UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM SAÚDE

Monica Couto Guedes Sejanes da Rocha

Avaliação do desempenho do estudante de medicina pela medida da proficiência: descrição de uma metodologia de análise dos resultados do Teste de Progresso por meio da Teoria de Resposta ao Item.

Juiz de Fora
2023

Monica Couto Guedes Sejanes da Rocha

Avaliação do desempenho do estudante de medicina pela medida da proficiência: descrição de uma metodologia de análise dos resultados do Teste de Progresso por meio da Teoria de Resposta ao Item.

Tese apresentada ao Programa de Pós-Graduação em Saúde da Faculdade de Medicina da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutora em Saúde - Área de concentração: Saúde Brasileira.

Orientador: Prof. Dr. Julio Maria Fonseca Chebli

Coorientador: Profa. Dra. Sandra Helena Cerrato Tibiriçá

Juiz de Fora

2023

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Guedes Sejanes da Rocha, Monica Couto.
Avaliação do desempenho do estudante de medicina pela medida da proficiência: : descrição de uma metodologia de análise dos resultados do Teste de Progresso por meio da Teoria de Resposta ao Item. / Monica Couto Guedes Sejanes da Rocha. -- 2023.
79 f. : il.

Orientador: Julio Maria Fonseca Chebli
Coorientadora: Sandra Helena Cerrato Tibiriçá
Tese (doutorado) - Universidade Federal de Juiz de Fora, Faculdade de Medicina. Programa de Pós-Graduação em Saúde Brasileira, 2023.

1. Educação Médica. 2. Avaliação educacional. 3. Cognição. 4. Competência profissional. 5. Psicometria. I. Fonseca Chebli, Julio Maria, orient. II. Cerrato Tibiriçá, Sandra Helena, coorient. III. Título.

Monica Couto Guedes Sejanes da Rocha

Avaliação do desempenho do estudante de medicina pela medida da proficiência: descrição de uma metodologia de análise dos resultados do Teste de Progresso por meio da Teoria de Resposta ao Item.

Tese apresentada ao Programa de Pós-Graduação em Saúde da Faculdade de Medicina da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Doutora em Saúde - Área de concentração: Saúde Brasileira.

Aprovada em ____/____/____

BANCA EXAMINADORA

_____

Prof. Dr. Julio Maria Fonseca Cebli - Orientador
Universidade Federal de Juiz de Fora

_____

Profa. Dra. Sandra Helena Cerrato Tibiriçá
Universidade Federal de Juiz de Fora

_____

Prof. Dr. Djalma Rabelo Ricardo
Faculdade de Ciências Médicas e da Saúde – Suprema

_____

Profa. Dra. Oscarina da Silva Ezequiel

Universidade Federal de Juiz de Fora

_____

Prof. Dr. Wellington Silva

Centro de Políticas Públicas e Avaliação da Educação-CAED/UFJF

_____

Profa. Dra. Ana Lúcia de Lima Guedes

Universidade Federal de Juiz de Fora

Juiz de Fora, 07 / 07 / 2023.

Documento assinado eletronicamente por **Monica Couto Guedes Sejanes da Rocha**, **Usuário Externo**, em 10/07/2023, às 21:53, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Djalma Rabelo Ricardo**, **Usuário Externo**, em 14/07/2023, às 11:05, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ana Paula Ferreira**, **Usuário Externo**, em 14/07/2023, às 11:11, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Julio Maria Fonseca Chebli**, **Professor(a)**, em 11/08/2023, às 11:38, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Sandra Helena Cerrato Tibirica**, **Professor(a)**, em 11/08/2023, às 18:33, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **WELLINGTON SILVA**, **Usuário Externo**, em 14/08/2023, às 15:36, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ana Lucia de Lima Guedes**, **Chefe de Departamento**, em 16/08/2023, às 11:25, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **1344953** e o código CRC **66F02736**.

Dedico essa tese ao meu marido Cláudio, e aos meus filhos, Julia e Pedro, por darem sentido e iluminarem cada dia da minha vida.

AGRADECIMENTOS

A Deus, que cuidou de mim e me mostrou luz em momentos em que parecia inalcançável a conclusão desse projeto.

Ao meu marido Cláudio, meu amor, meu porto seguro, companheiro com quem escolhi compartilhar a vida. Obrigada pelo respeito, compreensão, apoio e cuidado em todos os momentos dessa nossa jornada.

Aos meus filhos Júlia e Pedro, agradeço por serem meus maiores mentores em metodologias ativas de aprendizagem para aquisição das competências essenciais previstas no meu projeto pedagógico de curso de mãe. Meu amor por vocês é incondicional, imensurável e atemporal. Estarei sempre com vocês, independentemente de onde a vida os levar.

Ao meu avô Félix e minha avó Otília, que jamais me abandonaram ou mediram esforços para minha educação e crescimento pessoal. Seu amor e a vida dedicados a mim são bens eternos. Ao meu avô, que partiu tão cedo, agradeço o amor, abnegação e luta para minha formação. Você foi e sempre será meu exemplo de pai e de caráter. À minha avó, que sempre transpões todas as dificuldades do caminho para que meus sonhos se realizassem, agradeço a felicidade de tê-la ao meu lado e poder cuidar e amar como sempre fez comigo. A vocês meu amor eterno.

À Laura, minha amada Laura, que escolheu ficar ao meu lado, cuidar de mim como filha e aos meus filhos como netos, dedicando a nós sua vida. Seu carinho e doação são inestimáveis. Agradeço por tê-la ao meu lado e por seu sorriso cada vez que me vê. Meu sincero amor e gratidão por toda a eternidade.

À Ana Maria e José Augusto, primos/irmãos, sempre ao meu lado em todos os momentos. Obrigada por possibilitar à Laura e Otília, dias mais felizes.

À Angélica, Cláudia e Isabella, que com tanto amor cuidam das minhas preciosas Otília e Laura, tornando nossa vida mais leve e alegre.

À Lúcia, que há anos cuida de nossa família como se fosse dela, sempre presente.

À Rosa, minha amiga/irmã, sou grata a Deus por sua presença na minha vida e por permitir que faça parte da sua. Amiga leal, de todos os momentos, um sentimento para o qual não tenho palavras. Que sigamos juntas para sempre!

Ao meu orientador professor doutor Júlio Maria Fonseca Chebli, profissional ético e compromissado com a Ciência, pela oportunidade oferecida e por compartilhar

seu conhecimento. Agradeço a disponibilidade, compreensão e apoio na conclusão desse projeto.

À professora doutora Sandra Helena Cerrato Tibiriçá, exemplo de ética, competência e comprometimento com a pesquisa científica, agradeço por ter acreditado e enfrentado ao meu lado, os muitos obstáculos à realização desse projeto. Obrigada por sua consistente colaboração para o meu crescimento como docente e pesquisadora, mas principalmente agradeço por ser a Tibi, a amiga que estimo, confio e sempre serei leal. Sem sua persistência e apoio nada teria sido possível.

Aos prezados professores da banca, por generosamente aceitarem o convite e dividirem comigo seu conhecimento. À professora doutora Oscarina Ezequiel, que tenho por referência como educadora e gestora na Educação em Saúde, agradeço o carinho, a generosidade e a honra de poder compartilhar seus ensinamentos. Ao professor doutor Djalma Rabelo, agradeço a confiança e ressalto minha profunda admiração e respeito por sua gestão ética, fundamentada no compromisso com a educação e respeito pela instituição no seu todo. Ao professor doutor Wellington Silva, por sua generosidade e disponibilidade, que por meio de sua vasta experiência em avaliações educacionais em larga escala, me deu suporte com a análise dos resultados do estudo. À professora doutora Ana Lúcia Guedes, Aninha para mim, agradeço por todos esses anos de amizade sincera, pelos bons momentos juntas na assistência e na docência, por seu carinho e incentivo.

Um agradecimento especial ao Prof. Raimundo Bechara, Coordenador do Curso de Medicina da FCMS/JF- Suprema, exemplo de gestor e ser humano. Obrigada por lá no início, ter confiado em mim e dado a oportunidade de trabalhar com avaliação educacional, que é o motivo pelo qual essa tese existe. Obrigada pelo carinho, amizade e apoio incondicionais. Conte sempre com minha amizade e lealdade.

Aos professores, amigos da Pediatria, pelo apoio para que eu tivesse tranquilidade para finalizar esta etapa da minha vida profissional.

À Agna Fagundes, que deixei intencionalmente para o final, agradeço por me resgatar, orientar minha busca e ajudar a reconstruir quem sou, tornando possível chegar aqui. Seu trabalho ético e competente transformou minha vida.

De repente tudo vai ficando tão simples
que assusta.
A gente vai perdendo algumas necessidades,
antes fundamentais e que hoje chegam a ser insignificantes.
Vai reduzindo a bagagem e
deixando na mala apenas as cenas e pessoas que valem a pena.
As opiniões dos outros são unicamente dos outros
e mesmo que sejam sobre nós, não têm a mínima importância.
Vamos abrindo mão das certezas,
pois com o tempo já não temos mais certeza de nada.
E de repente isso não faz a menor falta.
Paramos de julgar, pois já não existe certo ou errado,
mas sim a vida que cada um escolheu experimentar.
Por fim entendemos que tudo que importa é ter paz e sossego.
É viver sem medo
e simplesmente fazer algo que alegra o coração naquele momento.
É ter fé.
E só.

Mario Quintana

RESUMO

Objetivo: Avaliar, por meio da Teoria de Resposta ao Item, os resultados do Teste de Progresso em função dos níveis de proficiência e desempenho dos estudantes. Métodos: Foi realizado um estudo transversal com 1559 estudantes de Medicina brasileiros. A estimativa da proficiência e calibração dos parâmetros dos itens foi realizada pelo software BILOG MG 3.0. Para a construção da escala de proficiência, foram selecionados itens de discriminação a > 1, que quando submetidos a técnica de agrupamento por meio da metodologia k-means, formaram seis subconjuntos com níveis crescentes de proficiência. A interpretação pedagógica da escala foi obtida pela relação dos níveis de proficiência com os domínios cognitivos testados. Calculado o intervalo de confiança de 95% e aplicado o teste de Tukey, para analisar a diferença entre as médias de proficiência. Resultados: A frequência de alunos que realizou o teste foi de 94,2%. A proficiência variou de 470 a 580 pontos, com média de 500 pontos e aumentou progressivamente na graduação, um indicador de validade das medidas obtidas. A curva de ajuste dos estudantes ao teste, indicou um teste difícil e a discriminação foi moderada a alta em 61,5% dos itens. Conclusões: A análise dos resultados do Teste de Progresso realizada pela Teoria de Resposta ao Item, demonstrou ser uma ferramenta robusta para avaliar o desempenho de estudantes de Medicina

Palavra-chave: Educação médica. Avaliação educacional. Cognição. Competência profissional. Psicometria.

ABSTRACT

Objective: To evaluate, through Item Response Theory, the results of the Progress Test according to the students' proficiency and performance levels. Methods: A cross-sectional study was conducted with 1559 Brazilian medical students. The estimation of proficiency and calibration of item parameters was performed using BILOG MG 3.0 software. For the construction of the proficiency scale, items with discrimination a > 1 were selected, which when submitted to the clustering technique using the k-means methodology, formed six subsets with increasing levels of proficiency. The pedagogical interpretation of the scale was obtained by relating proficiency levels to the cognitive domains tested. The 95% confidence interval was calculated and Tukey's test was applied to analyze the difference between the proficiency means. Results: The frequency of students who took the test was 94.2%. Proficiency ranged from 470 to 580 points, with a mean of 500 points, and increased progressively in graduation, an indicator of the validity of the measures obtained. The students' adjustment curve to the test indicated a difficult test and discrimination was moderate to high in 61.5% of the items. Conclusions: The analysis of the results of the Progress Test performed by Item Response Theory proved to be a robust tool to assess the performance of medical students.

Keywords: Medical education. Educational assessment. Cognition. Professional competence. Psychometrics.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ABEM	Associação Brasileira de Educação Médica
AFC	Análise Fatorial Confirmatória
AFE	Análise Fatorial Exploratória
ANOVA	Análise de Variância
b_esc	Ponto da CCI que indica o auge do aprendizado de um domínio cognitivo
CCI	Curva Característica do Item
D.M.S.	Diferença Mínima Significativa
EAP	*Expected A Posteriori*
FCMS/JF	Faculdade de Ciências Médicas e da Saúde de Juiz de Fora
i	Ponto da CCI que indica o início aprendizado de um domínio cognitivo
IC	Intervalo de confiança
MG	Minas Gerais
MMAP	*Maximum Marginal A Posteriori*
PRF	Proficiência
s	Ponto da CCI que indica a consolidação do aprendizado de um domínio cognitivo
SUPREMA	Sociedade Universitária para o Ensino Médico Assistencial
TCT	Teoria Clássica do Teste
TepMinas I	1º Consórcio Mineiro de Teste de Progresso
TP/TEP	Teste de Progresso
TRI	Teoria de Resposta ao Item
UEL	Universidade Estadual de Londrina
UFJF	Universidade Federal de Juiz de Fora
UFTM	Universidade Federal do Triângulo Mineiro
UFU	Universidade Federal de Uberlândia
UFV	Universidade Federal de Viçosa
UNIFENAS	Universidade José do Rosário Vellano

LISTA DE SÍMBOLOS

$>$	Maior
$<$	Menor
$\bar{x}$	Média da amostra
$Z\alpha/2$	Valor associado ao nível de significância $\alpha/2$
s	Desvio padrão da amostra
n	Tamanho da amostra
$Z\alpha/2 * s/\sqrt{n}$	Variação

SUMÁRIO

-

1 INTRODUÇÃO

1.1 EDUCAÇÃO MÉDICA

É impossível falar sobre educação médica sem a referência Flexner e seu relatório, grande responsável pela mais importante reforma das escolas médicas nos Estados Unidos da América (EUA), com significativas repercussões para a formação médica e a medicina mundial (PAGLIOSA; DA ROS, 2008). De fato, para o contexto e época, as suas importantes contribuições para a educação médica com ênfase no modelo biomédico centrado na doença e no hospital tiveram seu valor . Nos dias atuais, com a evolução do conhecimento, não cabe nos programas educacionais médicos uma visão reducionista, exclusivamente hospitalocêntrica com pequeno espaço para as dimensões sociais, psicológicas e econômicas da saúde, sendo necessário a inclusão do amplo espectro da saúde, que vai muito além da medicina e seus médicos (PAGLIOSA; DA ROS, 2008).

