DOI: 10.1590/0100-6991e-20233636

Avaliação do conhecimento de estudantes de medicina na área de Cirurgia a partir do Teste de Progresso

Assessment of medical students' Surgery knowledge based on Progress Test

Pedro Tadao Hamamoto Filho TCBC-SP[1]; Angélica Maria Bicudo[2]; Gerson Alves Pereira-Júnior TCBC-SP[3].

R E S U M O

O Teste de Progresso (TP) é uma ferramenta de avaliação cujo uso tem crescido em todo o Brasil na última década. O TP permite avaliar o ganho de conhecimento dos estudantes ao longo do curso de graduação e, para que suas interpretações sejam válidas, é preciso que seus itens (questões) tenham qualidade adequada do ponto de vista de validade de conteúdo e confiabilidade de resultados. Neste estudo, analisamos as características psicométricas dos itens e o desempenho dos estudantes na área de cirurgia do TP de 2017 a 2023. Para as análises, usamos os pressupostos da Teoria Clássica dos Testes, a taxonomia de Bloom e o coeficiente de fidedignidade alfa de Cronbach. Os itens se mostraram fáceis (índice de dificuldade média entre 0,3-0,4), com discriminação de regular a boa (índice de discriminação entre 0,3-0,4) e com predomínio de questões de média a alta taxonomia. A confiabilidade se manteve substancial ao longo dos anos (>0,6). O ganho de conhecimento dos estudantes em cirurgia é progressivo e mais importante a partir do 3º ano do curso de graduação, chegando a aproximadamente 70-75% no 6º ano. Este arcabouço de aferições pode ser replicado em outros contextos para melhor compreensão do aprendizado dos estudantes e para qualificação dos processos avaliativos.

Palavras-chave: Cirurgia. Avaliação Educacional. Educação Médica. Psicometria.

INTRODUÇÃO

O Teste de Progresso (TP) é uma avaliação aplicada aos estudantes de medicina visando a análise do ganho consecutivo de conhecimento ao longo do curso. A mesma prova é aplicada a todos os estudantes, de primeiro a sexto ano, sendo provas diferentes a cada aplicação, que tem uma periodicidade fixa e cujo conteúdo é direcionado para o nível do médico recém-formado[1].

O TP foi criado na Holanda e nos Estados Unidos na década de 1970 e visava mudar a cultura de avaliação do processo de ensino-aprendizagem, com o princípio de avaliação longitudinal e acompanhamento da efetividade do processo[2,3]. Hoje, o TP é reconhecido pelo seu potencial de feedback detalhado para estudantes, professores e para a própria escola médica, provendo informações sobre desempenho pessoal, de grupo, de currículo e de instituição[4]. Além disso, o TP reduz o efeito de endogenia das avaliações realizadas dentro de uma mesma escola, pois, trabalhando com consórcios de escolas, há múltiplas fontes de origem de itens (questões)[5]. Por isso mesmo, o TP tem se mostrado útil como preditor do desempenho em exames de certificação profissional ou para residência médica[6,7].

No Brasil, o TP é utilizado desde o final da década de 1990 e início dos anos 2000[8]. Com quase 20 anos de experiência, o Núcleo Interinstitucional de Estudos e Práticas de Avaliação em Ensino Médico (NIEPAEM) é o consórcio que reúne as escolas públicas de medicina do Estado de São Paulo e suas práticas têm sido base para replicação do modelo em todo o Brasil[9].

Tradicionalmente, o TP divide suas questões nas áreas estipuladas para os exames de residência médica: clínica, pediatria, cirurgia, tocoginecologia e saúde coletiva[10]. Essa divisão tem sido questionada por conferir pesos iguais a áreas com extensão de conteúdo heterogênea, o que, em última análise, pode comprometer a confiabilidade da prova em subáreas e, portanto, da própria prova (unpublished data). Ainda assim, dada a série histórica de aplicação do TP, seria possível inferir informações sobre o ensino de cirurgia no Brasil hoje, particularmente nas escolas paulistas. Portanto, o objetivo deste estudo foi analisar a característica dos itens e o desempenho dos estudantes na área de cirurgia do TP de 2017 a 2023.

