https://doi.org/10.18222/eae.v34.9220_port

AVALIAÇÃO LONGITUDINAL DE ESTUDANTES DE MEDICINA: O TESTE DE PROGRESSO É APROPRIADO?

CARLOS EDUARDO ANDRADE PINHEIRO[I]
DIOGO ONOFRE DE SOUZA[II]
Tradução de: Fernando Effori de Mello[III]

[I] Universidade Federal de Santa Catarina (UFSC), Florianópolis-SC, Brasil; *ceapinheiro1@gmail.com*
[II] Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre-RS, Brasil; *diogo.bioq@gmail.com*
[III] Tradutor freelancer, São Paulo-SP, Brasil; *feffori@gmail.com*

RESUMO

O artigo objetiva aferir se o Teste de Progresso é apropriado para avaliar cursos e estudantes em diferentes fases da graduação de medicina. Analisam-se as características das questões e a confiabilidade de três testes de progresso já aplicados. Constatou-se que, para os estudantes do 2º ano, 76,4% das questões se mostraram de qualidade pobre (bisserial < 0,2); diminuindo para 47,7% no 4º ano e para 25,3% no 6º ano. A confiabilidade dos testes, pelo alfa de Cronbach, foi de somente 0,60 para os alunos do 2º ano, aumentando para 0,76 para os do 4º ano e 0,87 para os alunos do 6º ano. A forma atual do Teste de Progresso mostrou confiabilidade baixa e inaceitável para os estudantes do 2º ano, razoável para os do 4º e ótima para os estudantes do 6º ano. Um aperfeiçoamento dessa forma de avaliação longitudinal é proposto.

PALAVRAS-CHAVE AVALIAÇÃO DA EDUCAÇÃO • EDUCAÇÃO MÉDICA • AVALIAÇÃO EXTERNA.

COMO CITAR:

Pinheiro, C. E. A., & Souza, D. O. de. (2023). Avaliação longitudinal de estudantes de medicina: O Teste de Progresso é apropriado? *Estudos em Avaliação Educacional*, *34*, Artigo e09220. https://doi.org/10.18222/eae.v34.9220_port

LONGITUDINAL ASSESSMENT OF MEDICAL STUDENTS: IS PROGRESS TEST APPROPRIATE?

ABSTRACT

The purpose of this article is to assess whether the Progress Test is appropriate to evaluate programs and students during different stages of medical studies. The characteristics of the items and reliability of three previously applied progress tests were analyzed. For second-year students, 76.4% of the questions demonstrated poor quality (biserial < 0.2). This percentage decreased to 47.7% in the forth year and to 25.3% in the sixth year. The test's reliability, measured by Cronbach's alpha, was only 0.6 for second-year students and increased to 0.76 in the fourth year and to 0.87 for sixth-year students. The current form of the Progress Test showed low and unacceptable reliability for second-year students, reasonable for the fouth year, and excellent for the sixth year. An improvement of this longitudinal assessment is proposed.

KEYWORDS EDUCATION EVALUATION • MEDICAL EDUCATIONAL MEASUREMENT • EXTERNAL EVALUATION.

EVALUACIÓN LONGITUDINAL DE ESTUDIANTES DE MEDICINA: ¿LA PRUEBA DE PROGRESO ES APROPIADA?

RESUMEN

El artículo tiene el propósito de verificar si la Prueba de Progreso es apropiada para evaluar cursos y estudiantes en distintas fases del curso de medicina. Se analizan las características de las preguntas y la confiabilidad de tres pruebas de progreso ya aplicadas. Se constató que, para los estudiantes de 2º año, el 76,4% de las preguntas se mostraron de baja calidad (biserial < 0,2), reduciéndose a 47,7% en 4º año y a un 25,3% en 6º año. La confiabilidad de las pruebas, por alfa de Cronbach, fue de tan solo 0,60 para los alumnos de 2º año, y aumentó a 0,76 para los de 4º año y a 0,87 para los estudiantes de 6º año. La forma actual de la Prueba de Progreso mostró confiabilidad baja e inaceptable para los alumnos de 2º año, razonable para los de 4º y excelente para los de 6º. Se propone que se perfeccione dicha forma de evaluación longitudinal.

PALABRAS CLAVE EVALUACIÓN DE LA EDUCACIÓN • EDUCACIÓN MÉDICA • EVALUACIÓN EXTERNA.

Recebido em: 16 DEZEMBRO 2021
Aprovado para publicação em: 22 MARÇO 2023

INTRODUÇÃO

Embora haja concordância a respeito da vantagem da avaliação longitudinal de conhecimentos para a educação médica (Albanese & Case, 2016; Cecilio-Fernandes et al., 2021; Vleuten et al., 2018; Wrigley et al., 2012), um dos grandes desafios enfrentados pela educação médica atual é a preparação de avaliações capazes de, simultaneamente: I) mensurar os conhecimentos adquiridos por estudantes ao longo de seus estudos (avaliação somativa); II) ser um instrumento para que os estudantes ajustem seus estudos (avaliação formativa); e III) ajudar o docente e a instituição de ensino a entender (avaliação informativa) e aprimorar a aquisição de conhecimentos (Pugh & Regehr, 2016). O Teste de Progresso (TP) é um tipo de avaliação extensiva de conhecimentos cognitivos capaz de realizar essas três funções que tem sido utilizado em faculdades de medicina. Ele é realizado periodicamente, aplicado no mesmo dia para todos os estudantes de graduação em medicina, e é direcionado ao nível de conhecimento esperado para estudantes do último ano do curso. Entretanto há incertezas quanto à adequação da atual estrutura do TP para avaliar os estudantes nos estágios iniciais e intermediários do programa (Albanese & Case, 2016; Henning et al., 2017).

O TP foi desenvolvido na esteira da teoria pedagógica construtivista, que tem como um de seus princípios o conceito de que os estudantes são construtores do próprio conhecimento, de maneira que o professor não é mais aquele que concede o conhecimento, mas um parceiro em sua busca. Essa grande mudança resultou no uso de novas formas de abordar o ensino e a aprendizagem na formação médica, chamadas de metodologias ativas, como a aprendizagem baseada em problemas (ABP), aprendizagem baseada em equipes (ABE), etc. Essas novas formas de ensino têm gerado a necessidade de novas formas de avaliação das faculdades e seus estudantes (Cecilio-Fernandes et al., 2021; Heeneman et al., 2017; Reberti et al., 2020; Vleuten & Schuwirth, 2019).

