Scientia Medica
EDUCAÇÃO EM CIÊNCIAS DA SAÚDE
e-ISSN: 1980-6108 | ISSN-L: 1806-5562
10.15448/1980-6108.2019.3.34432
Open Access

Implementando o teste adaptativo computadorizado

Implementing computerized adaptative test

Dario Cecilio-Fernandes [1]

[1] Departamento de Psicologia Médica e Psiquiatria, Faculdade de Ciências Médicas, Universidade Estadual de Campinas. Campinas, São Paulo, Brasil.

Como citar este artigo (How to cite this article):

Cecilio-Fernandes D. Implementando o teste adaptativo computadorizado (*Implementing computerized adaptative test*). Sci Med. 2019;29(2):e34432. http://doi.org/10.15448/1980-6108.2019.3.34432

RESUMO

Tradicionalmente, a avaliação de conhecimento é composta por itens dos quais os alunos respondem ao mesmo tempo e os mesmos itens, como por exemplo a prova de uma disciplina. Essa avaliação pode ser considerada muito fácil ou difícil pelo aluno. Em ambos os casos, a prova será entediante e fornecerá pouca informação sobre o nível de conhecimento do aluno. Uma forma de resolver esse problema, é criar provas customizadas para cada aluno, sendo que a próxima questão será selecionada baseada no desempenho anterior do aluno. Essa forma de avaliação é conhecida por teste adaptativo computadorizado. O teste adaptativo computadorizado apresenta tanto vantagens educacionais quanto psicométricas quando comparado às provas de formato tradicional. O teste adaptativo computadorizado requer menos itens do que o teste no formato tradicional, algo que consequentemente diminuirá a fadiga dos alunos, aumentando a aprendizagem do aluno. Ademais, o teste adaptativo computadorizado é desenhado especificamente para cada aluno, considerando a dificuldade de cada item. Isso torna o teste mais atrativo e autêntico, pois os itens sempre estarão alinhados ao conhecimento do aluno. Por ser necessária a informação sobre a dificuldade do item e a habilidade do aluno, o teste adaptativo computadorizado utiliza a Teoria de Resposta ao Item, que estabelece uma relação entre a habilidade do sujeito, a dificuldade do item e a probabilidade de acertar o item, essencial para o funcionamento do teste adaptativo computadorizado. Apesar da complexidade técnica para a implementação do teste adaptativo computadorizado, o mesmo traz um maior padrão para a avaliação tanto de um ponto de vista psicométrico como do ponto de vista do alinhamento às teorias modernas de aprendizagem. Por causa de sua alta complexidade, a implementação e a utilização do teste adaptativo computadorizado geralmente estão associadas a provas de decisão de alto impacto e em grande escala. No entanto, com o novo entendimento educacional, onde se faz necessário respeitar a individualidade e o ritmo de cada aluno, o teste adaptativo computadorizado será cada vez mais utilizado.

DESCRITORES: Educação médica; avaliação educacional; teste adaptativo computadorizado

ABSTRACT

Traditionally, the assessment of knowledge consists of items, who students answered the same items at the same time, such as test of a specific subject. This assessment may be considered too easy or difficulty by the student. In both cases, the test is likely to be boring by the students and it may provide little information on students' knowledge level. One way of solving this problem is by creating tailored tests for each student, considering that the next question will be selected based on students' performance on previous items. This type of test is known as computerized adaptative test. Computerized adaptative test provides both educational and psychometrics advantages compared to the traditional paper-pen testing. Computerized adaptative test requires less items than the traditional test, which in turns will decrease students' fatigue, and optimizing learning. Furthermore, computerized adaptative test is designed for each student, considering the level of difficulty of each item. This makes the teste more attractive and authentic, since the items will be always aligned with the level of students' knowledge. Since computerized adaptative test requires both the difficulty of the item and students' ability, it requires the use of Item Response Theory, which establish a relation between difficulty of the item, students' ability and the probability of answering a question correctly. Although the implementation of computerized adaptative test is complex, computerized adaptative test has a higher standard in both psychometric point of view and the alignment with modern theories of learning. Because of the high complexity, the implementation of computerized adaptative test is usually in high-stakes test and large scale. However, the new educational paradigm in which requires tailored-made education respecting the pace of each student, the computerized adaptative test will be more used over time.