1 - UNESP - Universidadde Estadual Paulista, Faculdade de Medicina de Botucatu - Botucatu - SP - Brasil 2 - UNICAMP - Universidade Estadual de Campinas, Faculdade de Ciências Médicas - Campinas - SP - Brasil 3 - FOB/USP - Curso de Medicina de Bauru - Bauru - SP - Brasil

MÉTODO

Desenho do estudo

Este é um estudo analítico observacional transversal realizado a partir de informações da base de dados do Núcleo Interinstitucional de Estudos e Práticas de Avaliação em Educação Médica (NIPAEM). Trata-se de um consórcio das seguintes escolas médicas: Universidade Estadual Paulista (UNESP), Universidade de São Paulo (USP, cursos de Ribeirão Preto e Bauru), Universidade Estadual de Campinas (UNICAMP), Universidade Federal de São Paulo (UNIFESP), Universidade Federal de São Carlos (UFSCar), Universidade Estadual de Londrina (UEL), Faculdade de Medicina de São José do Rio Preto (FAMERP) e Faculdade de Medicina de Marília (FAMEMA). Até 2022 o grupo contou com a participação da Universidade Regional de Blumenau (FURB). Trata-se de estudo a partir de base de dados agregados com informações individualizadas em nível de itens (questões). Não há, portanto, informação individualizada de estudantes (sexo/idade). Como código de condutas do NIEPAEM, também não há identificação de desempenho por instituição, evitando-se comparações que levem à classificação das escolas.

Foram incluídos dados das provas aplicadas anualmente a partir de 2017, incluindo a primeira aplicação da prova de 2023 (quando a prova passou a ser aplicada bianualmente). Para os dados psicométricos das questões, foram consideradas apenas as notas dos estudantes do 6º ano de graduação, já que a prova é formulada para o nível do recém-formado. Para análise do progresso de desempenho, foi considerado o desempenho de todos os estudantes. Até 2022, a seção de cirurgia do TP contava com 20 itens por prova. A partir de 2023, a seção passou a contar com 23 itens. A Tabela 1 contém a matriz dos temas das questões. Essa matriz é seguida para elaboração da prova, de modo a garantir semelhança de conteúdos em diferentes aplicações.

Tabela 1 - *Matriz de conhecimentos utilizada para elaboração das questões.*

Subárea	Temas
Princípios gerais da cirurgia	Cuidados perioperatórios Feridas operatórias Conceitos de fios, suturas e nós
Cirurgia geral	Apendicite Hérnias
Cirurgia do aparelho digestivo	Lesões esofágicas, gástricas e colo-retais Fígado, pâncreas e vias biliares Hemorragias digestivas / obesidade
Cirurgia pediátrica	Malformações do trato gastrintestinal Afecções testiculares / fimose
Cirurgia vascular	Doenças arteriais e venosas
Cirurgia torácica	Neoplasias / infecções / afecções pleurais
Urologia	Nefrolitíase Neoplasias do trato gênito-urinário
Ortopedia	Fraturas / lesões articulares, músculo-esqueléticas, ligamentares / lombalgia
Neurocirurgia	Hipertensão intracraniana / trauma crânio-encefálico / trauma raquimedular / malformações do sistema nervoso central
Cirurgia plástica	Queimaduras / enxertos e retalhos
Cirurgia de cabeça e pescoço	Trauma de face / lesões neoplásicas
Oftalmologia	Trauma ocular / olho vermelho / redução de acuidade visual
Anestesiologia	Tipos de anestesia / farmacologia / avaliação pré-anestésica / complicações anestésicas / dor
Medicina de urgência	Princípios do ATLS, ressuscitação cardio-pulmonar Trauma torácico, abdominal, pélvico e vascular

ATLS: Advanced Life Trauma Support.

Para as análises, foram investigados o índice de dificuldade e discriminação dos itens (de acordo com a Teoria Clássica dos Testes), a classificação taxonômica das questões, e o coeficiente de confiabilidade da prova (medido pelo coeficiente alfa de Cronbach).

Considerações éticas

Por e tratar de estudo de base de dados disponibilizados de forma agregada sem possibilidade de identificação individual de estudantes, este estudo dispensa de apreciação por comitê de ética em pesquisa, de acordo com a legislação da Comissão Nacional de Ética em Pesquisa com Seres Humanos (CONEP)[11].