Na década de 1960, as críticas recorrentes ao setor da saúde, evidenciaram o descompromisso com a realidade e com as necessidades da população e a partir dos anos 1980, seguindo diferentes cenários socioeconômicos e políticos começaram os processos de reforma do setor da saúde.

Desde então, a escola médica tem realizado intensos debates sobre as necessidades de mudança curricular, visando ao cumprimento das Diretrizes Curriculares Nacionais e à transformação do processo de formação dos profissionais de saúde. Os conhecimentos e práticas da saúde coletiva, o entendimento e a participação na construção das políticas públicas e na organização dos serviços de saúde tornam-se competências necessárias e imprescindíveis ao desempenho dos profissionais da saúde (PAGLIOSA; DA ROS, 2008).

Segundo as Diretrizes Curriculares Nacionais (DCN) de 2001 e 2014 (BRASIL; MINISTÉRIO DA EDUCAÇÃO, 2001; 2014), o currículo médico deve garantir ao final da graduação, a formação de um profissional generalista, com domínio de conhecimentos, habilidades e atitudes para atuar com a competência necessária ao exercício da prática médica, nas áreas de Atenção, Gestão e Educação em Saúde, contemplando o sistema de saúde vigente no País.

Deseja-se do profissional médico, um cidadão humanizado, crítico-reflexivo, com formação ético-política e comprometido com a transformação social (BRASIL; MINISTÉRIO DA EDUCAÇÃO, 2014). Para isso, é necessário um modelo pedagógico que aproxime a formação e a rotina nos serviços de saúde e na comunidade, a fim de que esse profissional esteja preparado para atuar em situações reais, lidar com as incertezas e agir com consciência em sua prática médica (TSUJI; AGUILAR-DA-SILVA, 2004).

1.1.1 Currículo integrado

Esse novo perfil exigido ao graduado em medicina, requer mudanças no processo de formação, um modelo curricular que articule teoria e prática, integrando ensino-serviço-comunidade. Professores e estudantes precisam repensar seu papel nesse processo e as instituições, por sua vez, necessitam rever sua estrutura organizacional de forma a prover um cenário adequado aos objetivos educacionais propostos pelo currículo (TSUJI; AGUILAR-DA-SILVA, 2004).

Um currículo integrado, busca conjugar o conteúdo teórico à sua aplicação em situações da vivência profissional, nos serviços de saúde e na comunidade, a partir de experiências vividas num determinado cenário de ensino-aprendizagem. As sucessivas exposições, auxiliam o estudante na identificação e solução dos problemas propostos, preparando-o para intervir efetivamente em situações reais como as vividas pelos profissionais de saúde, nas áreas de Atenção, Gestão e Educação em Saúde.

A implementação do currículo é um projeto construído coletivamente, ocorre por meio da integração ensino-serviço e comunidade, de forma que a prática tenha compromisso e responsabilidade com o usuário do sistema de saúde e promova mudanças na realidade onde se situa. A transformação das práticas profissionais se apoia na reflexão crítica sobre as atitudes profissionais nos serviços de saúde, ou seja, a problematização dos processos de trabalho (CAMPOS, 2000).

O currículo integrado possibilita a associação do desenvolvimento de competências profissionais, definindo-se competência segundo (TSUJI; AGUILARDA-SILVA, 2004): “a competência profissional em saúde é a capacidade circunstancial de mobilizar, articuladamente, os recursos cognitivos, psicomotores e afetivos, visando à

abordagem ou à resolução de uma situação complexa de vigilância de saúde individual ou coletiva, e gestão do trabalho".

Segundo os autores, esse conceito se relaciona a entender a competência como uma capacidade do sujeito mobilizar recursos cognitivos, psicomotores e afetivos necessários à abordagem e resolução de situações complexas nas áreas de atenção, gestão e educação em saúde, objetivando a formação de um profissional capaz de desempenhar sua função na sociedade, na perspectiva da educação inclusiva. Tais recursos serão adquiridos por meio de vivências e práticas coletivas e a articulação adequada dos mesmos nas diferentes situações possibilita verificar a aquisição das competências esperadas.

1.1.2 Currículo por competências

As "competências" tornaram-se a unidade de planejamento do ensino médico ((ALBANESE *et al.*, 2008)). Estruturas de competência como CanMEDS (FRANK; DANOFF, 2007) , o Outcome Project of the (US) Accreditation Council for Graduate Medical Education (ACGME 2001) e o Scottish Doctor ((SIMPSON *et al.*, 2002)), passaram a formar a base do treinamento para a maioria dos estudantes de medicina no mundo ocidental. Enquanto os currículos tradicionais, organizados em torno de objetivos de conhecimento, enfatizavam o processo institucional, a nova proposta era de currículos orientados pelos resultados, os processos curriculares seriam secundários (HARDEN, 2000). Os novos currículos devem garantir que os graduados sejam competentes em todos os domínios essenciais, pois nas profissões de saúde, os escores de avaliação não devem ser compensatórios de um domínio para outro, ou seja, conhecimento excelente não compensa habilidades de comunicação ruins. Devem enfatizar as habilidades e atitudes a serem adquiridas e sua síntese em competências observáveis. A proposta de centralização no estudante, promove um maior envolvimento do mesmo ,que passa a coordenar seu aprendizado, sendo responsável por seu progresso e desenvolvimento (FRANK *et al.*, 2010).

1.1.2 Métodos Ativos de Ensino-Aprendizagem

Segundo as DCN 2014 para a graduação em Medicina (BRASIL, 2001; 2014): o curso deverá ser centrado no estudante como sujeito da aprendizagem

e apoiado no professor como facilitador e mediador do processo, com vistas à formação integral e adequada do estudante. [...] Para isso, o curso deve utilizar metodologias que privilegiem a participação ativa do estudante na construção do conhecimento e na integração entre os conteúdos, assegurando a indissociabilidade do ensino, pesquisa e extensão, bem como promover a integração e a interdisciplinaridade em coerência com o eixo de desenvolvimento curricular.

Logo, é primordial pensar na utilização de recursos didáticos que redirecionem o ensino para uma educação alinhada à formação de um médico generalista, que compreenda o processo saúde-doença, pautando-se em uma postura humana, crítica, reflexiva e ética, com responsabilidade social e compromisso com a cidadania (CARABETTA JR, 2016).

Em alinhamento com as DCN, as metodologias ativas de aprendizagem (MAA) possibilitam ao estudante ser o principal responsável por sua aprendizagem, ou seja, aprender a aprender, sendo o professor apenas o facilitador desse processo, estimulando o pensamento reflexivo, raciocínio crítico, as habilidades de comunicação e orientando as discussões das situações-problema com base em referenciais teóricos fundamentados na Medicina Baseada em Evidências (MBE), na ética e na moral. A MBE constitui um importante suporte ao objetivo de desenvolvimento da avaliação crítica e tomada de decisão (DELORS *et al.*, 1998).

1.2 AVALIAÇÃO NA EDUCAÇÃO MÉDICA

A avaliação é um momento no processo ensino-aprendizagem. Para que o estudante alcance os atributos essenciais à aquisição das competências desejadas, a avaliação deve ser planejada, coerente com a proposta curricular e usar diferentes instrumentos, ajustados aos objetivos educacionais do projeto pedagógico. A ferramenta escolhida para cada avaliação e o método de análise, devem considerar não apenas as propriedades psicométricas de validade e confiabilidade, mas também o impacto educacional, alinhamento com as competências avaliadas, aceitação acadêmica e viabilidade financeira e logística (COLLARES; LOGULO; GREC, 2012).

A avaliação deve cumprir seu importante papel na regulação dos processos de aprendizagem e sua qualidade está relacionada à definição de seus objetivos. Missão,

competências e habilidades a serem adquiridas pelo profissional em formação devem estar claramente definidas pela instituição.

1.2.1 Avaliação programática

Avaliação programática é um conjunto de métodos avaliativos planejados a fim de otimizar a qualidade da avaliação. Sua estrutura é organizada em torno dos objetivos da avaliação, que por sua vez, devem estar alinhados aos objetivos educacionais da instituição. Promover o aprendizado, melhorar as práticas educacionais e dar suporte a situações de tomada de decisão, representam os principais objetivos da avaliação (Vianna 2003).

Na avaliação programática, a responsabilidade pelas diferentes avaliações do estudante, em momentos variados, é da escola médica, como instituição, sendo planejada e executada por instância central, contando com a contribuição das várias unidades curriculares. Esta instância deve cuidar para que os vários métodos de avaliação do estudante sejam empregados de maneira uniforme, nas diferentes áreas de formação, nas melhores condições possíveis. É esta instância institucional centralizada que integra as informações fornecidas pelos métodos, para que sejam utilizadas para cumprir as funções da avaliação do estudante (TRONCON, 2016).

A estrutura da avaliação programática envolve operacionalização, suporte, documentação, melhoria e justificativa (DIJKSTRA; VAN DER VLEUTEN; SCHUWIRTH, 2010; TRONCON, 2016).

Operacionalização é a coleta de informações, combinação e valoração das informações colhidas e tomada de decisão. A coleta da informação são os métodos e instrumentos utilizados na avaliação das competências adotadas.

Suporte é a qualidade da informação, obtido pela revisão dos testes e instrumentos antes de sua aplicação, análise das propriedades psicométricas dos testes, bem como capacitação docente voltada à qualidade da construção e aplicação dos instrumentos de avaliação. É importante o envolvimento das lideranças educacionais e dos protagonistas do currículo na elaboração do sistema de avaliação, melhorando a aceitabilidade do programa e a qualidade da avaliação.

A documentação diz respeito às regras e regulamentos do programa e é composta pela descrição dos elementos e procedimentos do “programa em ação”, dos contextos de aplicação dos testes e dos mecanismos de adequação dos

conteúdos aos métodos de avaliação (blueprinting). A documentação confere clareza e transparência às práticas avaliativas (DIJKSTRA; VAN DER VLEUTEN; SCHUWIRTH, 2010; TRONCON, 2016).

1.3 FUNÇÕES DA AVALIAÇÃO

A avaliação do estudante de Medicina serve a várias finalidades e deve cumprir diferentes funções, entre as quais três adquirem maior importância: estimular o aprendizado (função formativa), embasar a tomada de decisões (função somativa) e prover elementos para o controle da qualidade do currículo (função informativa) (TRONCON, 2016).

1.3.1 Avaliação Formativa

Consiste em uma autorregulação permanente da aprendizagem pelo estudante, permitindo identificar suas fortalezas e fragilidades e com isso, potencializar o aprendizado e rever os pontos frágeis, entendendo que o conhecimento é construído ao longo do processo (PETITJEAN, 1994). O docente, por sua vez, por meio das informações colhidas , identifica as falhas e reorienta suas atividades (TRONCON, 2016).

1.3.2 Avaliação Somativa

A avaliação somativa é realizada ao término do processo com intenção de verificação de resultados. É reconhecida como avaliação diagnóstica, pois, permite inferir a competência profissional nos atributos cognitivo, afetivo e psicomotor, por meio da observação do desempenho, determinando a condição do estudante em avançar (SCRIVEM; W.; R.M., 1967).

1.3.3 Avaliação informativa

Seus resultados fornecem informações sobre a qualidade dos processos educacionais, para a escola médica e os órgãos responsáveis pela regulação da educação e do exercício profissional. Esta função permite que a avaliação do

estudante (assessment), junto com outras informações, sirva à avaliação do processo educacional (evaluation) e, eventualmente, conduza ao seu aperfeiçoamento (TRONCON, 2016).

1.4 INSTRUMENTOS DE AVALIAÇÃO

A avaliação do estudante compreende processos de obtenção de informações sobre o seu desempenho em diferentes domínios, sendo necessário escolher o instrumento adequado de acordo com os objetivos educacionais a serem avaliados. O objetivo desse estudo foi trabalhar com avaliação do domínio cognitivo da competência, sendo escolhido como instrumento o TP.

1.4.1 Teste de Progresso

No final da década de 1970, com a mudança de paradigma na educação médica, e a introdução de uma nova metodologia de aprendizagem, o ensino baseado em problemas (VAN DER VLEUTEN, 1996; TOMIC, Eliane R *et al.*, 2005), as escolas médicas perceberam a necessidade de avaliar os estudantes e o processo pedagógico frente às mudanças curriculares implementadas.

