UNIVERSIDADE MUNICIPAL DE SÃO CAETANO DO SUL
PROGRAMA DE MESTRADO PROFISSIONAL INOVAÇÃO NO ENSINO SUPERIOR EM SAÚDE - MIESS

BRUNNA CLEMENTE DE OLIVEIRA

ANÁLISE DE QUESTÕES DE MÚLTIPLA ESCOLHA E PROPOSTA DE ESTRATÉGIAS PARA A MELHORIA DA ELABORAÇÃO DE AVALIAÇÕES COGNITIVAS DO CURSO DE MEDICINA DA UNIVERSIDADE MUNICIPAL DE SÃO CAETANO DO SUL - CAMPUS SÃO PAULO

São Paulo - SP
2020

BRUNNA CLEMENTE DE OLIVEIRA

ANÁLISE DE QUESTÕES DE MÚLTIPLA ESCOLHA E PROPOSTA DE ESTRATÉGIAS PARA A MELHORIA DA ELABORAÇÃO DE AVALIAÇÕES COGNITIVAS DO CURSO DE MEDICINA DA UNIVERSIDADE MUNICIPAL DE SÃO CAETANO DO SUL - CAMPUS SÃO PAULO

Dissertação apresentada à Universidade Municipal de São Caetano do Sul para obtenção do título de Mestre em Inovação do Ensino Superior para a Área da Saúde (MIESS).

Linha de pesquisa 1: Currículo Integrado em Saúde.

Orientadora: Profª. Drª. LENA VÂNIA CARNEIRO PERES

São Paulo - SP

2020

FICHA CATALOGRÁFICA

O48a

Oliveira, Brunna Clemente de.

Análise de questões de múltipla escolha e proposta de estratégias para a melhoria da elaboração de avaliações cognitivas do curso de medicina da Universidade Municipal de São Caetano do Sul - campus São Paulo. / Brunna Clemente de Oliveira. – 2020.

48 f.: il.

Orientadora: Profª. Dra. Lena Vânia Carneiro Peres

Dissertação (mestrado) – Programa de mestrado profissional Inovação no Ensino Superior em Saúde - (MIESS), Universidade Municipal de São Caetano do Sul - USCS, São Paulo, 2020.

1. Avaliação da educação. 2. Educação médica. 3. Ensino superior (medicina). 4. Psicometria. I. Peres, Lena Vânia Carneiro. II. Título.

BRUNNA CLEMENTE DE OLIVEIRA

ANÁLISE DE QUESTÕES DE MÚLTIPLA ESCOLHA E PROPOSTA DE ESTRATÉGIAS PARA A MELHORIA DA ELABORAÇÃO DE AVALIAÇÕES COGNITIVAS DO CURSO DE MEDICINA DA UNIVERSIDADE MUNICIPAL DE SÃO CAETANO DO SUL - CAMPUS SÃO PAULO

Dissertação apresentada à Universidade Municipal de São Caetano do Sul para obtenção do título de Mestre em Inovação do Ensino Superior para a Área da Saúde (MIESS).

Linha de pesquisa 1: Currículo Integrado em Saúde.

Aprovada em: _____

BANCA EXAMINADORA

_____

Prof. Dra. Lena Vânia Carneiro Peres – Orientadora

Universidade Municipal de São Caetano do Sul

_____

Prof. Dr. José Lúcio Martins Machado

Centro Universitário de Belo Horizonte

_____

Prof. Dr. Joaquim Edson Vieira

Universidade de São Paulo

Aos meus pais, **Ivanildo de Oliveira** e **Edna Antônia Capeli da Silva**, meus primeiros professores, que, com amor e apoio incondicionais foram primordiais para o desenvolvimento da minha autonomia e coragem para enfrentar os desafios da vida.

Ao meu irmão, **João Pedro Macene de Oliveira**, que diariamente me ensina sobre doação, respeito e preocupação com o próximo mais do que consigo mesmo.

Ao meu marido, **Raul Clemente Neto**, e à minha filha, **Elis Clemente de Oliveira**, por darem propósito à minha vida e serem o meu lugar de acolhimento, escuta e paz.

AGRADECIMENTOS

Cultivo imensa gratidão por todos aqueles que me acompanharam nesta jornada pessoal e profissional, em especial:

À minha orientadora, Profª. Drª. Lena Vânia Carneiro Peres, pela confiança, por me enxergar como pessoa e para além dos deveres quando eu mais precisava, mas principalmente pelo exemplo de como educar com amorosidade e assertividade.

Aos membros da banca do exame de qualificação, Professores Doutores Carlos Alexandre Felício Brito e Valéria Menezes Peixoto, pelas valiosas sugestões oferecidas.

Aos colegas da Faculdade de Medicina da USCS – Campus São Paulo, pela oportunidade e pelo acolhimento, apesar da minha tão pouca experiência na docência, e pelos ensinamentos diários que me moldaram nesta arte tão desafiadora.

A todos os colegas do Departamento de Radiologia do Instituto de Radiologia (InRad) do HCFMUSP, pela contribuição primordial na minha formação profissional como médica radiologista e também como educadora, que a mim confiaram a importante tarefa de preceptora da graduação em 2018.

A todos os meus professores, sempre tão dedicados e humildes no seu ofício, dando subsídios para que eu pudesse escrever minha própria história de aprendizagem.

Aos meus avós: José Macene (*in memoriam*), Antônia Capeli, Mário de Oliveira (*in memoriam*) e Neusa Boaroto por serem o alicerce da nossa família, pelo amor incondicional e pelo exemplo de força e determinação para educar os filhos frente a tantas adversidades.

Às famílias Clemente e Leite, em especial ao Raul Clemente Júnior e à Valéria Leite, por me tratarem como filha e me apoiarem de tantas formas, confiando a mim o seu bem mais precioso.

Aos meus amigos da Medicina UFSCar, Raphaela Moreira, Larissa Yaegaschi, Thaís Colacioppo, Nádia Korkischko, Edison Almeida e Thiago Taffo, por

acompanharem de perto a minha trajetória há tanto tempo, vibrarem com as minhas conquistas e me apoiarem nos momentos de dificuldade.

Às minhas queridas amigas Natally Horvat e Camila Tavares, pela amizade sólida, sem julgamentos, e pela parceria em tantos projetos pessoais e profissionais, por serem exemplos de mulheres fortes e dedicadas diante das inúmeras exigências do mundo atual e tornarem meus dias mais leves e felizes.

Aos alunos, por serem fonte de motivação e a razão da busca incansável pelo conhecimento, ensinando-me mais do que eu a eles.

Às minhas queridas amigas Cláudia Silva e Márcia Santos, por cuidarem de mim e da minha família com tanto carinho.

Ao meu querido Joaquim, amigo leal, pela companhia durante a escrita deste trabalho e por tornar os meus dias mais divertidos.

“Quando se sonha tão grande a realidade aprende.”

(Walter Hugo Mãe. O Filho de Mil Homens, 2011)

RESUMO

Os processos avaliativos são um pilar fundamental do ensino. Contudo, a elaboração de avaliações de qualidade é desafiadora. Muitos estudos identificaram falhas em questões de múltipla escolha (QME) em avaliações na área da saúde. Para atingir todo o seu potencial educativo, estas devem ser confiáveis e válidas. Por meio da análise psicométrica, este trabalho se propõe a interpretar com precisão os resultados obtidos em algumas avaliações e identificar oportunidades de melhoria na sua elaboração, bem como no processo ensino-aprendizagem. Para tanto, foi realizado um estudo observacional retrospectivo utilizando as QME aplicadas nas avaliações do 2º ano de um curso de graduação em Medicina em 2019, analisando-as também segundo os critérios propostos pelo National Board of Medical Examiners (NBME) e pela Taxonomia de Bloom revisada. Os dados obtidos possibilitaram a análise dos índices de dificuldade e de discriminação e da eficiência dos distratores. Foi também realizada a avaliação da confiabilidade da avaliação como um todo por meio do KR-20. Todas as avaliações apresentaram coeficientes de KR-20 < 0,8. A média dos índices de dificuldade e de discriminação e da eficiência dos distratores foi de 72,7%, 0,23 e 50,3%, respectivamente. O índice de discriminação e a eficiência dos distratores apresentaram correlação direta estatisticamente significativa, enquanto ambos apresentaram correlação inversa com o índice de dificuldade. Da amostra total, 37,1% das questões não estavam de acordo com as recomendações do NBME. Dentre outras coisas, este estudo mostrou que as avaliações e as questões de múltipla escolha avaliadas são pouco confiáveis ou válidas, e que apresentam oportunidades de melhoria nos critérios de qualidade.

Descritores: Avaliação educacional, Educação médica, Psicometria

ABSTRACT

Assessment processes are a fundamental pillar of teaching. However, the development of quality assessments is challenging. Many studies have identified shortcomings in multiple choice questions (MCQ) in health assessments. To achieve their full educational potential, they must be reliable and valid. By means of psychometric analysis, this work aims to accurately interpret the results obtained in assessments and identify opportunities for improvement in its elaboration, as well as in the teaching-learning process. For that, a retrospective observational study was performed using the QME applied in the assessments of the 2nd year of an undergraduate medical course in 2019, evaluating than also according to the criteria proposed by the National Board of Medical Examiners (NBME) and the revised Bloom Taxonomy. The data obtained made it possible to analyze the difficulty and discrimination indices, and the efficiency of the distractors. The reliability of the assessment as a whole was also performed using the KR-20 formula. All evaluations presented KR-20 coefficients < 0.8. The means of the difficulty and discrimination indices and the distractor efficiency were 72.7%, 0.23 and 50,3%, respectively. The discrimination index and the distractor efficiency showed a statistically significant direct correlation, while both presented an inverse correlation with the difficulty index. Of the total sample, 37.1% of the questions were not in line with the NBME recommendations. Hence, this study showed that the assessments and multiple choice questions evaluated are not reliable or valid, and that they present opportunities for improvement in quality criteria.