Análise de dados

O nível de dificuldade de cada item foi calculado como porcentagem de erros em cada item (ou seja, quanto mais próximo de 1, mais difícil é a questão). Para classificação do grau de dificuldade de cada item, adotaram-se os seguintes valores: acima de 0,8 – difícil; entre 0,4 e 0,8 – média; abaixo de 0,4 – fácil.

O índice de discriminação foi calculado pela diferença de acerto para cada item entre os 27% dos estudantes de desempenho superior na prova e os 27% de desempenho inferior. Assim, o índice pode variar de -1 a 1, sendo que, quanto mais próximo de 1, melhor a discriminação. Adotaram-se os seguintes valores para classificação das questões: ≥0,4 – boa; ≥0,3 e abaixo de 0,4 – regular; ≥0,2 e abaixo de 0,3 – fraca; <0,2 – deficiente.

O coeficiente alfa de confiabilidade de cada prova foi calculado de acordo com a fórmula proposta por Cronbach[12]. Refere-se à consistência interna da medida, ou seja, à extensão pela qual os itens medem um mesmo construto. Adotou-se a seguinte classificação: >0,8 – quase perfeita; de 0,8 a 0,61 – substancial; de 0,6 a 0,41 – moderada; de 0,4 a 0,21 – razoável; <0,21 – pequena.

De acordo com a Taxonomia de Bloom, modificada posteriormente por Anderson e Krathwol, os domínios educacionais cognitivos podem ser classificados de acordo com a complexidade de processos cognitivos em: conhecimento, compreensão, aplicação, análise, síntese e avaliação[13,14]. Para classificação taxonômica dos itens, eles foram classificados de acordo com o repertório cognitivo envolvido em sua resolução como de baixa (memorização), média (compreensão) ou alta taxonomia (aplicação/análise).

O desempenho dos estudantes foi calculado como função do percentual de acertos médio para cada ano de graduação.

Para análise da tendência temporal dos indicadores psicométricos, foi realizada regressão linear simples. Na análise da diferença de desempenho dos estudantes, foi realizado teste ANOVA de uma via seguido do teste de Tukey para comparações pareadas entre anos subsequentes do curso de graduação. Considerou-se $p<0,05$ para determinação de significância estatística.

As análises foram realizadas com os softwares GraphPad v. 9.5.0 (GraphPad Software Inc. San Diego, CA, EUA) e SPSS (Statistical Package for Social Sciences, IBM Corp., Armonk, NY, EUA).

RESULTADOS

Com relação à dificuldade dos itens, observamos que na média anual, os itens foram fáceis, com índice de dificuldade média variando entre 0,3 e 0,4 (Figura 1). Apenas a prova de 2019 apresentou média de dificuldade mais alta, próxima de 0,6. Já com relação à discriminação dos itens, a média apontou para discriminação regular (entre 0,3 e 0,4), sendo as provas de 2019, 2021 e 2023 as provas com índice próximo ou superior a 0,4 (boa discriminação, Figura 1). A análise de confiabilidade da prova medida pelo coeficiente de fidedignidade (alfa de Cronbach) mostrou que, exceto a prova de 2017, todas tiveram valor maior ou igual a 0,6 (consistência interna substancial, Figura 1). Na análise de tendência temporal dos indicadores, todos demonstraram estabilidade: coeficientes angulares baixos e estatisticamente não significativo (Tabela 2).

Na análise da classificação taxonômica dos itens, observamos predomínio de questões de taxonomia média a alta (Figura 2), ou seja, há poucos itens que enfatizam memorização de conteúdos e conceitos e mais itens que requerem maior complexidade cognitiva, com compreensão, análise e aplicação de conteúdos.

Sobre o desempenho dos estudantes, observamos que houve ganhos progressivos a cada

ano de graduação, partindo de uma média de acertos de 25 a 35% no primeiro ano, chegando a 70-75% no sexto ano. Na comparação de anos subsequentes de graduação, a diferença de desempenho foi diferente para praticamente todos os anos, à exceção da comparação de 1º e 2º anos, que se mostrou não diferente nas aplicações da prova em 2017, 2018, 2021 e 2022 (Figura 3).