KEYWORDS: Medical education; educational assessment; computerized adaptive test.

Recebido: *10/06/2019*
Aceito: *10/08/2019*
Publicado: *16/10/2019*

Correspondência: dario.fernandes@gmail.com
Rua Tessália Vieira de Camargo, 126 – Cidade Universitária
3083-887, Campinas, São Paulo, Brasil

Abreviatura: CAT, teste adaptativo computadorizado.

INTRODUÇÃO

A avaliação na medicina vem se tornando cada vez mais complexa [1], principalmente com a implementação do ensino baseado em competências ao redor do mundo e no Brasil, especialmente, após a mudança nas diretrizes nacionais curriculares, com a inclusão do ensino de competências. O aumento da complexidade da avaliação ocorre em paralelo ao reconhecimento da necessidade de se aferir não só o conhecimento teórico do estudante adquirido durante o curso, mas também suas competências, habilidades e comportamento profissional. Para responder a esta demanda, nos últimos anos, as pesquisas em avaliação médica vêm focando na criação de novos métodos de avaliação, assim como na sua aplicabilidade. Esses novos métodos levam em consideração o impacto da avaliação na aprendizagem do aluno [2, 3], sendo que a título exemplificativo pode-se citar a inclusão do feedback como parte do processo de avaliação. Ademais, os métodos acima citados visam, ainda, possibilitar a formação de médicos competentes que além de cuidarem de seus pacientes, serão capazes de contribuir para o funcionamento do sistema de saúde em que estão insertos. De forma mais direta, a avaliação impacta o comportamento do aluno enquanto discente, norteando-o sobre *o que*, *como, quando* e com que *frequência* estudará. Portanto, constata-se que a avaliação tem papel primordial na aprendizagem dos alunos.

A prática médica requer não apenas a reprodução do conhecimento, mas também sua aplicação. A reprodução e aplicação de conhecimentos requerem processos cognitivos diferentes. Enquanto a reprodução apenas exige relembrar ou um mínimo entendimento do conhecimento, a aplicação requer um entendimento mais profundo deste conhecimento [4]. A avaliação, por sua vez, pode ser utilizada como uma ferramenta que facilite a habilidade de aplicar o conhecimento por parte do aluno. Em um recente estudo, Cecilio-Fernandes et al. demonstraram que alunos iniciantes, quando avaliados, responderam corretamente as questões relacionadas à reprodução do conhecimento, enquanto que alunos no final do curso respondem corretamente questões relacionadas à aplicação do conhecimento [5]. Nota-se também que questões relacionadas à aplicação do conhecimento estão associadas a casos clínicos, enquanto questões relacionados a reprodução do conhecimento está relacionado a questões sem casos clínicos [6] . Portanto, além de ajudar os alunos a desenvolver a habilidade de aplicação do conhecimento, é importante considerar o contexto que o aluno está inserido (por exemplo, treinamento pré-clínico e clínico) para a avaliação.

