetência.

Importante destacar que, nesses estudos hierárquicos, quando as informações em nível grupal são obtidas pela agregação dos resultados de testes realizados por indivíduos – procedimento usual nas análises de regressão multinível – evidências de validade e estimativas de fidedignidade devem ser relatadas para o nível de agregação em que os resultados são descritos. As pontuações não devem ser relatadas para indivíduos ou grupos sem evidências apropriadas para apoiar as interpretações para os usos pretendidos (AERA et al., 2014).

## Importância dos Modelos Bifactor para Instrumentos com Fatores de Segunda Ordem

Quando um instrumento de AP ou AE é composto de vários fatores correlacionados, é recomendado analisar a existência de um fator de segunda ordem a partir da utilização de análise fatorial hierárquica (Canivez & Watkins, 2010; Thompson, 2006). No entanto, a interpretação de fatores de ordem superior apresenta algumas dificuldades especiais. Fatores são abstrações de variáveis mensuradas. Fatores de segunda ordem, por sua vez, são originados a partir dos fatores de primeira ordem e, então, podem ser considerados abstrações de abstrações ainda mais distantes das variáveis mensuradas. De alguma forma, gostaríamos de interpretar os fatores de segunda ordem em termos das variáveis mensuradas, e não como uma manifestação dos fatores das variáveis mensuradas (Thompson, 2004).

Nesse contexto, Schmid e Leiman (1957) propuseram um método eficiente para expressar os fatores de primeira ordem em termos de variáveis medidas, mas também remover toda a variância nos fatores de primeira ordem que está presente nos fatores de segunda ordem. Isso permite que o pesquisador determine qual variância, se houver, é exclusiva para um determinado nível de análise ou perspectiva. Assim, Schmid e Leiman (1957) sugeriram que tal solução não apenas preserva as características de interpretação desejadas da rotação oblíqua, mas também revela a estrutura hierárquica das variáveis.

Especificamente, os modelos *bifactor* ganharam popularidade na comunidade científica nas últimas décadas. Tais modelos consideram um fator geral que explica todas as variáveis observáveis (escores de testes ou itens, por exemplo), bem como fatores específicos que explicam grupos determinados de variáveis observáveis. Nos modelos *bifactor*, os fatores são ortogonais, ou seja, as correlações entre as dimensões são fixadas em zero. Essa característica permite avaliações sobre a relevância e a precisão de uma variável latente após controlar o efeito explicativo das demais variáveis (Canivez, 2016; Reise et al., 2013; Rios & Wells, 2014; Valentini et al., 2015).

Por exemplo, Gomes et al. (2018) objetivaram estimar a fidedignidade do ENEM por meio da utilização dos escores de acerto e erro dos 180 itens do teste, aplicando a modelagem bifatorial na matriz de correlação desses itens, bem como calculando a fidedignidade composta dos domínios, a partir dos betas obtidos por essa análise. Os autores concluíram que o modelo bifatorial e a confiabilidade composta, combinados em uma mesma análise, são uma estratégia apropriada e bastante efetiva para se investigar a fidedignidade de escores de variáveis latentes, como é o caso do ENEM.

## Uso Adequado de Métodos para Tratamento de Dados Ausentes (*Missing Values*)

Nas avaliações, a ocorrência de dados ausentes (*missing values*) é comum. Para reduzir o impacto de dados ausentes nas análises estatísticas, o uso de métodos adequados torna-se fundamental (Vinha & Laros, 2018). Segundo Rubin (1987), os valores ausentes são gerados por três mecanismos distintos que relacionam a propensão de ausência aos dados observados: 1. valores ausentes completamente ao acaso (MCAR, *Missing Completely at Random*); 2. valores ausentes ao acaso (MAR, *Missing at Random*) e 3. valores ausentes não ao acaso (MNAR, *Missing not at Random*).

O MCAR é identificado quando a ocorrência não está relacionada a qualquer variável observada no estudo ou à própria variável que apresenta os valores faltantes. Os dados faltantes MAR, por sua vez, são identificados quando a ocorrência está relacionada aos valores observados de outras variáveis, mas independe do valor da variável em questão. Por fim, quando a ocorrência dos dados faltantes está relacionada aos valores da própria variável analisada os dados são identificados como MNAR. Essa classificação de dados ausentes proposta por Rubin (1987) está diretamente relacionada ao impacto da ausência de informação e à escolha da abordagem mais apropriada para a análise dos dados.