As escolas com pedagogia tradicional, que não utilizam metodologias ativas, tendem a avaliar a aprendizagem dos estudantes em cada um dos componentes curriculares ou áreas de conhecimento separadamente. Nos exames de fim de ano, é comum que os estudantes façam muitas provas durante a mesma semana. Eles frequentemente estudam para cada uma delas no último momento, adotando estratégias cognitivas de ordem inferior, como o estudo intensivo de última hora ou a memorização mecânica (Pugh & Regehr, 2016). Esse tipo de avaliação promove uma aprendizagem descontextualizada e fragmentada, que em pouco tempo é esquecida. O TP, por sua vez, por cobrir um espectro muito mais amplo de conhecimento, é integrativo e proporciona *feedbacks* oportunos, desestimulando as estratégias acima referidas e estimulando um conhecimento mais aprofundado, contextualizado e duradouro (Vleuten et al., 2004). As avaliações de amplo conteúdo aplicadas de forma

regular e sistemática ao longo do programa permitem aos estudantes verificar seus pontos fortes e fracos nos conhecimentos, potencialmente proporcionando motivação e direcionamento para futuros estudos (Epstein, 2007; Heeneman et al., 2017; Tavakol & Dennick, 2017).

Utilizado em vários países desde a década de 1970, o TP é rotineiramente aplicado em faculdades de medicina de países da Europa e da América do Norte, duas a quatro vezes ao ano, com números variáveis de questões (Blake et al., 1996; Vleuten et al., 2004). Ele foi desenvolvido como uma avaliação mais alinhada com o currículo baseado em problemas, e também como ferramenta de comparação entre os conhecimentos adquiridos utilizando metodologias ativas e os conhecimentos adquiridos utilizando currículos tradicionais (Cecilio-Fernandes et al., 2021). Atualmente, nos Países Baixos, o teste é composto de 200 questões de múltipla escolha, com duas a cinco alternativas e penalidades para respostas incorretas.

Os itens são criados e revisados por docentes de diferentes universidades, com o objetivo de evitar que representem a visão e a experiência de apenas um docente. Na maioria dos casos, as questões são vinhetas clínicas, não muito longas, que exigem um tipo de conhecimento e raciocínio lógico envolvendo processos cognitivos de ordem superior. Elas constituem uma matriz que compreende todas as áreas de conhecimento esperadas ao fim do curso. Em data e horário estabelecidos, todos os estudantes de medicina realizam a mesma prova. Depois, eles recebem a pontuação de seu teste individualmente, com seus conhecimentos adquiridos comparados aos do(s) ano(s) anterior(es), além de diferentes formas de *feedback*.

No Brasil, desde 2001, vários consórcios de faculdades de medicina têm sido criados para elaborar avaliações com TP. No entanto, apenas um teste de 120 questões de múltipla escolha (QME) simples, com quatro alternativas, sem penalização por respostas incorretas, é aplicado a cada ano. Com apoio da Associação Brasileira de Educação Médica (Abem), há atualmente 13 consórcios regionais que administram o TP em suas faculdades de medicina. Em 2015, 40% das faculdades de medicina brasileira já haviam entrado para algum consórcio de TP (Bicudo et al., 2019; Rosa et al., 2017; Sakai et al., 2008; Sartor et al., 2020). A organização das faculdades sob a forma de consórcios para o TP promove a integração entre elas, capacita os docentes para a elaboração dos itens (questões) e reduz os custos associados ao processo de elaboração, impressão e correção dos testes (Hamamoto & Bicudo, 2020a; Vleuten et al., 2018).

Em 2020, havia no Brasil 357 faculdades de medicina, com 37.823 vagas para novos estudantes; em 2001, por sua vez, o número de novas vagas era consideravelmente menor: 11.541 (Scheffer et al., 2020). Esse grande aumento gerou a discussão a respeito da criação de exames de habilitação para o exercício da medicina, pela necessidade de desenvolvimento de novos formatos de avaliação externa para garantir

a qualidade das faculdades e dos graduados (Bica & Kornis, 2020; Troncon, 2019). No Brasil, o TP é uma forma de avaliação externa longitudinal e essencialmente formativa, oferecendo diferentes oportunidades de *feedback*. Tanto em faculdades com currículo tradicional como nas que utilizam metodologias ativas, as pontuações do TP quase não interferem nas notas regulares dos estudantes; além disso, não influenciam na aprovação ou reprovação do estudante ao final do curso ou período letivo.

Em algumas faculdades europeias e norte-americanas, os resultados do TP acrescentam pontuações que determinarão se o estudante pode avançar no curso (avaliação somativa) ou classificam aqueles que fizeram o teste de acordo com seu desempenho no exame (Albanese & Case, 2016; Blake et al., 1996; Heeneman et al., 2017). Por esse motivo, os TPs precisam ser confiáveis (Downing, 2004; Kibble, 2017). Na Holanda, quatro TPs nacionais são aplicados a cada ano a todos os estudantes de medicina, para aumentar a confiabilidade da avaliação (Wrigley et al., 2012); essa avaliação longitudinal de conhecimentos ao longo do programa também substitui o exame de habilitação ao fim dos estudos médicos (Vleuten et al., 2004). No entanto, esses testes nacionais longitudinais com esse tipo de intervalo demandam custos com os quais muitos países, como o Brasil, dificilmente poderiam arcar.

No Brasil e no exterior, após o TP, as faculdades recebem as pontuações gerais de seus estudantes e as pontuações para as diferentes áreas de conhecimento abrangidas no currículo médico, como ciências básicas, saúde coletiva, pediatria, clínica, etc. Ao comparar as pontuações dos estudantes – tanto nos escores gerais quanto nas diferentes áreas de conhecimento – com a média obtida pelo grupo de faculdades participantes do TP, as instituições têm a possibilidade de verificar quais áreas correspondem a seus pontos fortes e fracos. Com essas informações disponíveis, as faculdades podem alterar seus currículos, processos ou estratégias de ensino e aprendizagem, e, nos anos seguintes, avaliar o impacto de tais mudanças (Rosa et al., 2017).

