RBEM Revista Brasileira de Educação Médica

ARTIGO ORIGINAL

DOI: https://doi.org/10.1590/1981-5271v46.supl.1-20220302

Onze anos de Teste de Progresso na Unicamp: um estudo sobre a validade do teste

Eleven years of Progress Testing at Unicamp: a test validity study

Ruy Guilherme Silveira de Souza[1] ruysouza28@gmail.com
Angélica Maria Bicudo[2] angelicabicudo@gmail.com

RESUMO

Introdução: O curso de Medicina da Universidade Estadual de Campinas (Unicamp) vem pondo à prova o aquisição cumulativa cognitiva de seus alunos por meio do Teste de Progresso (TP) há mais de uma década, de modo a possibilitar a análise da utilidade do exame como estratégia de apoio a decisões pedagógicas e apontar principais ameaças à validade dele.

Objetivo: Este estudo teve como objetivos oferecer a análise de validade do TP e explicitar as oportunidades de utilização do teste especialmente para a determinação de padrões de suficiência cognitiva para a progressão no curso e ao final deste, e a identificação de estudantes em risco.

Método: Trata-se de estudo observacional retrospectivo de uma série histórica de sucessivos testes escritos realizados para analisar o acúmulo cognitivo no período de 2006 a 2016, totalizando 11 anos e seis turmas consecutivas. Em cada momento de medida (aplicação do teste), o estudo utilizou um modelo misto, em que a exposição (realização do teste) e o desfecho (escore do teste) foram avaliados no mesmo ponto de tempo, o que caracteriza um estudo transversal (cross-sectional) cujos resultados sucessivos originarão as curvas de crescimento cognitivo

Resultado: Observou-se um acúmulo cognitivo em torno de 6 pontos percentuais a cada nova testagem. Os estudantes ao completarem o sexto ano obtiveram um acerto de cerca de 65,7% (± 9,1). A cada testagem, determinou-se um "efeito piso" para identificar alunos com rendimento abaixo da média, que em geral se situou em cerca de 1,5 DP abaixo da média da respectiva turma.

Conclusão: O TP-Unicamp oferece dados confiáveis para apoiar importantes decisões pedagógicas, tais como identificação de alunos em risco acadêmico por baixa performance, critérios para progressão e desempenho cognitivo ao final do curso. Como confiabilidade sofre influência da amostragem, e o aumento do número de itens de cada teste e o aumento da frequência de testagem podem ser estratégias a serem tomadas para superar essas limitações.

Palavras-chave: Educação Médica; Avaliação de Aprendizagem; Validade; Teste de Progresso.

ABSTRACT

Introduction: *The Medical School of the State University of Campinas (UNICAMP) has been testing the cumulative cognitive acquisition among its students through the Progress Test (PT) for over a decade, making it possible to analyze the utility of the test as a strategy to support pedagogical decisions and to point out the main threats to its validity.*

Objective: *To provide an analysis of the validity of the PT, explaining opportunities for its use, especially in determining standards of cognitive sufficiency for progression, standards of cognitive sufficiency at the end of the course, and identification of students at risk.*

Method: *Retrospective observational study of historical series of successive written tests performed to analyze cognitive accumulation, covering a period from 2006 to 2016, totaling 11 years and 6 consecutive classes. At each instance of measurement (test application), the study uses a mixed model where exposure (test performance) and outcome (test score) are evaluated at the same time point, characterizing a cross-sectional study, the successive results of which will generate the cognitive growth curves.*

Result: *A cognitive accumulation of around 6 percentage points was observed with each new test. Students, upon completing the 6th year, scored around 65.7% (± 9.1). A "floor effect" was determined for each test to identify students with below-average performance, which in general was about 1.5 SD below the average of the respective class.*

Conclusion: *TP-Unicamp offers reliable data to support important pedagogical decisions, such as the identification of students at academic risk for low performance, criteria for progression and cognitive performance at the end of the course. As reliability is influenced by sampling, increasing the number of items in each test and increasing the frequency of tests could represent strategies to overcome potential limitations.*

Keywords: *Medical Education; Learning Assessment; Validity; Progress Test.*

[1]Universidade Federal de Roraima, Boa Vista, Roraima, Brasil.
[2]Universidade Estadual de Campinas, Campinas, São Paulo, Brasil.

Editor: Aristides Augusto Palhares Neto.

Recebido em 18/09/22; Aceito em 17/10/22.