Os que menos influenciam os resultados das análises estatísticas são os dados ausentes do tipo MCAR, uma vez que a amostra de valores completos pode ser vista como representativa da população. Quando os dados faltantes são do tipo MAR, a ausência pode ser considerada não ignorável, pois se faz necessária a modelagem adicional do mecanismo de ausência de dados no processo de estimação. Ainda, os dados MNAR também são chamados de não ignoráveis, dado que o mecanismo gerador de ausência deve ser modelado para que sejam obtidas boas estimativas dos parâmetros de interesse (Vinha & Laros, 2018).

Abordagens clássicas, como ignorar valores ausentes ou tratá-los como respostas incorretas, são atualmente aplicadas em estudos de grande escala, enquanto abordagens recentes baseadas em modelos que podem explicar a falta de respostas não ignoráveis ainda são incipientes no Brasil. As estimativas de parâmetros de item e da pessoa demonstraram ser tendenciosas para abordagens clássicas quando os dados ausentes são do tipo MNAR (Köhler et al., 2017).

O impacto da ausência de dados depende do tipo de ausência. No estudo realizado por Vinha e Laros (2018), os desvios observados nas estimativas dos coeficientes da regressão foram pequenos quando 10% dos dados da variável resposta estavam ausentes (menor para MCAR e maior para MNAR) e aumentou quando o percentual de ausência foi maior. A substituição simples pela média mostrou resultados insatisfatórios em todos os cenários. O procedimento *listwise deletion* (LD), por sua vez, apresentou resultados semelhantes aos procedimentos baseados na máxima verossimilhança (MV) e imputação múltipla (IM) nos cenários simulados e, ainda, as variáveis auxiliares foram importantes para redução dos desvios. Os autores recomendaram a utilização dos procedimentos baseados na MV e IM. Tais procedimentos além de apresentarem bons resultados na estimação dos coeficientes, são também mais apropriados para a estimação de erros padrão, o que resulta em testes de hipóteses mais confiáveis. Por fim, o método da imputação simples pela média também deve ser evitado.

## Considerações Finais

Ao longo do presente artigo apresentamos diretrizes para pesquisadores no que se refere à AP e AALE. Em relação à AP, especificamente, indicamos os estudos de equivalência das versões dos testes lápis e papel e versões remotas/*on-line* como necessários em decorrência da pandemia da Covid-19. O distanciamento e isolamento social parecem ter intensificado a necessidade da elaboração e obtenção de parâmetros psicométricos para testes remotos/*on-line*. Nesse contexto, os estudos de equivalência devem ser utilizados a fim de verificar se o tipo de aplicação do instrumento não ocasiona interpretações diferentes dos itens ou estímulos.

Em relação à AALE, discutimos uma série de aspectos metodológicos e analíticos. Longe de serem inovadores, tais aspectos precisam ser considerados na prática de pesquisadores e avaliadores educacionais. Aspectos como necessidade de utilização de pesos amostrais, consideração da estrutura hierárquica dos dados e avaliação da dimensionalidade das provas, entre outros, devem ser considerados em conjunto. Por que trabalhamos com AALE? Porque o objetivo central da educação é que os(as) estudantes aprendam igualmente bem em todas as escolas de um país, mesmo que essas escolas sejam diferentes em termos de composição dos estudantes (Steinmann & Olsen, 2022). Além dos escores provenientes das AALEs serem utilizados como critérios para admissões em instituições de ensino superior, tais avaliações assumem grande importância na medida em que identificam lacunas da aprendizagem e sugerem caminhos para uma avaliação formativa. Por avaliação formativa entende-se o uso diagnóstico da avaliação para fornecer *feedback* a professores e alunos(as) ao longo dos estudos (Boston, 2002).

Concluímos que a AP e AALE se relacionam na medida em que ambas cumprem a função social de identificar lacunas que merecem atenção, bem como identificar aspectos funcionais que devem ser mantidos e incentivados. Outra importante característica é a exigência de constante aprimoramento técnico por parte dos avaliadores e pesquisadores.