Além de diferentes tipos de questões, o formato da avaliação também pode influenciar a aprendizagem do aluno. Geralmente, as avaliações de conhecimento se restringem a uma mesma prova para todos os alunos. No entanto, avaliações também podem ser individualizadas respeitando a habilidade de cada aluno, sendo que essa forma de avaliação é conhecida como teste adaptativo computadorizado (*Computerized adaptative test* – CAT) [7]. O CAT é um teste onde a escolha da questão subsequente é baseada no desempenho do aluno na questão anterior [7-10]. Por exemplo, caso o aluno acerte a questão anterior, ele receberá uma questão subsequente com um nível de dificuldade maior do que a anterior. O CAT poderá favorecer uma avaliação para aprendizagem por diversos motivos. Primeiro, os alunos responderão apenas questões que sejam compatíveis com suas habilidades. Isso aumentará o grau de autenticidade do teste, assim como evitará que este se torne entediante, seja por causa de questões muito difíceis ou muito fáceis [7]. Além disso, se comparado ao teste tradicional, o CAT permite a utilização de menos itens, mantendo ou até mesmo elevando a precisão do teste [11]. A diminuição do número de itens, consequentemente, diminui a fadiga dos alunos, aumentando a possibilidade de aprendizagem. Em relação ao teste tradicional, o CAT apresenta outra vantagem. Enquanto no primeiro é utilizado um número fixo de itens, o CAT pode ser programado para testar áreas específicas do conhecimento teórico onde o aluno apresente maior dificuldade. Por último, o CAT é desenhado especificamente para cada aluno, respeitando tanto o nível de conhecimento como as dificuldades de cada um. Desta forma, pode-se, inclusive, modular a carga cognitiva, ou seja, o nível de recursos cognitivos utilizados para executar a tarefa [12]. A título exemplificativo, explicita-se que questões muito fáceis utilizarão baixa carga cognitiva enquanto questões muito difíceis utilizarão alta carga cognitiva. Tanto uma quanto outra situação não são adequadas à aprendizagem. No CAT a carga cognitiva pode ser programada para o ideal de aprendizagem. Entretanto, pesquisas são necessárias para a identificação do ponto que otimiza a carga cognitiva para a aprendizagem.

Mesmo reduzindo o número de itens, o CAT é mais fidedigno do que o teste tradicional [11]. Além disso, o CAT pode aumentar a autenticidade das questões,

por exemplo com a inclusão de exames de uma forma interativa, medir o (ou a falta do) conhecimento específico de cada aluno e mensurar a habilidade de cada aluno, sendo todas essas questões relacionadas a validade do teste. Por exemplo, com o CAT pode-se verificar se o aluno se encontra na fase de reprodução ou aplicação do conhecimento, sendo possível ajustar o teste para a fase em que o aluno se encontra. Ademais, o CAT está alinhado às teorias modernas de aprendizagem, tanto cognitivas quanto construtivistas. Portanto, além das vantagens para a aprendizagem do aluno, o CAT também possui melhores propriedades psicométricas do que os testes tradicionais.

Apesar das diversas vantagens apresentadas para o uso do CAT, a sua implementação requer diversos passos que serão apresentados em seguida [13].

CRIAÇÃO DE BANCO DE ITENS

Para a implementação do CAT é imperativo que exista um banco de itens. Tal banco de itens deve conter itens de diferentes níveis de dificuldade, variando desde o mais fácil até o mais difícil. Também, é importante que os itens sejam baseados em uma matriz de conhecimentos definida a priori. A quantidade de itens necessária para a realização do CAT é maior do que no teste tradicional. Por exemplo, digamos que o teste tenha quatro formatos diferentes com 100 questões cada. Destas 100 questões, 20 são utilizadas como âncora. Nesse caso, seriam necessários 340 itens para a aplicação desses quatro testes. Caso fosse utilizado o CAT, estudos de simulação de Monte Carlo apontam para a necessidade de pelo menos 500 itens em seu banco de itens [13]. Apesar de um número maior de itens inicialmente, os mesmos itens podem ser utilizados para diferentes alunos e testes, diminuindo a escrita de itens a longo prazo. Finalmente, todos os itens nesse banco de dados devem ser previamente calibrados, para que o algoritmo possa escolher o próximo item.