Avaliado pelo processo de *double blind review.*

INTRODUÇÃO

O Teste de Progresso (TP) vem se tornando uma das formas mais populares de avaliação do ensino médico em todo o mundo[1]. A escola de Medicina da Universidade Estadual de Campinas (Unicamp) vem testando a aquisição cumulativa cognitiva de seus alunos há mais de uma década, e o volume de dados gerados já possibilita oferecer uma análise consistente da validade do exame no âmbito da formação médica na graduação. Inicialmente é necessário apresentar a justificação epistêmica que fundamenta a visão de validade neste trabalho. Validade é o atributo mais importante para qualquer tipo de teste e sob a qual todas as demais qualidades ficam em estado de dependência[2-6]. No entanto, esse conceito no campo da pesquisa em avaliação de aprendizagem na educação médica ainda é confuso e fragmentado, colocando a validade no mesmo patamar de outros aspectos da testagem ou até mesmo em posição inferior, que na verdade lhe são subalternos. Uma das razões para isso está no rápido crescimento que a pesquisa em avaliação no ensino médico experimentou desde a década de 1970. Apenas para que se possa ter uma noção desse crescimento, o estudo pioneiro sobre exame clínico estruturado para avaliação de competência publicado em 1979 já foi seguido de mais de 1.800 artigos sobre o mesmo método[7]. Esse crescimento se caracterizou por um foco excessivo na testagem e em seus resultados imediatos, desviando a atenção dos pesquisadores na direção dos testes padronizados e das novas técnicas psicométricas voltadas para a obtenção de escores fidedignos fortemente ancorados em análises estatísticas cada vez mais complexas e inacessíveis em um grande número de instituições[8], transformando a fidedignidade da medida na base para a generalização da utilização de um teste. Com isso, um teste que apresentasse por exemplo um alfa de Cronbach com uma boa consistência interna ganhava credenciais para ser replicado em diversas escolas, de diferentes contextos, sem nenhum outro tipo de reflexão. Royal[9], ao realizar uma busca geral no banco de dados PubMed sobre estudos de avaliação do ensino em estudantes de Medicina, encontrou a expressão "instrumento confiável" mais de duas mil vezes em um período de apenas cinco anos. Souza et al.[10], em uma análise bibliométrica sobre o atual conceito de validade na educação médica, concluíram que 96% dos estudos ainda apresentam um conceito fragmentado de validade, limitando-se basicamente aos resultados do processo de testagem, sem explicitar uma análise das consequências e utilidade dos testes. Essa fragmentação do conceito decorre da confusão que frequentemente acontece entre os termos "validação" e "validade". Validação representa o processo de coleta de dados (evidências) obtidos pelos resultados de um teste[3], mas isso é somente parte do processo que só se completa quando, após a análise de todas as fontes de evidência, define-se como elas apoiam a interpretação e a utilidade de um teste; em outras palavras, sem a explicitação da utilidade e das consequências do teste, não é possível um julgamento de validade. Messick, um dos pontas de lança do movimento pela volta de um conceito unificado de validade, propunha um processo de análise de validade composto por duas etapas: a primeira, de natureza científica, deveria se preocupar com as propriedades psicométricas do teste, e a segunda, de natureza ética, deveria focar a análise extensa das consequências do teste "em termos de valores humanos"[11].

O futuro da educação médica, especialmente diante dos desafios criados pela pandemia da *coronavirus disease 2019* (Covid-19), aponta para novas e necessárias transformações no campo da avaliação. Faz-se necessário que a pesquisa em avaliação do ensino médico resgate a importância dessa "natureza ética" proposta por Messick, desviando um pouco o foco dos resultados numéricos para a "narrativa" do processo, de modo a possibilitar uma concepção ampla de avaliação que só termina na explicitação da utilidade e das consequências da testagem, em resumo: "uma mudança de números para palavras"[12].

Neste trabalho, propõe-se uma análise de validade do TP ancorada em uma visão unificada que explicite oportunidades de utilização do teste no dia a dia do ensino de graduação, tais como a definição de parâmetros que possibilitem a determinação de padrões de suficiência cognitiva para a progressão do estudante ao longo do curso e ao final deste, e a identificação de discentes em risco.

MÉTODO

Trata-se de um estudo retrospectivo de uma série histórica de sucessivos testes escritos de oferta anual única, realizados para analisar o acúmulo cognitivo de estudantes do curso de Medicina da Unicamp, no período de 2006 a 2016, totalizando 11 anos e seis turmas consecutivas de Medicina, computando um total de 574 alunos que concluíram sua formação médica entre 2011 e 2016. As curvas de crescimento cognitivo de cada turma foram consolidadas para uma visão global do crescimento cognitivo da escola, a fim de permitir uma melhor generalização das interpretações.