Por fim, quando falamos em necessidade de aprimoramento, alguns temas emergentes não foram aqui tratados. Por exemplo, como recurso analítico, a utilização da abordagem *Bayesiana* na Psicologia apresenta vantagens em relação à inferência frequentista. Entre tais vantagens, citamos a possibilidade de incorporar conhecimento anterior, bem como a capacidade de monitorar e atualizar essas evidências à medida que os dados são obtidos (Franco & Andrade, 2021; Wagenmakers et al., 2018). Ainda, os métodos de inteligência artificial como *deep learning* e *machine learning* buscam, automaticamente, descobrir padrões nos dados, além de criar representações úteis que permitam a previsão de uma variável relevante (Primi, 2018). Tais avanços são importantes para coleta de grandes fluxos de dados e, certamente, permitirão grandes avanços na AP e AALE.

**Agradecimentos**

Não há menções.

**Financiamento**

O segundo autor agradece ao apoio concedido pela Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF).

**Contribuições dos autores**

Declaramos que todos os autores participaram da elaboração do manuscrito. Especificamente, os autores Jacob Arie Laros e Josemberg Moura de Andrade

participaram da escrita inicial do estudo – conceituação, investigação, visualização, análise de dados, escrita final do trabalho – revisão e edição.

**Conflito de interesses**

Os autores declaram que não há conflitos de interesse.

## Referências

Addey, C. (2016). O PISA para o desenvolvimento e o sacrifício de dados com relevância política. *Educação & Sociedade, 37*(136), 685-706. https://doi.org/10.1590/ES0101-73302016166001

Agasisti, T., & Cordero, J. M. (2017). The determinants of repetition rates in Europe: Early skills or subsequent parents' help? *Journal of Policy Modeling, 39*(1), 129-146. https://doi.org/10.1016/j.jpolmod.2016.07.002.

Alves, M. T. G., & Soares, J. F. (2013). Contexto escolar e indicadores educacionais: Condições desiguais para a efetivação de uma política de avaliação educacional. *Educação e Pesquisa, 39*(1), 177-194. https://doi.org/10.1590/S1517-97022013000100012

Ambiel, R. A. M., Hauck-Filho, N., Barros, L. de O., Martins, G. H., Abrahams, L., & De Fruyt, F. (2018). 18REST: A short RIASEC-interest measure for large-scale educational and vocational assessment. *Psicologia: Reflexão e Crítica, 31*(6). https://doi.org/10.1186/s41155-018-0086-z

American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Amorim-Gaudêncio, C., Andrade, J. M. & Gouveia, V. V. (2013). Avaliação psicológica na atualidade: Processo, metodologia e área de aplicação. In N. T. Alves, J. M. Andrade, I. F. Rodrigues, & J. B. Costa (Eds.). *Psicologia: Reflexões para ensino, pesquisa e extensão* (pp. 181-209). Editora Universitária UFPB.

Andrade, J. M., & Laros, J. A. (2007). Fatores associados ao desempenho escolar: Estudo multinível com dados do SAEB/2001. *Psicologia: Teoria e Pesquisa, 23*(1), 33-42. https://doi.org/10.1590/S0102-37722007000100005

Andrade, J. M., Laros, J. A., & Gouveia, V. V. (2010). O uso de Teoria de Resposta ao Item em avaliações educacionais: Diretrizes para pesquisadores. *Avaliação Psicológica, 9*(3), 421-435.

Andrade, J. M., Laros, J. A., & Lima, K. S. (2021). Teoria de Resposta ao Item Paramétrica e Não Paramétrica. Em C. Faiad, M.N. Batista, & R. Primi (Orgs.), *Tutoriais em análise de dados aplicada à psicometria* (pp. 183-204). Editora Vozes.

Andrade, J. M., & Sales, H. F. S. (2017). A diferenciação entre avaliação psicológica e testagem psicológica: questões emergentes. In M. R. C. Lins & J. C. Borsa (Eds.), *Avaliação psicológica: Aspectos teóricos e práticos* (pp. 9-22). Editora Vozes.

Andrade, J. M., & Valentini, F. (2018). Diretrizes para a construção de testes psicológicos: A resolução CFP nº 009/2018 em destaque. *Psicologia: Ciência e Profissão, 38*(spe), 28-39. https://doi.org/10.1590/1982-3703000208890

Baker, F. B., & Kim, S. (2004). Item response theory: Parameter estimation techniques. Marcel Dekker.