Keywords: Educational assessment, Medical Education, Psychometrics

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AC	Avaliações Cognitivas
AMEE	*Association for Medical Education in Europe*
APA	Avaliação do Processo de Aprendizagem
DNF	Distratores Não Funcionantes
KR-20	Fórmula de Kuder-Richardson
NBME	*National Board of Medical Examiners*
QME	Questões de Múltipla Escolha
SUS	Sistema Único de Saúde
TCT	Teoria Clássica do Teste
TG	Teoria da Generalização
TRI	Teoria da Resposta ao Item
UC	Unidades Curriculares
USCS	Universidade de São Caetano do Sul

SUMÁRIO

1 INTRODUÇÃO

1.1 Contexto

1.1.1 História da Avaliação

A avaliação das intervenções sociais é uma atividade muito antiga. Os primeiros registros nos remetem à China, que há mais de quatro mil anos usava métodos avaliativos para recrutar seus funcionários públicos (BROUSSELLE, 2011). No Brasil, o modelo de exame que conhecemos hoje foi sistematizado nos séculos XVI e XVII por meio das pedagogias religiosas (jesuítica e comeniana) (LUCKESI, 2011).

Para Luckesi (2011), a avaliação é um ato democrático e inclusivo que busca acolher a realidade como ela é, satisfatória ou não, sem julgamentos, e, a partir dela, criar estratégias de superação dos limites e promoção da aprendizagem. Tendo em vista as diferenças observadas entre a teoria e a prática nas nossas escolas, para esse autor, praticamos mais exames do que avaliações, sendo aqueles uma atividade que privilegia a classificação dos alunos entre aprovados ou reprovados, deixando de lado o processo reflexivo que subsidiaria possíveis intervenções educacionais (CAVALCANTI NETO, 2009; LUCKESI, 2011).

No campo da saúde, as inúmeras mudanças no século XX na atenção à saúde e o rápido desenvolvimento da Medicina tornaram necessárias adequações na forma de educar. A hegemonia das metodologias tradicionais foi questionada a ponto de fazer emergir as metodologias ativas de ensino-aprendizagem, atualmente incorporadas pelas Diretrizes Curriculares Nacionais do Curso de Graduação em Medicina (BRASIL, 2014). Estas ensejam uma formação médica contemporânea, direcionada por competências, na qual o processo é centrado no estudante, capacitando-o a construir ativamente conhecimentos que o tornem resolutivo frente aos problemas do processo saúde-doença. Por conseguinte, essas estratégias inovadoras pressupõem também mudanças nas práticas de avaliação, favorecendo a avaliação inclusiva e transformadora ao invés dos exames.

Entretanto, são muitos ainda os obstáculos que dificultam essa transição da forma de educar e também de avaliar, tais como a replicação psicológica por parte do educador do que ele viveu, as relações de poder ainda presentes no ambiente

acadêmico, os efeitos da nossa herança histórica e o modelo autoritário e excludente da sociedade na qual vivemos (LUCKESI, 2011; PERRENOUD, 1999). As mudanças necessárias para tal são muito mais atitudinais do que a modificação do uso de técnicas e práticas metodológicas (LUCKESI, 2011).

1.1.2 Importância, Validade e Confiabilidade das Avaliações Cognitivas

Os processos avaliativos são um pilar fundamental do ensino, uma vez que por meio deles é possível acompanhar o desenvolvimento discente e docente, bem como a efetividade dos métodos de ensino adotados, permitindo a implementação de estratégias de melhoria (COUGHLIN; FEATHERSTONE, 2017; TAVAKOL; DENNICK, 2017). Apesar disso, muitas podem ser as finalidades de uma avaliação, como, por exemplo, formativa e somativa. As avaliações formativas são processuais, realizadas para monitorar o aprendizado durante o ensino (TAVAKOL; DENNICK, 2017). Já a avaliação somativa tem o papel de determinar se os estudantes adquiriram conhecimentos, habilidades e atitudes necessários para progredir na carreira acadêmica ou profissional (COOMBES; ROBERTS; ZAHRA; BURR, 2016; TAVAKOL; DENNICK, 2017), o que não exclui a reflexão sobre os seus resultados e a proposição de intervenções para promover o aprendizado.

Contudo, a elaboração de avaliações de qualidade é desafiadora e requer competências específicas dos docentes (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; ALI; RUIT, 2015). Uma avaliação ruim pode levar à insatisfação por parte dos docentes e discentes, subutilizando o seu potencial como agente transformador do processo de ensino aprendizagem. Por outro lado, avaliações bem elaboradas alteram o comportamento do aluno frente ao aprendizado, somando entusiasmo para que o mesmo retenha melhor as informações explícitas e implícitas dos recursos educacionais disponíveis (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; CILLIERS; SCHUWIRTH; HERMAN; ADENDORFF et al., 2012).

Neste contexto, durante a elaboração de qualquer avaliação, deve-se atentar para a sua validade, que reflete sua capacidade de medir o que pretende (TAVAKOL; DENNICK, 2017). Os diversos instrumentos de avaliação apenas são válidos para determinados construtos, devendo, ainda assim, ser submetidos constantemente a experimentos e observações críticas que garantam a validade e a generalização ao longo do tempo (SCHUWIRTH; VAN DER VLEUTEN, 2011).

O conteúdo é uma das facetas dessa característica primordial das avaliações, determinando que não sejam cobrados assuntos que não foram abordados nas atividades de ensino (TAVAKOL; DENNICK, 2017). Em um artigo publicado na *Medical Teacher* (CODERRE; WOLOSCHUK; MCLAUGHLIN, 2009), estudiosos da Universidade de Calgary sugerem o uso do *Blueprinting* para garantir a validade de conteúdo das avaliações, como demonstrado abaixo (Quadro 1).

Tema	Impacto *	Frequência **	I x F	Peso	N. de itens	Diagnóstico	Tratamento	Ciências básicas
Febre	1	3	3	0,33	3	1	1	1
Sepse	3	3	6	0,66	7	2	3	2
Total	-	-	9	1	10	3	4	4

*Impacto: 1 = não é urgente, pouco potencial de prevenção; 2 = grave, mas não determina risco iminente; 3 = emergência, alto potencial de prevenção. **Frequência: 1 = raro; 2 = relativamente comum; 3 = muito comum.

Quadro 1 - *Blueprinting*

Nessa tabela, o impacto (I) e a frequência (F) da febre e da sepse são determinados pelos docentes segundo a legenda. Em seguida, o produto dessas variáveis para cada apresentação clínica é dividido pela soma dos produtos (9 no nosso exemplo), definindo o peso e o número de questões de cada um desses temas na avaliação. Essas questões podem, em seguida, ser subdivididas entre os aspectos que se esperam abordar de cada assunto de acordo com o nível dos alunos (CODERRE; WOLOSCHUK; MCLAUGHLIN, 2009).

Além do *Blueprinting*, a tomada conjunta de decisão sobre a composição da prova possibilita também uma maior garantia da sua validade de conteúdo (COOMBES; ROBERTS; ZAHRA; BURR, 2016; TAVAKOL; DENNICK, 2017). A elaboração e a revisão das questões por mais de um docente amplia o olhar e pode auxiliar a remover falhas técnicas que somam dificuldade irrelevante ou beneficiam examinandos mais experientes (NBME, 2016).

Para atingir todo o seu potencial educativo, as avaliações devem ser também confiáveis, ou seja, precisas, o que garante a reprodutibilidade da medida no tempo e espaço e/ou por diferentes avaliadores (TAVAKOL; DENNICK, 2017). A confiabilidade

é uma das abordagens da generalização, passo necessário para garantir a validade (SCHUWIRTH; VAN DER VLEUTEN, 2011).

A falta de confiabilidade implica não somente que um aluno competente possa ter tido um desempenho ruim (falso negativo), mas também que um aluno despreparado possa ter sido bem sucedido (falso positivo) (SCHUWIRTH; VAN DER VLEUTEN, 2011). No primeiro caso, o aluno será submetido a novas oportunidades de comprovar sua capacidade, e o erro da avaliação poderá ser remediado. Já no segundo caso, o erro de julgamento não poderá ser remediado (SCHUWIRTH; VAN DER VLEUTEN, 2011).

1.1.3 Questões de Múltipla Escolha e a Taxonomia de Bloom Revisada

Dentre os inúmeros formatos de avaliação disponíveis na atualidade (EPSTEIN, 2007; SHUMWAY; HARDEN; EUROPE, 2003), as questões de múltipla escolha (QME) permanecem um método mundialmente aceito de avaliação do aprendizado em educação médica, especialmente das competências cognitivas (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; ALI; RUIT, 2015; COUGHLIN; FEATHERSTONE, 2017). Elas são objetivas, fáceis de corrigir, altamente confiáveis e válidas, além de poderem ser aplicadas a um grande grupo de alunos simultaneamente, abordando uma ampla variedade de conteúdos (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; SHUMWAY; HARDEN; EUROPE, 2003; VAN DER VLEUTEN; SCHUWIRTH, 2005).

Uma QME bem construída consiste em um enunciado, preferencialmente um caso clínico, um questionamento claro e direto, seguidos por quatro alternativas, sendo uma a melhor resposta, e as três restantes, distratores (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; NBME, 2016). O *National Board of Medical Examinations* (NBME, 2016) fez as seguintes recomendações:

1. Evitar questões de verdadeiro ou falso;
2. Os enunciados devem ser claros e objetivos;
3. Após a leitura do enunciado, o examinado deve ser capaz de saber a resposta correta antes da leitura das alternativas;
4. Assim como o aprendizado, as avaliações também devem ser contextualizadas através da utilização de casos clínicos nos enunciados das questões,

facilitando a abordagem de níveis cognitivos mais altos (SCHUWIRTH; VAN DER VLEUTEN, 2011);

5. As opções de resposta devem ser homogêneas e estar relacionadas a uma única dimensão do conhecimento;
6. As opções incorretas podem estar parcialmente ou totalmente incorretas;
7. Evitar opções longas ou complicadas demais;
8. Dados numéricos devem ser apresentados em uma sequência lógica e em um único formato;
9. Evitar termos vagos, como "frequentemente" e "normalmente", ou absolutos, como "nunca";
10. Evitar dar dicas gramaticais, opções excludentes ou que apresentem convergência entre si;
11. Não utilizar a opção "nenhuma das anteriores";
12. Não utilizar enunciados formulados de forma negativa.