Designa-se "testagem 1" a primeira avaliação por meio do TP de cada turma testada e assim sucessivamente até a "testagem 6". No período compreendido entre 2006 a 2016, seis turmas completaram sua formação, com seis momentos de testagem. O quadro 1 demonstra as sucessivas turmas testadas no período de 2006 a 2016.

 DOI: https://doi.org/10.1590/1981-5271v46.supl.1-20220302

Quadro 1. Número de testagens e turmas testadas no período de 2006 a 2016

Ano	Testagem 1	Testagem 2	Testagem 3	Testagem 4	Testagem 5	Testagem 6	Turmas
2006	95						
2007	89	81					
2008	92	89	88				
2009	100	88	94	93			
2010	98	102	91	92	92		
2011	88	96	101	96	90	89	**Turma 1** (2006-2011)
2012		86	99	92	91	92	**Turma 2** (2007-2012)
2013			91	98	93	89	**Turma 3** (2008-2013)
2014				96	91	97	**Turma 4** (2009-2014)
2015					98	90	**Turma 5** (2010-2015)
2016						100	**Turma 6** (2011-2016)

Fonte: Teste de Progresso Unicamp.

Como se trata de uma série histórica e, portanto, sem a possibilidade de determinação prévia de critérios para cada prova, a fim de minimizar as diferenças entre os diversos graus de dificuldade dos diferentes testes, os escores individuais de cada turma foram organizados a partir de normorreferenciamento[13]. A partir da consolidação da distribuição de frequências, foi possível determinar um "efeito teto" para alunos com rendimento de 1,5 a 2 desvios padrões acima da média, bem como um "efeito piso" para alunos com rendimento de 1,5 a 2 desvios padrões abaixo da média.

O TP-Unicamp utiliza como instrumento uma prova objetiva, do tipo múltipla escolha, composta de 120 itens. Quanto ao número de alternativas, até 2011 o TP utilizava cinco, e, a partir de 2012, cada item passou a ter quatro alternativas no formato "melhor resposta individual" e sem estratégias para acertos casuais ("chutes"). O conteúdo do exame abrange todo a área de conhecimento esperado para um aluno concluinte do curso e aplicado simultaneamente para todas as séries.

Em cada teste, obteve-se uma média global do teste geral e de cada área de conhecimento clínico (isto é, cirurgia, clínica Médica, ginecologia e obstetrícia, e pediatria).

O escore é a métrica da prova, como a soma dos pontos dos acertos, ou como a soma relativa, ou ainda como a soma média.

Por tratar-se de um estudo com medidas repetidas ao longo de uma escala de tempo, adotou-se um modelo de regressão linear de efeitos mistos para o escore relativo da prova. Para esse modelo, considerou-se como variável dependente (desfecho) o escore médio relativo de cada uma das seis turmas. As testagens (cada construção de prova, desde 2006 até 2016) foram consideradas como variáveis (ordinais) preditoras (efeito fixo).

A correlação de Pearson foi utilizada para determinar a contribuição de cada área de conhecimento analisada em relação à prova toda e nas relações interáreas. O alfa de Cronbach foi utilizado como medida de fidedignidade (confiabilidade), em termos dos domínios e do construto geral, a fim de analisá-los quanto à sua capacidade de produzir resultados precisos.

Para redução dos erros de medida na construção da curva de crescimento cognitivo, utilizou-se a variação do percentual de acertos em todo o teste, convertidos para um escore Z, definido como o valor de cada indivíduo em cada ano subtraindo a média e dividindo pelo desvio padrão da turma respectiva, como estratégia de controle da dificuldade do exame em diferentes testes e turmas.

O estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de Roraima (UFRR):

Certificado de Apresentação para Apreciação Ética (CAAE): nº 09314919.8.0000.5302, nº 3.206.008.

RESULTADOS

A Tabela 1 apresenta os escores relativos das seis turmas testadas ao longo de 11 anos.

A Tabela 2 apresenta os valores do alfa de Cronbach obtidos para cada momento de testagem (do primeiro ao sexto ano). Os valores de alfa de Cronbach são inferiores a 0,7, porém apresentam crescimento e ficam limítrofes após a terceira testagem.