A Faculdade de Medicina da Universidade de Maastricht, na Holanda e a Faculdade de Medicina da Universidade do Missouri, Kansas, foram as primeiras escolas a utilizar uma avaliação longitudinal, a qual chamaram “The Quartely Profile Examination “(QPE). O QPE foi projetado com a intenção de avaliar o ganho de conhecimento dos estudantes e por meio de feedback adequado, promover a recuperação dos pontos de fragilidade. Na mesma época, a Faculdade de Medicina da Universidade de McMaster, Canadá, criou um teste semelhante, o Personal Progress Índex (PPI) (DREES; ARNOLD; JONAS, 2007).

Atualmente com o nome de TP, é utilizado por diversas escolas médicas no mundo (BLAKE 1996; VAN DER VLEUTEN 1996 ; SAKAI *et al.*, 2008; SAKAI; FERREIRA FILHO; MATSUO, 2011; BICUDO *et al.*, 2019). (Anexo C)

No Brasil, a introdução do Teste do Progresso (TP) aconteceu em 1996, no curso de Medicina da Universidade Federal de São Paulo, UNIFESP (UNIFESP, 2017); em 1998, no Curso de Medicina da Universidade Estadual de Londrina (UEL) (SAKAI *et al.*, 2008) e em 2001, na Faculdade de Medicina da Universidade de São

Paulo (TOMIC, Eliane R. *et al.*, 2005). Em 2005, foi fundado em São Paulo, o Núcleo Interinstitucional de Estudos e Práticas de Avaliação em Educação Médica, o primeiro núcleo para aplicação do TP. As escolas do grupo passaram a utilizar então, um teste único, elaborado por uma comissão de docentes das mesmas (SAKAI *et al.*, 2008).

Desde 2012, a ABEM trabalha na consolidação de um modelo de avaliação nacional, o TP ABEM. Atualmente, no Brasil, cerca de 50% das escolas médicas, formando 18 núcleos regionais, aplicam o TP. Em 2015, a ABEM, em conjunto com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) organizou o primeiro Teste de Progresso Interinstitucional Nacional, com a participação de mais de 80 escolas médicas de todas as regiões do país (OLIVEIRA *et al.*, 2022).

Os núcleos são responsáveis pelo planejamento e principalmente por garantir os critérios de qualidade do teste.

O TP é construído com itens de múltipla escolha, que seguem uma matriz de referência, construída com os objetivos educacionais do final do curso. A preparação e revisão de itens transdisciplinares vinculados ao currículo médico, visando condições e emergências comumente encontradas, requer a participação de especialistas de todas as áreas. O número de itens, aplicações e análise dos resultados é definido pelo núcleo (ALAMRO *et al.*, 2023).

Em Minas Gerais, em 2013, foi criado o primeiro Consórcio Mineiro de Escolas Médicas para a aplicação do TP, o TepMinas (EZEQUIEL *et al.*, 2016), que conta atualmente com 10 escolas. São realizadas anualmente duas reuniões, com docentes das diferentes escolas e áreas, para construção do teste. Inicialmente, são elaboradas as “encomendas” dos itens, seguindo uma matriz construída pelo núcleo (Anexo B), que determina conteúdo, cenário, faixa etária, gênero e dificuldade dos itens. Na segunda reunião, são escolhidos para constituir o teste,120 itens originais, de múltipla escolha de resposta única, com quatro alternativas, nas áreas de Ciências Básicas, Clínica Médica, Clínica Cirúrgica, Pediatria, Saúde Coletiva, Ética, Ginecologia e Obstetrícia, abordando o conteúdo cognitivo final do curso (EZEQUIEL *et al.*, 2016). O TP é presencial e obrigatório; aplicado uma vez no ano e considerado uma avaliação formativa.

O movimento maciço em direção à aprendizagem baseada em competências, mostrou uma grande necessidade de formas mais autênticas de avaliação de

desempenhos, o que traria um impacto enorme e conduziria a educação em uma direção muito desejável (VAN DER VLEUTEN; FREEMAN; COLLARES, 2018).

Com todo o conhecimento sobre TP, pensar em um exame de licenciamento tradicional, com todos os possíveis efeitos colaterais, parece uma estratégia quase ultrapassada (VAN DER VLEUTEN; FREEMAN; COLLARES, 2018).

A utopia dos grandes núcleos internacionais de TP, baseado em suas experiências, é que as escolas além de trabalharem de forma colaborativa para desenvolver e administrar o TP , compartilhem seus resultados e experiências, proporcionando uma enorme contribuição individual e coletiva para a educação médica (VAN DER VLEUTEN; FREEMAN; COLLARES, 2018).

1.5 MÉTODOS DE ANÁLISE

Medir com qualidade os resultados de uma avaliação é fundamental para a tomada de decisão. Na avaliação educacional, tais medidas são importantes para identificar o conhecimento já adquirido pelos estudantes para os domínios testados, acompanhar sua evolução ao longo do processo e fazer as comparações necessárias. Além disso, pode ser útil para entender melhor a realidade educacional, mapear os problemas e formular políticas educacionais que contribuam para um melhor desempenho do estudante. A qualidade da medida de um constructo, em nosso caso específico, a proficiência do estudante, é garantida pela qualidade do instrumento e da metodologia utilizada na medição.

A qualidade do instrumento de medição é verificada por meio de dois parâmetros, a validade e a fidedignidade. A validade ocorre quando o instrumento mede aquilo que se pretende medir (PASQUALI, 2009). A fidedignidade tem a ver com a variabilidade das medidas e está relacionada com a calibração do instrumento, de forma a minimizar os erros de medição.

No entanto, como atribuir valores e ter uma medida para o conhecimento, no caso de uma variável latente ?

Segundo Pasquali (2009), etimologicamente, psicometria representa a teoria e a técnica de medida dos processos mentais, especialmente aplicada na área da Psicologia e da Educação.

Dois são os modelos psicométricos mais utilizados na mensuração dos comportamentos e aptidões dos estudantes através das respostas fornecidas aos

itens dos testes: a Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI).

1.5.1 **Teoria Clássica do Teste**

Segundo a TCT, a dificuldade de um item é medida pela proporção ou porcentagem de estudantes que o acertam e o poder de discriminação é a característica que lhe permite oferecer informação sobre a proficiência de um estudante ou compará-la com a de outro, que também está sendo avaliado (ANDRADE, 2001).

Por essa metodologia, o cálculo do percentual de acertos do estudante no teste gera uma nota ou escore. Sob essa interpretação, o teste será considerado fácil, mediano ou difícil, dependendo da aptidão do grupo de respondentes que se submeteu ao teste. Ainda, examinandos que acertam a mesma quantidade de itens, porém de propriedades psicométricas (discriminação, dificuldade e probabilidade de acerto ao acaso) diferentes, apresentam o mesmo escore total ou desempenho (Andrade, Tavares & Valle, 2000; Pasquali, 2003; Pasquali, 2007).

Apesar de algumas limitações, a TCT não tem sido abandonada, e sim, utilizada em combinação com a TRI, a fim de oferecer informações adicionais. A TRI pode ser considerada como uma extensão da TCT e os conceitos das duas teorias estão relacionados uns com os outros.

1.5.2 Teoria de Resposta ao Item

A TRI é um conjunto de modelos matemáticos que relacionam a probabilidade de acerto de um item do teste com a proficiência da pessoa testada, colocando a proficiência na mesma escala métrica da dificuldade dos itens (Pasquali e Primi, 2003) e estimando ainda a capacidade de determinado item ter sido acertado ao acaso. É importante definir que a proficiência da qual estamos falamos, é a medida do conhecimento do estudante, uma variável não observável (ANDRADE; VALLE, 2000).

Estudos na literatura (COLLARES; LOGULO; GREC, 2012; PASQUALI, LUIZ ;PRIMI, 2003) enfatizam amplamente a TRI, como modelo recomendado para análise de avaliações em larga escala ,com testes objetivos de múltipla escolha, como

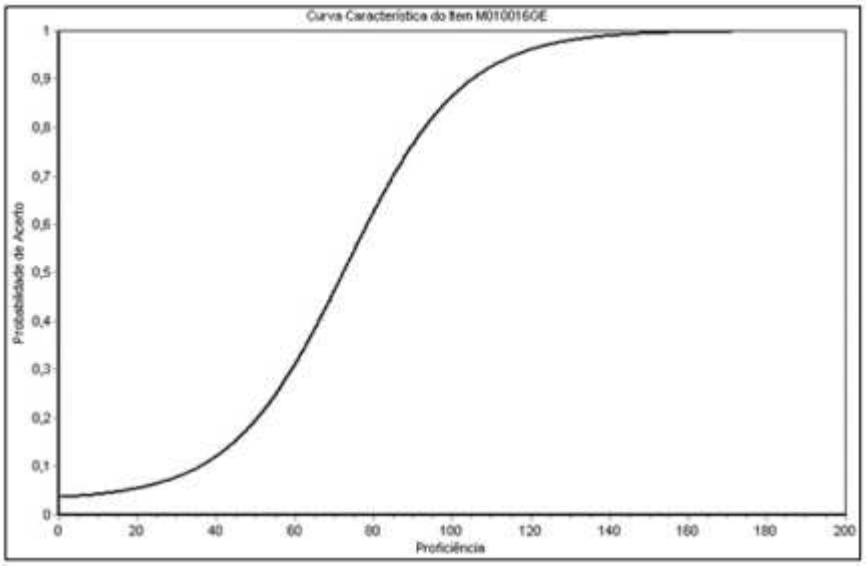
no caso do TP, em que utilizamos itens de múltipla escolha de resposta única, ou dicotômicos.

Quando comparada à TCT, a TRI apresenta como vantagens: (i) a possibilidade de comparação longitudinal de resultados de diferentes avaliações, desde que se incluam itens comuns aos testes e se conservem os mesmos critérios na construção e organização dos mesmos e na análise dos resultados (Thissen, 2003; Baker 2004); (ii) a possibilidade de avaliar com alto grau de precisão e abrangência uma determinada área do conhecimento, sem que cada estudante precise responder a longos testes; (iii) a possibilidade de comparação entre diferentes estágios de aprendizado, viabilizada pela construção de uma escala única de resultados para todos os estágios (ANDRADE; VALLE, 2000). A interpretação da escala fornece o diagnóstico do desenvolvimento gradual e progressivo das competências cognitivas demonstradas pelos estudantes.

Porém, a principal e mais importante distinção entre a Teoria da Resposta ao Item e a Teoria Clássica do Teste é a propriedade de invariância, característica da TRI. Os parâmetros do item não dependem da distribuição dos estudantes avaliados, bem como, o nível de proficiência, e os escores de proficiência dos estudantes avaliados, não dependem do conjunto de itens utilizados para estimá-los (ANDRADE; VALLE, 2000).

Na TRI, a relação entre a proficiência e a probabilidade do estudante acertar o item é descrita por uma curva denominada Curva Característica do Item (CCI). As CCIs (Gráfico 1) podem ser especificadas por meio de três parâmetros: parâmetro a, parâmetro b e parâmetro c. O parâmetro a, de discriminação, corresponde à inclinação da curva medida no ponto b. Quanto maior for esse parâmetro, maior a capacidade do item em discriminar estudantes de proficiências diferentes. O parâmetro b, de dificuldade, mede a dificuldade de um determinado item, correspondendo à proficiência necessária, para que o percentual de acerto de um item seja em torno de 60%. O parâmetro c, chamado de acerto casual, representa o acerto de um item, mesmo quando o estudante possui uma proficiência relativamente baixa para tal (OLIVEIRA & FRANCO Jr. 2008).

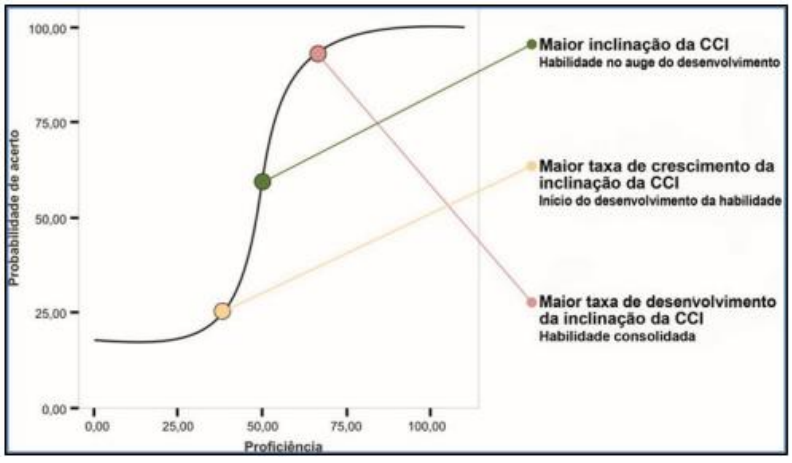
Gráfico 1 – CCI segundo um modelo 3 parâmetros.



Fonte: CAEd/UFJF (2018).

A interpretação da relação entre as características do item e as proficiências dos estudantes, podem ser observadas no gráfico da figura 1, que apresenta as três fases dos processos de aprendizagem (OLIVEIRA & FRANCO Jr. 2008).

Figura 1 – Estágios do processo de aprendizagem modelados segundo a CCI.



Fonte: OLIVEIRA & FRANCO Jr. (2008).

O ponto amarelo, indica o nível de proficiência em que os estudantes passam a ter maiores condições de desempenhar determinado atributo cognitivo (ponto i). O segundo ponto, assinalado pela cor verde (ponto b_esc), indica o parâmetro de dificuldade do item. Em torno desse ponto a aprendizagem está em rápido desenvolvimento, a CCI atinge a mais elevada inclinação. Esse ponto é um delimitador

(Threshold), ou seja, consegue separar os estudantes que desenvolveram a capacidade de desempenhar determinado atributo cognitivo (acima do ponto) daqueles que ainda não atingiram essa etapa (abaixo do ponto). O terceiro ponto, rosa (ponto s), sinaliza a consolidação da aprendizagem, quando observamos a estabilização (não discriminação) da curva. O entendimento desta relação entre a curva do item a aprendizagem é essencial para a interpretação das escalas de proficiência.

