Teste de Concordância de Scripts: uma proposta para a avaliação do raciocínio clínico em contextos de incerteza

Script Concordance Test: an approach to the evaluation of clinical reasoning in uncertain contexts

Ronaldo Delmonte Piovezan[I]
Osvladir Custódio[I]
Maysa Seabra Cendoroglo[I]
Nildo Alves Batista[I]

RESUMO

A avaliação do raciocínio clínico em situações de incerteza é pouco pesquisada na educação médica. Os testes escritos mais aplicados são de múltipla escolha, capazes de avaliar como se lida com problemas bem definidos. Porém, a maioria das situações contém incertezas. Um método de avaliação do raciocínio clínico em contextos de incerteza foi desenvolvido a partir da teoria de *scripts*, com situações em geriatria. Um grupo de especialistas e um grupo de estudantes de graduação resolveram o teste. A comparação entre os resultados trouxe indícios da validade do instrumento, capaz de diferenciar o raciocínio relacionado ao nível de experiência profissional. A média dos escores dos especialistas (80,41) foi superior à dos estudantes (70,71), $p < 0,001$. As análises de consistência interna e um estudo G forneceram resultados que estão de acordo com metodologias que buscam avaliar uma competência profissional. Concluiu-se que uma proposta de teste de concordância de *scripts* em língua portuguesa aplicado em uma instituição de ensino brasileira pode ser uma alternativa para a avaliação do raciocínio clínico em contextos de incerteza.

PALAVRAS-CHAVE
- Avaliação Educacional
- Competência Profissional
- Educação Médica

ABSTRACT

Little research has been done in Brazilian medical education on the evaluation of clinical reasoning in situations of uncertainty. The most common tests are still multiple-choice, which are capable of evaluating skills when dealing with well-defined problems. However, in practice the majority of situations involve uncertainties. A method for the evaluation of clinical reasoning in contexts of uncertainty was developed on the basis of the cognitive script theory in relation to professional reasoning. The objectives of the research were to develop, apply, and analyze this methodology in a Brazilian educational setting, based on clinical situations in Geriatrics that involved diagnostic, therapeutic, or ethical dilemmas. A group of specialists in this area and a group of undergraduate students that were completing their training in the Geriatrics internship took the test. Comparison of the results led to evidence of the instrument's validity, capable of distinguishing clinical reasoning according to the participants' level of experience. The mean score for the specialists (80,41) was higher than that of students (70,71) ($p < 0,001$). In addition, analyses of the internal consistency and a G study design furnished results that are consistent with a scoring system that seeks to evaluate a professional skill. In conclusion, a proposal for a script concordance test in the Portuguese language, applied in a Brazilian teaching institution, may be a viable alternative for evaluating clinical reasoning in contexts of uncertainty.

KEY WORDS
- Educational Measurement
- Professional Competence
- Education, Medical

Recebido em: 07/01/2009
Reencaminhado em: 25/04/2009
Aprovado em: 02/06/2009

[I] *Universidade Federal de São Paulo, São Paulo, SP, Brasil.*

INTRODUÇÃO

A educação médica não se atém somente a objetivos técnicos e precisos. A prática profissional exige múltiplas competências. O conhecimento técnico, a cognição, os aspectos emocionais estão entre eles, sendo que muitos são de difícil mensuração. O reconhecimento desses domínios contribui para a definição desse essencial e complexo objeto, ou seja, a competência profissional.

O raciocínio clínico é uma das principais competências médicas[1]. Embora uma parcela da capacidade para este raciocínio recaia sobre soluções para problemas bem definidos, reconhece-se que, na prática, muitas situações são mal delimitadas. Em parte, as dificuldades encontradas no ensino dessa competência poderiam originar-se na carência de instrumentos capazes de avaliá-la.