O Gráfico 1 apresenta os perfis (dispersão) obtidos na análise longitudinal e a curva de crescimento obtida a partir dos escores relativos para cada uma das seis turmas que realizaram sua formação no período analisado.

Na Tabela 3, são apresentados os ajustes para a análise utilizando o modelo de regressão linear de efeitos mistos para o escore relativo da prova. Com exceção das turmas de 2013 e 2014, todas as variáveis foram significativamente associadas com o escore.

Com base no modelo longitudinal linear misto, observou-se um acúmulo cognitivo em torno de 6 pontos percentuais a cada nova testagem. Pediatria e ginecologia e obstetrícia aumentaram, em média, 9 pontos a cada nova testagem, e clínica médica e cirurgia podem aumentar, em média, 8 pontos a cada nova testagem.

Os estudantes ao completarem o sexto ano obtiveram um acerto de cerca de 65,7% (± 9,1), com um intervalo de confiança de 95% de 65,6545 a 65,8206, com variações de desempenho perto de 30% de acerto até, aproximadamente de 93% de acerto.

O "efeito piso/efeito teto" é utilizado como uma ideia de um alarme para os estudantes que necessitem de remediação.

Tabela 1. Medidas descritivas do escore relativo das provas realizadas pelas seis turmas, pelo ano de formatura, durante o período de 2006 a 2016

Testagem	Ano de testagem	Ano de formatura	Média	Desvio padrão da média	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo
1	2006	2011	32	4,78	23	29	32	36	50
	2007	2012	39	4,85	25	36	38	43	48
	2008	2013	33	5,32	18	28	33	37	44
	2009	2014	31	4,78	20	28	31	34	44
	2010	2015	31	4,60	21	28	31	34	44
	2011	2216	36	4,44	27	33	37	39	46
2	2007	2011	43	5,47	31	40	43	47	55
	2008	2012	33	6,29	21	28	33	38	48
	2009	2013	34	5,94	18	29	35	38	49
	2010	2014	32	5,15	20	28	32	36	43
	2011	2015	38	5,92	23	34	38	42	54
	2012	2216	34	4,48	24	31	33	36	45
3	2008	2011	42	7,16	17	38	43	46	58
	2009	2012	39	6,91	22	34	38	44	53
	2010	2013	39	6,92	22	33	38	44	56
	2011	2014	45	7,63	27	39	45	50	60
	2012	2015	39	6,01	25	36	39	43	58
	2013	2216	44	7,77	26	38	44	50	62
4	2009	2011	49	0,76	33	43	48	54	66
	2010	2012	45	0,70	26	40	45	48	61
	2011	2013	52	0,68	34	47	52	57	68
	2012	2014	44	0,73	28	39	45	50	62
	2013	2015	50	0,85	27	45	51	56	69
	2014	2216	57	0,84	36	52	57	63	71
5	2010	2011	52	6,41	26	48	53	53	68
	2011	2012	58	6,72	42	53	58	58	71
	2012	2013	53	9,28	33	46	53	53	77
	2013	2014	57	6,49	34	53	58	58	71
	2014	2015	60	7,04	38	56	61	61	77
	2015	2216	63	8,17	45	57	63	63	82
6	2011	2011	64	8,17	33	60	65	68	79
	2012	2012	58	9,81	37	50	57	67	82
	2013	2013	64	7,54	43	58	63	69	81
	2014	2014	68	6,27	48	65	68	73	81
	2015	2015	71	9,17	44	65	73	78	88
	2016	2216	69	7,54	44	64	70	74	84

Fonte: Teste de Progresso Unicamp.

Tabela 2. Medidas de confiabilidade do escore relativo dos seis momentos de testagem

Testagem	Alfa de Cronbach
1	0,4609
2	0,5810
3	0,5851
4	0,6546
5	0,6251
6	0,6904

Fonte: Teste de Progresso Unicamp.