As medidas da TRI são produzidas por meio de um software especializado na modelagem utilizada, no caso dessa tese, o BILOG-MG 3.0. Para garantir a qualidade das medidas produzidas, é necessário um ajuste do modelo ao constructo avaliado por técnicas estatísticas específicas. Inicialmente os dados são analisados por TCT e posteriormente pela TRI, obtendo-se os parâmetros dos itens e as medidas de proficiência, representadas em uma escala, através de uma curva normal padronizada variando de menos infinito a mais infinito com média zero e desvio-padrão igual a “1”. Para não trabalhar com números negativos, na prática, multiplicamos e somamos a proficiência de cada estudante por duas constantes, de forma a termos uma escala somente com valores positivos. Os níveis de proficiência são apresentados na escala em ordem crescente e cumulativa.

É importante lembrar novamente, que proficiência é a medida do conhecimento e que não existe nulidade de conhecimento. O resultado da medida de proficiência de cada estudante deve ser analisado com base no valor médio obtido pelos estudantes que realizaram o teste e não a partir do zero.

As escalas de proficiência devem ser qualitativamente interpretadas, para que adquiram significância pedagógica. Os resultados, além de produzirem um diagnóstico das competências desenvolvidas pelos estudantes avaliados, são importantes na melhora da qualidade do ensino, por meio do direcionamento de professores e gestores em suas práticas pedagógicas.

Importantes programas nacionais e internacionais de avaliação educacional utilizam a interpretação de escalas de proficiência para basear seus resultados, tais como: Sistema Nacional de Avaliação da Educação Básica – SAEB; o National Assessment for Educational Progress – NAEP; o Programa Internacional de Avaliação de Alunos – PISA; o Trends in International Mathematics and Science Study – TIMSS; o Projeto GERES/2005 – Estudo Longitudinal sobre a Qualidade e Equidade no Ensino Fundamental Brasileiro.

1.6 RACIOCÍNIO CLÍNICO

A inteligência humana opera como uma estrutura de conjunto, não sendo possível destacar uma única operação mental como responsável por um determinado desempenho ou aprendizagem, mas um conjunto ou agrupamento delas, que atuam simultaneamente na construção do raciocínio clínico.(MAMEDE *et al.*, 2014). A esse conjunto de ações e operações mentais, chamamos de competências cognitivas.um importante atributo a ser desenvolvido durante o curso médico, a fim de formar profissionais capazes de elaborar diagnósticos corretos e definir a melhor conduta.

O processo do desenvolvimento da competência diagnóstica ao longo da formação médica ocorre em sucessivos estágios. Inicialmente os estudantes relacionam dados semiológicos aos conhecimentos biomédicos aprendidos. Reconhecem sinais e sintomas de forma isolada, sem os relacionar a um grupo de doenças.

Em um segundo estágio do desenvolvimento da expertise, a exposição repetida aos casos, faz com que os conhecimentos biomédicos fiquem “encapsulados” em padrões diagnósticos.

Por fim, o conhecimento previamente encapsulado, será reorganizado na memória de longo prazo como scripts ou modelos mentais de doenças, que serão aprimorados com a experiência clínica.

2 JUSTIFICATIVA

Segundo as Diretrizes Curriculares Nacionais (DCN) do Curso de Graduação em Medicina, o currículo médico tem como missão, garantir a formação de um profissional de saúde com as competências essenciais para atuar de forma segura e eficaz em situações nas áreas de Atenção, Gestão e Educação em Saúde. Para que tal objetivo se cumpra, é necessário que ao longo da graduação, o processo ensino-aprendizagem inclua momentos para verificação da aquisição dessas competências, mediante avaliação do desempenho do estudante em suas tarefas, de acordo com os objetivos educacionais propostos.

A avaliação dos atributos cognitivos, psicomotores e afetivos das competências, exige uma abordagem programática, por meio de diferentes instrumentos e métodos , cuja escolha se sustenta nas propriedades psicométricas, impacto educacional, alinhamento com as competências avaliadas, aceitação acadêmica e viabilidade financeira e logística dos mesmos.

A proposta do estudo em questão, foi demonstrar a utilização da TRI como método de análise do TP e sua efetividade em avaliar as competências esperadas.

O TP é aplicado em grande parte das escolas médicas internacionais e nacionais e é validado na literatura como instrumento de avaliação longitudinal do desempenho. As aplicações repetidas do teste, promovem o conhecimento de longo prazo e os resultados possibilitam aos estudantes acompanhar seu ganho de conhecimento e às escolas, identificar fragilidades no currículo e rever a proposta pedagógica do curso.

A TRI, como método de análise, é o modelo mais recomendado para avaliações educacionais em larga escala, com testes objetivos de múltipla escolha. Permite a construção de uma escala numérica representativa do conhecimento do estudante em relação aos conteúdos testados, diferente da TCT que avalia apenas percentual de acertos.

O que gostaríamos de salientar com esse estudo, é que no Brasil, embora cerca de metade das escolas médicas do país apliquem o TP, a metodologia utilizada para análise dos resultados é a TCT e com isso, os resultados obtidos não representam uma medida robusta do desempenho do estudante ao longo do curso, que garanta que ao final da graduação, o mesmo terá adquirido, de acordo com as resoluções previstas nas DCN 2014, as competências necessárias ao exercício da profissão. A

proposta deste estudo é que essa análise seja ampliada por meio da TRI, possibilitando a elaboração e interpretação de uma escala de proficiência para o curso de Medicina.

3. OBJETIVOS

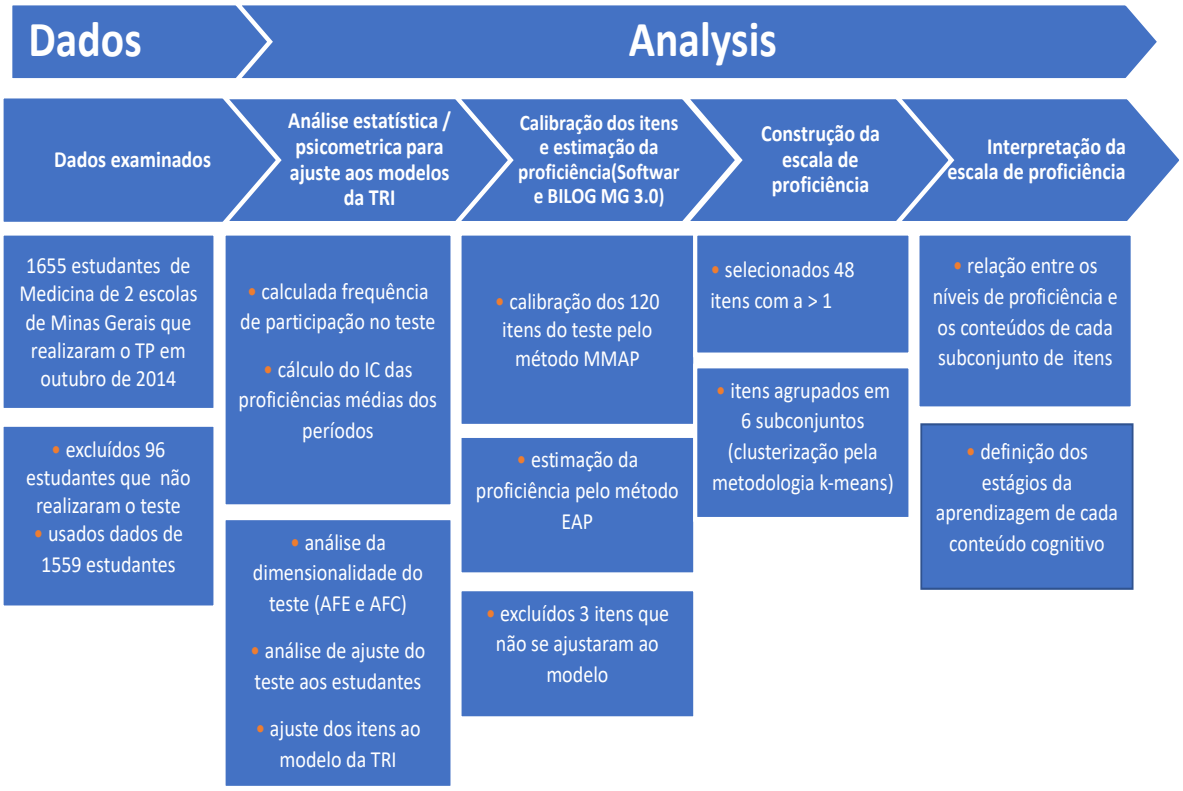
- Avaliar por meio da TRI, os resultados do TP, em função do desempenho cognitivo de estudantes de medicina.
- Discutir as vantagens da TRI em relação à TCT na análise do TP, em referência aos objetivos educacionais de um currículo construído por competências, para o Curso de Medicina.
- Fortalecer o TP como ferramenta de avaliação dos atributos cognitivos indispensáveis à formação do médico.
- Interpretar pedagogicamente a relação entre os níveis de proficiência obtidos no TP 2014 e os subgrupos de desempenhos testados.

4 MÉTODOS

4.1 DESENHO DO ESTUDO

Este estudo transversal, avaliou por meio da TRI, o desempenho cognitivo de estudantes de Medicina no TP, aplicado pelo Consórcio Mineiro de Teste de Progresso (TepMinas I), em 2014. A figura 2 fornece uma visão geral do fluxo de análise.

Figura 2- Fluxograma das etapas de pré-processamento e análise de dados.



Fonte: Elaborado pelo autor

4.2 CONTEXTO

Em 2013 foi criado em Minas Gerais o primeiro consórcio de escolas médicas para aplicação do TP, o TepMinas I. Em 2014, ano do estudo, o TP foi aplicado por seis, das sete escolas que integravam o consórcio naquele ano e totalizavam 4557 estudantes matriculados no curso de Medicina. Uma das escolas não participou, pois havia acabado de ingressar e seus docentes ainda se encontravam em treinamento. O teste foi presencial, obrigatório, e considerado uma avaliação formativa.

Os itens do teste foram elaborados segundo uma matriz construída pelo consórcio, que teve por modelo a matriz da Associação Brasileira de Educação Médica (ABEM). Foi constituído por 120 itens originais, múltipla escolha de resposta única, com quatro alternativas, nas áreas de Ciências Básicas, Clínica Médica, Clínica Cirúrgica, Pediatria, Saúde Coletiva, Ética, Ginecologia e Obstetrícia.

A análise dos resultados do teste fornecida ao consórcio, foi realizada por uma empresa de tecnologia da informação, com expertise em avaliação educacional e a metodologia utilizada foi a TCT. Cada escola recebeu os resultados de seus estudantes e do consórcio em geral, sendo preservado o sigilo de identificação individual das escolas.

O acesso ao banco de dados do teste, necessário à realização das análises do estudo, foi solicitado oficialmente à coordenação do curso de Medicina das seis escolas do TepMinasI, porém, apenas a Universidade Federal de Juiz de Fora – UFJF (instituição pública) e a Faculdade de Ciências Médicas e da Saúde de Juiz de Fora (FCMS/JF) – SUPREMA (instituição privada), cederam seus resultados.

4.3 PARTICIPANTES

O estudo envolveu 1655 estudantes de Medicina, sendo 973 matriculados em uma instituição pública e 682 em uma instituição privada. Foram excluídos por motivo de não comparecimento ao teste, 42 estudantes (4,32%) e 54 (7,92%) de cada instituição. Os 1559 estudantes incluídos, correspondiam a 46% do total de estudantes que realizaram o teste. O tamanho amostral não pode ser calculado, uma vez que o acesso aos dados da população total de estudantes não era conhecido.

4.4 ASPECTOS ÉTICOS

Aprovado pelo Comitê de Ética e Pesquisa da FCMS/JF sob o parecer nº 2.076.924.

A pesquisadora assinou o Termo de Compromisso de Utilização de Dados – TCUD com as duas escolas participantes do estudo.

4.5 PROCEDIMENTO DE ANÁLISE DOS DADOS

Determinada frequência de participação dos estudantes das escolas avaliadas por análise descritiva.

As proficiências médias de cada período foram comparadas, para verificar o ganho ao longo da graduação. Foi calculado o intervalo de confiança que contém o valor verdadeiro da média do período e comparado com as médias dos outros intervalos, para determinar se houve mudança significativa entre elas. O cálculo do intervalo de confiança de 95% para as médias de proficiência dos estudantes, utilizou a seguinte fórmula (OGLIARI; ANDRADE, 2005):

$$IC = \bar{x} \pm Z\alpha/2 * s/\sqrt{n} \quad (1)$$

Onde,

$\bar{x}$ = Média da amostra, nesse caso PRF_TEP.

$Z\alpha/2$ = Valor associado ao nível de significância $\alpha/2$, nesse caso 1,96.

s = Desvio padrão da amostra, nesse caso PRF_TEP_sd.

n = Tamanho da amostra, nesse caso ALUNOS

Variação = $Z\alpha/2 * s/\sqrt{n}$

Utilizado o teste de Tukey para verificar se as diferenças dos valores de proficiência entre os períodos de ensino foram significativas. A distribuição da amplitude estudentizada, o quadrado médio dos resíduos da ANOVA e o tamanho amostral dos grupos, foram calculadas a fim de determinar a diferença mínima significativa (D.M.S.), considerando os percentis do grupo.