Os organizadores do TP, em todos os lugares do mundo onde ele é aplicado, após empregarem a prova, analisam sua validade e confiabilidade, bem como os aspectos pedagógicos e psicométricos de cada item. Eles selecionam algumas questões e modificam outras, de modo a formar um banco de dados de questões. O surgimento dos testes adaptativos computadorizados passou a permitir a elaboração de avaliações em que as questões se tornem mais ou menos difíceis à medida que o estudante responde a prova, dependendo da habilidade e dos conhecimentos do avaliado. A existência de um banco de itens, com as questões previamente testadas e calibradas, é fundamental para essa nova forma de avaliação. Um TP internacional informatizado já está disponível e pode oferecer resultados com elevado nível de confiabilidade, utilizando um número menor de questões (Cecilio-Fernandes, 2019; Collares & Cecilio-Fernandes, 2019).

No Brasil, o Ministério da Educação vem aplicando um exame nacional, o Exame Nacional de Desempenho dos Estudantes (Enade), para todos os estudantes no último ano da graduação, incluindo os que cursam medicina, desde 1996. Esse exame foi mudado em 2004, quando passou a incluir estudantes do primeiro ano, com o intuito de determinar a contribuição da graduação para a aprendizagem discente. No entanto, com a universalização do Enem, o exame nacional aplicado a estudantes do ensino médio, utilizado em todo o país para a admissão em universidades públicas – que também funciona como referência de qualidade para uma série de políticas públicas de inclusão para o ensino superior –, essa iniciativa foi abandonada, e atualmente exige-se que apenas estudantes formandos realizem a prova. Como esse exame de 40 questões, com apenas 30 delas tratando diretamente do campo profissional, é aplicado para cada área (com seus cursos/habilitações) uma vez a cada três anos, ele tem pouco impacto formativo, e é claramente inadequado para avaliar o progresso da aprendizagem do estudante (Ristoff, 2022). Isso levou as faculdades de medicina a avançar para outras formas de avaliação, como o TP.

A fim de melhorar a confiabilidade dos TPs e superar dificuldades econômicas relacionadas à realização de múltiplos testes a cada ano, o desenvolvimento de um TP de melhor qualidade é uma abordagem promissora para aprimorar a avaliação dos conhecimentos dos estudantes de medicina (Aubin et al., 2020; Sahoo & Singh, 2017). Assim, este estudo teve como objetivo verificar a confiabilidade dos TPs aplicados no Brasil para estudantes em diferentes fases de estudos (2º, 4º e 6º ano). Com base em nossos resultados, propomos um novo formato de avaliação longitudinal de conhecimentos, chamado de Teste de Progresso Customizado (TPC), capaz de avaliar as faculdades de medicina e o progresso dos estudantes ao longo da graduação, servindo como alternativa a exames de habilitação aplicados apenas ao fim do programa.

OBJETIVOS

Este estudo objetiva avaliar se os TPs brasileiros aplicados na atualidade são apropriados para avaliar as faculdades/cursos e os conhecimentos dos estudantes nas etapas iniciais, intermediárias e finais dos programas de medicina. Além disso, tem o objetivo de gerar uma base de conhecimento para subsidiar uma proposta de melhoria do TP, colaborando para criar um tipo de avaliação longitudinal que aprimore a avaliação dos estudantes de medicina em todos os anos letivos em todo o Brasil.

METODOLOGIA

Desenho do estudo, local e participantes

Os resultados foram obtidos a partir de três TPs já realizados, com foco nos conhecimentos esperados de estudantes do 6º ano, e aplicados a alunos em três anos diferentes do programa de medicina. Foram calculadas as características dos itens (dificuldade e índice de discriminação) e a confiabilidade dos testes.

Os TPs analisados foram preparados por um consórcio de dez faculdades de medicina, com o apoio da Associação Brasileira de Educação Médica (Abem) – Regional Sul II. Os testes foram aplicados entre 2015 e 2017, realizados simultaneamente por todos os estudantes das dez faculdades. No entanto, apenas os resultados de alunos de 2º, 4º e 6º ano dos programas de medicina foram analisados, em vez de todos os anos, em razão de existir uma proposta, ainda não implementada, de aplicação de avaliação longitudinal de conhecimentos dos estudantes em nível nacional somente aos alunos desses anos de curso (Ministério da Educação, 2016).

Nos TPs analisados, o número de estudantes do 2º ano foi o maior, e o de alunos do 6º ano foi o menor, uma vez que três das faculdades participantes eram novas, e outras duas haviam aumentado o número de vagas logo antes dos testes (Scheffer & Dal Poz, 2015). No Brasil, os programas de medicina de todas as instituições de ensino superior têm duração de seis anos.

Caracterização dos testes de progresso aplicados

Os três TPs foram elaborados para avaliar todos os estudantes com base nos conhecimentos esperados dos alunos do último (6º) ano, sendo que o conteúdo dos testes se baseou em uma matriz de temas das Diretrizes Curriculares Nacionais do Curso de Graduação em Medicina, do Ministério da Educação (Resolução n. 3, 2014). Cada TP continha 120 QMEs simples com quatro alternativas, das quais apenas uma era correta, sem pontuação negativa para respostas incorretas. Os três TPs continham 20 itens de cada uma das seis grandes áreas a seguir: ciências básicas, medicina clínica, cirurgia, ginecologia/obstetrícia, pediatria e saúde pública.

Variáveis estimadas e análise estatística

Foram analisadas, nos três TPs, as seguintes variáveis:

i) Dificuldade do item.
ii) Índice de discriminação.
iii) Confiabilidade do teste.

A dificuldade do item foi determinada pelo percentual de respostas corretas para cada questão: os itens com menos de 45% de respostas corretas foram considerados difíceis, aqueles com percentual entre 45% e 80% foram considerados médios, e os com mais de 80% foram considerados fáceis.

O índice de discriminação do item foi determinado pelo coeficiente de correlação ponto-bisserial, que mede a correlação entre o acerto ou erro em cada item (questão) e a pontuação final no teste, indicando a qualidade de cada questão. Questões com um coeficiente de correlação ponto-bisserial < 0,2 são consideradas de baixa capacidade de discriminação, indicando que tais itens devem ser revistos ou eliminados, enquanto aqueles com coeficiente > 0,2 são considerados aceitáveis ou bons (De Champlain, 2010; Walsh et al., 2018).

A confiabilidade do teste foi medida utilizando o coeficiente alfa de Cronbach, que afere a consistência interna do teste. A confiabilidade é a probabilidade de que a pontuação no teste represente os conhecimentos reais da pessoa avaliada: ela deve ser de, no mínimo, 0,9 para uma avaliação com consequências importantes para os avaliados (*high-stakes assessment*), valores entre 0,89 e 0,8 são bons, e 0,7 é considerado o mínimo aceitável para uma avaliação educacional de amplo espectro (Downing, 2004; Kibble, 2017; Vleuten & Schuwirth, 2005).