Os testes escritos mais empregados para avaliar a aprendizagem dos estudantes na educação médica são os de múltipla escolha. A capacidade para a resolução de situações mal definidas e duvidosas não pode ser avaliada completamente por tais testes. A avaliação padronizada do raciocínio em contextos de incerteza[2], baseada na teoria cognitiva de *scripts*, parece ser uma alternativa para analisar a tomada de decisões nessas situações.

Na essência da competência profissional, estão contidos a capacidade de julgamento e o *insight,* provenientes do conhecimento tácito[3]. Trata-se de um conhecimento difícil de ser observado ou mensurado, mas que em grande parte diferencia o raciocínio de médicos com *expertise* e com decisões mais apuradas em determinada área de atuação. Esta espécie de conhecimento somente pode ser exposta em situações reais, práticas, que envolvem dúvidas ou incertezas.

Este potencial de raciocínio provém de "redes" ou interligações de conhecimentos e dados diretamente ligados a tarefas práticas e regulares do exercício profissional. Estas interligações são conhecidas como *scripts*, cujos conceitos são provenientes de princípios da psicologia cognitiva sobre o raciocínio profissional. Estes *scripts* começam a aparecer quando os estudantes são confrontados com seu primeiro caso clínico e serão refinados ao longo de toda a vida profissional.

Portanto, a avaliação padronizada do raciocínio em contextos de incerteza, ou o teste de concordância de *scripts*, parece ser uma proposta interessante para o estudo e a avaliação do processo de aprendizagem do raciocínio genuinamente profissional. Esse instrumento tem sido desenvolvido em diversos ambientes educacionais, em diferentes países e línguas, e se baseia na apresentação escrita de casos clínicos, seguida de opções de escolha sobre decisões diagnósticas e terapêuticas. O formato das respostas é uma escala do tipo Likert, refletindo como uma informação é processada para a tomada de decisões. A mensuração dos resultados considera a variabilidade do processo de resolução entre especialistas na área de aplicação das questões.

Seria importante desenvolver, aplicar e analisar um instrumento de avaliação do raciocínio clínico em situações de incerteza, em língua portuguesa, baseado no teste de concordância de *scripts*, em escolas médicas brasileiras. Espera-se, com isso, contribuir criticamente em determinadas questões que ainda cercam o processo de desenvolvimento desta proposta. Por se tratar de um método de avaliação educacional ainda experimental, poderia também ser útil a busca de conclusões sobre a sua validade, com sugestões para o seu aprimoramento.

OBJETIVO

O objetivo desta pesquisa foi buscar alguns dos indícios de viabilidade, aplicabilidade e validade de um instrumento escrito de avaliação educacional para o raciocínio clínico em contextos de incerteza num contexto médico brasileiro.

METODOLOGIA

Desenvolvimento do instrumento e elaboração de seu sistema de pontuação

O teste baseou-se no método de concordância de *scripts*[3], que consiste em apresentar uma série de problemas sob a forma de descrições de casos clínicos breves e, então, questioná-los quanto a elementos de diagnósticos, investigações ou decisões terapêuticas, após a apresentação de uma nova informação em cada questão.

O sistema de pontuação foi elaborado para aferir a concordância entre as respostas dos estudantes e as dos especialistas, sendo que estes últimos formaram o painel de referência para o valor de cada opção de resposta.

Uma equipe de três especialistas em Geriatria foi responsável pela elaboração do instrumento. Foram solicitadas propostas por escrito que incluíssem tópicos relevantes para a construção de situações clínicas problemáticas, mal definidas ou duvidosas na área de atuação dos mesmos.

Para cada uma dessas situações, foram especificados os seguintes tópicos: hipóteses diagnósticas consideradas, estratégias de investigação e opções de tratamento. Além disso, cada um deles deveria especificar as perguntas que fariam para solucionar cada problema, o exame físico que realizariam e o que esperariam para cada hipótese, além dos exames que solicitariam. Por final, deveriam especificar também as informações clínicas, positivas ou negativas, que buscariam em cada caso, além de eventuais opções de tratamento.