Embora essa e outras literaturas forneçam orientação prática (COUGHLIN; FEATHERSTONE, 2017), muitos estudos identificaram deficiências e falhas em questões de múltipla escolha de avaliações na área da saúde (ALI; RUIT, 2015; DEEPAK; AL-UMRAN; AI-SHEIKH; DKOLI et al., 2015; DOWNING, 2005; TARRANT; WARE, 2008).

A Taxonomia de Bloom revisada classifica os domínios cognitivos em seis categorias hierarquicamente ordenadas em termos de complexidade dos processos mentais: lembrar, entender, aplicar, analisar, avaliar e criar (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; ANDERSON; KRATHWOHL, 2000; TAVAKOL; DENNICK, 2011). No seu trabalho, Abdulghani e colaboradores (2017) usaram a taxonomia simplificada em dois níveis diferentes; o primeiro nível representa o conhecimento básico e a compreensão, enquanto o segundo compreende aplicação e análise.

As QME devem estar alinhadas com os objetivos educacionais guiados pelo *Blueprinting*. Estudos recentes indicam que, quando bem elaboradas, elas são capazes de abordar, além do conhecimento e da compreensão, o raciocínio crítico e a aplicação da informação adquirida nas atividades curriculares, atingindo níveis cognitivos mais altos (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; PALMER; DEVITT, 2007; VANDERBILT; FELDMAN; WOOD, 2013).

O aprendizado reforçado pelo teste tem sido discutido recentemente, uma vez que, além de motivar a recordar o que foi aprendido, estimula a ressignificação e a aplicação em um contexto diferente, levando a uma maior retenção e recuperação mais flexível do conhecimento adquirido (SCHUWIRTH; VAN DER VLEUTEN, 2011). Salienta-se também a sua capacidade de estimular os alunos a estudar, destacando os assuntos que são mais importantes e identificando áreas de deficiência com necessidade de aprendizagem adicional.

1.1.4 Análise Psicométrica das Avaliações e das Questões de Múltipla Escolha

Partindo do princípio de que a confiabilidade é mais fácil de se medir e que sem confiabilidade não há validade, existem três classes de teorias usadas para se estabelecerem as generalizações por meio da confiabilidade: a Teoria Clássica do Teste (TCT), a Teoria da Generalização (TG) e a Teoria da Resposta ao Item (TRI) (TAVAKOL; DENNICK, 2013). A TCT é a mais antiga, mais usada e mais fácil de se compreender (SCHUWIRTH; VAN DER VLEUTEN, 2011; TAVAKOL; DENNICK, 2012).

Ela é útil para avaliações diretas e normorreferenciadas, como as compostas por QME. A confiabilidade se baseia na consistência interna por meio da correlação teste-reteste, podendo ser medida através do coeficiente alpha de Cronbach para variáveis não-dicotômicas (p.ex., escala Likert) ou da fórmula Kuder-Richardson 20 (KR-20) para variáveis dicotômicas, como em QME em que as possibilidades de resposta são correto ou incorreto (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; SCHUWIRTH; VAN DER VLEUTEN, 2011; TAVAKOL; DENNICK, 2012; 2017). Um KR-20 alto (> 0,90) indica uma avaliação com testes homogêneos, enquanto que valores abaixo de 0,8 sugerem baixa confiabilidade da avaliação. Portanto, valores em torno de 0,8 são desejáveis para avaliações somativas de alto impacto (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017). Tarrant e Ware (2008), entretanto, consideram que valores acima de 0,5 são suficientes para fins educativos.

Essa análise psicométrica permite interpretar de forma precisa os resultados obtidos em uma avaliação, garantindo que a tomada de decisão seja justa tanto para

os alunos quanto para a instituição e possibilitando a identificação de oportunidades de melhoria nas avaliações elaboradas, nos recursos educacionais utilizados e no aprendizado dos alunos (COOMBES; ROBERTS; ZAHRA; BURR, 2016; TAVAKOL; DENNICK, 2017).

As questões de múltipla escolha podem ser avaliadas quanto ao índice de dificuldade, índice de discriminação e eficiência dos distratores (TAVAKOL; DENNICK, 2017). O índice de dificuldade indica a porcentagem de alunos que responderam corretamente ao teste em relação à população total. Portanto, um teste fácil apresenta um índice de dificuldade alto (TAVAKOL; DENNICK, 2017). Já o índice de discriminação avalia a capacidade do teste de discriminar os alunos que apresentaram um bom desempenho na avaliação daqueles que apresentaram um desempenho ruim (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017).

Por último, a eficiência dos distratores identifica a quantidade de distratores (alternativas incorretas) escolhidos por menos de 5% dos alunos que responderam ao teste, sendo então chamados de distratores não funcionantes (DNF) (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; TAVAKOL; DENNICK, 2017). Em um teste com quatro alternativas, sendo uma a correta, temos três distratores; portanto, a eficiência dos distratores poderá ser 0% (3 DNF), 33,3% (2 DNF), 66,6% (1 DNF) ou 100% (0 DNF). Quanto maior o número de distratores eficientes, mais bem elaborado é o teste (TAVAKOL; DENNICK, 2017).

Alguns estudos mostraram que erros na elaboração das QME geralmente resultam em alterações indesejáveis nesses índices (GAJJAR; SHARMA; KUMAR; RANA, 2014).

1.1.5 O Curso de Medicina da Universidade Municipal de São Caetano do Sul - Campus São Paulo e os Processos Avaliativos

A primeira turma do Curso de Medicina da Universidade Municipal de São Caetano do Sul (USCS) - Campus São Paulo teve início no ano de 2016. Uma das Unidades Curriculares (UC) desse curso chama-se Necessidades e Cuidados em Saúde; seu papel é explorar capacidades cognitivas, atitudinais e psicomotoras, utilizando metodologias ativas de ensino-aprendizagem por meio de disparadores de aprendizado prevalentes na prática médica no contexto do Sistema Único de Saúde (SUS) nos diferentes ciclos de vida, considerando o perfil de competência do

estudante segundo a sua série, bem como as diversas dimensões do processo saúde-doença e terapêutica (PADILHA et al., 2016).

As avaliações do desempenho discente nessa UC são predominantemente critério-referenciadas, sendo realizadas tanto avaliações formativas quanto somativas ao longo do semestre. No ano de 2019, as capacidades psicomotoras e atitudinais foram avaliadas por meio da auto-avaliação e dos encontros de Avaliação do Processo de Aprendizagem (APA) e Portfólio; as capacidades cognitivas, por meio de questões de múltipla escolha; e as questões abertas de resposta longa, ao final de cada Unidade Temática por meio das chamadas Avaliações Cognitivas (AC), numeradas em ordem temporal crescente desde o início do curso (PADILHA et al., 2016).

Desde a sua criação, as avaliações cognitivas têm sido tema constante de discussão nas reuniões de colegiado, tanto por parte dos docentes, que se queixam de falta de capacitação e de tempo para sua adequada elaboração, quanto por parte dos discentes, que referem, por exemplo, uma validade de conteúdo inadequada. Por esse motivo, ao longo de 2019, algumas estratégias foram implementadas visando à melhoria das avaliações nessa instituição, tais como a realização de uma oficina de avaliação na semana de capacitação docente, a criação de um Núcleo Permanente de Avaliação e a análise psicométrica de algumas avaliações (PADILHA et al., 2016).

No ano de 2019, a terceira e a quarta etapa do Curso de Medicina do Campus São Paulo foram submetidas às AC 05, 06, 07, 08 e 09. O *Blueprinting* e as questões de cada aspecto da unidade temática eram elaborados em conjunto por pelo menos dois docentes e apresentados posteriormente a todo o grupo de professores. Nessas ocasiões, eram feitas adequações nos enunciados e nas alternativas de acordo com as falhas identificadas pelos pares, e as questões que iriam compor a avaliação final eram escolhidas conjuntamente.

1.2 Hipóteses

Hipotetizamos que as avaliações e as questões de múltipla escolha elaboradas no Curso de Medicina da USCS apresentam baixa confiabilidade e baixo índice de discriminação com poucos distratores funcionantes, sendo compostas predominantemente por questões fáceis.

A criação de um Núcleo Permanente de Avaliação, a implantação de educação permanente e continuada oferecendo oficinas de capacitação docente e a análise

psicométrica das avaliações com *feedback* aos docentes podem auxiliar o processo de melhoria na elaboração de QME (ABDULGHANI; AHMAD; IRSHAD; KHALIL et al., 2015; ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; ALFARIS; NAEEM; IRFAN; QURESHI et al., 2015; KHAFAGY; AHMED; SAAD, 2016; VAN DER VLEUTEN; SCHUWIRTH, 2005).

2 OBJETIVOS

2.1 Objetivo geral

Analisar a confiabilidade de avaliações cognitivas e a qualidade de questões de múltipla escolha aplicadas na Faculdade de Medicina da USCS – Campus São Paulo.

2.2 Objetivos específicos

- Analisar a confiabilidade de avaliações cognitivas realizadas na Faculdade de Medicina da USCS.
- Analisar os índices de dificuldade, discriminação e a eficiência dos distratores das questões de múltipla escolha das avaliações cognitivas.
- Analisar a Taxonomia de Bloom Revisada das questões de múltipla escolha, bem como a presença de erros na sua elaboração.
- Relacionar as estatísticas da análise de cada questão de múltipla escolha com a qualidade do seu formato, de acordo com os critérios propostos pelo *National Board of Medical Examiners* (NBME).