4.6 ANÁLISE PSICOMÉTRICA DO TESTE

Foram realizadas análises para verificar a possibilidade de usar o modelo unidimensional da TRI nos dados do TP 2014 e em seguida as análises de ajuste do

teste aos estudantes e ajuste dos itens ao modelo da TRI.

4.6.1 Análise de dimensionalidade

Realizado inicialmente uma análise de dimensionalidade por meio de técnicas fatoriais sob perspectiva exploratória, para verificar se os dados do TEPMINAS I possuíam característica unidimensional, condição essencial para utilização do modelo unidimensional da TRI para itens dicotômicos.

A análise fatorial exploratória (AFE), verificou que dois fatores contribuíram significativamente para explicação da variância nos itens, podendo ser extraídos. Esta análise está ancorada na literatura (MATOS; RODRIGUES, 2019) e nos arquivos gerados (ScreePlot e Variância Acumulada), considerando a base de dados do teste. Em seguida, assumindo que os fatores 1 e 2 explicam a variância nos itens, por análise fatorial confirmatória (AFC) verificou-se que os dois fatores extraídos eram auto correlacionados, o que permitiu inferir que as respostas aos itens poderiam ser ajustadas por um modelo TRI unidimensional. Esta análise está ancorada na literatura (BROWN, 2007; HAIR JUNIOR *et al.*, 2005) e nos arquivos (Fator ITEM, Fator Correlação e Goodness fit) gerados considerando a base de dados do teste.

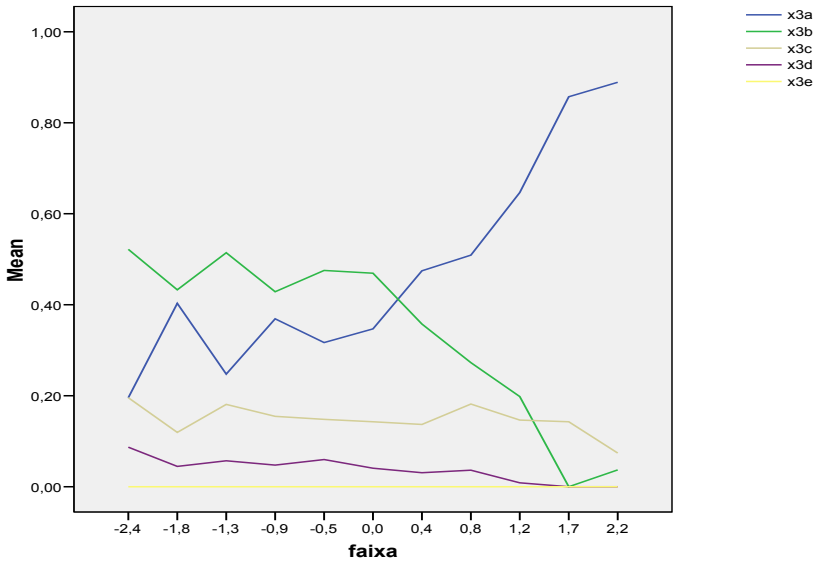
4.6.2 Análise de ajuste do teste aos estudantes

A fim de avaliar a qualidade do teste, foi verificado seu ajuste aos estudantes. A escala foi dividida em intervalos de meio desvio padrão e verificado os percentuais de estudantes (proficiência) e itens (parâmetro de dificuldade) em cada um dos intervalos construídos.

4.6.3 Ajuste dos itens ao modelo da TRI

Os ajustes gráficos de cada item (Gráfico 2) permitiram avaliar o comportamento de cada alternativa do item, possibilitando identificar problemas de elaboração e de conteúdo dos itens. Excluídos 3 itens por inadequação ao modelo.

Gráfico 2 – Representação gráfica de ajuste de um item do teste

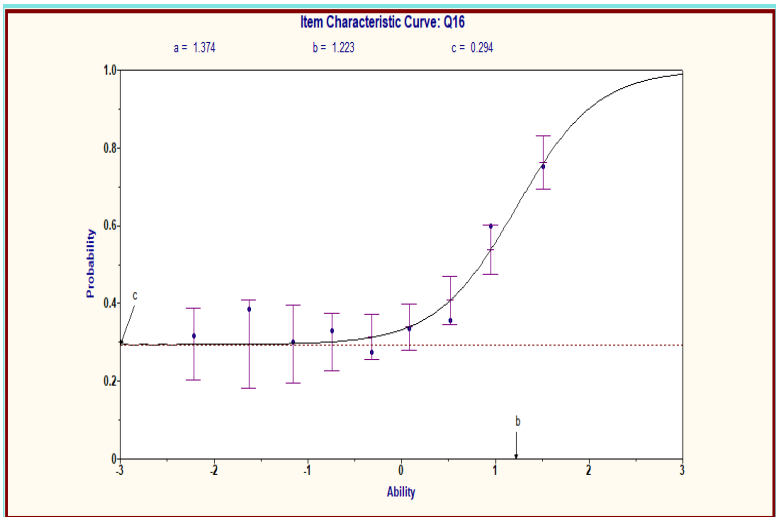


Fonte: Software BILOG MG 3.0.

4.6.4 Calibração de itens e estimação da proficiência

Os itens foram calibrados pelo método MMAP (Maximum Marginal A Posteriori) e obtidos os parâmetros de discriminação (a), dificuldade (b) e acerto ao acaso (c) para cada item (ANDRADE; VALLE, 2000). As proficiências foram estimadas pelo método EAP (Expected A Posteriori), com base no padrão de acertos do examinado (ANDRADE; VALLE, 2000). A relação entre a proficiência do estudante (variável latente medida) e a probabilidade de responder corretamente o item (ANDRADE; VALLE, 2000) foi graficamente representada pelas CCIs (Gráfico 3). A inclinação da curva de cada item, informou os pontos de início, auge e consolidação da aprendizagem (i,b_esc e s) de cada domínio testado. O software BILOG MG 3.0 forneceu ainda os gráficos das curvas de informação de cada item e do teste.

Gráfico 3 – Representação gráfica da curva característica de um item



Fonte: Software BILOG MG 3.0.

4.7 CONSTRUÇÃO DA ESCALA DE PROFICIÊNCIA

Os valores obtidos como parâmetros dos itens e como pontos de início, auge e consolidação da aprendizagem de cada desempenho testado, possibilitaram a construção de uma escala, com média 0 e desvio padrão 1, representativa do constructo proficiência para o teste aplicado. Selecionados como representativos da escala de proficiência, itens com parâmetro alto de discriminação a > 1(TREVISAN *et al.*, 2019) e baixa possibilidade de acerto casual, c < 0,5. Por estarem em uma mesma escala métrica, os valores do parâmetro b são considerados representativos da proficiência na escala criada e para evitar valores muito pequenos ou negativos, estes foram multiplicados e somados a valores arbitrários, procedimento matemático chamado transformação linear, criando-se então, uma escala arbitrária, com 28 intervalos de distância de 5 pontos entre si.

4.7.1 Determinação dos níveis de proficiência

A interpretação pedagógica da escala, no entanto, depende da relação da proficiência com os desempenhos cognitivos abordados no teste. Para estabelecer essa relação ,os itens da escala foram agrupados por proximidade ou semelhança em subconjuntos (clusters),de acordo com a medida da distância entre os pontos b e s de sua CCI, procedimento realizado por meio da metodologia K-means. O nível de proficiência associado a cada subconjunto de itens, foi definido pela média dos

pontos referentes ao auge (b) e a consolidação da aprendizagem (s) dos itens desse subgrupo. Cada nível foi representado por intervalos de proficiência com valores limite semelhantes aos valores determinados nos intervalos da escala.

4.8 INTERPRETAÇÃO PEDAGÓGICA DA ESCALA

Em cada subconjunto de itens, o desempenho cognitivo foi avaliado por meio de quatro intervalos de aprendizagem, considerados significativos: aquém do ponto que permite acertar os itens; o máximo crescimento da aprendizagem necessária ao acerto dos itens; início da consolidação da aprendizagem e aprendizagem consolidada (KÁTIA MESQUITA DE OLIVEIRA; FRANCO; SOARES, 2007).

A definição dos desempenhos cognitivos testados em cada item foi realizada pela pesquisadora e revisada por duas professoras doutoras, com expertise em educação médica, por meio da análise dos itens e gabaritos do teste.

A interpretação sobre a relação dos níveis de proficiência com os desempenhos testados em cada subconjunto de itens foi realizada com base nas teorias de construção do raciocínio clínico (MAMEDE *et al.*, 2014; PEIXOTO; SANTOS; FARIA, 2018).

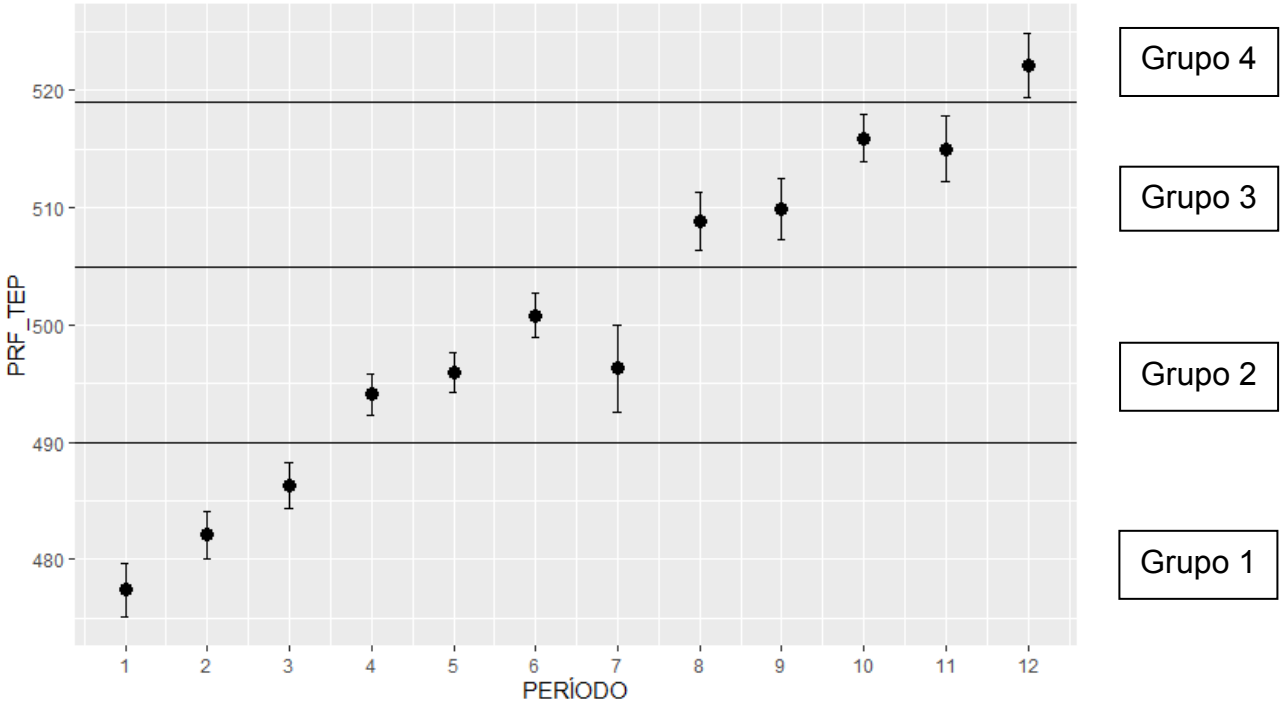
5 RESULTADOS

A frequência de estudantes que realizaram o teste foi 94,2%, sendo 59,7% e 40,3% de cada escola.

Apresentamos no gráfico 4, as médias por período, os respectivos IC e os resultados obtidos pelo teste de Tukey. Verificamos uma divisão em quatro grupos com proficiência crescente ao longo da graduação (linhas horizontais).

Os períodos 1 a 3 (grupo 1) se posicionaram na mesma faixa de proficiência, o que também ocorreu para os períodos 4 a 7 (grupo 2) e 8 a 11 (grupo 3). Os valores de proficiência aumentaram progressivamente na ordem crescente dos grupos, atingindo o auge no período 12 (grupo 4), ou seja, os estudantes estão no ápice de seus respectivos níveis de conhecimento nesse último período. Por meio dessa análise, constatamos que os resultados foram sensíveis à evolução da escolaridade, um indicador de validade das medidas obtidas.

Gráfico 4 – Teste de Tukey para as médias dos períodos avaliados

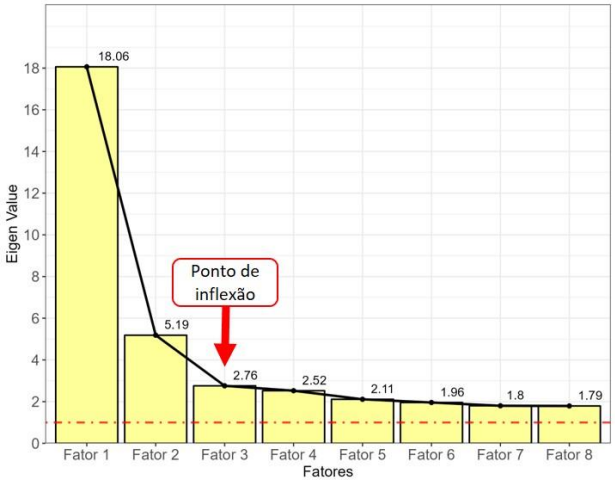


Fonte: Software BILOG MG 3.0.