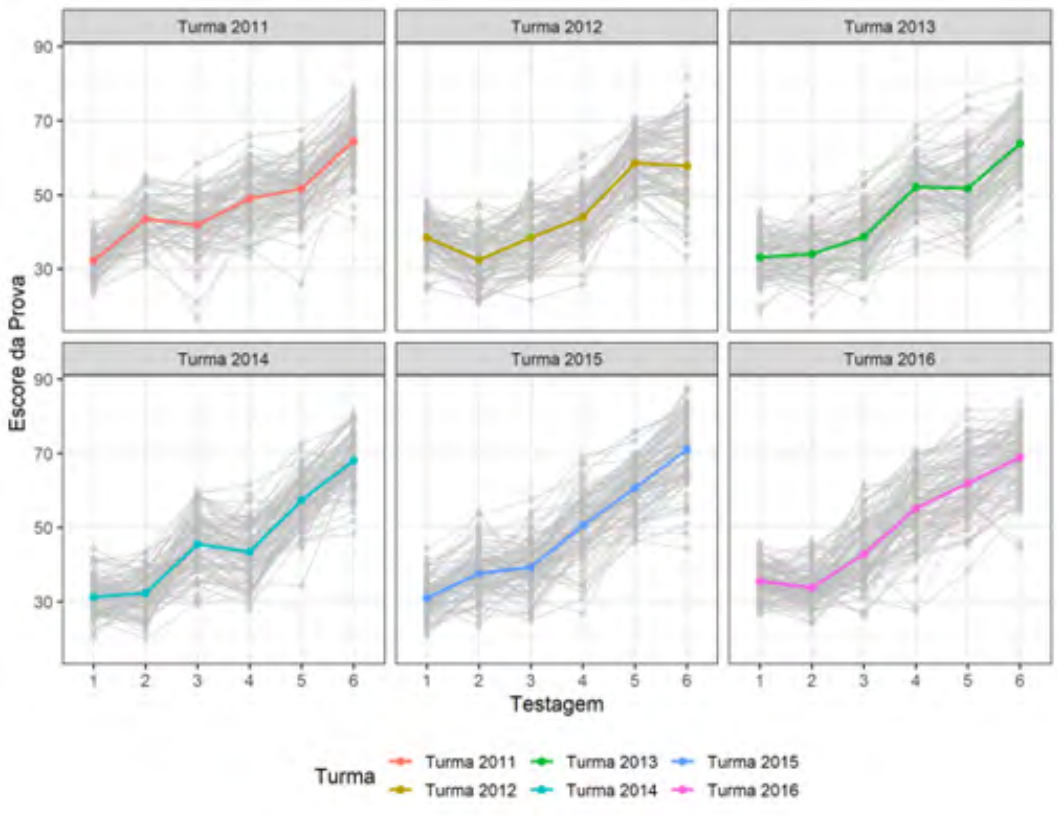
Assim sendo, observou-se uma variação de performance em torno de 18% a 28% da prova na primeira testagem; de 18% a 29% na segunda testagem; de 17% a 35% na terceira testagem; de 26% a 43% na quarta testagem; de 26% a 50% na quinta testagem; e de 33% a 58% na sexta testagem; pela tabela de normatização construída, os valores inferiores estão abaixo da média, entre o primeiro e o terceiro desvio padrão negativo.

DISCUSSÃO

A literatura sobre educação médica já apresenta amplas evidências tanto de "plausibilidade teórica" como de "viabilidade prática" do TP em medir o crescimento cognitivo de uma turma de aprendizes em diversos modelos curriculares e em uma diversidade de cenários, oferecendo dados valiosos para o desempenho de um currículo escolar[13,14].

O modelo que a escola adota para seus itens é o de múltipla escolha, já amplamente referendado pela literatura como o modelo de avaliação escrita mais confiável e válido, a ponto de ter diminuído a importância dos demais formatos de testes escritos[15]. Mesmo o teste do consórcio holandês tido frequentemente como referência mudou seu formato original que remontava à década de 1970 para o formato de "múltipla escolha com melhor alternativa" ou "múltipla escolha tipo A"[16]. O modelo de teste de múltipla escolha tipo A apresenta ainda como vantagem oferecer melhor consistência interna e confiabilidade, além de exigir um grau de análise mais profunda para sua resolução[17]. Quanto ao número de itens, no entanto, a prova apresenta uma ameaça à consistência dos resultados. É conhecido o efeito da amostragem sobre a fidedignidade da medida, e a literatura sobre TP demonstra que o número de itens utilizados em testes com comprovada fidedignidade varia de 180 a 400[18]. No caso do TP-Unicamp, observou-se que a consistência interna do teste só se tornou mais adequada a partir da quarta testagem, o que em parte pode ser melhorado com o aumento do número de itens, uma estratégia a ser adotada caso a escola pretenda utilizar seus resultados para fins somativos.

Gráfico 1. Perfil do escore relativo das turmas nas seis testagens do TPI



Fonte: Teste de Progresso Unicamp.