3 MÉTODOS

Uma vez que esta pesquisa trabalha com banco de dados de informações agregadas, sem possibilidade de identificação individual, não foi necessária a avaliação pelo Comitê de Ética em Pesquisa.

3.1 Tipo de estudo

Observacional retrospectivo.

3.2 Local

Curso de Medicina da USCS - Campus São Paulo.

3.3 Amostra

3.3.1 Critérios de inclusão e exclusão

Foram incluídas no estudo as QME das avaliações cognitivas realizadas nas atividades de Necessidades e Cuidados em Saúde, no período de janeiro a dezembro de 2019, da terceira e da quarta etapa do curso de graduação em Medicina. Foram excluídas as QME que foram anuladas ou que admitiram duas respostas corretas após recursos dos alunos.

3.3.2 Amostragem

Não foi realizado cálculo amostral para este estudo. Foram incluídos dados de todas as QME com quatro alternativas das avaliações cognitivas de Necessidades e Cuidados em Saúde aplicadas aos alunos da terceira e da quarta etapa do curso de graduação em Medicina da USCS - Campus São Paulo, no ano de 2019.

Nesse ano foram realizadas cinco avaliações cognitivas, totalizando 136 questões de múltipla escolha com quatro alternativas. Destas, apenas quatro se enquadravam nos critérios de exclusão. A população final do estudo foi composta

então por 132 questões de múltipla escolha (Figura 1). O número de participantes em cada avaliação variou de 115 a 117 alunos.

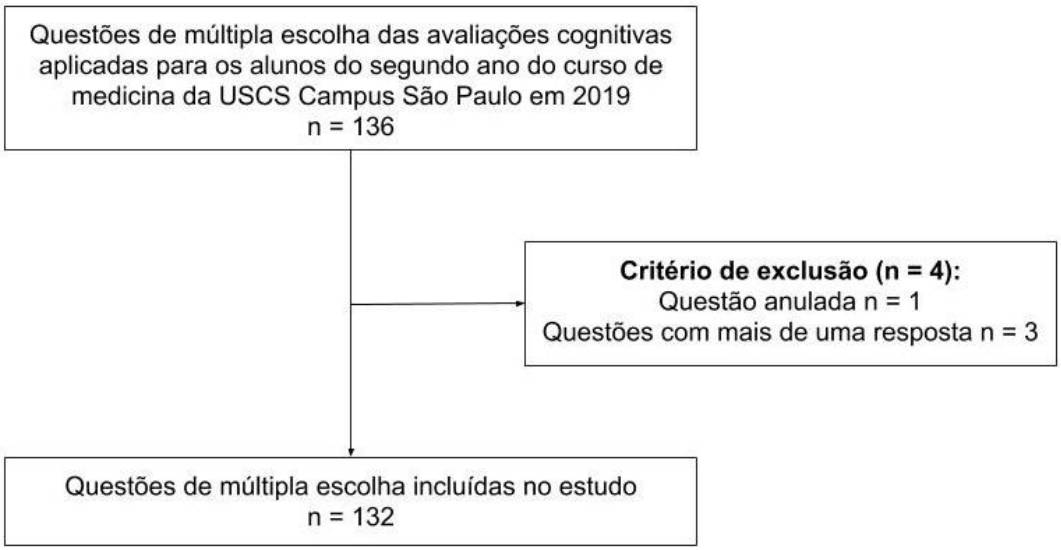

Figura 1- Fluxograma dos critérios de inclusão e exclusão da população do estudo.

3.4 Procedimentos

A coleta de dados teve início em abril de 2019, após a realização da primeira avaliação cognitiva (AC 5) para a terceira etapa do curso de Medicina da USCS – Campus São Paulo. Os dados necessários para a realização deste estudo foram disponibilizados anonimizados pela secretaria com a aprovação da coordenação do curso de Medicina. As QME foram individualmente analisadas quanto à qualidade do seu formato, de acordo com os critérios propostos pelo NBME, e a Taxonomia de Bloom foi revisada pelo pesquisador principal. De acordo com o desempenho de cada QME, foram calculados os índices de dificuldade, discriminação e eficiência dos distratores. Foi realizada também a avaliação da confiabilidade das avaliações como um todo por meio do KR-20.

3.5 Variáveis

3.5.1 Variável primária

a) Confiabilidade das avaliações cognitivas realizadas na atividade curricular de Necessidades e Cuidados em Saúde do Curso de Medicina da USCS - Campus São Paulo.

3.5.2 Variáveis secundárias

a) Índice de dificuldade das questões de múltipla escolha;
b) Índice de discriminação das questões de múltipla escolha;
c) Eficiência dos distratores das questões de múltipla escolha;
d) Taxonomia de Bloom das questões de múltipla escolha;
e) Falhas na elaboração das questões de múltipla escolha;

3.6 Análise estatística

A análise estatística foi realizada de forma quantitativa. A confiabilidade, mais precisamente a consistência interna das avaliações, foi realizada por meio da análise do KR-20.

Cada questão de múltipla escolha foi avaliada por meio dos índices de dificuldade, discriminação e eficiência dos distratores, sendo esses resultados correlacionados entre si e com a Taxonomia de Bloom revisada e a presença de erros na elaboração das questões de múltipla escolha.

De acordo com o resultado obtido no cálculo do índice de dificuldade, as QME foram classificadas em fáceis, moderadas ou difíceis, correspondendo, respectivamente, a índices de dificuldade maiores que 70%, entre 70 e 30% e inferiores a 30% neste estudo (TAVAKOL; DENNICK, 2011). Com relação ao índice de discriminação, foram categorizadas em baixo (inferior a 0,15), moderado (entre 0,15 e 0,25) (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017) e alto (superior 0,25) (TAVAKOL; DENNICK, 2017).

Foram descritos os critérios de avaliação das questões com uso de frequências absolutas e relativas, e os índices com uso de medidas-resumo (média, desvio padrão, mediana, mínimo e máximo) para todas as questões avaliadas.

Foram calculadas as correlações de Pearson entre os índices ou de Spearman (KIRKWOOD; STERNE, 2006) para a eficiência dos distratores, para avaliar a existência de correlações entre os índices. Os índices foram também descritos segundo a Taxonomia de Bloom e segundo a presença de falhas na questão e comparados com uso dos testes *t*-Student ou Mann-Whitney para a eficiência dos distratores (KIRKWOOD; STERNE, 2006).

Foram descritas as falhas nas questões segundo a Taxonomia de Bloom e avaliada a associação entre as características com uso de teste qui-quadrado (KIRKWOOD; STERNE, 2006).

Para a realização das análises foi utilizado o *software* IBM-SPSS *for Windows* versão 22.0, e para a tabulação dos dados foi utilizado o *software* Microsoft Excel 2010. Os testes foram realizados com nível de significância de 5%.

4 ORÇAMENTO

Este estudo não pressupôs gastos financeiros. Eventuais despesas foram custeadas pelo pesquisador principal.

5 ASPECTOS ÉTICOS

5.1 Análise dos riscos e benefícios

a) Riscos

Os riscos atribuídos a esta pesquisa podem ser considerados mínimos, uma vez que utilizou um banco de dados. Não houve exposição do desempenho individual dos alunos.

b) Benefícios

Com os resultados desta pesquisa, foi possível traçar estratégias para a melhoria das avaliações cognitivas e das questões de múltipla escolha elaboradas pelo corpo docente da Faculdade de Medicina da USCS - Campus São Paulo.

5.2 Medidas para minimização dos riscos

Os arquivos contendo os dados da pesquisa serão protegidos por medidas de segurança e compartilhados apenas com os pesquisadores deste estudo.

5.3 Medidas para proteção da confidencialidade

Os dados pessoais dos alunos que realizaram as avaliações utilizadas para análise foram mantidos em sigilo, evitando a exposição do seu desempenho individual.

6 RESULTADOS

6.1 Confiabilidade das Avaliações Cognitivas

Todas as avaliações apresentaram coeficientes de KR-20 abaixo de 0,8, variando entre 0,35 e 0,57 (Tabela 1), que representam baixa confiabilidade e, por conseguinte, ausência de validade.

Tabela 1 - Valor do coeficiente KR-20 de cada avaliação cognitiva

Kr-20	
AC 5	0,41
AC 6	0,57
AC 7	0,44
AC 8	0,38
AC 9	0,35

6.2 Análise psicométrica das questões de múltipla escolha

O índice de dificuldade variou de 5% a 100%, sendo as avaliações majoritariamente compostas por QME fáceis, e raras as questões difíceis. A média do índice de dificuldade nessa amostra foi de 72,7% (± 18,9) (Tabela 2).

Tabela 2 - Descrição dos índices avaliados para todas as QME

Variável	Descrição (N = 132)
Índice de dificuldade (%)	
média ± DP	72,7 ± 18,9
mediana (mín.; máx.)	73,7 (5; 100)
Índice de discriminação	
média ± DP	0,23 ± 0,14
mediana (mín.; máx.)	0,23 (0; 0,64)
Eficiência dos distratores (%)	
média ± DP	50,3 ± 31,8
mediana (mín.; máx.)	50 (0; 100)

A distribuição de questões fáceis, moderadas e difíceis em cada avaliação, segundo o índice de dificuldade, variou como mostra a Tabela 3:

Tabela 3 - Classificação das QME segundo os índices de dificuldade em cada avaliação

Índice de dificuldade	Baixo (difícil)	Médio (moderada)	Alto (fácil)
AC 5	2 (5,9%)	16 (47,1%)	16 (47,1%)
AC 6	1 (2,5%)	14 (35,0%)	25 (62,5%)
AC 7	0 (0,0%)	6 (35,3 %)	11 (64,7%)
AC 8	0 (0,0%)	9 (45,0%)	11 (55,0%)
AC 9	0 (0,0%)	5 (23,8%)	16 (76,2%)
Total	3 (2,3%)	50 (37,9%)	79 (59,8%)

O índice de discriminação variou de 0 a 0,64 em números absolutos com média de 0,23 (± 0,14) (Tabela 2). A porcentagem de questões com baixo índice de discriminação oscilou entre 19,0% e 47,1% em cada prova, correspondendo a 34,8% do total de QME (Tabela 4). Nenhuma questão apresentou índice de discriminação negativo.