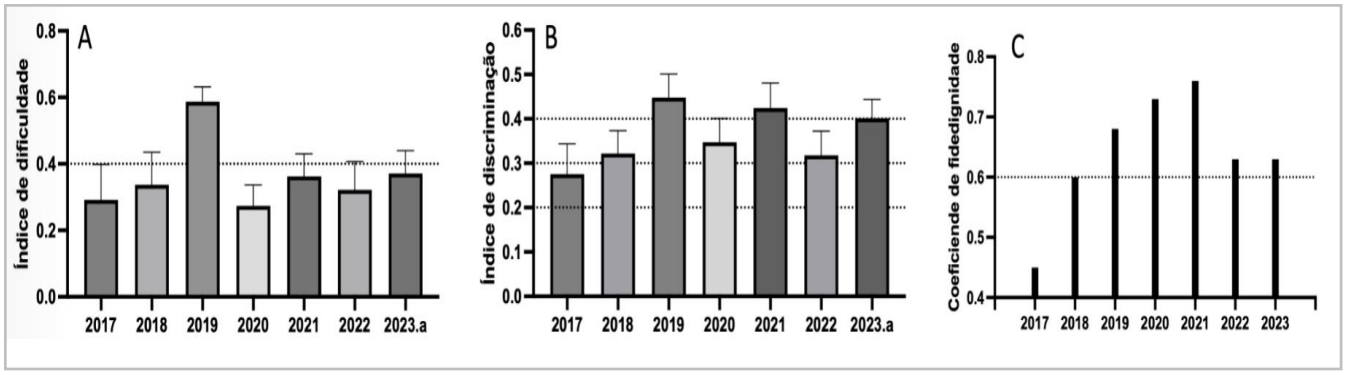

Figura 1. *Valores de média com respectivos intervalos de confiança a 95% para os índices de dificuldade (A) e discriminação (B) dos itens de cirurgia de cada aplicação do teste de progresso nos anos de 2017 a 2023. C: valores do coeficiente de fidedignidade da área de cirurgia para os mesmos anos.*

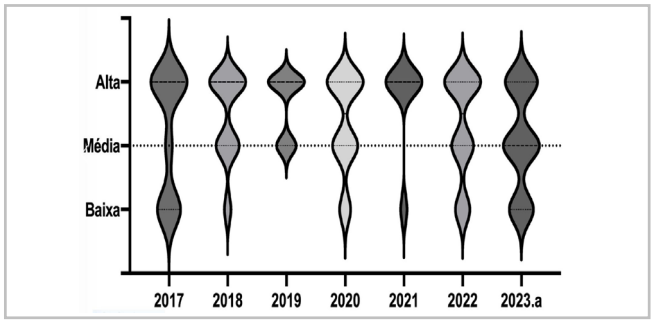

Figura 2. *Gráfico de violino para a quantidade de itens por classificação taxonômica. O maior diâmetro do violino indica maior concentração de itens naquela classificação.*

DISCUSSÃO

O TP tem sido utilizado de forma crescente nas escolas médicas brasileiras. Frente a inúmeras discussões sobre a importância de avaliações externas e seriadas dos estudantes de medicina, o TP se mostra como uma ferramenta útil por permitir diagnósticos sobre o desempenho de estudantes e o comportamento do currículo e em última análise, da efetividade do processo ensino-aprendizagem[15,16].