O formato das questões dependeria dos objetivos a serem alcançados. As questões foram agrupadas de acordo com uma meta de avaliação que, em geral, foi de investigação, diagnóstico ou tratamento. Houve também questões relacionadas a prognóstico e dilemas éticos.

tipo Likert, com cinco pontos, que representava o espectro de possibilidades de decisão diante de cada nova informação. A Figura 1 demonstra o modelo estrutural do teste.

Figura 1
Formato do teste.

Descrição de caso clínico: Um homem de 72 anos procura atendimento médico por episódios de perda de consciência nos últimos 12 meses. Refere ter realizado tomografia de crânio, eletrocardiograma e holter de 24 horas há seis meses, com resultados dentro da normalidade.

Exemplo de item ou questão para hipótese diagnóstica

Se você está pensando em	E você encontra	A hipótese se torna
Hipotensão postural	Perda de consciência não precedida de outros sintomas	1 2 4 5

1: praticamente descartada
2: menos provável
3: nem mais, nem menos pr
4: mais provável
5: praticamente certa

Legenda com o significado das respostas

Escala para ser respondida

Exemplo de item ou questão para investigação

Se você está considerando solicitar	E você encontra	A investigação se torna
Mapa (monitorização ambulatorial de pressão arterial)	Varizes calibrosas em membros inferiores	1 2 3 4 5

1: contraindicada totalmente
2: possivelmente prejudicial
3: nem mais, nem menos útil
4: útil
5: absolutamente necessária

Legenda diferente, de acordo com o objetivo de cada questão

Portanto, para um mesmo caso clínico poderia haver mais de um grupo de questões. Cada questão apresentaria três partes. A primeira teria uma hipótese diagnóstica, uma ação para investigação ou uma opção terapêutica. A segunda, uma nova informação clínica – um sinal, sintoma, condição, exame de imagem, resultado de testes de laboratório, etc. – relevante para a resolução do problema. A última parte seria formada por uma escala do

Em uma etapa anterior à aplicação final do teste, este foi resolvido por um grupo com cinco geriatras, em uma espécie de aplicação piloto do mesmo. Esses especialistas poderiam opinar sobre as questões, que poderiam ser reformuladas ou descartadas de acordo com esta primeira opinião e análise.

Em seguida, o instrumento deveria ser aplicado em um grupo de especialistas para a formação de um painel de referência

para a pontuação do teste, com o objetivo de comparação com as respostas de outros grupos. Não se buscariam apenas respostas corretas ou erradas, diametralmente opostas. Seria utilizado o método de escore agregado[4,5]. Para qualquer resposta de um especialista, haveria um valor próprio e unitário, mesmo que os outros especialistas não concordassem com ele. O escore de cada questão foi composto segundo as frequências de respostas para cada ponto da escala Likert, oferecido por este grupo.

Como exemplo, caso em determinada questão 8 especialistas entre 10 escolhessem a opção de resposta 3 da escala Likert, 1 respondesse a opção 2 e outro a opção 4, o valor intrínseco da opção 3 seria 8/8 (1,0), das opções 2 e 4, 1/8 (0,125), e das opções 1 e 5, 0/8 (0). O escore máximo para essa questão, portanto, seria alcançado ao se assinalar a resposta 3, obtendo-se a pontuação de 1,0.

Com isso, o escore máximo para cada questão seria sempre igual a um, de acordo com a frequência de respostas oferecidas pelos especialistas que formaram o painel de referência. A pontuação total do teste seria obtida por meio do somatório do escore obtido em cada questão.

Formação dos grupos a serem avaliados

Como etapa inicial de validação, o instrumento também seria aplicado num grupo de estudantes no final de primeiro ano de internato, após o término de sua formação em Geriatria na graduação. Com base na comparação entre as respostas oferecidas pelos especialistas e pelos estudantes, poderia se concluir se o padrão de respostas está de acordo com as teorias sobre o raciocínio clínico que norteiam a construção do instrumento.