Tabela 4 - Classificação das QME segundo os índices de discriminação em cada avaliação

Índice de discriminação	Baixo	Médio	Alto
AC 5	16 (47,1%)	6 (17,6%)	12 (35,3%)
AC 6	17 (42,5%)	12 (30,0%)	11 (27,5%)
AC 7	4 (23,5%)	4 (23,5%)	9 (52,9%)
AC 8	5 (25,0%)	4 (20,0%)	11 (55,0%)
AC 9	4 (19,0%)	4 (19,0%)	13 (61,9%)
Total	46 (34,8%)	30 (22,7%)	56 (42,4%)

A eficiência dos distratores variou de 0% (3 DNF) a 100% (0 DNF) (Tabela 2). Nessa amostra, 50% das questões apresentavam dois ou mais distratores não funcionantes (Tabela 5). Dos 396 distratores que compunham as QME das avaliações, 199 eram funcionantes, e 197(49,7%) eram não funcionantes.

Tabela 5 - Classificação das QME segundo a eficiência dos distratores em cada avaliação

Eficiência dos distratores	3 DNF* (0%)	2 DNF (33%)	1 DNF (66%)	0 DNF (100%)
AC 5	5 (14,7%)	10 (29,4%)	11 (32,4%)	8 (23,5%)
AC 6	6 (15,0%)	13 (32,5%)	13 (32,5%)	8 (20,0%)
AC 7	4 (23,5%)	7 (41,2%)	5 (29,4%)	1 (5,9%)
AC 8	3 (15,0%)	6 (30,0%)	7 (35,0%)	4 (20,0%)
AC 9	3 (14,3%)	9 (42,9%)	8 (38,1%)	1 (4,8%)
Total	21 (15,9%)	45 (34,1%)	44 (33,3%)	22 (16,7%)

*Distrator não funcionante

6.3 Correlação entre os índices das questões de múltipla escolha

Houve correlação estatisticamente significativa entre todos os índices calculados ($p < 0,05$), sendo que o índice de dificuldade apresentou correlação inversa com o índice de discriminação e com a eficiência dos distratores ($r = -0,507$ e $r = -0,771$, respectivamente) (Tabela 6), ou seja, quanto maior o índice de dificuldade e, portanto, mais fácil a questão, menores os demais índices (Gráficos 1 e 2). Já o índice de discriminação e a eficiência dos distratores apresentaram correlação direta (Gráfico 3), ou seja, quanto maior a eficiência dos distratores, maior o índice de discriminação ($r = 0,545$).

Tabela 6 - Correlações entre os índices de avaliação das QME

Correlação		Índice de dificuldade	Índice de discriminação
Índice de discriminação	r	-0,507	
	p	**<0,001**	
Eficiência dos distratores*	r	-0,771	0,545
	p	**<0,001**	**<0,001**

Correlação de Pearson; * Correlação de Spearman

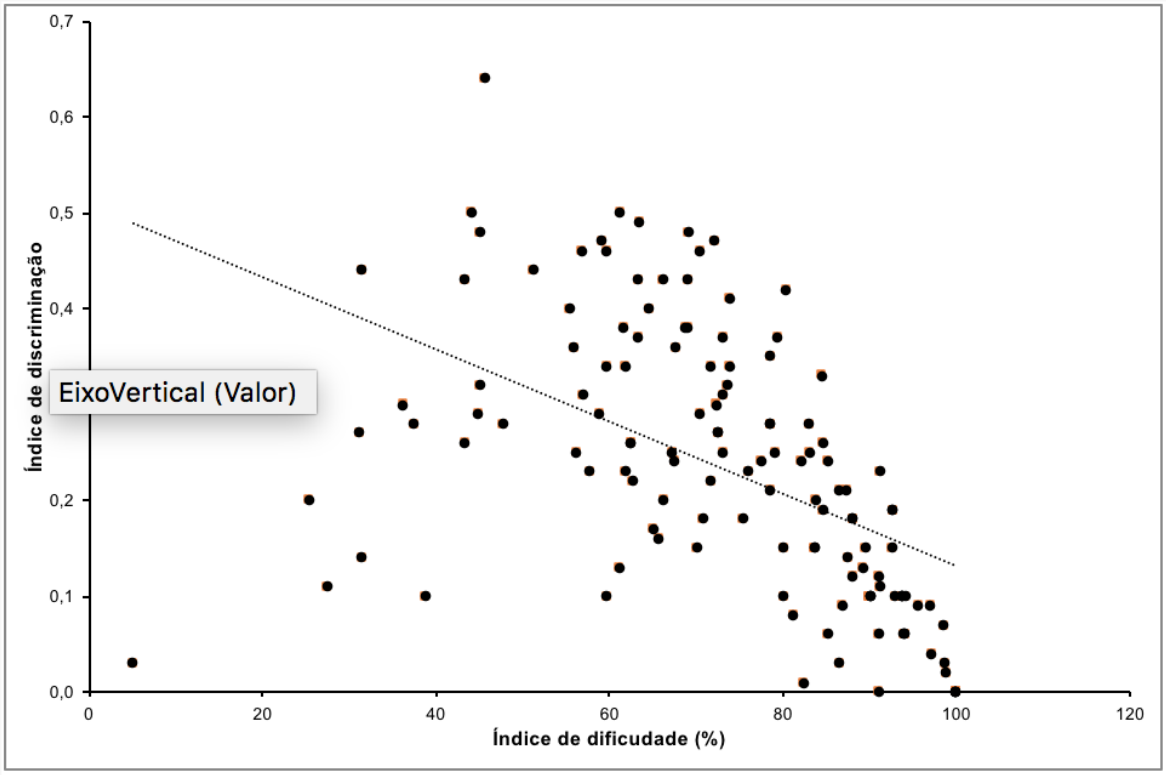

Gráfico 1 - Diagrama de dispersão entre o índice de dificuldade e o índice de discriminação

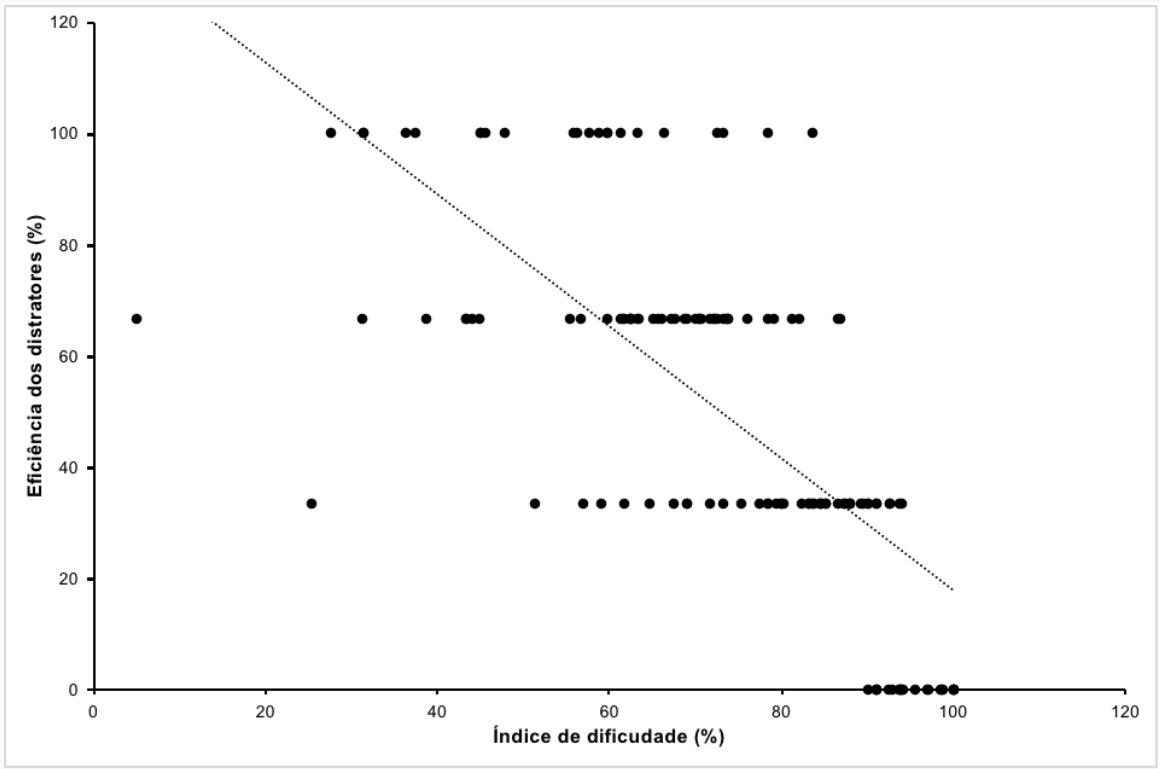

Gráfico 2 - Diagrama de dispersão entre o índice de dificuldade e a eficiência dos distratores