Análise estatística

A dificuldade do item e o índice de discriminação para cada questão foram mensurados na fase 1 do *software* BILOG-MG (versão 3.0.2327.2), desenvolvido pela Scientific Software International, Inc., para Windows. A comparação entre as médias de distribuição de dificuldade e discriminação foi realizada por meio de uma análise de variância com um fator (ANOVA), e a confiabilidade dos TPs, por meio do alfa de Cronbach e da ANOVA com dois fatores sem replicação. Essas comparações e o intervalo de confiabilidade ($\alpha < 0,05$) foram calculados utilizando o Microsoft Excel Professional 2016.

As pontuações apresentadas são os percentuais de acertos para cada estudante no teste.

O consórcio que organiza os TPs permitiu este estudo sob a condição de manter a confidencialidade das identidades das faculdades e dos estudantes.

RESULTADOS

O número de estudantes participantes dos TPs (do 2º, 4º e 6º ano) e as pontuações por eles obtidas nos três anos analisados (2015, 2016 e 2017) são apresentados na Tabela 1.

TABELA 1

Número de estudantes (N) em diferentes anos dos estudos e pontuações médias obtidas nos três testes de progresso realizados em dez faculdades de medicina na região Sul do Brasil – 2015, 2016 e 2017

ANO DOS EXAMES	ESTUDANTES					
	2° ANO		4° ANO		6° ANO	
	N	PONTUAÇÃO	N	PONTUAÇÃO	N	PONTUAÇÃO
2015	716	42,5	517	52,0	386	60,4
2016	822	39,8	539	49,5	411	58,3
2017	821	43,0	635	53,8	502	62,7
MÉDIAS*		41,8		51,8		60,5
IC 95%		(40,2-43,4)		(49,8-53,7)		(58,4-62,5)

Fonte: Elaboração dos autores.

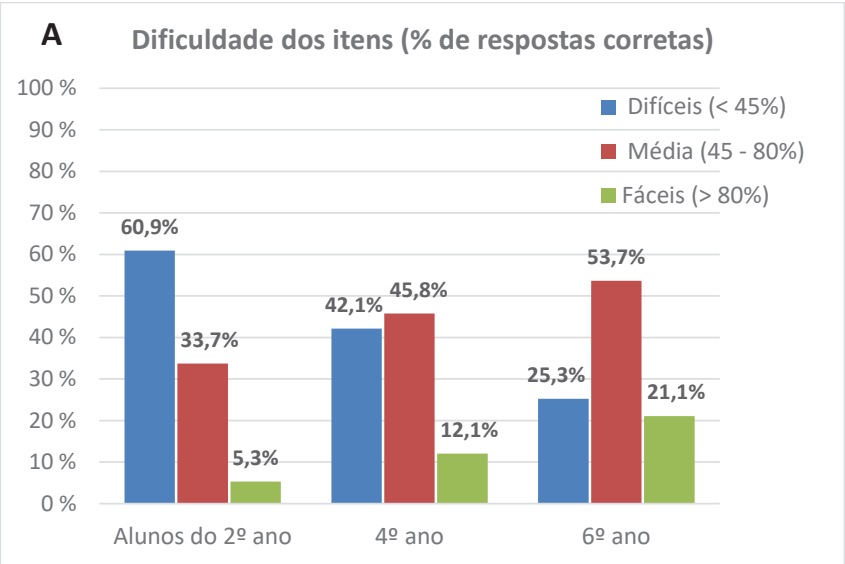
* ANOVA: $p < 0,0001$.

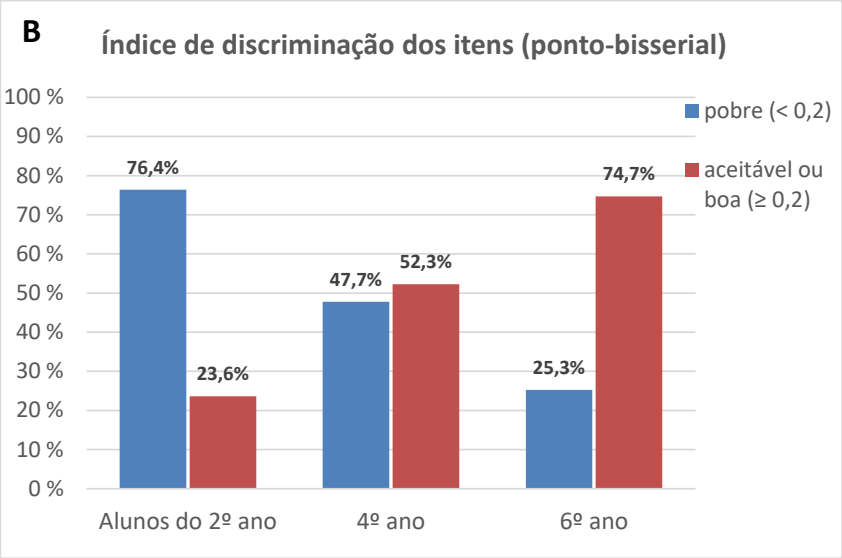
As pontuações médias dos estudantes aumentaram progressivamente do 2º ano (41,8) para o 4º ano (51,8) e para o 6º ano (60,5); as diferenças entre as médias eram estatisticamente significativas.

Os valores de dificuldade e discriminação dos itens são apresentados na Figura 1.

FIGURA 1

Médias da dificuldade e discriminação dos itens do Teste de Progresso mensurados para estudantes em diferentes anos dos estudos em dez faculdades de medicina da região Sul do Brasil (2015, 2016 e 2017)





Fonte: Elaboração dos autores.

O percentual de itens considerados difíceis diminuiu de 60,9% (estudantes do 2º ano) para 42,1% e 25,3% (estudantes do 4º e do 6º ano, respectivamente), enquanto o das questões consideradas fáceis aumentou de 5,3% (estudantes do 2º ano) para 12,1% e 21,1% (estudantes do 4º e 6º ano, respectivamente) (Figura 1A).

O percentual de itens com baixa discriminação (má qualidade) diminuiu dos estudantes do 2º ano (76,4%) para os do 4º e 6º ano (47,7% e 25,3%, respectivamente). Todas essas comparações entre as características das questões para alunos em diferentes anos dos cursos se mostraram significativas ($p < 0,001$) pela ANOVA (Figura 1B).