Bandeira, D. R., Andrade, J. M., & Peixoto, E. M. (2021). O uso de testes psicológicos: formação, avaliação e critérios de restrição. *Psicologia: Ciência e Profissão, 41*(spe). https://doi.org/10.1590/1982-3703003252970

Bekman, R. M. (2001). Aplicação dos blocos incompletos balanceados na teoria de resposta ao item. *Estudos em Avaliação Educacional, 24*, 119-138. https://doi.org/10.18222/eae02420012202

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research, and Evaluation, 8*(Article 9). https://doi.org/10.7275/kmcq-dj31

Carias, I. A., Gondim, S. M. G., Andrade, J. M., & Brantes, C. dos A. A. (no prelo). Medida de competências socioemocionais de docentes de ensino fundamental: Evidências de validade. *Avaliação Psicológica.*

Canivez, G. L. (2016). Bifactor modeling in construct validation of multifactorial tests: Implications for understanding multidimensional constructs and test interpretation. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction* (Vol. 3, pp. 247-271). Hogrefe.

Canivez, G. L., & Watkins, M. W. (2010). Exploratory and higher-order factor analyses of the Wechsler Adult Intelligence Scale-Fourth Edition (WAIS-IV): Adolescent subsample. *School Psychology Quarterly, 25*(4), 223-235. https://doi.org/10.1037/a0022046

Conselho Federal de Psicologia [CFP] (2019). *Nota técnica nº 7/2019/GTEC/CG.* CFP.

Conselho Federal de Psicologia [CFP] (2018). *Resolução nº 009, de 25 de abril de 2018.* CFP.

Conde, F. N., & Laros, J. A. (2007). Unidimensionalidade e a propriedade de invariância das estimativas da habilidade pela TRI. *Avaliação Psicológica, 6*(2), 205-215.

Damásio, B. F. (2013). Contribuições da análise fatorial confirmatória multigrupo (AFCMG) na avaliação de invariância de instrumentos psicométricos. *Psico-USF, 18*(2), 211-220. https://doi.org/10.1590/S1413-82712013000200005

Ferrão, M. E. (2003). *Introdução aos modelos de regressão multinível em educação.* Komedi.

Ferrão, M. E., Costa, P. M., & Matos, D. A. S. (2017). The relevance of the school socioeconomic composition and school proportion of repeaters on grade repetition in Brazil: A multilevel logistic model of PISA 2012. *Large-Scale Assessments in Education, 5*(7), 1-13. https://doi.org/ 10.1186/s40536-017-0036-8

Franco, V. R., & Andrade, J. M. (2021). Aplicações da psicometria bayesiana: Do básico ao avançado. Em C. Faiad, M. N. Batista, & R. Primi (Orgs.), *Tutoriais em análise de dados aplicada à psicometria* (pp. 225-245). Editora Vozes.

Glas, C. A. W. (2019). Reliability issues in high-stakes educational tests. In B. P. Veldkamp & C. Sluijter (Eds.). *Theoretical and practical advances in computer-based educational measurement* (pp. 213-230). Springer. https://doi.org/10.1007/978-3-030-18480-3

Gomes, C. M. A., Golino, H. F., & Peres, A. J. S. (2018). Análise de fidedignidade composta dos escores do ENEM por meio da análise fatorial de itens. *European Journal of Education, 5*(8), 331-344. https://doi.org/10.5281/zenodo.2527904

Hernández-Torrano, D., & Courtney, M. G. R. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-Scale Assessments in Education, 9*(17), 1-33. https://doi.org/10.1186/s40536-021-00109-1