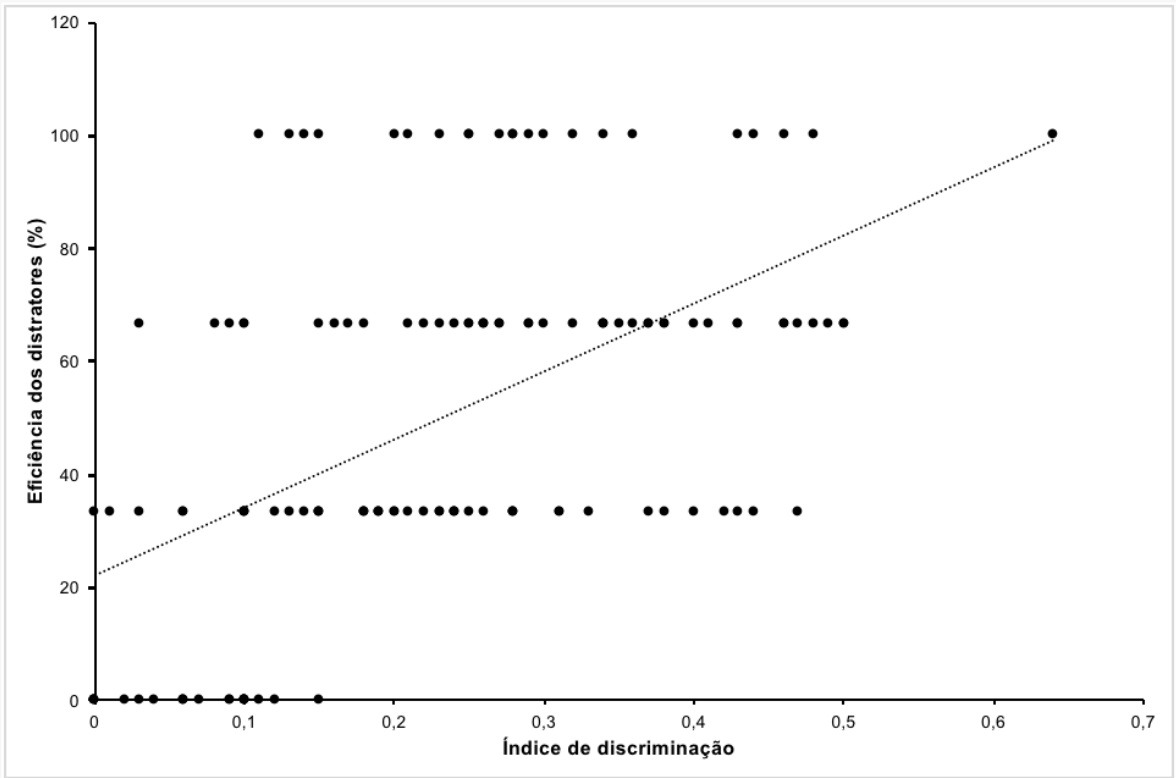

Gráfico 3 - Diagrama de dispersão entre o índice de discriminação e a eficiência dos distratores

6.4 Classificação das questões de múltipla escolha de acordo com a categoria da Taxonomia de Bloom Revisada e a presença de falhas

A porcentagem de QME de baixo nível cognitivo variou de 40,0% a 58,8%, e a de alto nível cognitivo variou de 41,2% a 60,0% em cada avaliação. Da amostra total, 65 questões apresentavam baixo nível cognitivo e 67 apresentavam alto nível cognitivo (Tabela 7).

Tabela 7 - Descrição das QME segundo a categoria da Taxonomia de Bloom revisada em cada avaliação

Taxonomia de Bloom Revisada	Baixo nível	Alto nível
AC5	20 (58,8%)	14 (41,2%)
AC6	16 (40,0%)	24 (60,0%)
AC7	10 (58,8%)	7 (41,2%)
AC8	8 (40,0%)	12 (60,0%)
AC9	11 (52,4%)	10 (47,6%)
Total	65 (49,2%)	67 (50,8%)

A porcentagem de falhas nas QME variou de 14,3% a 57,5% em cada avaliação. Da amostra total, 49 (37,1%) questões não estavam de acordo com as recomendações do NBME 2016 (Tabela 8).

Tabela 8 - Descrição da presença de falhas nas QME em cada avaliação

Falhas nas QME	Não	Sim
AC 5	26 (76,5%)	8 (23,5%)
AC 6	17 (42,5%)	23 (57,5%)
AC 7	11 (64,7%)	6 (35,3%)
AC 8	11 (55,0%)	9 (45,0%)
AC 9	18 (85,7%)	3 (14,3%)
Total	83 (62,9%)	49 (37,1%)

6.5 Correlação entre os índices das questões de múltipla escolha e a Taxonomia de Bloom Revisada

Os índices calculados para as QME das avaliações aplicadas não diferiram estatisticamente entre as categorias da Taxonomia de Bloom Revisada ou segundo presença de falhas nas questões (Tabelas 9 e 10).

Tabela 9 - Descrição dos índices segundo a categoria da Taxonomia de Bloom revisada e resultado dos testes comparativos

Variável	Taxonomia de Bloom Revisada		p
	Baixo nível (N = 65)	Alto nível (N = 67)	
Índice de dificuldade (%)			0,292
média ± DP	71 ± 18,5	74,5 ± 19,2	
mediana (mín.; máx.)	72,5 (25,4; 100)	78,5 (5; 100)	
Índice de discriminação			0,188
média ± DP	0,25 ± 0,14	0,22 ± 0,14	
mediana (mín.; máx.)	0,26 (0; 0,5)	0,2 (0; 0,64)	
Eficiência dos distratores (%)			0,896*
média ± DP	50,8 ± 30,1	49,8 ± 33,5	
mediana (mín.; máx.)	33,3 (0; 100)	66,7 (0; 100)	

Teste t-Student; * Teste Mann-Whitney

Tabela 10 - Descrição dos índices segundo falhas nas questões e resultado dos testes comparativos

Variável	Falhas da questão		p
	Não (N = 83)	Sim (N = 49)	
Índice de dificuldade (%)			0,482
média ± DP	73,6 ± 18,1	71,2 ± 20,2	
mediana (mín.; máx.)	73,9 (5; 100)	72,1 (25,4; 100)	
Índice de discriminação			0,566
média ± DP	0,23 ± 0,14	0,24 ± 0,14	
mediana (mín.; máx.)	0,23 (0; 0,64)	0,24 (0; 0,48)	
Eficiência dos distratores (%)			0,194*
média ± DP	47,4 ± 30,4	55,1 ± 33,7	
mediana (mín.; máx.)	33,3 (0; 100)	66,7 (0; 100)	

Teste t-Student; * Teste Mann-Whitney

6.6 Correlação entre a categoria da Taxonomia de Bloom Revisada e a presença de falhas nas QME

Não houve associação estatisticamente significativa entre a categoria da Taxonomia de Bloom Revisada e a presença de falhas nas questões (p = 0,500) (Tabela 11).

Tabela 11 - Descrição das falhas nas questões segundo a Taxonomia de Bloom Revisada e resultado do teste de associação

Taxonomia de Bloom Revisada	Falhas na questão		Total	p
	Não n (%)	Sim n (%)		
Baixo nível	39 (60)	26 (40)	65	0,500
Alto nível	44 (65,7)	23 (34,3)	67	
Total	**83 (62,9)**	**49 (37,1)**	**132**	

Teste qui-quadrado

7 DISCUSSÃO

Todas as avaliações apresentaram baixa confiabilidade. Algumas estratégias podem ser tomadas para melhorar esse parâmetro. Podemos observar que, apesar da grande proporção de questões fáceis com baixo índice de discriminação, neste estudo a AC 06 apresentou o mais alto coeficiente KR-20, o que poderia até ser considerado adequado para fins educativos (TARRANT; WARE, 2008), provavelmente relacionado à maior quantidade de QME (TAVAKOL; DENNICK, 2011). O Guia No. 57 da *Association for Medical Education in Europe* (AMEE) deixa claro que não podemos confiar em avaliações curtas ou com uma quantidade reduzida de casos clínicos diferentes para tomar decisões de alto risco, pois são inerentemente falhas quanto à confiabilidade e, por conseguinte, à validade (SCHUWIRTH; VAN DER VLEUTEN, 2011). Além disso, estudos que demonstraram coeficientes maiores incluíram avaliações com uma gama muito maior de QME (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017). Dado o caráter somativo das avaliações cognitivas incluídas neste estudo, caso ainda se opte pelo uso de QME, seria, portanto, importante aumentar o número de questões por avaliação (TAVAKOL; DENNICK, 2011).

Embora a média do índice de dificuldade e a porcentagem de questões fáceis possam parecer preocupantes a princípio, devemos lembrar que muitas avaliações educacionais, mesmo as somativas, são desenvolvidas com a finalidade de determinar se os alunos estão suficientemente aptos a progredir de nível, e não a de selecionar apenas os mais brilhantes, como, por exemplo, os exames de admissão para a residência médica (RAYMOND; STEVENS; BUCAK, 2019). Assim, é comum encontrarmos avaliações com média dos índices de dificuldade em torno de 80-90% (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; RAYMOND; STEVENS; BUCAK, 2019).

Abdulghani e colaboradores (2017) encontraram um valor semelhante à média do índice de dificuldade, mas obtiveram maiores porcentagens de distratores não funcionantes e menores porcentagem de questões com baixo índice de discriminação. Distratores não funcionantes ameaçam a validade dos valores obtidos pelas QME (ALI; RUIT, 2015; DOWNING; HALADYNA, 2004). Estudos demonstraram que há mínimo impacto na dificuldade média e na capacidade discriminatória do item ao se eliminar o distrator menos funcionante de QME com quatro alternativas (TARRANT;

WARE, 2010), permitindo que os demais distratores sejam mais escolhidos e discriminantes (RODRIGUEZ, 2005). QME com menos alternativas deveriam, na teoria, reduzir o tempo de elaboração, revisão e edição pelos docentes e de leitura por item pelos discentes, permitindo que mais questões sejam incluídas na avaliação (RAYMOND; STEVENS; BUCAK, 2019).