A confiabilidade dos três TPs nos três anos de curso e nos anos em que foram aplicados está apresentada na Tabela 2.

TABELA 2

Confiabilidade dos três testes de progresso, mensurada pelo alfa de Cronbach, para os diferentes anos de curso em dez faculdades de medicina na região Sul do Brasil – 2015, 2016 e 2017

ANO DOS TESTES	ESTUDANTES		
	2º ANO	4º ANO	6º ANO
2015	0,64	0,75	0,86
2016	0,52	0,76	0,87
2017	0,65	0,77	0,87
MÉDIAS*	0,60	0,76	0,87
IC 95%	(0,54-0,67)	(0,75-0,77)	(0,86-0,87)

Fonte: Elaboração dos autores.

* ANOVA: $p < 0,001$.

A confiabilidade média dos três TPs, mensurada pelo alfa de Cronbach, aumentou dos estudantes do 2º ano (0,60) para os do 4º e do 6º ano (0,76 e 0,87, respectivamente). As diferenças entre os valores médios eram significativas ($p < 0,001$), e não houve intersecção de intervalos de confiança nos valores médios para nenhuma das comparações da confiabilidade dos TPs entre os estudantes nos diferentes anos de curso.

DISCUSSÃO

As diferenças nas pontuações médias entre estudantes de 2º, 4º e 6º ano, em todos os TPs descritos na Tabela 1, mostram um aumento semelhante nos três anos analisados, e estão de acordo com a literatura nacional (Bicudo et al., 2019) e internacional (Wrigley et al., 2012).

Para contribuir com o entendimento da abordagem avaliativa atual do TP, este estudo avaliou: I) a dificuldade do item; II) o índice de discriminação; e III) a confiabilidade dos testes dos TPs já aplicados a estudantes de 2º, 4º e 6º ano de dez faculdades de medicina brasileiras nos anos de 2015, 2016 e 2017. O percentual de itens considerados difíceis diminuiu, e o dos considerados fáceis aumentou do 2º para o 4º e 6º ano (Figura 1A). O percentual do índice de discriminação de questões consideradas de baixa qualidade diminuiu, e das consideradas aceitáveis ou boas aumentou do 2º para o 4º e 6º ano (Figura 1B). O índice de discriminação determinado por meio da correlação ponto-bisserial é um bom indicador da qualidade do item (Tavakol & Dennick, 2017). Valores ponto-bisseriais < 0,2 sugerem que a questão não é capaz de discriminar estudantes com diferentes níveis de conhecimentos, portanto devem ser considerados com cautela (De Champlain, 2010; Kibble, 2017; Pasquali, 2004; Tavakol & Dennick, 2013). O percentual de 76,4% de itens de baixa qualidade (ponto-bisserial < 0,2) para estudantes do 2º ano encontrado neste estudo provavelmente ocorre porque o TP geralmente é elaborado para avaliar o nível de conhecimentos esperado para estudantes do último ano, o que interfere na confiabilidade do teste.

A confiabilidade dos TPs é influenciada pela qualidade e pelo número de questões, bem como pela frequência com que o teste é aplicado (Kibble, 2017; Wrigley et al., 2012). A confiabilidade média dos TPs neste estudo (Tabela 2), mensurada pelo alfa de Cronbach, foi de 0,60 para estudantes do 2º ano, considerada baixa e inaceitável para avaliações com consequências reais moderadas ou mesmo baixas (Downing, 2004; Kible, 2017). A confiabilidade média aumentou para 0,76 nos estudantes do 4º ano, um número considerado razoável, e para 0,87 nos estudantes do 6º ano, um número considerado próximo do ideal. Essa elevada confiabilidade para o 6º ano sugere que os TPs estão alcançando a consistência esperada para esse tipo de

avaliação para os estudantes do último ano de graduação em medicina (Downing, 2004; Kibble, 2017; Pasquali, 2004). Em um estudo anteriormente descrito sobre TP (Wrigley et al., 2012), foram observados valores semelhantes para estudantes de 2º e 4º ano. No presente estudo, a confiabilidade do TP de 0,60 para estudantes do 2º ano indica que as pontuações no TP apresentaram maiores erros aleatórios para esse ano do curso (Tavakol & Dennick, 2011). Esse número tão elevado de erros mensurado em um TP de aplicação anual única com 120 QMEs, como neste estudo, torna a utilidade do teste inviável para a avaliação atual de estudantes do 2º ano de medicina no Brasil (Albanese & Case, 2016; Downing, 2004; Kibble, 2017). Assim, essa baixa confiabilidade confirma a dificuldade de utilizar esse formato de TP no Brasil como avaliação longitudinal que objetiva ser uma alternativa a um exame nacional de habilitação para o exercício profissional da medicina.

Os valores de desempenho padrão utilizados neste estudo para dificuldade do item, índice de discriminação do item e confiabilidade do teste podem ser questionados. Alguns autores consideram difíceis itens com menos de 20% (Aubin et al., 2020) ou 30% (De Champlain, 2010; Kheyami et al., 2018; Primi, 2012; Sahoo & Singh, 2017) de acertos. Esses valores são geralmente utilizados quando há penalidade para respostas erradas (*formula scoring*), para evitar acertos por sorte. Com relação ao índice de discriminação, alguns autores consideram bons os itens com coeficiente ponto-bisserial ≥ 0,25 ou superior (Tavakol & Dennick, 2017). No entanto, é quase unânime que um teste com valor de confiabilidade < 0,70, como se observou em estudantes do 2º ano neste estudo, não pode ser utilizado para avaliar o desenvolvimento acadêmico dos alunos.

Em alguns lugares da Europa, onde o TP surgiu há mais de 50 anos, o intuito foi o de avaliar mudanças curriculares e criar exames alinhados com os currículos que adotam ABP. É interessante notar que, inicialmente, ele era utilizado somente como avaliação formativa, passando gradualmente a substituir as formas tradicionais de avaliação somativa. Com o incremento da frequência dos exames anuais para aumentar a confiabilidade dos resultados, em algumas localidades o TP tornou-se a única ferramenta para avaliar conhecimentos ou sua aquisição. O TP é um dos componentes de um sistema de avaliação, que também utiliza outros métodos para analisar habilidades e competências. No entanto, isso não é praticado atualmente no Brasil, onde o TP começou a ser empregado há menos de 20 anos, sendo utilizado, principalmente, como ferramenta formativa.