O grupo de especialistas seria convidado segundo uma lista de geriatras vinculados a instituições de ensino, pesquisa e assistência, envolvidos com a área de Geriatria, no Estado de São Paulo. Para a análise do desempenho dos estudantes de graduação, seriam considerados os estudantes de Medicina da Unifesp, no final da formação em Geriatria na graduação desta instituição. Estes aprendem sobre Geriatria do primeiro ao quinto ano do curso médico.

Análise de resultados

Inicialmente, seria importante comparar as médias dos dois grupos. Na análise dos escores das questões, seria calculado também o coeficiente de correlação parcial-total, que correlaciona cada um dos escores de cada item com o escore total, incluindo todos os participantes da pesquisa. Com isso, se buscaria analisar a consistência interna dos resultados da aplicação do instrumento, o que se traduz em sua confiabilidade, precisão e potencial de replicação[6].

Considerou-se utilizar o teste t de Student na comparação dos escores entre os grupos, em virtude de os dados a serem obtidos seguirem o padrão de normalidade em sua distribuição. Por fim, buscamos aprimorar a análise das fontes de erro do escore obtido por meio de uma proposta de desenho de estudo G, que pode também fornecer indícios a respeito da validade do instrumento. Assim, pretendemos obter as proporções de três fontes de variabilidade (examinandos, itens e "*p* x *i*") dos escores para os itens com melhor consistência interna para ambos os grupos.

RESULTADOS

Processo de construção e aprimoramento do instrumento

O grupo responsável por essa etapa construiu 13 casos clínicos e desenvolveu algumas questões relevantes para a resolução dos mesmos, envolvendo raciocínio para hipóteses diagnósticas, investigação, tratamento ou decisões éticas. Finalmente, obtivemos 115 questões para a aplicação do instrumento num grupo piloto. De acordo com a relevância das questões para cada caso, o numero mínimo de questões por caso foi de quatro. Nem todos os casos apresentaram questões sobre investigação diagnóstica ou terapêutica. Todos os 13 casos clínicos apresentaram questões sobre hipóteses diagnósticas. Os casos com todos os tipos de questão apresentaram até o número máximo de questões por caso, que foi de 15.

Na resolução pelo grupo piloto, foram registradas as frequências de resposta em cada questão, para cada um dos integrantes deste grupo. Realizaram-se análises de concordância entre as respostas para cada questão ou item[7]. Consideraram-se as concordâncias denominadas "justas ou boas", cujos valores calculados se situavam entre 0,70 e 0,85. Foram julgadas impróprias as concordâncias abaixo ou acima destes valores.

Níveis de concordância abaixo de 0,70 (baixa concordância) entre especialistas de uma mesma área podem revelar falta de tendência nas respostas. Essas questões podem ter sido mal interpretadas durante a resolução. Em contrapartida, as questões com respostas com concordâncias muito elevadas (acima de 0,85) podem não ter gerado incerteza, parecendo-se com questões de um teste de múltipla escolha e, portanto, também foram excluídas ou modificadas. Após esse ajuste, o teste em sua versão final apresentou 104 questões para os 13 casos clínicos desenvolvidos, ou seja, uma média de oito questões por caso clínico.

Características dos grupos para a aplicação da versão final do teste

Todos os 21 especialistas do painel de referência responderam ao teste individualmente. As explicações sobre a resolução do teste e sobre a teoria de *scripts* foram feitas antes do início da resolução. Havia também um capítulo introdutório, por escrito, com instruções sobre sua resolução. Os 41 estudantes que aceitaram participar da pesquisa, formando o grupo a ser comparado com o painel de referência, também estavam sujeitos às mesmas condições de resolução dos especialistas e às mesmas instruções sugeridas para esta resolução.