CALIBRAÇÃO DOS ITENS

Como a escolha do próximo item depende da habilidade do sujeito e da dificuldade do item, os itens dever ser calibrados utilizando a Teoria de Resposta ao Item. A Teoria de Resposta ao Item consiste em um modelo matemático que estabelece uma relação entre a habilidade do sujeito, a dificuldade do item e a probabilidade de acertar o item. Por exemplo, quando a habilidade do sujeito é igual a dificuldade do item, a probabilidade de o sujeito acertar o item será de 50%. Apesar de existirem diversos estudos utilizando o CAT com diferentes modelos de Teoria de Resposta ao Item, a utilização do modelo de Rasch traz diversas vantagens em relação aos outros modelos. Primeiramente, a quantidade de sujeitos para estimar os parâmetros do modelo de Rasch é bem menor, sendo necessário por volta de 100 sujeitos [14]. O modelo de Rasch também é mais simples e rigoroso do que os outros modelos. Consequentemente, Rasch é um modelo mais suscetível a violações, permitindo identificar os itens problemáticos mais facilmente quando comparados aos modelos mais complexos. Por ser um modelo mais simples, Rasch também requer um poder computacional menor, tornando-o mais rápido e prático para a aplicação em larga escala [15]. Finalmente, o modelo de Rasch permite diversas alternativas para equalização de diferentes testes.

ESPECIFICAÇÃO DO TESTE ADAPTATIVO COMPUTADORIZADO

O CAT é uma ferramenta de testagem extremamente flexível, por isso é de extrema importância alinhar sua especificação com o objetivo do teste e com a percepção dos envolvidos (*stakeholders*). Em linhas gerais, a especificação do CAT pode ser divido em 5 partes, a saber, banco de itens, ponto inicial, seleção dos itens, ponto final e pontuação [13].

Banco de itens: A decisão relacionada a quantidade de itens necessária para o CAT. O por exemplo, o banco de itens pode ter 1000 itens. No entanto, para aplicar o CAT seriam necessários 400 itens. Dessa forma, não é necessária a utilização de todos os itens, evitando a exposição dos itens. A segunda especificação se refere ao ponto inicial do teste.

Ponto inicial: Tradicionalmente, fixa-se um número de itens para todos os respondentes para que o algoritmo deduza a habilidade de cada participante. A vantagem de utilizar um número fixo de itens que não é preciso ter informações prévias sobre os respondentes. No entanto, isso leva a superexposição dos itens iniciais, algo que precisa ser considerado, principalmente quando as aplicações são feitas em diferentes dias. Uma alternativa, é utilizar a performance anterior do respondente como ponto de partida. Essa alternativa é bastante utilizada no ambiente universitário, onde essa informação é de fácil acesso. Diversos pontos iniciais podem ser escolhidos, dependendo da quantidade de informações disponíveis de cada sujeito.

Seleção dos itens: por ser extremamente flexível, existem diversas formas de selecionar os itens. Por

exemplo, se é essencial seguir a matriz de conhecimento do teste, é possível selecionar a quantidade itens máxima e mínima para cada categoria. Ademais, deve-se considerar a decisão que será feita após o teste. Se é necessário ter uma grande fidedignidade em um conteúdo, pode-se escolher a utilização de um número maior de itens de determinado conteúdo. Também, é necessário decidir se o teste terá uma quantidade fixa de item ou quando atingir a especificação do ponto final.

Ponto final: Quando terminar o teste é um ponto essencial. CAT mais flexíveis se baseiam quando a fidedignidade da decisão for alta com pouco erro associado. Consequentemente, o número de itens difere por respondente, algo que nem sempre é percebido como justo. Por isso, diversos testes também utilizam uma quantidade fixa de itens para o término do teste.

Pontuação: Apesar de ser possível utilizar a Teoria Clássica dos Testes para definir a pontuação final do aluno, a maioria dos CAT definem a pontuação utilizando a Teoria de Resposta aos Itens. No entanto, é utilizado uma transformação para a padronização da pontuação, como por exemplo, transformando em uma escala tendo 500 pontos como média.

APLICAÇÃO DO TESTE ADAPTATIVO COMPUTADORIZADO

Após a criação do banco de itens, calibração dos itens e a definição das especificações, o CAT está pronto para ser aplicado. Durante a aplicação do CAT, também é possível verificar se os respondentes estão trapaceando, houve vazamento de questões entre outras possibilidades. Apesar do CAT utilizar itens previamente calibrados, análises psicométricas após o teste são necessárias para verificar se os itens corresponderam a expectativa, se houve mudança na dificuldade dos itens, a precisão do teste e se um item favoreceu algum subgrupo de respondentes em detrimento de outros.