Tabela 3. Modelo linear misto ajustado para o escore relativo das turmas nas seis testagens do TPI

	Estimativa	Erro padrão	Graus de liberdade	Valor T	Valor p
Intercepto	23,53	0,55	2767	42,51	< 0,001*
Momento	6,68	0,07	2767	96,72	< 0,001*
Turma 2012	-1,99	0,70	573	-2,85	0,005*
Turma 2013	-1,29	0,70	573	-1,84	0,066*
Turma 2014	-0,55	0,69	573	-0,79	0,432*
Turma 2015	1,43	0,70	573	2,03	0,043*
Turma 2016	2,71	0,68	573	3,98	< 0,001*

* valor p < 0,05
Fonte: Teste de Progresso Unicamp.

Outro aspecto do TP-Unicamp que merece ser analisado é a ausência de estratégia para acertos casuais (chute). Como uma das características do TP é a oferta de uma prova de conteúdo abrangente mesmo para alunos do primeiro ano, sempre foi uma preocupação a influência dos acertos casuais na confiabilidade dos resultados. Na verdade, essa é uma questão que remete ao início do século XX e à popularização dos testes padronizados, mas que também demonstra como aspectos psicométricos podem muitas vezes dominar uma discussão a ponto de obscurecer aspectos fundamentais no ensino. Muitas escolas penalizam candidatos que marcam respostas erradas para desencorajar marcações baseadas em palpites[19], e outras incluem uma alternativa "não sei", que tira a penalidade no caso de erro, e o resultado é obtido por meio da utilização de uma fórmula para o escore (*formula scoring*)[17]. No entanto, a literatura tem demonstrado que não somente tais estratégias apresentam menos confiabilidade[20], como também podem ter um efeito prejudicial sobre aspectos fundamentais do processamento de um problema por um ser humano. Bliss[23], em um estudo com crianças do ensino básico, demonstrou que a utilização da *fomula scoring* penalizou alunos mais capazes, o que poderia inclusive levantar questões éticas sobre essa prática, e, da mesma forma, Muitjens et al.[25] demonstraram que alunos de Medicina que são menos predispostos a adivinhar e, portanto, procuram desenvolver um entendimento mais profundo da questão tendem a ter um escore mais baixo. Especialmente no caso da utilização dos testes de múltipla escolha tipo A, em que todas as alternativas são plausíveis, ao penalizarmos alunos que não escolham a opção "não sei", negamos a possibilidade de que um conhecimento parcial possa contribuir para a resolução de problemas.

Um aspecto crítico quanto à execução do teste diz respeito ao número de testagens. Nesse sentido, o número de edições do TP-Unicamp por turma é bem inferior àqueles utilizados por outras escolas que utilizam o TP para fins somativos. As escolas de Medicina de McMaster e Utrecht realizam três testagens anuais; e as escolas de Maastricht e de Missouri-Kansas, quatro testagens anuais[14]. Sabe-se que é justamente a repetição desses testes que permite a construção da curva cognitiva. Este estudo demonstrou que os escores alfa das testagens só alcançaram valores limítrofes a 7 após a terceira testagem, o que demonstra que, caso a escola deseje utilizar os resultados como avaliação somática para certificar a progressão de seus alunos, precisará aumentar o número de edições anuais para pelo menos três testes anuais.

O crescimento cognitivo observado revelou escore médio percentual que variou de 34% para a primeira testagem no primeiro ano do curso a 65,7% para testagem ao final do curso, o que está acima da média nacional[22]. Para a construção da curva de crescimento cognitivo, o estudo distribuiu os dados por meio de normorreferenciamento, já que não foi possível definir um critério prévio por meio de um ponto de corte. Nesse sentido, os consórcios holandês e alemão também utilizam uma estratégia normorreferenciada, apoiados em evidências de que a determinação de um ponto de corte prévio produz resultados menos confiáveis[23]. Essa estratégia seria particularmente importante em escolas sem consultores psicométricos, já que adotam estratégias estatísticas tradicionais que não necessitam de programas complexos e historicamente mais utilizadas como o escore Z[24].