Entre os especialistas que aceitaram fazer parte do painel de referência, 5 (23,8%) eram do sexo feminino e 16 (76,2%) do sexo masculino. A média de idade desse grupo foi de 45,6 anos, sendo a menor idade de 33 anos e a maior de 58 anos. A média de anos de exercício da medicina foi de 22,5 anos, com o mínimo de 11 anos e o máximo de 33 anos. Quanto ao tempo de especialização em Geriatria, a média foi de 15,4 anos, com o mínimo de 6 anos e o máximo de 22 anos.

Entre os estudantes que participaram desta pesquisa, 17 (41,5%) eram do sexo feminino e 24 (58,5%) do sexo masculino. A média de idade desse grupo foi de 26,7 anos.

Análise dos resultados

Por meio de análise estatística, foram obtidas as médias, as medianas, os desvios padrões (D. P.), os valores mínimos (mín.) e máximos (máx.), entre outros, dos escores dos grupos participantes (Tabela 1). A variação dos escores foi maior no grupo de estudantes (32,96) do que no grupo de especialistas (21,82). A média dos escores do grupo de especialistas (81,41) foi superior à média dos escores do grupo de estudantes (70,71), com $p < 0,001$.

O coeficiente de alfa de Cronbach, considerando-se todos os participantes, para o total de itens, foi de 0,842. O coeficiente correlação parcial-total foi utilizado para selecionar os itens que poderiam distinguir melhor os indivíduos que obtêm pontuações altas daqueles que obtêm pontuações baixas. Para ser mantido, um item deveria apresentar este coeficiente com valor de 0,19 ou mais. Quarenta e três itens apresentavam baixa correlação e foram excluídos. Com os 61 itens restantes (58,7% do total), o coeficiente alfa de Cronbach elevou-se para 0,882.

Por meio de um desenho de teste G, os grupos de especialistas e estudantes foram analisados separadamente. A Tabela 2 indica as fontes de variância para o estudo G proposto para o escore obtido nos 61 itens com melhor consistência interna pelo grupo a ser avaliado, ou seja, os estudantes. A Tabela 3 compara os grupos quanto à participação dos componentes das fontes de erro no escore, considerando o teste com 61 itens selecionados pela análise de consistência interna.

Esse desenho de estudo foi obtido a partir de cálculos de estudo G com três fontes de variância: os participantes (p), os itens (i) e a interação p x i, a partir de um desenho com facetas cruzadas. As percentagens de variâncias para esses componentes foram, respectivamente, 4,4% (p), 1,6% (i) e 94% (p x i) para o grupo de especialistas, e 6,1% (p), 10% (i) e 83,9% (p x i) para o grupo de estudantes.

DISCUSSÃO

Este estudo apresentou uma proposta inédita e multicêntrica para a formação do painel de referência, diferenciando-se das pesquisas anteriores com este método de avaliação[2]. O número de 21 integrantes foi mais do que suficiente para demonstrar a variabilidade de opiniões e decisões entre os especialistas e está adequado, segundo estudos que buscaram encontrar o melhor número de integrantes para este grupo[8].

O desenho de pesquisa e a análise consideraram que as avaliações em educação médica precisam de evidências de validade para serem interpretadas significativamente[9]. O que pode ter mais ou menos evidência para ser validado é o escore obtido pela aplicação do teste, de acordo com os objetivos de avaliação. Em nosso caso, o objetivo foi interpretar as diferenças de desempenho entre os indivíduos, segundo suas competências para o raciocínio clínico em situações de incerteza em Geriatria, de acordo com as situações propostas. Buscar as diferenças entre es-

Tabela 1
Análise do escore obtido pelos grupos

Grupos	N	Média*	Mediana	D.P.	Mín.	Máx.	Variação
Especialistas	21	81,41	80,91	5,46	70,27	92,09	21,82
Estudantes	41	70,71	72,17	7,28	51,16	84,12	32,96