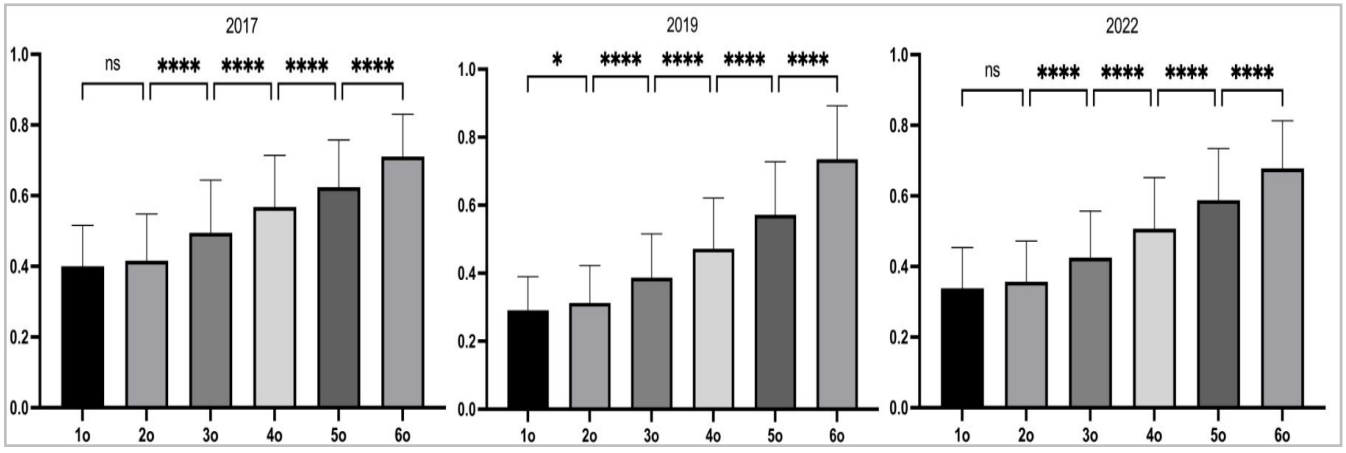

Figura 3. *Desempenho dos estudantes na área de cirurgia do teste de progresso exemplificado nas provas de 2017, 2019 e 2022. A comparação do desempenho entre séries subsequentes sempre é significativa, à exceção da comparação entre 1º e 2º anos. *p<0,05. ***p<0,0001. ns: não significativo.*

Tabela 2 - *Resultados da regressão linear para tendência de comportamento dos indicadores psicométricos da área de cirurgia do teste de progresso de 2017 a 2023.*

Indicador	Coeficiente angular*	p-valor
Índice de dificuldade	≅0,00%	0,978
Índice de discriminação	1,24%	0,346
Coeficiente de fidedignidade	2,43%	0,236

**Altos valores do coeficiente angular indicam tendência de aumento ao longo do tempo. Valores negativos indicam tendência de redução. Valores próximos de zero sugerem estabilidade.*

Hás válidas críticas ao TP com relação à sua limitação para inferências a cerca de disciplinas específicas. Por exemplo, o erro do estudante em uma única questão de anatomia não significa que o aluno não tenha as competências cognitivas necessárias sobre anatomia. Por isso, para análises mais acuradas a cerca de uma área em particular, são necessárias abordagens mais aprofundadas sobre os resultados do teste, ou aumento da amostragem da área em questão aumentando o número de seus itens na prova[17].

Assim, acreditamos que este trabalho, analisando provas equivalentes na área de cirurgia (que compõe de 1/6 a 1/5 do TP) ao longo sete aplicações, mantendo fixa a matriz de conhecimentos abordados, permite algumas conclusões.

Primeiramente, observamos que a prova de cirurgia não é uma prova difícil. As questões têm índices de dificuldade média baixos (abaixo de 0,4) e que têm se mantido estável ao longo dos anos. Sabendo-se que os conteúdos cobrados na prova estão no nível do médico recém-formado, trata-se de um bom indicador.

Essa observação é reforçada pelo índice médio de discriminação, que também tem se mantido estável e acima de 0,3, portanto, com itens de regular a boa discriminação. A interpretação deste indicador é a de que os itens têm sido adequados em detectar estudantes de bom ou mau desempenho. Portanto, mesmo em questões fáceis, podem-se identificar estudantes com desempenho insatisfatório e que merecem atenção de suas escolas e mentores. Devemos ressaltar que as provas de 2019, 2021 e 2022 foram provas que utilizaram itens pré-testados – ou seja, eram itens já utilizados em versões anteriores do TP e foram escolhidos com base em seu bom comportamento psicométrico. Não por acaso, portanto, estas foram as provas em que se observaram melhores médias de discriminação dos itens.