A construção da curva também permite uma estratégia educacional formativa importante e frequentemente negligenciada em educação médica, que é o da identificação de alunos com baixo rendimento e que necessitem de suporte pedagógico (remediação). Na curva obtida no TP-Unicamp, é possível observar três padrões que podem ser combinados na tomada de decisões quanto à progressão do estudante: 1. uma expectativa anual de crescimento de 6 pontos percentuais, 2. um desempenho em relação à média não inferior a 1,5 desvio padrão e 3. um desempenho cognitivo ao final do

curso, em termos de acertos, em torno de 65%. Nenhuma decisão pedagógica de alta aposta deveria ser baseada em um único parâmetro, porém a triangulação desses indicadores pode sustentar pareceres mais fidedignos que os atualmente utilizados. Na escola de Medicina de McMaster, os alunos são identificados quando seu desempenho cai abaixo de 1,5 e 2 desvios padrão da média de sua turma[14]. Embora essa definição de até 2 desvios padrão seja frequentemente a mais utilizada em termos estatísticos, o mais importante é a consolidação de uma curva de crescimento consistente, a partir da qual se possam identificar variações de desempenho[27]. No caso da escola de Medicina da Unicamp, o efeito piso ficou compreendido entre 1 e 3 desvios padrão da média, de tal maneira que, com base no presente estudo, um desempenho a partir de 1,5 desvio padrão abaixo da média poderia servir como um sinalizador para alunos que necessitem de remediação. Destaca-se então aquele que seja o aspecto de maior valor utilitário do TP: a possibilidade de oferecer estratégias formativas e somativas para apoiar decisões pedagógicas tanto na identificação e remediação de estudantes em risco (avaliação formativa) como para atestar com qualidade e confiabilidade a progressão do aprendiz ao longo do curso e definição de parâmetros para a sua conclusão (avaliação somativa).

Um aspecto necessário a ser discutido diz respeito à escolha do método utilizado para a medida dos construtos propostos. Vivemos uma época na qual a popularidade da Teoria da Resposta ao Item (TRI) cresceu a ponto de alguns estudiosos questionarem mesmo a relevância da Teoria Clássica dos Testes (TCT) e até mesmo, segundo as palavras de Zickar et al.[29], tratarem a TRI como "uma panaceia para todos os problemas psicométricos". No entanto, existem ainda situações em que a TCT pode ser suficiente e até mesmo preferida[26]. Diante da conhecida complexidade dos programas computacionais e da necessidade de uma amostragem bem maior, muitas escolas podem apresentar dificuldades em desenvolver uma análise de seus testes utilizando a TRI, recorrendo a consultorias e ao mercado de banco de itens, que oferecem soluções fidedignas, mas distantes dos objetivos pedagógicos da escola. Dessa forma, como o TP é uma ferramenta que valoriza o acúmulo de conhecimento ao longo do tempo, um estudo longitudinal com normorreferenciamento e utilizando a TCT apresenta uma lógica inerente e que pode ser facilmente instituído em qualquer escola.

CONCLUSÃO

O TP no Brasil não tem sido utilizado para apoiar decisões pedagógicas de "alta aposta" (*high stakes*), ou seja, não representa uma avaliação com consequências mais significativas na vida acadêmica do aprendiz, especialmente na forma de atestar progressão ao longo do curso. Com isso, corre o risco de se transformar em uma mera data no calendário escolar, à qual o corpo estudantil dará pouco valor. Além disso, os gestores poderão questionar o custo-benefício do TP. Dessa forma, se a escola pretender utilizar o TP como instrumento somativo válido desde o primeiro ano do ensino médico terá que promover pelo menos três testes anuais, a exemplo de outras escolas.

Mesmo com um único teste anual, o TP revelou-se um instrumento valioso para atestar a progressão do aluno ao longo do curso e principalmente para identificar os estudantes em risco e que necessitam de um programa de remediação.

A pandemia da Covid-19 trouxe para as escolas médicas inúmeros desafios, mas, acima de tudo, várias possibilidades de avanço. No atual contexto, o TP torna-se mais relevante do que nunca, já que a curva de crescimento cognitivo poderá trazer dados valiosos até mesmo sobre o impacto do ensino remoto na educação médica. O isolamento também forçou as escolas médicas a adotar estratégias de avaliação remota baseadas no computador, o que implicará significativa redução de custos, possibilitando a adequação de recursos que possibilitem um maior número de testagens.

CONTRIBUIÇÃO DOS AUTORES

Ruy Guilherme Silveira de Souza participou da pesquisa do projeto e da redação e revisão do manuscrito. Angélica Maria Bicudo participou da revisão do manuscrito.

CONFLITO DE INTERESSES

Declaramos não haver conflito de interesses.

FINANCIAMENTO

Declaramos não haver financiamento.