CONSIDERAÇÕES FINAIS

A implementação e a utilização do CAT geralmente estão associadas a provas de decisão de alto impacto e de grandes escalas. Apesar da dificuldade técnica da sua utilização, o CAT traz um maior padrão para a avaliação tanto de um ponto psicométrico como do alinhamento às teorias modernas de aprendizagem, possibilitando uma avaliação da aprendizagem e uma avaliação para aprendizagem, sendo que a devolutiva do resultado pode ocorrer tanto durante a prova, imediatamente após a prova ou após um período [7].

NOTAS

Apoio financeiro

Esta pesquisa foi parcialmente financiada pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), com auxílio à pesquisa jovem pesquisador, número 2018/15642-1, concedido a Dario Cecilio-Fernandes. As opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do autor e não necessariamente refletem a visão da FAPESP.

Declaração de conflito de interesses

O autor declara não haver conflitos de interesses relevantes ao conteúdo deste estudo.

REFERÊNCIAS

1. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357(9260):945-9. https://doi.org/10.1016/S0140-6736(00)04221-5

2. Wood T. Assessment not only drives learning, it may also help learning. Med Educ 2009;43(1):5-6. https://doi.org/10.1111/j.1365-2923.2008.03237.x

3. Cecilio-Fernandes D, Cohen-Schotanus J, Tio RA. Assessment programs to enhance learning. Phys Ther Rev. 2018;23(1):17-20. https://doi.org/10.1080/10833196.2017.1341143

4. Bloom BS. Taxonomy of educational objectives: the classification of education goals. New York: Longman; 1956. v. 1.

5. Cecilio-Fernandes D, Kerdijk W, Jaarsma ADC, Tio RA. Development of cognitive processing and judgments of knowledge in medical students: analysis of progress test results. Med Teach. 2016;38(11):1125-9. https://doi.org/10.3109/0142159X.2016.1170781

6. Cecilio-Fernandes D, Kerdijk W, Bremers AJ, Aalders W, Tio RA. Comparison of level of cognitive process between case-based items and non-case-based items of the interuniversity progress test of medicine in the Netherlands. J Educ Eval Health Prof. 2018;15:28. https://doi.org/10.3352/jeehp.2018.15.28

7. Collares CF, Cecilio-Fernandes D. When I say ... computerised adaptive testing. Med Educ. 2019;53(2):115-6. https://doi.org/10.1111/medu.13648

8. Van der Linden WJ, Glas CAW, editors. Computerized adaptive testing: theory and practice. Dordrecht: Kluwer Academic; 2000. https://doi.org/10.1007/0-306-47531-6

9. Weiss DJ. Computerized adaptive testing for effective and efficient measurement in counseling and education. Meas Eval Couns Dev. 2004;37(2):70-84. https://doi.org/10.1080/07481756.2004.11909751

10. Chang H. Psychometrics behind computerized adaptive testing. Psychometrika. 2015;80(1):1-20. https://doi.org/10.1007/s11336-014-9401-5

11. Martin AJ, Lazendic G. Computer-adaptive testing: implications for students' achievement, motivation, engagement, and subjective test experience. J Educ Psychol. 2018;110(1):27-45. http://dx.doi.org/10.1037/edu0000205

12. Beatty J. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol Bull. 1982;91(2):276-92. http://dx.doi.org/10.1037/0033-2909.91.2.276

13. Thompson NA, Weiss DJ. A framework for the development of computerized adaptive tests. Pract Asses Res Evaluation. 2011;16(1):1-9.

14. Linacre J. Sample Size and Item Calibration Stability. Rasch Meas Trans. 1994;7(4):328.

15. Cecilio-Fernandes D, Medema H, Collares CF, Schuwirth L, Cohen-Schotanus J, Tio RA. Comparison of formula and number-right scoring in undergraduate medical training: a Rasch model analysis. BMC Med Educ. 2017;17:192. http://dx.doi.org/10.1186/s12909-017-1051-8