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd edition). Routledge.
Hwang, J. (2019). Relationships among locus of control, learned helplessness, and mathematical literacy in PISA 2012: Focus on Korea and Finland. *Large-Scale Assessments in Education,* 7(4), 1-19. https://doi.org/10.1186/s40536-019-0072-7
International Test Commission [ITC] (2005). *International guidelines on computer-based and internet delivered testing* [www.intestcom.org]. ITC.
Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC: Likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement, 72*(3), 469-492. https://doi.org/10.1177/0013164411427395
Kirisci, L., Hsu, T. & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*(2), 146-162. https://doi.org/10.1177/01466210122031975
Kish, L. (1987). *Statistical design for research.* Wiley.
Köhler, C., Pohl, S., & Carstensen, C. H. (2017). Dealing with item nonresponse in large-scale cognitive assessments: The impact of missing data methods on estimated explanatory relationships. *Journal of Educational Measurement, 54*(4), 397-419. https://doi.org/10.1111/jedm.12154
Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259-269. https://doi.org/10.1177/2515245918770963
Laros, J. A., Marciano, J. L. P., & Andrade, J. M. (2010). Fatores que afetam o desempenho na prova de Matemática do SAEB: Um estudo multinível. *Avaliação Psicológica, 9*(2), 173-186.
Laros, J. A., Marciano, J. L. P., & Andrade, J. M. (2012). Fatores associados ao desempenho escolar em Português: Um estudo multinível por regiões. *Ensaio: Avaliação e Políticas Públicas em Educação, 20*(77), 1-9. https://doi.org/10.1590/S0104-40362012000400002
Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*(1), 21–39. https://doi.org/10.1037/1082-989X.10.1.21
Magis, D., & Barrada, J. R. (2017). Computerized adaptive testing with R: Recent updates of the package catR. *Journal of Statistical Software, 76*, 1-19. https://doi.org/10.18637/jss.v076.c01
Martinková, P., Drabinová, A., Liaw, Y., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE – Life Sciences Education, 16*(2). https://doi.org/ 10.1187/cbe.16-10-0307. PMID: 28572182; PMCID: PMC5459266.
Meyners, M. (2012). Equivalence tests – a review. *Food Quality and Preference, 26*(2), 231-245. https://doi.org/10.1016/j.foodqual.2012.05.003.
Mullis, I. V. S., Martin, M. O., & Von Davier, M. (Eds.) (2021). TIMSS 2023 assessment frameworks. TIMSS & PIRLS International Study Center / International Association for the Evaluation of Educational Achievement (IEA).
Osborne, J. (2011). Best practices in using large, complex samples: The importance of using appropriate weights and design effect compensation. *Practical Assessment, Research, and Evaluation, 16*(Article 12). https://doi.org/10.7275/2kyg-m659
Peres, A. J. S. (2019). Testagem adaptativa por computador (CAT): Aspectos conceituais e um panorama da produção brasileira. *Revista Examen, 3*(3), 66-85.
Primi, R. (2018). Avaliação psicológica no Século XXI: De onde viemos e para onde vamos. *Psicologia: Ciência e Profissão, 38*(spe), 87-97. https://doi.org/10.1590/1982-3703000209814
Puente-Palacios, K. E., & Laros, J. A. (2009). Análise multinível: Contribuições para estudos sobre efeito do contexto no comportamento individual. *Revista Estudos de Psicologia, 26*(3), 349-362. https://doi.org/10.1590/S0103-166X2009000300008
Raudonyte, I., (2019). *Use of learning assessment data in education policy-making. IIEP-UNESCO Working papers*. International Institute for Educational Planning. http://www.unesco.org/open-access/terms-use-ccbysa-en
Reise, S. P., Scheines R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*(1), 5-26. https://doi.org/10.1177/0013164412449831
Reppold, C. T., Zanini, D. S., & Noronha, A. P. P. (2019). O que é avaliação psicológica. Em M. N. Baptista, M. Muniz, C. T. Reppold, C. H. S. S. Nunes, L. F. Carvalho, R. Primi, A. P. P. Noronha, A. G. Seabra, S. M. Wechsler, C. S. Hutz; & L. Pasquali (Orgs.), *Compêndio de Avaliação Psicológica* (pp. 15-28). 1ª edição. Vozes.
Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Wiley.
Russell, M., & Kaplan, L. (2021). An intersectional approach to differential item functioning: Reflecting configurations of inequality. *Practical Assessment, Research, and Evaluation, 26*(Article 21). https://doi.org/10.7275/20614854
Sales, H. F. S., Andrade, J. M., & Asfora, V. F. O. (2018). Desenvolvimento de um banco de itens para avaliar o transtorno depressivo maior. *Avaliação Psicológica, 17*(4), 451-461. https://doi.org/10.15689/ap.2018.1704.5.05
Santos, R. G. (2015). *ECCOs 4/10: Do papel ao teste adaptativo computadorizado* (Tese de Doutorado Não-Publicada). Universidade Federal de Pernambuco.
Schoot, R. V., Lugtig, P. & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486-492. http://dx.doi.org/10.1080/17405629.2012.686740
Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53–61.
Snijders, T., & Bosker, R. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Sage Publishers.
Steinmann, I., & Olsen, R. V. (2022). Equal opportunities for all? Analyzing within country variation in school effectiveness. *Large-scale Assessments in Education, 10*(2), 1-34. https://doi.org/10.1186/s40536-022-00120-0
Thompson, B. (Eds.) (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* American Psychological Association.
Thompson, B. (2006). *Foundations of Behavioral statistics: An insight-based approach.* The Guilford Press.
Unesco (2019). *A promessa das avaliações de aprendizagem em larga escala: Reconhecer os limites para desbloquear oportunidades.* Unesco.
Valentini, F., Gomes, C. M. A., Muniz, M., Mecca, T. P., Laros, J. A., & Andrade, J. M. (2015). Confiabilidade dos índices fatoriais da WAIS-III adaptada para a população brasileira. *Psicologia: Teoria e Prática, 17*(2), 123-139. https://doi.org/10.15348/1980-6906/psicologia.v17n2p123-139
Veldkamp, B. P., & Sluijter, C. (Eds.). (2019). Theoretical and practical advances in computer-based educational measurement. *Springer*. https://doi.org/10.1007/978-3-030-18480-3
Vinha, L. G. A., Karino, C. A., & Laros, J. A. (2016). Factors associated with mathematics performance in Brazilian basic education. *Psico-USF, 21*(1), 87-100. https://doi.org/10.1590/1413-82712016210108