A discussão a respeito da necessidade de empregar um TP a cada semestre (duas vezes ao ano) e formas de integrar os resultados nas avaliações somativas já se iniciou em nossos consórcios. No entanto, tanto nas faculdades com currículos modernos, pedagogicamente alinhados com o TP, quanto nas de currículo tradicional, essa discussão precisa ser aprofundada. Além disso, parece haver a necessidade

de aumentar a adesão e o compromisso por parte dos estudantes, principalmente nas fases iniciais, e evitar o comportamento de recorrer ao acaso ("chutar") nas respostas. No entanto, essa questão ainda não foi estudada ou descrita na literatura com relação ao TP na educação médica no Brasil.

Em geral, os TPs são desenvolvidos e empregados por um grupo (consórcio) de faculdades. Essa iniciativa reduz o custo da utilização e da análise do TP e eleva a qualidade das questões. Além disso, o TP é vantajoso para os estudantes, que podem retificar seu percurso acadêmico desde o início, assim como o TP oferece um parâmetro comparativo para que as instituições possam verificar seus pontos fortes e fracos. No Brasil, a preocupação com os resultados e seu uso como ferramenta para mudanças nas faculdades parece despertar mais o interesse das faculdades privadas do que das públicas, como aparenta já acontecer com o Enade (Damas & Miranda, 2019).

A fim de evitar os riscos do Enade, que cria um *ranking* e uma disputa por colocação entre as faculdades, podendo servir como um parâmetro que pode encobrir problemas, no TP cada faculdade recebe somente os próprios resultados, e o resultado médio do grupo de faculdades participantes serve como fator de comparação. Esse ambiente, de colaboração sem disputa, além de integrar as faculdades do consórcio, fortalece as regionais da Abem. Ademais, estimula a criação de uma área de pesquisas sobre avaliação e tem o potencial de melhorar as avaliações de conhecimentos e a educação médica (Hamamoto & Bicudo, 2020b).

A elaboração de testes com questões de múltipla escolha bem estruturadas é considerada adequada para avaliar a evolução dos conhecimentos teóricos dos estudantes, também chamada de avaliação de domínio cognitivo (Aubin et al., 2020; Epstein, 2007; Vleuten & Schuwirth, 2019). Isso deve explicar parcialmente por que o TP é cada vez mais aplicado em um número maior de países (Vleuten et al., 2018) e em diferentes programas de graduação, como odontologia (Ali et al., 2016; Oliveira et al., 2020), farmácia (Albekairy et al., 2021) e veterinária (Herrmann et al., 2020), bem como na formação médica em nível de pós-graduação (Rutgers et al., 2018), no Brasil e no exterior (Alkhalaf et al., 2021; Sá et al., 2021).

Em razão de seu amplo conteúdo, a utilização do TP desestimula o estudo de última hora e estimula o estudo contínuo, representando um avanço nas avaliações educacionais no campo da saúde (Pugh & Regehr, 2016). No entanto, os conceitos mais modernos de avaliação sugerem que cada faculdade deve criar não apenas um ou dois tipos de avaliação, mas sistemas de avaliação que abordem todos os domínios teóricos, habilidades e atitudes (Norcini et al., 2018). Os mesmos conceitos devem ser aplicados a avalições em larga escala.

Outros domínios fundamentais para a prática profissional – como habilidades comunicativas, habilidades clínicas e atitudes – também devem ser avaliados

(Epstein, 2007; Vleuten, 2016). No entanto, esses domínios exigem formatos avaliativos muito mais complexos, trabalhosos e caros, como o Exame Clínico Objetivo Estruturado (OSCE, sigla em inglês de Objective Structured Clinical Examination), o Mini Exercício Clínico Avaliativo (Mini-Cex, sigla em inglês de Mini Clinical Evaluation Exercise), o Feedback 360 Graus. A aplicação dessas metodologias em avaliações educacionais externas de larga escala merece mais estudos e análises para se tornarem viáveis no Brasil.

Se um sistema avaliativo longitudinal amplo e de larga escala é inadequado para determinar se um profissional tem conhecimentos suficientes e se ele é ou não capaz de exercer a medicina, ainda mais inadequado seria realizar um único exame teórico, como o exame de habilitação profissional, aplicado ao fim do programa de graduação em medicina. Tal exame não impactaria diretamente as faculdades, tampouco melhoraria a educação médica, pois analisaria os alunos com relação a terem ou não adquirido o conhecimento teórico durante seus estudos de graduação, sem lhes ter permitido a possibilidade de remediação ainda durante o programa. Daí a necessidade de avaliações longitudinais confiáveis para estudantes em diferentes estágios do curso.

Assim, considerando apenas a avaliação do conhecimento teórico, este estudo apoia a percepção de que o formato atual do TP, isto é, 120 QMEs, sendo o teste aplicado uma vez por ano no Brasil, voltado para os conhecimentos esperados para alunos do 6º ano, é inadequado para avaliar os conhecimentos adquiridos por estudantes de medicina desde os primeiros anos do curso. O TP pode e deve ser melhorado para os estudantes nos estágios iniciais e intermediários.

Este estudo é o primeiro de uma série que se destina a investigar e propor uma avaliação longitudinal confiável que mensure de forma adequada a aquisição de conhecimentos pelos estudantes em todos os anos de curso no programa de medicina. Para isso, pretendemos aplicar um novo tipo de TP chamado de Teste de Progresso Customizado (TPC). Nessa abordagem, os itens serão divididos da seguinte forma:

- Para estudantes do 2º ano: 25% das questões abordarão o nível de conhecimento esperado para seu ano; 25% para o nível de 4º ano; e 50% para o nível de 6º ano.
- Para estudantes do 4º ano: 50% das questões abordarão o nível de conhecimento esperado para seu ano; e 50% para o nível de 6º ano.
- Para estudantes do 6º ano: todas as questões abordarão o nível de conhecimento esperado para seu ano, como é feito atualmente com excelente confiabilidade (Wrigley et al., 2012).

Espera-se que essa abordagem ajude a oferecer o tipo de avaliação capaz de: I) ter boa confiabilidade para todos os estágios do programa; e II) ser utilizada em situações com moderadas a elevadas consequências para os examinados.

Da maneira como se encontra, o atual TP longitudinal anual em programas de medicina não é confiável para avaliar os conhecimentos de estudantes de medicina em todos os anos do curso. Esperamos que este estudo reforce a relevância de propor o TPC como estratégia inovadora para avaliações somativas e formativas, potencialmente contribuindo para melhorar significativamente a avaliação da educação médica no Brasil. Uma avaliação como a proposta pelo TPC promete ser uma boa alternativa à proposta de exame nacional de habilitação ao fim dos programas de medicina.