Entretanto, o critério utilizado neste e nos estudos descritos no parágrafo anterior para determinar os distratores não funcionantes, apesar de amplamente reconhecido, depende da dificuldade geral do item. Questões fáceis terão poucos alunos disponíveis para selecionar distratores (RAYMOND; STEVENS; BUCAK, 2019). Por exemplo, uma questão que foi respondida corretamente por 90% dos alunos, caso a alternativa escolhida pelos 10% restantes seja igualmente distribuída entre os distratores, nenhum deles será funcionante, podendo, ainda assim, ser efetiva. Logo, para avaliações que visam a avaliar a suficiência, predominantemente compostas por questões fáceis, esse critério não parece ser muito apropriado.

Recentemente, Raymond e colaboradores (2019) sugeriram um refinamento para o cálculo da eficiência dos distratores, de forma que o ponto de corte para determinar um DNF varie de acordo com o índice de dificuldade da questão. Neste trabalho, a porcentagem de DNF passou de 58,9% para 34,7% após a aplicação do novo critério (RAYMOND; STEVENS; BUCAK, 2019), e, assim, alguns distratores antes considerados irrelevantes foram considerados importantes para a qualidade do item, justificando o posicionamento dos autores a favor da manutenção de quatro alternativas para as QME.

De posse dessas informações, a reestruturação dos DNF parece ser a medida mais adequada para aumentar a dificuldade e a capacidade de discriminação das QME, dada a correlação estatisticamente significativa encontrada entre os índices, contribuindo também para a maior confiabilidade e validade da avaliação (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017).

Uma desvantagem da Teoria Clássica do Teste utilizada neste estudo é que não há como diferenciar os efeitos da dificuldade dos testes dos efeitos do grupo de alunos no qual foram aplicados (SCHUWIRTH; VAN DER VLEUTEN, 2011; TAVAKOL; DENNICK, 2012; 2013). Em outras palavras, tomando como exemplo a alta porcentagem de questões consideradas fáceis nas avaliações analisadas, podemos inferir que essas QME eram realmente fáceis ou que os alunos estavam muito bem preparados. Outro aspecto negativo é que seus resultados estatísticos são

dependentes do tamanho da amostra, podendo ser altos ou baixos a despeito da qualidade da questão (TAVAKOL; DENNICK, 2013). A Teoria da Resposta ao Item tenta superar esses problemas estimando a dificuldade do item independentemente da habilidade do aluno, e vice-versa, podendo ser uma ferramenta para futuros estudos (SCHUWIRTH; VAN DER VLEUTEN, 2011; TAVAKOL; DENNICK, 2012; 2013).

A proporção de questões envolvendo baixo nível cognitivo e alto nível cognitivo poderia ser mais bem equilibrada em algumas avaliações, mas não está inadequada para aplicação em alunos do ciclo básico do curso de Medicina. Estudos nessa área encontraram maior proporção de questões envolvendo baixo nível cognitivo mesmo em ano posterior da educação médica (PALMER; DEVITT, 2007). O uso de casos clínicos aliado às outras orientações para a construção de boas QME, bem como o maior conhecimento acerca das taxonomias do domínio cognitivo e suas aplicações, como a de Bloom modificada ou a de Marzano, podem auxiliar no aumento das questões de alto nível cognitivo (ANDERSON; KRATHWOHL, 2000; IRVINE, 2017; MARZANO, 2008).

Identificamos uma porcentagem consideravelmente alta de QME com erros, sobretudo se levarmos em consideração que haviam sido revisadas pelos docentes responsáveis que estavam cientes das recomendações para a elaboração de boas questões. Pouca compreensão dos objetivos educacionais, dos princípios da avaliação, da construção de QME e poucas oportunidades de treinamento e tempo reduzido para elaboração podem justificar esse achado (TAVAKOL; DENNICK, 2011).

Considerando-se todas as avaliações, a porcentagem de falhas nas QME está condizente com o que foi encontrado por Downing (2002; 2005) ao analisar a qualidade das avaliações aplicadas a alunos de Medicina nos Estados Unidos; na sua amostra, 33-46% das questões apresentavam erros (DOWNING, 2002; 2005). Como consequência, 10-25% dos alunos reprovados teriam sido aprovados se estas questões fossem removidas (DOWNING, 2002; 2005). Apesar de não termos identificado correlação entre o nível cognitivo e a presença de falhas, Tarrant e Ware (2008) demonstraram que QME que testavam funções cognitivas baixas eram mais propensas a apresentar erros na sua construção.

A análise psicométrica das QME auxilia na tomada de decisão, uma vez que permite a identificação de questões com índices de dificuldade extremos ou com índices de discriminação negativos, podendo ser descartadas ou reestruturadas, por

exemplo, por meio da modificação dos DNF (TAVAKOL; DENNICK, 2013). Além disso, possibilita a criação de um banco de questões, um investimento de curto prazo de tempo e esforço que pode, no futuro, reduzir o ônus da elaboração de novas avaliações, uma vez que os itens podem ser compartilhados entre cursos e até mesmo entre escolas médicas (CODERRE; WOLOSCHUK; MCLAUGHLIN, 2009; TAVAKOL; DENNICK, 2017).

A literatura demonstra que atividades de educação continuada são capazes de melhorar a qualidade das QME no que tange aos índices de discriminação e dificuldade, ao nível da Taxonomia de Bloom do domínio cognitivo, redução de erros e aumento de distratores funcionantes, melhorando a confiabilidade e a validade das avaliações (ABDULGHANI; IRSHAD; HAQUE; AHMAD et al., 2017; JOZEFOWICZ; KOEPPEN; CASE; GALBRAITH et al., 2002; TARRANT; WARE, 2008).

7.1 Pontos fortes do estudo

Por meio deste estudo, experiências foram adquiridas quanto à elaboração das avaliações cognitivas e, sobretudo, das questões de múltipla escolha, bem como quanto à análise psicométrica das mesmas, que poderão ser compartilhadas com todos os docentes do departamento de Medicina para aplicação nos demais anos do curso. Espera-se que, desta forma, ocorra a melhoria gradual dos processos avaliativos na instituição pesquisada.

7.2 Limitações do estudo

A generalização desses achados é limitada por múltiplos fatores. Este foi um estudo descritivo, realizado em apenas uma turma do Curso de Medicina da USCS por um período limitado de tempo. Além disso, como as avaliações não foram selecionadas aleatoriamente, a análise pode apresentar viés de seleção e não refletir adequadamente a realidade das avaliações e, por conseguinte, das QME aplicadas nessa instituição.

A análise das QME de acordo com a Taxonomia de Bloom revisada e a presença de erros foi realizada por apenas uma pessoa. Estudos demonstram que há grande variabilidade e baixas confiabilidade e acurácia da Taxonomia de Bloom quando a classificação é realizada por avaliadores diferentes, havendo melhor concordância quando feitas por especialistas em avaliação (KARPEN; WELCH, 2016;

NÄSSTRÖM, 2009). Entretanto, a reorganização dos seis níveis em três, ou menos, categorias se mostrou uma estratégia eficiente para melhorar a acurácia (KARPEN; WELCH, 2016).

7.3 Perspectivas Futuras

Novas teorias, ou extensões das existentes, estão surgindo, a maioria relacionada à mudança de visão da “avaliação do aprendizado” para “avaliação para o aprendizado”, uma vez que a educação médica não é somente psicologia cognitiva ou apenas psicométrica (SCHUWIRTH; VAN DER VLEUTEN, 2011). Essas mudanças são ainda mais significativas nas metodologias construtivistas (VAN DER VLEUTEN; SCHUWIRTH, 2019).

Avaliar para o aprendizado exige que informações sejam extraídas dos vários instrumentos de avaliação para responder às seguintes perguntas: (1) Consigo compreender globalmente o processo de aprendizado deste aluno ou necessito de informação adicional? (questão diagnóstica); (2) Qual intervenção educacional é mais indicada para este aluno neste momento? (questão terapêutica); e (3) Este estudante está no caminho certo para se tornar um bom profissional no tempo proposto? (questão prognóstica) (SCHUWIRTH; VAN DER VLEUTEN, 2011).

Fica claro que essas informações não podem ser alcançadas por meio de poucos métodos de avaliação, mas sim de um conjunto de métodos com seus pontos fortes e fracos, alguns mais objetivos, como as QME, e outros mais subjetivos, compondo, juntos, o programa de avaliação da instituição (SCHUWIRTH; VAN DER VLEUTEN, 2011). Surge assim a ideia de avaliação programática (SCHUWIRTH; VAN DER VLEUTEN, 2011), na qual as informações obtidas pelos métodos avaliativos são trianguladas com as competências esperadas para aquele aluno (VAN DER VLEUTEN; SCHUWIRTH, 2019). As decisões de alto risco (promoção, graduação) são tomadas por comitês de competência e as decisões intermediárias são tomadas com o objetivo de informar o aluno sobre seu progresso, que tem reuniões de aprendizado recorrentes com mentores usando uma auto-análise de todos os dados das avaliações realizadas (VAN DER VLEUTEN; SCHUWIRTH, 2019).

Entretanto, a garantia da qualidade do programa de avaliação depende, apesar de não exclusivamente, da qualidade dos seus componentes. Assim, novos estudos podem ser realizados para analisar a qualidade dos diversos instrumentos de

avaliação utilizados no Curso de Medicina da USCS – Campus São Paulo, bem como do programa de avaliação como um todo. Uma dimensão importantíssima, por exemplo, seria avaliar a percepção dos alunos e do corpo docente sobre o feedback dos processos avaliativos, para compreender os fatores que influenciam seu bom funcionamento, ou não, e propor mudanças na forma como são realizados (SCHUWIRTH; VAN DER VLEUTEN, 2011).

8 CONCLUSÃO

Este estudo demonstrou que as avaliações e as QME analisadas são pouco confiáveis e válidas. Muitas QME apresentam falhas na sua elaboração com baixos índices de discriminação e poucos distratores funcionantes.

A manutenção das atividades de educação permanente e continuada oferecendo oficinas de capacitação docente e a análise psicométrica das avaliações com *feedback* aos docentes podem auxiliar o processo de melhoria na elaboração de QME.