Esta observação tem uma importante implicação prática para as escolas médicas. Habitualmente, professores repetem questões de prova – seja por falta de organização de um banco de itens, seja por limitação da criatividade para redação de itens inéditos[18-20]. É recomendável que itens sejam reutilizados com intervalo de tempo suficiente para minimizar vieses de memorização de questões entre estudantes e seus pares. Com nossos dados, chamamos a atenção para que os professores, além de garantir intervalo na aplicação de itens repetidos, estudem o comportamento psicométrico das questões após seu uso, identificando questões muito fáceis, muito difíceis ou que não tenham boa discriminação, ou seja, que não contribuem para uma avaliação adequada[21]. A análise da psicometria de itens pode ser feita de modo automático nas plataformas de avaliação online, cujo uso ganhou notoriedade com a pandemia de COVID-19[22].

O coeficiente de fidedignidade também é um indicador de que a prova de cirurgia do TP é de boa qualidade, com índices que têm se mantido acima de 0,6 e, portanto, com consistência interna substancial e comparável aos valores obtidos internacionalmente[23]. Obviamente, o conjunto de todos os itens que compõem a prova eleva o coeficiente para valores acima de 0,8-0,9 (unpublished data). Deve-se reconhecer, porém, que este coeficiente é influenciado pela variância do score e, portanto, pelo número de respondentes[21]. Como temos estudantes de nove escolas, o tamanho amostral é grande e eleva naturalmente, a variância de respostas e o valor do alfa de Cronbach. No entanto, o conjunto dos diversos indicadores psicométricos tomados em conjunto sugerem que a avaliação tem tido uma qualidade satisfatória.

A este conjunto de indicadores, acrescentamos a categoria taxonômica dos itens, com predomínio de

questões de média a alta taxonomia. Anteriormente, já foi demonstrado que itens de maior taxonomia têm melhor índice de discriminação do que itens de baixa taxonomia[24,25]. Assim, os itens de cirurgia do TP têm cobrado dos estudantes mais raciocínio clínico do que memorização de fatos ou conceitos, aproximando-se, possivelmente, de domínios cognitivos mais próximos da prática clínica do médico recém-formado.

Finalmente, sobre o desempenho dos estudantes de medicina, observamos ganhos conhecimento ao longo do curso, como esperado. O dado interessante é o de que o ganho é significativo a partir do 3º ano, o que reflete a realidade do currículo da maioria das escolas médicas brasileiras, no qual o ensino de técnica e clínica cirúrgica ocorrem mais frequentemente no 3º ano. Recentemente, demonstrou-se que a exposição curricular de estudantes a conteúdos de cirurgia melhora seu desempenho em itens de cirurgia no TP, embora o desempenho ao final do curso seja semelhante entre estudantes, independente do desenho curricular[26]. O fato de os estudantes do 6º ano apresentarem taxa de acerto média próxima a 70-75% é comparável ao das outras áreas do conhecimento e ao que se reporta na literatura internacional sobre o TP[2,27,28].

Este estudo não está isento de limitações. Pela própria natureza do TP, não podemos inferir conclusões sobre o aprendizado específico em cada especialidade cirúrgica. Obviamente, por se tratar de uma avaliação de conhecimentos, também não é possível fazer qualquer inferência sobre o ensino e aprendizagem de habilidades cirúrgicas básicas que todo médico deve ter e nem sobre atitudes profissionais. Há que se ressaltar, também, que os dados da prova de 2023 correspondem à prova aplicada ao final do primeiro semestre, e não ao final do ano, já que neste ano a periodicidade da prova foi aumentada para duas vezes ao ano. Além disso, não dispomos de informação a nível individual dos estudantes para fazer outras inferências a partir de covariáveis como sexo, idade e instituição. Como código de condutas do NIEPAEM, as informações de desempenho de cada estudante são disponíveis apenas à própria escola do estudante, e não ao grupo de escolas.

A despeito dessas limitações, este estudo fornece informações úteis sobre a qualidade do TP para avaliação de conhecimentos de cirurgia e fornece mais evidências sobre a curva de ganho de conhecimentos dos estudantes de medicina. Em conjunto, apresentamos um arcabouço de aferições de qualidade de avaliação que podem ser repetidos em outros contextos para qualificar a avaliação do estudante de medicina.