REFERÊNCIAS

1. Freeman A, van der Vleuten C, Nouns Z, Rickets Z. Progress testing internationally. Med Teach. 2010;32:451-5.
2. Messick S. Validity. In: Linen R, editor. Educational measurement. 3rd ed. Washington, DC: American Council on Education; 1989. p. 13-104.
3. Newton PE, Shaw S. Validity in educational & psychological assessment. London: Sage; 2014.
4. Sireci S. On the validity of useless tests. Assessment in Education: Principles, Policy & Practices. 2016;23(2):226-35.
5. Downing S. Validity: on the meaninful interpretation of assessment data. Med Educ. 2003;37:830-7.
6. Downing S. Validity threats: overcoming interference with proposed interpretation of assessment data. Med Educ. 2004;38(3):327-33.
7. Harden R, Lilley P, Patricio M. The definitive guide to the OSCE. Edinburgh: Elsevier; 2016.
8. Pasquali L, Primi R. Fundamentos da teoria da resposta ao item – TRI. Aval Psicol. 2003;2:99-110.

9. Royal K. Four tenets of modern validity theory for medical education assessment and evaluation. Advances in Medical Education and Practice. 2017;8(3):567-9.

10. Souza R, Bicudo A, Costa B, Martins A. Validity concept in medical education: a bibliometric analysis. Rev Bras Educ Med. 2020;44(4):e166.

11. Anderson S. Social competency in young children. Dev Psychol. 1974;10(2):282-93.

12. Govaerts M, van der Vleuten C. Validity in work-based assessment: expanding our horizons. Med Educ. 2013;47(12):1164-74.

13. Glaser AN. High-yield biostatistics. 2nd ed. Philadelphia: Lippincott Williams & Wilkins; 2001.

14. Tio R, Schutte B, Meiboom AA, Greidanus J, Dubois E, Bremers A, et al. Medicine, the progress test of medicine: the Dutch experience. Perspect Med Educ. 2016;5:51-5.

15. Albanese M, Case S. Progress testing: critical analysis and suggested practices. Adv Health Sci Educ. 2016;21(1):221-34.

16. Hift R. Should essays and other open-ended type questions retain a place in written summative assessment in clinical medicine? BMC Med Educ. 2014;14:249. doi: 10.1186/s12909-014-0249-2.

17. Schuwirth L, Bosman G, Henning R, Rinkel R, Wenink A. Collaboration on Progress Testing in medical schools in the Netherlands. Med Teach. 2009;32(6):476-9.

18. Wrigley W, van der Vleuten CP, Freeman A, Muijtjens A. A systemic framework for the progress test: strengths, constraints and issues. Med Teach. 2012;34(9):683-97.

19. Rademakers J, Ten Cate T, Bär P. Progress testing with short answer questions. Med Teach. 2005;27(7):578-82.

20. McHarg J, Bradley P, Chamberlain S, Ricketts C, Searle J, Mclachlan J. Assessment of progress tests. Med Educ. 2005;39:221-7.

21. CJ R, et al. The don't know option in Progress Testing. Adv Health Sci Educ. 2015;20:1325-38.

22. Muijtjens A, Mameren H, Hoogenboom R, Evers J, van der Vleuten C. The effect of a "don't know" option on test scores: number-right and formula scoring compared. Med Educ. 1999;33(4):267-75.

23. Bliss L. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. J Educ Meas. 1980;12(2):147-53.

24. Bicudo AM, Hamamoto Filho PT, Abbade JF, Hafner MLMB, Maffei CML. Teste de Progresso em consórcios para todas as escolas médicas do Brasil. Rev Bras Educ Med. 2019;43(4):151-6.

25. Muitjens A, Hoogenboom R, Verwijnen G, van der Vleuten C. Relative or absolute standards in assessing medical knowledge using Progress Tests. Adv Health Sci Educ. 1998;3(2):81-7.

26. Langer M, Swanson D. Practical considerations in equating progress tests. Med Teach. 2010;32(6):509-12.

27. Callegari-Jacques SM. Bioestística: princípios e aplicações. São Paulo: Artmed; 2003.

28. Harvey R, Hammer A. Item Response Theory. Couns Psychol. 1999;27:353-83.

29. Zickar J, Broadfoot A. The partial revival of a dead horse? Comparing Classical Test Theory and Item Response Theory. In: Lance C, Vandenberg R, editors. Statistical and methodological myths and urban legends: doctrine, verity and fable in the organizational and social sciences. New York: Routledge; 2009. p. 38-57.