A análise de dimensionalidade verificou que dois fatores contribuíram significativamente para variância nos itens. O gráfico 5 apresenta o diagrama de

inclinação (= Scree Test), método utilizado para definir o número de fatores a serem extraídos.

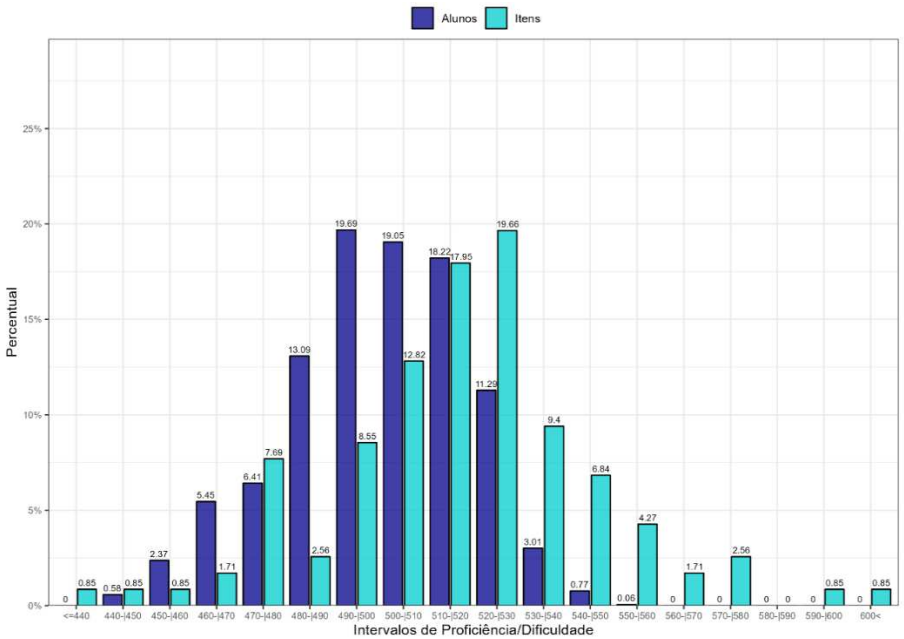
Gráfico 5 – Diagrama de inclinação



Fonte: Elaboração do próprio autor.

Na curva de ajuste do teste aos estudantes, foi observado alta concentração de itens em torno da proficiência média, 500 pontos e nos intervalos de 460 a 530 pontos. Verificou-se também que o teste está desalinhado para a direita da escala, ou seja, o teste estava difícil para os estudantes (Gráfico 6).

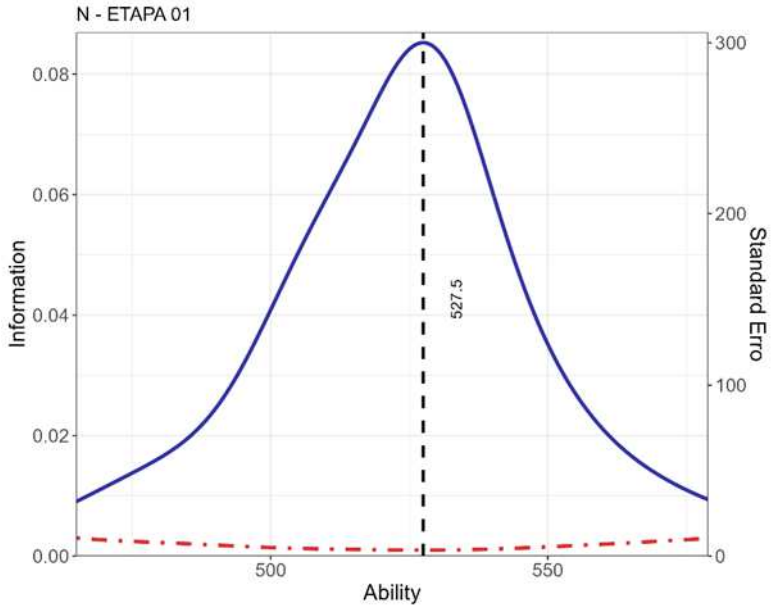
Gráfico 6 – Curva de ajuste do teste aos estudantes



Fonte: Software BILOG MG 3.0.

O teste apresentou boa estimativa da proficiência, conforme visualizado na curva de informação do teste, gráfico 7 . A máxima informação ocorreu no valor de proficiência de 527,5 (linha contínua) e coincidiu com o erro mínimo de medida (linha pontilhada).

Gráfico 7 – Curva de informação do teste



Fonte: Software BILOG MG 3.0.

A calibração dos itens mostrou discriminação moderada a alta (a > 0,65) (RABELO, 2013) em 61,5% dos itens . Em relação ao parâmetro dificuldade, 15,4% dos itens foram classificados como fáceis (b < -0,52), 23% com moderada dificuldade (-0,52 > b < 0,52), e 61,5% difíceis (b > 0,52) (RABELO, 2013). Como eram itens de múltipla escolha com quatro alternativas, a probabilidade de acerto ao acaso foi considerada para valores de c > 0,25 e foi de 46% (RABELO, 2013).

De acordo com os critérios descritos na metodologia, foram selecionados 48 itens como representativos da escala de proficiência. Os valores de proficiência (parâmetro b) desses itens foram multiplicados arbitrariamente por 20 e somados a 500, obtendo-se uma escala de proficiência que compreendeu um intervalo entre 455 e 590 pontos (Quadro 1).

Quadro 1 – Representação dos parâmetros que geraram a escala de proficiência do Teste de Progresso de 2014

n	item	relecao	a	b	c	i	b_erc	r
112,00	Q115	1,00	1,07209	-1,43495	0,13654	456,85	471,30	485,75
100,00	Q103	1,00	1,06162	-1,36239	0,12219	458,16	472,75	487,35
111,00	Q114	1,00	1,06872	-1,19840	0,09940	461,53	476,03	490,53
93,00	Q96	1,00	1,11879	-0,78365	0,26997	470,48	484,33	498,18
97,00	Q100	1,00	1,11529	-0,27975	0,25936	480,51	494,41	508,30
26,00	Q27	1,00	1,80258	-0,09574	0,50000	489,49	498,09	506,68
17,00	Q18	1,00	1,03110	-0,07780	0,32135	483,42	498,44	513,47
52,00	Q54	1,00	1,46529	-0,01134	0,35077	489,20	499,77	510,35
51,00	Q53	1,00	1,16576	0,01812	0,32113	487,07	500,36	513,65
73,00	Q75	1,00	1,19739	0,06026	0,26102	488,27	501,21	514,14
71,00	Q73	1,00	1,52577	0,11258	0,28005	492,10	502,25	512,41
31,00	Q33	1,00	1,69440	0,21393	0,20263	495,13	504,28	513,42
61,00	Q63	1,00	1,01225	0,23933	0,33084	489,48	504,79	520,09
23,00	Q24	1,00	1,46910	0,44985	0,14996	498,45	509,00	519,54
36,00	Q38	1,00	1,31334	0,48877	0,21166	497,98	509,78	521,57
91,00	Q94	1,00	1,13653	0,51320	0,26786	496,63	510,26	523,90
83,00	Q86	1,00	1,73023	0,53597	0,37242	501,76	510,72	519,67
14,00	Q14	1,00	1,51076	0,54121	0,40268	500,57	510,82	521,08
114,00	Q117	1,00	1,20042	0,62394	0,18096	499,57	512,48	525,39
53,00	Q55	1,00	1,40138	0,67989	0,17081	502,54	513,60	524,65
39,00	Q41	1,00	1,41256	0,72658	0,31553	503,56	514,53	525,50
42,00	Q44	1,00	1,31500	0,73734	0,40376	502,96	514,75	526,53
44,00	Q46	1,00	1,36209	0,74099	0,18598	503,44	514,82	526,19
86,00	Q89	1,00	1,59515	1,01257	0,32710	510,54	520,25	529,96
64,00	Q66	1,00	1,65806	1,14987	0,21133	513,65	523,00	532,34
60,00	Q62	1,00	1,79923	1,15705	0,23997	514,53	523,14	531,75
16,00	Q16	1,00	1,37425	1,22345	0,29418	513,19	524,47	535,74
11,00	Q11	1,00	1,03364	1,23069	0,43102	509,62	524,61	539,60
62,00	Q64	1,00	1,86579	1,23916	0,20141	516,48	524,78	533,09
102,00	Q105	1,00	1,90783	1,25494	0,17894	516,98	525,10	533,22
67,00	Q69	1,00	1,14419	1,30108	0,14756	512,48	526,02	539,56
66,00	Q68	1,00	1,58283	1,31845	0,12608	516,58	526,37	536,16
78,00	Q81	1,00	1,35283	1,34447	0,16056	515,44	526,89	538,34
25,00	Q26	1,00	1,61368	1,40929	0,28010	518,58	528,19	537,79
54,00	Q56	1,00	1,17053	1,42120	0,31523	515,19	528,42	541,66
89,00	Q92	1,00	1,19401	1,42599	0,21907	515,54	528,52	541,50
63,00	Q65	1,00	1,30729	1,49644	0,26265	518,08	529,93	541,78
80,00	Q83	1,00	2,10533	1,50162	0,21210	522,67	530,03	537,39
65,00	Q67	1,00	1,62431	1,52592	0,44034	520,98	530,52	540,06
20,00	Q21	1,00	1,02020	1,56806	0,23598	516,17	531,36	546,55
33,00	Q35	1,00	1,49300	1,64395	0,13314	522,50	532,88	543,26
68,00	Q70	1,00	2,16397	1,68605	0,19179	526,56	533,72	540,88
56,00	Q58	1,00	1,08439	1,82834	0,11867	522,28	536,57	550,85
50,00	Q52	1,00	1,14789	1,96935	0,17857	525,89	539,39	552,88
94,00	Q97	1,00	1,58347	2,06324	0,24183	531,48	541,26	551,05
104,00	Q107	1,00	1,35150	2,40284	0,24251	536,59	548,06	559,52
43,00	Q45	1,00	1,15788	3,17093	0,12083	550,04	563,42	576,80
24,00	Q25	1,00	1,07154	3,36558	0,33125	552,85	567,31	581,77

Fonte: Software BILOG MG 3.0.

A análise de conglomerados deu origem a seis subconjuntos de itens , que de acordo com a média dos pontos b e s dos seus itens, foram associados a seis níveis de proficiência.

Os níveis foram representados por intervalos de proficiência com valores limite semelhantes aos valores determinados na escala. Nível 1: até 490; Nível 2: 491 até 510; Nível 3: 511 até 525; Nível 4: 526 até 540; Nível 5: 541 até 560; Nível 6: 561 até 580. Como o último nível foi representado por apenas dois itens, optou-se por não o considerar.

Foi avaliada a relação entre o período do curso e o nível de proficiência definido em cada subconjunto de itens, observando-se aumento da proficiência de acordo com o avanço na graduação, como demonstrado no quadro 2.

Quadro 2 – Relação entre o período do curso e o nível de proficiência dos estudantes

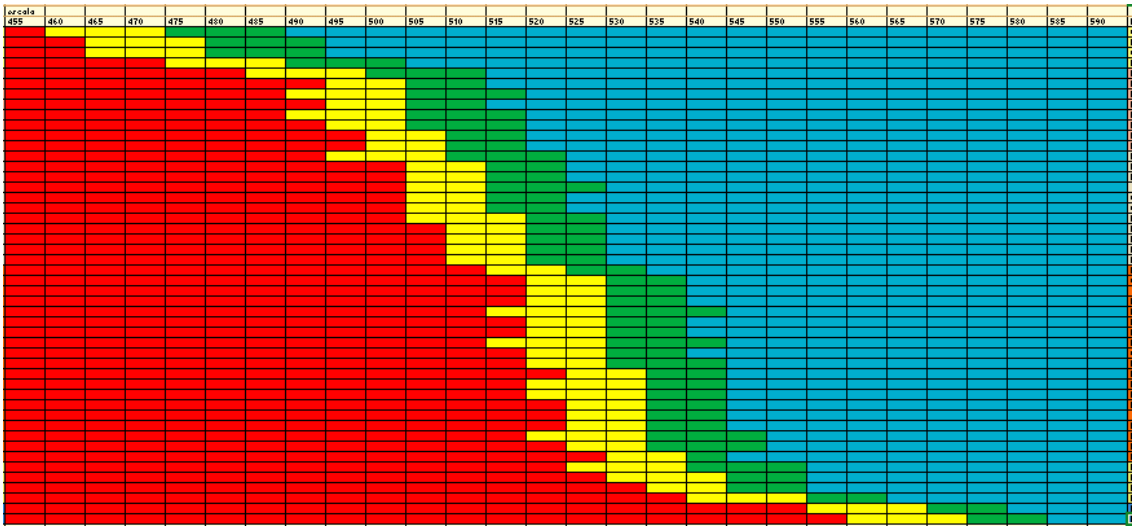
Período do Curso		Faixas de Proficiência por nível					Total de estudantes
		Nível 1 até 490	Nível 2 491 - 510	Nível 3 511 - 525	Nível 4 526 - 540	Nível 5 541- 560	
1	Nº	102	28	0	0	0	130
	%	78,5%	21,5%	,0%	,0%	,0%	100,0%
2	Nº	85	47	0	0	0	132
	%	64,4%	35,6%	,0%	,0%	,0%	100,0%
3	Nº	61	70	0	0	0	131
	%	46,6%	53,4%	,0%	,0%	,0%	100,0%
4	Nº	28	114	0	0	0	142
	%	19,7%	80,3%	,0%	,0%	,0%	100,0%
5	Nº	20	121	1	0	0	142
	%	14,1%	85,2%	,7%	,0%	,0%	100,0%
6	Nº	16	117	6	0	0	139
	%	11,5%	84,2%	4,3%	,0%	,0%	100,0%
7	Nº	39	69	10	7	0	125
	%	31,2%	55,2%	8,0%	5,6%	,0%	100,0%
8	Nº	15	78	29	19	0	141
	%	10,6%	55,3%	20,6%	13,5%	,0%	100,0%
9	Nº	10	64	24	29	0	127
	%	7,9%	50,4%	18,9%	22,8%	,0%	100,0%
10	Nº	2	43	22	41	2	110
	%	1,8 %	39,1%	20,0%	37,3%	1,8%	100,0%
11	Nº t	7	44	15	38	3	107
	%	6,5%	41,1%	14,0%	35,5%	2,8%	100,0%
12	Nº	5	28	8	72	20	133
	%	3,8%	21,1%	6,0%	54,1%	15,0%	100,0%
Total	Nº	390	823	115	206	25	1559
	%	25,0%	52,8%	7,4%	13,2%	1,6%	100,0%

Fonte: Software BILOG MG 3.0.