REFERÊNCIAS

Albanese, M., & Case, S. M. (2016). Progress testing: Critical analysis and suggested practices. *Advances in Health Sciences Education, 21*(1), 221-234. https://doi.org/10.1007/s10459-015-9587-z

Albekairy, A. M., Obaidat, A. A., Alsharidah, M. S., Alqasomi, A. A., Alsayari, A. S., Albarraq, A. A., Aljabri, A. M., Alrasheedy, A. A., Alsuwayt, B. H., Aldhubiab, B. E., Almaliki, F. A., Alrobaian, M. M., Aref, M. A., Altwaijry, N. A., Alotaibi, N. H., Alkahtani, S. A., Bahashwan, S. A., & Alahmadi, Y. A. (2021). Evaluation of the potential of national sharing of a unified progress test among colleges of pharmacy in the Kingdom of Saudi Arabia. *Advances in Medical Education and Practice, 12*, 1465-1475. https://doi.org/10.2147/amep.s337266

Ali, K., Coombes, L., Kay, E., Tredwin, C., Jones, G., Ricketts, C., & Bennett, J. (2016). Progress testing in undergraduate dental education: The Peninsula experience and future opportunities. *European Journal of Dental Education, 20*(3), 129-134. https://doi.org/10.1111/eje.12149

Alkhalaf, Z. S. A., Yakar, D., Groot, J. C. de, Dierckx, R. A. J. O., & Kwee, T. C. (2021). Medical knowledge and clinical productivity: Independently correlated metrics during radiology residency. *European Radiology, 31*(7), 5344-5350. https://doi.org/10.1007/s00330-020-07646-3

Aubin, A.-S., Young, M., Eva, K., & St-Onge, C. (2020). Examinee cohort size and item analysis guidelines for health professions education programs: A Monte Carlo simulation study. *Academic Medicine: Journal of the Association of American Medical Colleges, 95*(1), 151-156. https://doi.org/10.1097/acm.0000000000002888

Bica, R. B. da S., & Kornis, G. E. (2020). Exames de licenciamento em medicina: Uma boa ideia para a formação médica no Brasil? *Interface – Comunicação, Saúde, Educação, 24*, Artigo e180546. https://doi.org/10.1590/Interface.180546

Bicudo, A. M., Hamamoto, P., Filho, Abbade, J., Hafner, M. de L., & Maffei, C. (2019). Teste de Progresso em Consórcios para todas as escolas médicas do Brasil. *Revista Brasileira de Educação Médica, 43*(4), 151-156. https://doi.org/10.1590/1981-52712015v43n4RB20190018

Blake, J. M., Norman, G. R., Keane, D. R., Mueller, C. B., Cunnington, J., & Didyk, N. (1996). Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine, 71*(9), 1002-1007. https://doi.org/10.1097/00001888-199609000-00016

Cecilio-Fernandes, D. (2019). Implementando o teste adaptativo computadorizado. *Scientia Medica, 29*(3), Artigo e34432. https://doi.org/10.15448/1980-6108.2019.3.34432

Cecilio-Fernandes, D., Bicudo, A. M., & Hamamoto, P. T., Filho. (2021). Progress testing as a pattern of excellence for the assessment of medical students' knowledge – Concepts, history, and perspective. *Medicina, 54*(1), Article e-173770. https://doi.org/10.11606/issn.2176-7262.rmrp.2021.173770

Collares, C. F., & Cecilio-Fernandes, D. (2019). When I say... computerised adaptive testing. *Medical Education, 53*(2), 115-116. https://doi.org/10.1111/medu.13648

Damas, B. R., & Miranda, G. J. (2019). Preparação da Instituição para o Enade: Importa? In *Anais do 3. Congresso UFU de Contabilidade*. UFU. https://eventos.ufu.br/sites/eventos.ufu.br/files/documentos/030_artigo_completo.pdf

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109-117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*(9), 1006-1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x

Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine, 356*(4), 387-396. https://doi.org/10.1056/nejmra054784

Hamamoto, P. T., Filho, & Bicudo, A. M. (2020a). Improvement of faculty's skills on the creation of items for progress testing through feedback to item writers: a successful experience. *Revista Brasileira de Educação Médica, 44*(1), Article e018. https://doi.org/10.1590/1981-5271v44.1-20190130.ING

Hamamoto, P. T., Filho, & Bicudo, A. M. (2020b). Implementation of the Brazilian National Network for Practices and Research with Progress Testing – BRAZ-NPT. *Revista Brasileira de Educação Médica, 44*(3), Letter to the editor e074. https://doi.org/10.1590/1981-5271v44.3-20200089

Heeneman, S., Schut, S., Donkers, J., Vleuten, C. van der, & Muijtjens, A. (2017). Embedding of the progress test in an assessment program designed according to the principles of programmatic assessment. *Medical Teacher, 39*(1), 44-52. https://doi.org/10.1080/0142159x.2016.1230183

Henning, M., Pinnock, R., & Webster, C. (2017). Does Progress Testing violate the principles of constructive alignment? *Medical Science Educator, 27*(4), 825-829. https://doi.org/10.1007/s40670-017-0459-4

Herrmann, L., Beitz-Radzio, C., Bernigau, D., Birk, S., Ehlers, J. P., Pfeiffer-Morhenn, B., Preusche, I., Tipold, A., & Schaper, E. (2020). Status quo of progress testing in veterinary medical education and lessons learned. *Frontiers in Veterinary Science, 7,* Article 559. https://doi.org/10.3389%2Ffvets.2020.00559

Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item analysis of multiple choice questions at the department of paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos University Medical Journal, 18*(1), e68-e74. https://doi.org/10.18295/squmj.2018.18.01.011

Kibble, J. (2017). Best practices in summative assessment. *Advances in Physiology Education, 41*(1), 110-119. https://doi.org/10.1152/advan.00116.2016

Ministério da Educação. (2016). *Avaliação Nacional Seriada dos Estudantes de Medicina – Documento Básico*. Ministério da Educação.