* Estudantes *vs.* Especialistas: $p < 0,001$ (teste t de Student)

Tabela 2
Componentes de variância para o escore total obtido pelos estudantes

Fonte de Variância	Coeficiente de Variância	Erro Padrão	(%)
Participantes (p)	0,00894	0,00239	6,1
Itens (i)	0,01471	0,00318	10,0
Resíduo (p x i)	0,12318	0,00355	83,9

Tabela 3
Participação das fontes de erro na formação do escore

Fonte de Erro no Escore (%)	Especialistas	Estudantes
Items (i)	1,7	10,7
Resíduo (p x i)	98,3	89,3

pecialistas e estudantes e fundamentar alguns sentidos para essa descoberta são a etapa primária e essencial desse processo.

Claramente, todavia, a competência para o raciocínio clínico em situações de incerteza não é uma característica isolada. Pode ser considerada um elemento interdependente de múltiplos outros fenômenos que podem ser aferidos ou detectados, mas nem sempre diferenciados, isolados ou quantificados por um teste. Os conhecimentos prévios, as experiências pessoais, educacionais e profissionais anteriores, os aspectos emocionais, entre tantos outros, são elementos que se confundem com o processo de raciocínio e de tomada de decisões. Portanto, aferir de forma precisa e isolada essa competência seria uma proposta inverossímil.

Caberia, então, perguntar se esta proposta de avaliação é uma medida escrita de cognição ou de desempenho clínico. Provavelmente, nem uma coisa, nem outra[10]. Essa proposta parece ser única, na medida em que assume o desafio de ter um formato híbrido. Ela está no limite entre três características.

Primeiramente, ela se aproxima de um teste com um formato objetivo de respostas, ou seja, um teste escrito de múltipla escolha. Em segundo, seu método de julgamento, por meio de um escore numérico, formado a partir da organização e quantificação sistemática de decisões qualitativas e subjetivas de um grupo de especialistas, permite a comparação com decisões de um painel de referência, que representariam os avaliadores ou juízes do desempenho clínico. Todos os exames que pretendem dimensionar um desempenho prático devem ter seu escore baseado na opinião de especialistas[11]. E o formato de estímulo ou desafio para a sua resolução, ou seja, casos clínicos genuínos e envolvendo contextos de incerteza, é representativo das sugestões mais elaboradas de medida de desempenho profissional.

Esta singularidade na miscigenação de características provenientes de diferentes metodologias de avaliação faz com que as análises psicométricas deste teste busquem métodos mais complexos para a garantia da acurácia e da consistência de seus resultados. Os testes de múltipla escolha, por exemplo, precisam somente de medidas de consistência interna na análise da precisão de seus resultados. Os estudos G podem contribuir com o aprofundamento desta discussão[12].

O coeficiente de alfa de Cronbach para o teste completo foi de 0,842. Vários autores e escritores de livros-texto apresentam uma variedade de opiniões sobre o melhor valor de alfa[6]. Muitos profissionais da área de avaliação educacional sugerem alfas de no mínimo 0,90 para exames *high stakes*, ou seja, aqueles que devem aprovar os examinandos para ocupações ou obrigações de grande responsabilidade, como podem ser considerados o exercício da medicina, a obtenção de um título de especialista ou a aprovação em concurso para admissão profissional. Porém, para avaliações educacionais em final de cursos ou de módulos educacionais, alfas maiores ou iguais a 0,80 seriam suficientes.

Nesta pesquisa, de maneira similar à proposta utilizada anteriormente pelas primeiras publicações sobre este instrumen-

to[13,14], empregamos um desenho de estudo G somente com o intuito de identificar a principal fonte de erro no escore obtido. Considerando os 61 itens com adequada consistência interna, demonstrou-se uma predominância da participação do componente de resíduo ("p x i") como fonte de variabilidade do escore (98% para os especialistas e 89,3% para os estudantes). Esse componente de variância é constituído por uma série de fontes de erro. Porém, provavelmente, o mais importante deles é a interação entre os participantes e os itens.