Entre o total de estudantes, 22% obtiveram uma proficiência no teste superior ao nível 2, 490 a 510 pontos. Desses estudantes, 57,5% estavam no internato (9º a 10º período) e 13% no ciclo clínico (5º ao 8º período).

A proficiência foi relacionada ainda, à aprendizagem dos desempenhos cognitivos testados em cada subconjunto de itens, por meio de quatro intervalos considerados significativos: aquém do ponto que permitiria acertar os itens; auge da aprendizagem necessária ao acerto dos itens; início da consolidação da aprendizagem e aprendizagem já consolidada (Quadro 3).

Quadro 3 – Representação da relação da proficiência com a aprendizagem de domínio cognitivo testado.



Fonte: Software BILOG MG 3.0.

Legenda:

- Aquém
- Início do desenvolvimento
- Auge do desenvolvimento
- Consolidação

6 DISCUSSÃO

O TP é um instrumento de avaliação em larga escala, cuja análise dos resultados por meio dos modelos matemáticos da TRI, fornece dados relativos ao desempenho cognitivo do estudante de Medicina durante a graduação, bem como sobre o conteúdo e estrutura curricular, possibilitando diagnóstico precoce das fragilidades e tomada de decisão por parte de docentes e gestores da instituição (EMBRETSON, 1996).

No TP Minas 2014, cujos resultados foram analisados nesse estudo, o comparecimento de 94,2% dos estudantes das duas instituições envolvidas foi semelhante ao de 92% (SAKAI; FERREIRA FILHO; MATSUO, 2011; TOMIC, Eliane R *et al.*, 2005) descrito pela Universidade Estadual de Londrina (UEL), de 2004 a 2006, bem como o percentual de participação por período do curso (SAKAI; FERREIRA FILHO; MATSUO, 2011; TOMIC, Eliane R *et al.*, 2005).

Um estudo de caráter internacional publicado em 2010 por A. Freeman et al, apresentou contribuições de várias escolas e núcleos de escolas sobre seu uso do TP, porém ênfase foi dada ao formato, qualidade, sistema de aplicação e de feedback e não ao número de estudantes.

Alamro et al. (2023) relata participação em torno de 100% na aplicação do TP por um núcleo saudita de 20 escolas, nos anos de 2020 e 2021.O aumento na participação em relação aos anos anteriores foi atribuído à percepção adequada dos líderes universitários sobre os benefícios do teste. Nouns e Georg (2010) relatam uma participação de 85 a 95% no TP realizado por 13 escolas da Alemanha e Áustria e atribui principalmente ao conceito formativo sem decisão de aprovação/reprovação.

No estudo corrente, o desempenho dos estudantes foi avaliado pela medida de sua proficiência no teste. A proficiência média dos estudantes foi 500 pontos e apresentou o esperado aumento progressivo ao longo da graduação (477,3 a 522,8), contribuindo como medida de validade do teste (DIJKSTERHUIS *et al.*, 2009; GREEN; HEALES, 2023; TRONCON, 1996) Em estudos nacionais também podemos observar esse resultado, porém baseado no escore total de acertos na avaliação, calculado pela TCT (BICUDO *et al.*, 2019; SAKAI; FERREIRA FILHO; MATSUO, 2011; TOMIC, Eliane R *et al.*, 2005) e como itens de um instrumento de medida não são tratados individualmente, não é possível comparar. Núcleos como os da Holanda e Alemanha, com experiência em TP multicêntricos, utilizando TRI , apresentam resultado de

progressão semelhante aos deste estudo e, além da importância formativa, destacam o TP como fonte de informação para avaliação e monitoramento de mudanças curriculares e capacitação docente (ALAMRO *et al.*, 2023; TIO *et al.*, 2016). De acordo com o feedback programado pela escola, é permitido ao estudante compreender a sua pontuação geral e por disciplina e comparar com a média do seu grupo por momento de teste e longitudinalmente, permitindo reparar suas deficiências e promovendo um maior crescimento do conhecimento (TIO *et al.*, 2016).

O teste apresentou um grau alto de dificuldade, com 61,5% de itens difíceis (RABELO, 2013). A discriminação foi adequada, já que 61,5% dos itens tinham discriminação moderada a alta (a > 0,65) (RABELO, 2013). A definição de itens de qualidade se baseia em alguns critérios (WARE; VIK, 2009) como presença de pelo menos 50% dos itens em níveis cognitivos mais elevados (aplicação e raciocínio),com discriminação maior ou igual a 60%, o que ocorreu no teste estudado. A probabilidade de acerto ao acaso de 46%, foi considerada alta (RABELO, 2013). Quanto maior a discriminação e menor o acerto ao acaso, maior será a informação do item (RABELO, 2013).

O conjunto de informações do item gera a informação do teste (ANDRADE; VALLE, 2000). Em uma avaliação em larga escala, espera-se que muitos itens se localizem em torno da média, onde o erro de medida é mínimo e em menor número nas extremidades da escala (ANDRADE; VALLE, 2000) como observado no TP 2014, com alta concentração de itens em torno da proficiência média do teste (500) e nos intervalos de 460 a 530.

As avaliações em educação, costumam utilizar muitos itens para avaliar uma disciplina ou área curricular. Esta estratégia possibilita alcançar maior cobertura dos conteúdos previstos nas matrizes de referência e ainda apresentar itens com diferentes graus de dificuldade, possibilitando melhor caracterização dos itens representativos da escala. Quanto maior for o número e a qualidade dos itens de cada nível da escala, maior será o grau de representatividade em relação às proficiências avaliadas naquele nível (KÁTIA MESQUITA DE OLIVEIRA; FRANCO; SOARES, 2007).

Um estudo brasileiro de 2008 (OLIVEIRA & FRANCO 2008), sobre escalas de proficiência e suas interpretações, utilizando avaliações como a do Sistema Nacional de Avaliação da Educação Básica – SAEB; o National Assessment for Educational Progress – NAEP; o Programa Internacional de Avaliação de Alunos – PISA; o Trends

in International Mathematics and Science Study – TIMSS e o Projeto GERES/2005 – Estudo Longitudinal sobre a Qualidade e Eqüidade no Ensino Fundamental Brasileiro, relata que o número de itens selecionado foi de 168,169,168,74 e 166 itens respectivamente.

Normalmente os núcleos internacionais, aplicam TP com em torno de 200 itens, e mais vezes ao ano (ALAMRO *et al.*, 2023; NOUNS; GEORG, 2010; TIO *et al.*, 2016).

Em relação à qualidade dos itens, é necessário que estes abordem os desempenhos cognitivos contidos nas matrizes de referência e que possuam diferentes graus de dificuldade, fornecendo uma análise mais acurada do desempenho do estudante (KÁTIA MESQUITA DE OLIVEIRA; FRANCO; SOARES, 2007). Além disso, a fim de assegurar a qualidade da informação de itens não testados, como os usados no TP 2014, no núcleo holandês (TIO *et al.*, 2016), especialistas desenvolveram cinco critérios de relevância para melhorar a consistência e precisão da construção e revisão de itens: testar o conhecimento específico da especialidade médica; testar o conhecimento pronto (necessário como pré-requisito para funcionar em uma situação prática); ser um conhecimento importante para a prática médica; ter relevância prática para o tratamento de situações de alta prevalência ou alto risco; e o conhecimento deve formar a base de um ou mais conceitos importantes do currículo (SCHUWIRTH *et al.*, 2010; WRIGLEY *et al.*, 2012). No presente estudo, após ajuste inicial aos modelos da TRI e posteriormente aos critérios de bons itens, representativos da escala de proficiência, apenas 48 itens foram puderam ser utilizados na escala, prejudicando sua interpretação pedagógica, por conta dos fatores descritos acima.

Entre outros critérios, um item representativo da escala, deve ser discriminativo e ter a > 0,7 (MOREIRA JUNIOR, 2014). Na escala construída no estudo, os itens têm discriminação maior ou igual a 1.

No presente estudo, apenas 48 itens satisfizeram os critérios necessários para seleção de bons itens âncora, podendo de certa forma dificultar a interpretação pedagógica dos níveis de proficiência da escala.

A proficiência média do teste, 500 pontos, se posicionou no nível 2, relacionado ao segundo subgrupo de itens. Como no Brasil não temos outras análises do TP pela TRI e o currículo das escolas médicas internacionais não é semelhante ao nosso, não foram encontrados dados de literatura para comparação em relação ao valor

encontrado para proficiência média e para relação dos valores de proficiência dos estudantes com os períodos do curso e desempenhos testados.

Observou-se houve um aumento da proficiência durante a graduação. Até o quarto período (ciclo básico), as proficiências dos estudantes no teste ocuparam o primeiro e segundo níveis da escala, relacionados em grande parte a conceitos e conteúdos ministrados no ciclo básico, principalmente em farmacologia e epidemiologia. Como sabemos, o processo de desenvolvimento do raciocínio clínico ao longo da formação médica ocorre na forma de sucessivos estágios (PEIXOTO; SANTOS; FARIA, 2018) Durante os anos iniciais, são aprendidos conceitos biológicos e fisiopatológicos e informações semiológicas acerca de sinais e sintomas isoladamente, permitindo aos relacionar dados semiológicos ao conhecimento biomédico previamente aprendido, mas sem relacionar manifestações clínicas a um determinado grupo de doenças (PEIXOTO; SANTOS; FARIA, 2018). Além disso, foi observado que itens de teste bem construídos estão associados à aplicação do raciocínio clínico, resolução de problemas e pensamento criativo dos alunos (ALAMRO *et al.*, 2023).

Estudantes do quinto e sexto período (ciclo clínico), obtiveram médias de proficiência que se posicionaram até o terceiro nível. Neste nível, os dois itens iniciais ainda mantiveram as características do nível anterior, porém os itens seguintes, independente da área de conteúdo testada, necessitaram conhecimento de achados clínicos e laboratoriais referentes a doenças prevalentes na rotina diária de atendimento dos estudantes. Nessa etapa do curso, os estudantes têm maior exposição à casos reais e iniciam uma etapa do raciocínio clínico onde a interrelação das manifestações clínicas à fisiopatologia favorece o encapsulamento dos conceitos e a formação de scripts das doenças (Peixoto et al. 2018). O desempenho diagnóstico depende de ter na memória uma base de conhecimento extensa e bem organizada, com um rico acervo de representações mentais ou “scripts” de doenças (MAMEDE *et al.*, 2014).

A partir do sétimo, oitavo e nono períodos (ciclo clínico e estágio), os estudantes passam a dominar conhecimentos agrupados até o quarto nível; estudantes do décimo ao décimo segundo períodos (estágio),até o quinto nível. Nesses níveis, fica bem marcada a presença de itens com casos clínicos diversos, onde é necessário organizar o conhecimento prévio e relacionar com sinais e sintomas, mecanismos causais e condições em que a doença é provável de ocorrer.

Alguns itens situados mais ao final do quarto e no quinto nível abordam conteúdos mais complexos, relacionados a especialidades.

Os níveis de proficiência em uma escala de atributos cognitivos médicos e o raciocínio clínico estão intimamente relacionados. Uma vez que o raciocínio clínico envolve a capacidade de integrar informações de diversas fontes para formular um diagnóstico e elaborar o plano de tratamento para um paciente, quanto mais avançado o nível de proficiência, maior será a capacidade do estudante de realizar um raciocínio clínico preciso e eficiente. É importante identificar a etapa de construção do raciocínio clínico em que o estudante se encontra e sua relação com o momento atual do curso, apontar fortalezas e fragilidades e promover feedback individual apropriado e contínuo.

Sendo assim são necessários múltiplos testes para uma avaliação mais fidedigna do desempenho cognitivo e das etapas de construção do raciocínio clínico do estudante ao longo de sua formação , bem como é desejável que a matriz temática utilizada para a construção do TP como um todo, considere as etapas do raciocínio clínico intrinsecamente na construção dos itens.

Os resultados são fonte importante de material a ser relacionado com o projeto pedagógico da escola médica, estrutura curricular e trabalho educativo realizado, embora sejam necessárias outras análises semelhantes a fim de que seja possível construir uma escala de proficiência que seja representativa do conhecimento cognitivo esperado à um egresso do curso de Medicina.

7 CONSIDERAÇÕES FINAIS

A análise dos resultados de um TP por meio da TRI se mostrou efetiva para avaliar os níveis de proficiência

. É necessário, no entanto, assegurar a validade dos conteúdos do teste, utilizando uma matriz que garanta a elaboração de itens que contemplem os objetivos educacionais do projeto de ensino do curso, bem como, elementos do raciocínio clínico que podem ser direta ou indiretamente avaliados nos itens.