Norcini, J., Anderson, M. B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Hays, R., Palacios Mackay, M., Roberts, T., & Swanson, D. (2018). 2018 Consensus framework for good assessment. *Medical Teacher, 40*(11), 1102-1109. https://doi.org/10.1080/0142159x.2018.1500016

Oliveira, F. A. M., Martins, M. T., Ferraz, A. M. L., Júnior, Ribeiro, C. G., Oliveira, R. G. de, & Porto, F. R. (2020). Percepção dos acadêmicos de odontologia em relação ao Teste de Progresso. *Revista da Abeno, 20*(2), 26-37. https://doi.org/10.30979/rev.abeno.v20i2.821

Pasquali, L. (2004). *Psicometria dos testes na psicologia e na educação* (5a ed.). Vozes.

Primi, R. (2012). Psicometria: Fundamentos matemáticos da teoria clássica dos testes. *Avaliação Psicológica, 11*(2), 297-307.

Pugh, D., & Regehr, G. (2016). Taking the sting out of assessment: Is there a role for progress testing? *Medical Education, 50*(7), 721-729. https://doi.org/10.1111/medu.12985

Reberti, A. G., Monfredini, N. H., Ferreira, O., Filho, Andrade, D. F. de, Pinheiro, C. E., & Silva, J. C. (2020). Teste de Progresso na escola médica: Uma revisão sistemática acerca da literatura. *Revista Brasileira de Educação Médica, 44*(1), Artigo e015. https://doi.org/10.1590/1981-5271v44.1-20190194

Resolução n. 3, de 20 de junho de 2014. (2014). Institui Diretrizes Curriculares Nacionais do Curso de Graduação em Medicina e dá outras providências. *Diário Oficial da União*, Brasília, DF.

Ristoff, D. (2022). *Mitos e meias-verdades: A educação superior sob ataque*. Insular.

Rosa, M. I. da, Isoppoi, C., Cattaneoi, H., Madeirai, K., Adami, F., & Ferreira, O. F., Filho. (2017). O Teste de Progresso como indicador para melhorias em Curso de Graduação em Medicina. *Revista Brasileira de Educação Médica, 41*(1), 58-68. https://doi.org/10.1590/1981-52712015v41n1RB20160022

Rutgers, D., Raamt, F. van, Lankeren, W. van, Ravesloot, C., Gijp, A. van der, Ten Cate, T. J., & Schaik, J. van. (2018). Fourteen years of progress testing in radiology residency training: Experiences from The Netherlands. *European Radiology, 28*(5), 2208-2215. https://doi.org/10.1007%2Fs00330-017-5138-8

Sá, M. F. S. de, Romão, G. S., Fernandes, C. E., & Silva, A. L. da, Filho. (2021). The Individual Progress Test of Gynecology and Obstetrics Residents (TPI-GO): The Brazilian experience by Febrasgo. *Revista Brasileira de Ginecologia e Obstetrícia, 43*(6), 425-428. https://doi.org/10.1055/s-0041-1731803

Sahoo, D., & Singh, R. (2017). Item and distracter analysis of multiple choice questions (MCQs) from a preliminary examination of undergraduate medical students. *International Journal of Research in Medical Sciences, 5*(12), 5351-5355. http://dx.doi.org/10.18203/2320-6012.ijrms20175453

Sakai, M. H., Ferreira, O. F., Filho, Almeida, M., Mashima, D., & Marchese, M. (2008). Teste de Progresso e avaliação do curso: Dez anos de experiência da medicina da Universidade Estadual de Londrina. *Revista Brasileira de Educação Médica, 32*(2), 254-263. https://doi.org/10.1590/S0100-55022008000200014

Sartor, L. B., Rosa, L., Rosa, M. I. da, Madeirai, K., Uggioni, M. L., Ferreira, O., Filho. (2020). Undergraduate Medical Student's Perception about the Progress Testing. *Revista Brasileira de Educação Medica, 44*(2), Artigo e062. https://doi.org/10.1590/1981-5271v44.2-20190286.ING

Scheffer, M., Cassenote, A., Guerra, A., Guilloux, A. G., Brandão, A. P., Miotto, B. A., Almeida, C. de J., Gomes, J. de O., & Miotto, R. A. (2020). *Demografia médica no Brasil 2020*. Conselho Federal de Medicina.

Scheffer, M., & Dal Poz, M. (2015). The privatization of medical education in Brazil: Trends and challenges. *Human Resources for Health, 13*(1), Article 96. https://doi.org/10.1186/s12960-015-0095-2

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. https://doi.org/10.5116%2Fijme.4dfb.8dfd

Tavakol, M., & Dennick, R. (2013). Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE guide no. 72. *Medical Teacher, 35*(1), e838-e848. https://doi.org/10.3109/0142159X.2012.737488

Tavakol, M., & Dennick, R. (2017). The foundations of measurement and assessment in medical education. *Medical Teacher, 39*(10), 1010-1015. https://doi.org/10.1080/0142159x.2017.1359521

Troncon, L. E. (2019). Exames de licenciamento: Um componente necessário para avaliação externa dos estudantes e egressos dos cursos de graduação em Medicina. *Interface – Comunicação, Saúde, Educação, 24,* Artigo e190576. https://doi.org/10.1590/Interface.190576

Vleuten, C. van der. (2016). Revisiting 'Assessing professional competence: From methods to programmes'. *Medical Education, 50*(9), 885-888. https://doi.org/10.1111/medu.12632

Vleuten, C. van der, Freeman, A., & Collares, C. F. (2018). Progress test utopia. *Perspectives on Medical Education, 7*(2), 136-138. https://doi.org/10.1007/s40037-018-0413-1

Vleuten, C. van der, & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309-317. https://doi.org/10.1111/j.1365-2929.2005.02094.x

Vleuten, C. van der, & Schuwirth, L. W. T. (2019). Assessment in the context of problem-based learning. *Advances in Health Sciences Education, 24*(5), 903-914. https://doi.org/10.1007/s10459-019-09909-1

Vleuten, C. van der, Schuwirth, L. W. T., Muijtjens, A. M. M., Thoben, A. J. N. M., Cohen-Schotanus, J., & Boven, C. P. A. van. (2004). Cross institutional collaboration in assessment: A case on progress testing. *Medical Teacher, 26*(8), 719-725. https://doi.org/10.1080/01421590400016464

Walsh, J. L., Harris, B., Denny, P., & Smith, P. (2018). Formative student-authored question bank: Perceptions, question quality and association with summative performance. *Postgraduate Medical Journal, 94*(1108), 97-103. https://doi.org/10.1136/postgradmedj-2017-135018

Wrigley, W., Vleuten, C. van der, Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues: AMEE guide no. 71. *Medical Teacher, 34*(9), 683-697. https://doi.org/10.3109/0142159x.2012.704437