REFERÊNCIAS

ABDULGHANI, H. M.; AHMAD, F.; IRSHAD, M.; KHALIL, M. S. et al. Faculty development programs improve the quality of Multiple Choice Questions items' writing. **Sci Rep**, 5, p. 9556, Apr 2015.

ABDULGHANI, H. M.; IRSHAD, M.; HAQUE, S.; AHMAD, T. et al. Effectiveness of longitudinal faculty development programs on MCQs items writing skills: A follow-up study. **PLoS One**, 12, n. 10, p. e0185895, 2017.Disponível em: http://www.scielo.br/scielo.php?pid=S0104-530X2010000200015&script=sci_arttex. Acesso em 13 jun 2020.

ALFARIS, E.; NAEEM, N.; IRFAN, F.; QURESHI, R. et al. A One-Day Dental Faculty Workshop in Writing Multiple-Choice Questions: An Impact Evaluation. **J Dent Educ**, 79, n. 11, p. 1305-1313, Nov 2015.

ALI, S. H.; RUIT, K. G. The Impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. **Perspect Med Educ**, 4, n. 5, p. 244-251, Oct 2015.

ANDERSON, L.; KRATHWOHL, D. (Ed.) **A Taxonomy for learning, teaching, and assessing:** A revision of Bloom's Taxonomy of Educational Objectives**.** Boston: Longman. 2000.

BRASIL. Ministério da Educação. Câmara de Educação Superior. Conselho Nacional de Educação. Resolução nº 3, de 20 de junho de 2014. Institui Diretrizes Curriculares Nacionais do Curso de Graduação em Medicina e dá outras providências. 2014.

BROUSSELLE, A. et al. **Avaliação:** conceitos e métodos. 2. ed. Rio de Janeiro: Editora Fiocruz, 2011.

CAVALCANTI NETO, A. L. G.; AQUINO, J. L. F. A avaliação da aprendizagem como um ato amoroso: o que o professor pratica? **Educ. rev.**, Belo Horizonte, v. 25, n. 2, p. 223-240, ago. 2009. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-46982009000200010&lng=en&nrm=iso. Acesso em 28 out 2019.

CILLIERS, F. J.; SCHUWIRTH, L. W.; HERMAN, N.; ADENDORFF, H. J. et al. A model of the pre-assessment learning effects of summative assessment in medical education. **Adv Health Sci Educ Theory Pract**, 17, n. 1, p. 39-53, Mar 2012.

CODERRE, S.; WOLOSCHUK, W.; MCLAUGHLIN, K. Twelve tips for blueprinting. **Med Teach**, 31, n. 4, p. 322-324, Apr 2009.

COOMBES, L.; ROBERTS, M.; ZAHRA, D.; BURR, S. Twelve tips for assessment psychometrics. **Med Teach**, 38, n. 3, p. 250-254, 2016.

COUGHLIN, P. A.; FEATHERSTONE, C. R. How to Write a High Quality Multiple Choice Question (MCQ): A Guide for Clinicians. **Eur J Vasc Endovasc Surg**, 54, n. 5, p. 654-658, Nov 2017.

DEEPAK, K. K.; AL-UMRAN, K. U.; AI-SHEIKH, M. H.; DKOLI, B. V. et al. Psychometrics of Multiple Choice Questions with Non-Functioning Distracters: Implications to Medical Education. **Indian J Physiol Pharmacol**, 59, n. 4, p. 428-435, 2015 Oct-Dec 2015.

DOWNING, S. M. Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? **Acad Med**, 77, n. 10 Suppl, p. S103-104, Oct 2002.

DOWNING, S. M. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. **Adv Health Sci Educ Theory Pract**, 10, n. 2, p. 133-143, 2005.

DOWNING, S. M.; HALADYNA, T. M. Validity threats: overcoming interference with proposed interpretations of assessment data. **Med Educ**, 38, n. 3, p. 327-333, Mar 2004.

EPSTEIN, R. M. Assessment in medical education. **The New England Journal of Medicine**, v. 356, n. 4, p. 387–396, 25 Jan. 2007.

GAJJAR, S.; SHARMA, R.; KUMAR, P.; RANA, M. Item and Test Analysis to Identify Quality Multiple Choice Questions (MCQs) from an Assessment of Medical Students of Ahmedabad, Gujarat. **Indian J Community Med**, 39, n. 1, p. 17-20, Jan 2014.

IRVINE, J. A comparison of revised Bloom and Marzano's New Taxonomy of Learning. **Research in Higher Education Journal,** v. 33. Brock University. Disponível em: https://europe-creates.eu/wp-content/uploads/2020/03/172608.pdf. Acesso em 13 jun 2020.

JOZEFOWICZ, R. F.; KOEPPEN, B. M.; CASE, S.; GALBRAITH, R. et al. The quality of in-house medical school examinations. **Acad Med**, 77, n. 2, p. 156-161, Feb 2002.

KARPEN, S. C.; WELCH, A. C. Assessing the inter-rater reliability and accuracy of pharmacy faculty's Bloom's Taxonomy classifications. **Currents in Pharmacy Teaching and Learning,** v. 8, issue 6, pp 885-888, 2016. Disponível em: https://doi.org/10.1016/j.cptl.2016.08.003. Acesso em 13 jun 2020.

KHAFAGY, G.; AHMED, M.; SAAD, N. Stepping up of MCQs' quality through a multi-stage reviewing process. **Educ Prim Care**, 27, n. 4, p. 299-303, Jul 2016.

KIRKWOOD, B.R.; STERNE, J.A.C. **Essential Medical Statistics,** 2 ed. ISBN: 978-0-865-42871-3. Malden, USA: Blackwell Publishing. 512 pp. 2003.

LUCKESI, C. C. **Avaliação da aprendizagem componente do ato pedagógico.** 1. ed. São Paulo: Cortez, 2011.

MARZANO, R. J.; KENDALL, J. S. (Ed.) **Designing and assessing educational objectives**: applying the new taxonomy. 1. ed. Sage Publishing. 2008.

NÄSSTRÖM, G. Interpretation of standards with Bloom's revised taxonomy: a comparison of teachers and assessment experts. **International Journal of Research & Method in Education**, v. 32, no. 1, 39–51, April 2009.

NATIONAL BOARD OF MEDICAL EXAMINERS (NBME). **Constructing written test questions for the basic and clinical science.** 2016. Disponível em https://www.unmc.edu/facdev/_documents/ConstructingWrittenTestQuestions_WritingManual.pdf. Acesso em 13 jun 2020.

PADILHA, R. Q. et al. **Curso de graduação em Medicina, projeto pedagógico do curso - PPC:** Projeto em parceria USCS e IEP/HSL - São Caetano do Sul; São Paulo: Universidade Municipal de São Caetano do Sul; Instituto Sírio-Libanês de Ensino e Pesquisa, 2016.

PALMER, E. J.; DEVITT, P. G. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. **BMC Med Educ**, 7, p. 49, Nov 2007.

PERRENOUD, P. **Avaliação:** da excelência à regulação da aprendizagem entre duas lógicas. Porto Alegre: Artes Médicas, 1999.

RAYMOND, M. R.; STEVENS, C.; BUCAK, S. D. The optimal number of options for multiple-choice questions on high-stakes tests: application of a revised index for detecting nonfunctional distractors. **Adv Health Sci Educ Theory Pract**, 24, n. 1, p. 141-150, 03 2019.

RODRIGUEZ, M. C. **Three options are optimal for multiple-choice items:** A meta-analysis of 80 years of research. Educational Measurement: Issues and Practice. v. 24: pp 3-13. 2005; Disponível em: https://doi.org/10.1111/j.1745-3992.2005.00006.x. Acesso em 13 jun 2020.

SCHUWIRTH, L. W.; VAN DER VLEUTEN, C. P. General overview of the theories used in assessment: AMEE Guide No. 57. **Med Teach**, 33, n. 10, p. 783-797, 2011a.

SCHUWIRTH, L. W. T.; VAN DER VLEUTEN, C. P. M. Programmatic assessment: From assessment of learning to assessment for learning. **Medical Teacher**, v. 33, n. 6, p. 478–485, 2011b.

SHUMWAY, J. M.; HARDEN, R. M.; EUROPE, A. F. M. E. I. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. **Med Teach**, 25, n. 6, p. 569-584, Nov 2003.

TARRANT, M.; WARE, J. A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments. **Nurse Educ Today**, 30, n. 6, p. 539-543, Aug 2010.

TARRANT, M.; WARE, J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. **Med Educ**, 42, n. 2, p. 198-206, Feb 2008.

TAVAKOL, M.; DENNICK, R. Post-examination analysis of objective tests. **Med Teach**, 33, n. 6, p. 447-458, 2011.

TAVAKOL, M.; DENNICK, R. Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. **Med Teach**, 34, n. 3, p. e161-175, 2012.

TAVAKOL, M.; DENNICK, R. Psychometric evaluation of a knowledge based examination using Rasch analysis: an illustrative guide: AMEE guide no. 72. **Med Teach**, 35, n. 1, p. e838-848, 2013.

TAVAKOL, M.; DENNICK, R. The foundations of measurement and assessment in medical education. **Med Teach**, 39, n. 10, p. 1010-1015, Oct 2017.

VAN DER VLEUTEN, C. P.; SCHUWIRTH, L. W. Assessing professional competence: from methods to programmes. **Med Educ**, 39, n. 3, p. 309-317, Mar 2005.

VAN DER VLEUTEN, C. P. M.; SCHUWIRTH, L. W. T. Assessment in the context of problem-based learning. **Advances in health sciences education : theory and practice**, v. 24, n. 5, p. 903–914, 2 Oct. 2019.

VANDERBILT, A. A.; FELDMAN, M.; WOOD, I. K. Assessment in undergraduate medical education: a review of course exams. **Med Educ Online**, 18, p. 1-5, Mar 2013.