Essa interação se reflete em algumas interpretações sobre os resultados obtidos. Primeiramente, ela fornece indícios de que o desempenho de cada examinando ao longo do teste foi bastante variável ou instável, ou seja, imprevisível. Parece que a pontuação em determinada questão não foi capaz de gerar alguma previsão sobre o desempenho em questões subsequentes. Caso o raciocínio clínico em situações de incerteza seja, de fato, caso-específico, esse comportamento ao longo do teste confirmaria a hipótese de a medida desta competência ser realizada pelo teste por nós desenvolvido.

Além disso, um grande componente de interação entre pessoas e itens indica que os participantes utilizam frequentemente experiências educacionais e profissionais anteriores para a tomada de decisões durante a resolução do teste. Essa conclusão fundamenta a hipótese de que o instrumento é capaz de medir os *scripts* formados a partir de vivências individuais.

Entretanto, algumas questões ainda carecem de maior elucidação quanto ao teste proposto. Quando este estudo foi planejado, entre os anos de 2005 e 2006, não havia clareza quanto ao número de casos ou de questões por caso que deveriam ser construídos num teste de concordância de *scripts*. Sabia-se apenas que cerca de 80 questões seriam suficientes para garantir um coeficiente de confiabilidade de 0,80[2].

Não dispúnhamos também de guias de recomendação quanto à melhor maneira de construir o teste, material publicado somente em 2008[15]. Sendo assim, a escolha do número de casos e de questões por caso se baseou na disponibilidade de situações relevantes e diferentes envolvendo temas em Geriatria, com um número suficiente de questões para garantir os parâmetros de replicabilidade dos dados e a viabilidade de tempo para a resolução do teste, ou seja, cerca de 60 minutos.

A busca pela validade dos resultados de um instrumento de avaliação educacional é um processo multidimensional[16]. Nesse sentido, somente uma publicação mais recente[17] trouxe evidências diretas sobre o número mais adequado de casos clínicos e de questões por casos clínicos para um teste de *scripts* com bons níveis de confiabilidade em seus resultados. Em parte, pode-se reconhecer que o raciocínio clínico é caso-específico, já que o aumento no número de casos no teste aumenta a confiabilidade de seus resultados. Entretanto, parece ser o número de questões por caso que mais influencia as fontes de erro do escore.

Portanto, para obter resultados mais válidos, com melhores coeficientes de discriminação, no teste que aplicamos, poderíamos aumentar o número de casos clínicos para 15 a 20[17]. Utilizar números superiores à média de 8 questões por caso, como fizemos nesta proposta, parece não ser superior, por exemplo, ao número de 4 a 5 questões por caso no sentido de melhorar as qualidades psicométricas do escore obtido. Novas pesquisas são necessárias para elucidar melhor esta questão.

CONCLUSÃO

O teste de concordância de *scripts* pode ser mais uma opção para a avaliação da formação médica. Segundo as Diretrizes Curriculares Nacionais do Curso de Graduação em Medicina[18], uma das principais competências para o exercício da medicina é a capacidade para a tomada de decisões. Este elemento pode ser detectado e diferenciou os participantes por meio deste instrumento.

A construção deste teste parece ser relativamente simples, quando comparada à de outros testes escritos tradicionais. A aplicação desta proposta em dois momentos diferentes da formação profissional foi capaz de originar resultados, análises e discussões importantes para o aprofundamento da compreensão sobre essa nova metodologia.

Inicialmente, foi apresentada em nosso país uma promissora metodologia de avaliação, já em parte validada e aprimorada em outros países. Com isso, contribui-se com o complexo processo de análise deste instrumento, construindo-o, aplicando-o e analisando-o a partir da teoria cognitiva de *scripts* sobre o raciocínio clínico em situações de incerteza, num diferente idioma, noutro país, demonstrando a estabilidade dos princípios desta proposta no contexto educacional brasileiro.