CONCLUSÃO

Os itens de cirurgia que compõem o Teste de Progresso do NIEPAEM não são difíceis, têm boa discriminação, privilegiam raciocínio clínico e produzem bons indicadores de confiabilidade. O ganho de conhecimento dos estudantes é significativo a partir do 3º ano do curso de graduação e chega a aproximadamente 70-75% ao 6º ano.

ABSTRACT

Progress Testing (PT) is an assessment tool whose use has grown throughout Brazil in the last decade. PT makes it possible to assess the students' knowledge gain throughout the undergraduate course and, for their interpretations to be valid, their items (questions) must have adequate quality from the point of view of content validity and reliability of results. In this study, we analyzed the psychometric characteristics of the items and the performance of students in the content area of surgery from 2017 to 2023. For the analyses, we used the assumptions of Classical Test Theory, Bloom's taxonomy and Cronbach's alpha reliability coefficient. The items were easy (average difficulty index between 0.3-0.4), with fair to good discrimination (discrimination index between 0.3-0.4) and with a predominance of medium to high taxonomy. Reliability remained substantial over the years (>0.6). Students' knowledge gain in surgery was found to be progressive and more important from the 3rd year of the undergraduate course, reaching approximately 70-75% in the 6th year. This measurements framework can be replicated in other contexts for a better understanding of student learning and for qualification of evaluation processes.

Keywords: *Surgery. Educational Measurement. Medical Education. Psychometrics.*

REFERÊNCIAS

1. Schuwirth LW, van der Vleuten CP. The use of progress testing. Perspect Med Educ. 2012;1(1):24-30. doi: 10.1007/s40037-012-0007-2.
2. Van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem based learning curriculum. Med Teach. 1996;18(2):103-9. doi: 10.3109/01421599609034142.
3. Arnold L, Willoughby TL. The quarterly profile examination. Acad Med. 1990;65(8):515-6. doi: 10.1097/00001888-199008000-00005.
4. Coombes L, Ricketts C, Freeman A, Stratford J. Beyond assessment: feedback for individuals and institutions based on the progress test. Med Teach. 2010;32(6):486-90. doi: 10.3109/0142159X.2010.485652.
5. Muijtjens AM, Schuwirth LW, Cohen-Schotanus J, van der Vleuten CP. Origin bias of test items compromises the validity and fairness of curriculum comparisons. Med Educ. 2007;41(12):1217-23. doi: 10.1111/j.1365-2923.2007.02934.x.
6. Karay Y, Schauber SK. A validity argument for progress testing: Examining the relation between growth trajectories obtained by progress tests and national licensing examinations using a latent growth curve approach. Med Teach. 2018 Nov;40(11):1123-9. doi: 10.1080/0142159X.2018.1472370.
7. Hamamoto Filho PT, de Arruda Lourenção PLT, do Valle AP, Abbade JF, Bicudo AM. The Correlation Between Students' Progress Testing Scores and Their Performance in a Residency Selection Process. Med Sci Educ. 2019;29(4):1071-5. doi: 10.1007/s40670-019-00811-4.
8. Tomic ER, Martins MA, Lotufo PA, Benseñor IM. Progress testing: evaluation of four years of application in the school of medicine, University of São Paulo. Clinics (Sao Paulo). 2005;60(5):389-96. doi: 10.1590/s1807-59322005000500007.
9. Bicudo AM, Hamamoto Filho PT, Abbade JF, Hafner MLMB, Maffei CML. Consortia of Cross-Institutional Progress Testing for All Medical Schools in Brazil. Rev Bras Educ Med. 2019;43(4):151-6. doi: 10.1590/1981-52712015v43n4RB20190018.
10. Ministério da Educação. Resolução no 01, de 14 de agosto de 2000. [Acesso em 18 Jul 2023]. Disponível em: <https://www.gov.br/mec/pt-br/residencia-medica/ementario-da-legislacao-da-residencia-medica>.
11. Conselho Nacional de Saúde. Resolução no 674 de 06 de maio de 2022. [acesso em 05 Ago 2023]. Disponível em: <https://conselho.saude.gov.br/resolucoes-cns/2469-resolucao-n-674-de-06-de-maio-de-2022>.
12. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297-334. doi: 10.1007/BF02310555.
13. Bloom BS. Taxonomy of educational objectives: the classification of education goals. Cognitive domain. Handbook 1. New York: Longman; 1956. ISBN-10: 0679302093; ISBN-13: 978-0679302094
14. Anderson LW, Krathwohl DR, Airasian PW, et al. A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives. New York: Addison Wesley Longman; 2001.
15. Cecilio-Fernandes D, Bicudo AM, Hamamoto Filho PT. Progress testing as a pattern of excellence for the assessment of medical students' knowledge - concepts, history, and perspective. Medicina (Ribeirão Preto). 2021;54(1):e-173770. doi: 10.11606/issn.2176-7262.rmrp.2021.173770.
16. Troncon LEA, Elias LLK, Osako MK, Romão EA, Bollela VR, Moriguti JC. Reflections on the use of the Progress Test in the programmatic student assessment. Rev Bras Educ Med. 2023;47(2):e076. doi: 10.1590/1981-5271v47.2-2022-0334.ing.
17. Swanson DB, Case SM. Assessment in basic science instruction: directions for practice and research. Adv Health Sci Educ Theory Pract. 1997;2(1):71-84. doi: 10.1023/A:1009702226303.
18. Boulet JR, McKinley DW, Whelan GP, Hambleton RK. The effect of task exposure on repeat candidate scores in a high-stakes standardized patient assessment. Teach Learn Med. 2003;15(4):227-32. doi: 10.1207/S15328015TLM1504_02.
19. Wood TJ. The effect of reused questions on repeat examinees. Adv Health Sci Educ Theory Pract. 2009;14(4):465-73. doi: 10.1007/s10459-008-