Vinha, L. G. A., & Laros, J. A. (2018). Dados ausentes em avaliações educacionais: Comparação de métodos de tratamento. *Estudos em Avaliação Educacional, 29*(70), 156-187. https://doi.org/10.18222/eae.v0ix.3916

Wagenmakers, E., Marsman, M., Jamil, T. et al. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35–57. https://doi.org/10.3758/s13423-017-1343-3.

Wainer, H. (Ed.). (2015). *Computerized adaptive testing: A primer* (2nd ed.). Routledge. https://doi.org/10.4324/9781410605931

Wiberg, M., Ramsay, J. O., & Li, J. (2019). Optimal scores: An alternative to parametric item response theory. *Psychometrika, 84*(1), 310-322. https://doi.org/10.1007/s11336-018-9639-4

Zanini, D. S., Peixoto, E. M., Andrade, J. M., Campos, M. M., & Ribeiro, J. L. P. (2021). Social isolation in Brazil: Analysis of adherence, influence of personality, well-being and psychological distress. *Estudos de Psicologia (Natal), 26*(1), 23-32. https://dx.doi.org/10.22491/1678-4669.20210004

Zanini, D. S., Peixoto, E. M., Andrade, J. M., Tramonte, L. (2021). Practicing social isolation during a pandemic in Brazil: A description of psychosocial characteristics and traits of personality during covid-19 lockout. *Frontiers in Sociology, 6*, 615232. https://doi.org/10.3389/fsoc.2021.615232

Zopluoglu, C., & Davenport Jr., E. C. (2017). A note on using eigenvalues in dimensionality assessment. *Practical Assessment, Research, and Evaluation, 22*(Article 7). https://doi.org/10.7275/zh1k-zk32

recebido em julho de 2022
aprovado em outubro de 2022

## Sobre os autores

**Jacob Arie Laros** é doutor em Ciências Psicológicas, Sociais e Educacionais desde 1991 pela Rijksuniversiteit Groningen (RuG – Holanda). Atualmente, ele é professor no Instituto de Psicologia (IP) e professor permanente no PPG-PSTO na Universidade de Brasília.

**Josemberg Moura de Andrade** é psicólogo, especialista em Avaliação Psicológica e doutor em Psicologia Social, do Trabalho e das Organizações pela UnB (2008). Atualmente, ele é Professor Associado III no Instituto de Psicologia e professor permanente no PPG-PSTO na Universidade de Brasília.

## Como citar este artigo

Laros, J. A., & Andrade, J. M. (2022). Avaliação psicológica e avaliação da aprendizagem em larga escala: Diretrizes para pesquisadores. *Avaliação Psicológica, 21*(4), 397-406. http://dx.doi.org/10.15689/ap.2022.2104.24199.03