REFERÊNCIAS

1. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA. 2002; 287(2):226-35.
2. Charlin B, Van de Vleuten C. Standardized asessment o reasoning in contexts of uncertainty: the script concordance approach. Eval Health Prof. 2004; 27(3):304-319.
3. Charlin B, Roy L, Brailovsky C, Goulet F. The script concordance test: a toll to assess the reflective clinician. Teach Learn Med. 2000;12(4):189-195.

4. Norman GR. Objective measurement of clinical performance. Med Educ. 1985;19:43-47.
5. Norman GR, Neufeld VR, Walsh A, Woodward CA, McConvey GA. Measuring physicians' performances by using simulated patients. Med Educ. 1985;60(12):925–934.
6. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38:1006-1012.
7. Cicchetti DY, Showalter D, Rosenheck R. A new method for assessing interexaminer agreement when multiple ratings are made on a single subject: applications to the assessment of neuropsychiatric symptomatology. Psychiatry Res. 1997;72:51-63.
8. Gagnon R, Charlin B, Coletti M, Sauvé E, Van de Vleuten C. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? Med Educ. 2005;39:284-291.
9. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003; 37:830-837.
10. Schuwirth LWT, Van de Vleuten C. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ. 2004;38(9):974-979.
11. Clauser BE. Recurrent issues and recent advances in scoring performance assessments. Appl Psychol Meas. 2000;24(4):310-324.
12. Cronbach L, Linn RL, Brennan RL, Haertel EH. Generalizability analysis for performance assessment of student achievement or school effectiveness. Educ Psychol Meas. 1997;57(3): 373-399.
13. Charlin B, Brailovsky C, Leduc C, Blouin. The diagnosis script questionnaire: a new tool to assess a specific dimension of clinical competence. Adv Health Sci Educ Theory Pract. 1998;3(1): 51-58.
14. Charlin B, Brailovsky C, Brazeau-Lamontagne L, Salmon L, Leduc C, Van de Vleuten C. Script questionnaires: their use for assessment of diagnostic knowledge in radiology. Med Teach. 1998; 20(6):567-571.
15. Fournier JP, Demeester A, Charlin B. Script concordance test: guidelines for construction. BMC Med Inform Decis Mak. 2008;8:18.
16. Ericsson KA. An expert-perfomance perspective of research on medical expertise: the study of clinical performance. Med. Educ. 2007;41:1124-1130.
17. Gagnon R, Charlin B, Lambert C, Carrière B, Van de Vleuten C. Script concordance testing: more case or more questions? Adv Health Sci Educ Theory Pract. 2008;14(13):1382-4996.
18. Ministério da Educação. Conselho Nacional de Educação. Câmara de Educação Superior. Resolução CNE/CES Nº 4, de 7 de Novembro de 2001. Institui as Diretrizes Curriculares Nacionais do Curso de Graduação em Medicina. Brasília; 2001. [online]. [acesso em 25 fev. 2006]. Disponível em: http://portal.mec.gov.br/cne/arquivos/pdf/CES04.pdf.

CONTRIBUIÇÃO DOS AUTORES

Ronaldo D. Piovezan participou do planejamento de pesquisa, coleta de dados e redação do artigo, Osvladir Custódio participou da análise estatística e resultados, Maysa S. Cendoroglo participou do desenho do estudo e discussão dos resultados e Nildo Alves Batista, coordenou a pesquisa.

CONFLITO DE INTERESSES

Declarou não haver.

ENDEREÇO PARA CORRESPONDÊNCIA

Ronaldo Delmonte Piovezan
Av. Onze de Junho, 643
Vila Clementino – São Paulo
CEP. 04041-052 SP
E-mail: rdpiovezan@gmail.com