9129-z.

20. O'Neill TR, Sun L, Peabody MR, Royal KD. The Impact of Repeated Exposure to Items. Teach Learn Med. 2015;27(4):404-9. doi: 10.1080/10401334.2015.1077131.
21. Albanese M, Case SM. Progress testing: critical analysis and suggested practices. Adv Health Sci Educ Theory Pract. 2016;21(1):221-34. doi: 10.1007/s10459-015-9587-z.
22. Patael S, Shamir J, Soffer T, Livne E, Fogel-Grinvald H, Kishon-Rabin. Remote proctoring: Lessons learned from the COVID-19 pandemic effect on the large scale on-line assessment at Tel Aviv University. J Comput Assist Learn. 2022;38(6):1554–73. doi: 10.1111/jcal.12746.
23. Blake JM, Norman GR, Keane DR, Mueller CB, Cunnington J, Didyk N. Introducing progress testing in McMaster University's problem-based medical curriculum: psychometric properties and effect on learning. Acad Med. 1996;71(9):1002-7. doi: 10.1097/00001888-199609000-00016.
24. Rush BR, Rankin DC, White BJ. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. BMC Med Educ. 2016;16(1):250. doi: 10.1186/s12909-016-0773-3.
25. Hamamoto Filho PT, Silva E, Ribeiro ZMT, Hafner MLMB, Cecilio-Fernandes D, Bicudo AM. Relationships between Bloom's taxonomy, judges' estimation of item difficulty and psychometric properties of items from a progress test: a prospective observational study. Sao Paulo Med J. 2020;138(1):33-9. doi: 10.1590/1516-3180.2019.0459.R1.19112019.
26. Wearn A, Bindra V, Patten B, Loveday BPT. Relationship between medical programme progress test performance and surgical clinical attachment timing and performance. Med Teach. 2023;45(8):877-84. doi: 10.1080/0142159X.2023.2186205.
27. Nouns ZM, Georg W. Progress testing in German speaking countries. Med Teach. 2010;32(6):467-70. doi: 10.3109/0142159X.2010.485656.
28. Cecilio-Fernandes D, Aalders WS, Bremers AJA, Tio RA, de Vries J. The Impact of Curriculum Design in the Acquisition of Knowledge of Oncology: Comparison Among Four Medical Schools. J Cancer Educ. 2018;33(5):1110-4. doi: 10.1007/s13187-017-1219-2.

Recebido em: 06/08/2023
Aceito para publicação em: 14/10/2023
Conflito de interesses: não.
Fonte de financiamento: nenhuma.

Endereço para correspondência:
Pedro Tadao Hamamoto Filho
E-mail: pedro.hamamoto@unesp.br