Os resultados são fonte importante de material a ser relacionado com o projeto pedagógico da escola médica, estrutura curricular e trabalho educativo realizado, embora sejam necessárias outras análises semelhantes a fim de que seja possível construir uma escala de proficiência que seja representativa do conhecimento cognitivo esperado à um egresso do curso de Medicina.

REFERÊNCIAS

ALAMRO, A. S. *et al.* 10 years of experience in adopting, implementing and evaluating progress testing for Saudi medical students. **Journal of Taibah University Medical Sciences**, [*s. l.*], v. 18, n. 1, p. 175–185, 2023. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S1658361222001445.

ALBANESE, M. A. *et al.* Defining characteristics of educational competencies. **Medical Education**, [*s. l.*], v. 42, n. 3, p. 248–255, 2008.

ANDRADE, J. M.; LAROS, J. A.; GOUVEIA, V. V. O uso da Teoria de Resposta ao Item em avaliações educacionais: diretrizes para pesquisadores. **Avaliação Psicológica**, [*s. l.*], v. 9, n. 3, p. 421–435, 2010.

ANDRADE, D. F. De; VALLE, C. **QL-41lFJqD1.pdf**. [*S. l.: s. n.*], 2000.

ANDRADE, D. F. De; VALLE, C. **Teoria da Resposta ao Item: Conceitos e Aplicações**. [*S. l.: s. n.*], 2000.

BAIRD, J. A. *et al.* Assessment and learning: fields apart?. **Assessment in Education: Principles, Policy and Practice**, [*s. l.*], v. 24, n. 3, p. 317–350, 2017.

BELAY, L. M.; SENDEKIE, T. Y.; EYOWAS, F. A. Quality of multiple-choice questions in medical internship qualification examination determined by item response theory at Debre Tabor University, Ethiopia. **BMC Medical Education**, [*s. l.*], v. 22, n. 1, 2022.

BICUDO, A. M. *et al.* Teste de Progresso em Consórcios para Todas as Escolas Médicas do Brasil. **Revista Brasileira de Educação Médica**, [*s. l.*], v. 43, n. 4, p. 151–156, 2019.

BRASIL; MINISTÉRIO DA EDUCAÇÃO. Parecer CNE/CES nº 116/2014. **Diário Oficial da União**, [*s. l.*], n. D, p. 1–47, 2014. Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=15874-rces003-14&category_slug=junho-2014-pdf&Itemid=30192.

BROWN, T. A. Confirmatory factor analysis for applied research. **Choice Reviews Online**, [*s. l.*], v. 44, n. 05, p. 44-2769-44–2769, 2007.

CAMPOS, G. **Um método para análise e co-gestão de coletivos**. São Paulo: Hucitec, 2000.

CARABETTA JR, V. Metodologia ativa na educação médica. **Revista de Medicina**, [*s. l.*], v. 95, n. 3, p. 113, 2016.

COLLARES, C. F.; LOGULO, W.; GREC, P. Psicometria Na Garantia De Qualidade Da Educação Médica : Conceitos E Aplicações Psychometry and Medical Education Quality : Concepts and Application E Te O R Ia D E R Es P Os Ta Ao. **Science in Health**, [*s. l.*], v. 3, n. 1, p. 33–49, 2012. Disponível em:

https://www.researchgate.net/publication/230996347.

CRESO F JÚNIOR CO-ORIENTADOR, F. J.; MACHADO SOARES RIO DE JANEIRO, T. **Lina Kátia Mesquita de Oliveira TRÊS INVESTIGAÇÕES SOBRE ESCALAS DE PROFICIÊNCIA E SUAS INTERPRETAÇÕES**. [*S. l.: s. n.*], 2008. Disponível em: http://www.livrosgratis.com.br. .

DALTON F ANDRADE 2001. [*s. l.*],

DELORS, J. *et al.* **Educação: um tesouro a descobrir. Relatório para a UNESCO da Comissão Internacional sobre Educação para o Séc. XXI**. [*S. l.: s. n.*], 1998. *E-book*. Disponível em: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Educação:+um+tesouro+a+descobrir.+Relatório+para+a+UNESCO+da+Comissão+Internacional+sobre+Educação+para+o+século+XXI#0.

DIJKSTERHUIS, M. G. K. *et al.* Progress testing in postgraduate medical education. **Medical Teacher**, [*s. l.*], v. 31, n. 10, p. e464-8, 2009.

DIJKSTRA, J.; VAN DER VLEUTEN, C. P. M.; SCHUWIRTH, L. W. T. A new framework for designing programmes of assessment. **Advances in Health Sciences Education**, [*s. l.*], v. 15, n. 3, p. 379–393, 2010. Disponível em: http://link.springer.com/10.1007/s10459-009-9205-z.

EDUCAÇÃO, M. da. Diretriz Nacionais da Educação - Medicina. **Journal of Chemical Information and Modeling**, [*s. l.*], v. 53, n. 9, p. 1689–1699, 2001. Disponível em: http://portal.mec.gov.br/cne/arquivos/pdf/CES04.pdf.

EMBRETSON, S. E. **The New Rules of MeasurementPsychological Assessment**. [*S. l.: s. n.*], 1996.

EZEQUIEL, S. *et al.* Implantação do Teste de Progresso em uma instituição pública. [*s. l.*], n. April, 2016.

FRANCISCO DE ANDRADE, D.; TAVARES, H. R.; DA CUNHA VALLE, R. **Teoria da Resposta ao Item: Conceitos e Aplicações**. [*S. l.: s. n.*], [*s. d.*].

FRANK, J. R. *et al.* Competency-based medical education: Theory to practice. **Medical Teacher**, [*s. l.*], v. 32, n. 8, p. 638–645, 2010.

FRANK, J. R.; DANOFF, D. The CanMEDS initiative: Implementing an outcomes-based framework of physician competencies. **Medical Teacher**, [*s. l.*], v. 29, n. 7, p. 642–647, 2007.

GREEN, D. J.; HEALES, C. J. **Progress testing: An educational perspective exploring the rationale for progress testing and its introduction into a Diagnostic Radiography curriculum**. [*S. l.*]: Elsevier Inc., 2023.

HAIR JUNIOR, J. *et al.* **Análise multivariada de dados**. 5ªed. [*S. l.*]: Ed. Porto Alegre, 2005.

HARDEN, R. M. The integration ladder: A tool for curriculum planning and evaluation. **Medical Education**, [*s. l.*], v. 34, n. 7, p. 551–557, 2000.

KÁTIA MESQUITA DE OLIVEIRA, L.; FRANCO, C.; SOARES, T. M. **Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación**. [*S. l.: s. n.*], 2007.

MAMEDE, S. *et al.* How can students' diagnostic competence benefit most from practice with clinical cases? the effects of structured reflection on future diagnosis of the same and novel diseases. **Academic Medicine**, [*s. l.*], v. 89, n. 1, p. 121–127, 2014.

MATOS, D. A. S.; RODRIGUES, E. C. **Análise fatorial Metodologias**. [*S. l.: s. n.*], 2019.

MOREIRA JUNIOR, F. D. J. CONTRIBUIÇÕES DA TEORIA DA RESPOSTA AO ITEM NAS AVALIAÇÕES EDUCACIONAIS. **Ciência e Natura**, [*s. l.*], v. 36, n. 3, 2014.

NOUNS, Z. M.; GEORG, W. Progress testing in german speaking countries. **Medical Teacher**, [*s. l.*], v. 32, n. 6, p. 467–470, 2010.

OGLIARI, P. J.; ANDRADE, D. F. De. **Estatística básica para as ciências agronômicas e biológicas**. [*S. l.: s. n.*], 2005.

OLIVEIRA, S. S. de *et al.* Teste de Progresso da Abem: consolidando uma estratégia de avaliação para o ensino médico. **Revista Brasileira de Educação Médica**, [*s. l.*], v. 46, n. 1, 2022.

PAGLIOSA, F. L.; DA ROS, M. A. The Flexner Report: for Good and for Bad. **Revista Brasileira de Educação Médica**, [*s. l.*], v. 32, n. 4, p. 492–499, 2008.

PASQUALI, LUIZ ;PRIMI, R. Fundamentos da Teoria da Resposta ao Item. [*s. l.*], v. 2, n. 2, p. 99–110, 2003.

PASQUALI, L. Psicometria. **Revista da Escola de Enfermagem da USP**, [*s. l.*], v. 43, n. spe, p. 992–999, 2009.

PEIXOTO, J. M.; SANTOS, S. M. E.; FARIA, R. M. D. de. Processos de Desenvolvimento do Raciocínio Clínico em Estudantes de Medicina. **Revista Brasileira de Educação Médica**, [*s. l.*], v. 42, n. 1, p. 75–83, 2018.

PETITJEAN, B. **Pensar avaliação, melhorar a aprendizagem.** [*S. l.*]: LISBOA IEE, 1994.

RABELO, M. **Avaliação Educacional: fundamentos, metodologia**. [*S. l.: s. n.*], 2013.

SAKAI, M. H. *et al.* Teste de progresso e avaliação do curso: dez anos de experiência da medicina da Universidade Estadual de Londrina. **Revista Brasileira**

de Educação Médica, [*s. l.*], v. 32, n. 2, p. 254–263, 2008.

SAKAI, M. H.; FERREIRA FILHO, O. F.; MATSUO, T. Avaliação do crescimento cognitivo do estudante de Medicina: aplicação do teste de equalização no Teste de Progresso. **Revista Brasileira de Educação Médica**, [*s. l.*], v. 35, n. 4, p. 493–501, 2011.

SCHUWIRTH, L. *et al.* Collaboration on progress testing in medical schools in the Netherlands. **Medical Teacher**, [*s. l.*], v. 32, n. 6, p. 476–479, 2010.

SCHUWIRTH, L. W. T.; VAN DER VLEUTEN, C. P. M. Programmatic assessment and Kane's validity perspective. **Medical Education**, [*s. l.*], v. 46, n. 1, p. 38–48, 2012. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2923.2011.04098.x.

SCRIVEM, M.; W., T. R.; R.M., G. No Scriven, M. (1967). The Methodology of Evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), Perspectives of Curriculum Evaluation Chicago (pp. 39-83). Rand McNally.Title. [*s. l.*], p. 39–83, 1967.

SIMPSON, J. G. *et al.* The Scottish doctor - Learning outcomes for the medical undergraduate in Scotland: A foundation for competent and reflective practitioners. **Medical Teacher**, [*s. l.*], v. 24, n. 2, p. 136–143, 2002.

TIO, R. A. *et al.* The progress test of medicine: the Dutch experience. **Perspectives on Medical Education**, [*s. l.*], v. 5, n. 1, p. 51–55, 2016.

TOMIC, Eliane R. *et al.* Progress testing: evaluation of four years of application in the school of medicine, University of São Paulo. **Clinics (São Paulo, Brazil)**, [*s. l.*], v. 60, n. 5, p. 389–396, 2005.

TREVISAN, L. M. V. *et al.* Dimensionalidade e escala de proficiência em uma prova interdisciplinar. **Estudos em Avaliação Educacional**, [*s. l.*], v. 30, n. 74, p. 392, 2019.

TRONCON, L. E. de A. Avaliação do estudante de medicina. **Medicina (Ribeirao Preto Online)**, [*s. l.*], v. 29, n. 4, p. 429–439, 1996.

TRONCON, L. E. D. A. Avaliação Programática do Estudante: Estratégia Institucional para Melhor Cumprir as Funções da Avaliação Educacional. **Revista de Graduação USP**, [*s. l.*], v. 1, n. 1, p. 53, 2016. Disponível em: http://www.revistas.usp.br/gradmais/article/view/117725.

TSUJI, H.; AGUILAR-DA-SILVA, R. H. Experience with a problem-based curriculum implemented in the endocrine system unit in the 2nd grade medical course at Marília Medical School-FAMEMA. **Arquivos brasileiros de endocrinologia e metabologia**, [*s. l.*], v. 48, n. 4, p. 535–543, 2004.

VAN DER VLEUTEN 1996. [*s. l.*],

VAN DER VLEUTEN, C.; FREEMAN, A.; COLLARES, C. F. Progress test utopia. **Perspectives on Medical Education**, [*s. l.*], v. 7, n. 2, p. 136–138, 2018.

WARE, J.; VIK, T. Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. **Medical Teacher**, [*s. l.*], v. 31, n. 3, p. 238–243, 2009.

WRIGLEY, W. *et al.* A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. **Medical Teacher**, [*s. l.*], v. 34, n. 9, p. 683–697, 2012.

APÊNDICE A – Artigo Científico

TESTE DE PROGRESSO NA AVALIAÇÃO DO ESTUDANTE DE MEDICINA: o que vislumbrar na perspectiva da teoria de resposta ao item?

PROGRESS TEST IN MEDICAL STUDENT ASSESSMENT: what to glimpse from the perspective of item response theory?

Monica Couto Guedes Sejanes da Rocha*
Julio Maria Fonseca Chebli**
Sandra Helena Cerrato Tibiriça***
Wellington Silva****

RESUMO

Objetivo: Este estudo transversal, com 1559 estudantes brasileiros de Medicina, enfatizou o Teste de Progresso como instrumento de avaliação e descreveu uma metodologia para a análise dos resultados, pela Teoria de Resposta ao Item. Materiais e métodos: Calculado intervalo de confiança de 95% e aplicado o teste de Tukey para analisar a diferença entre as médias de proficiência. Realizada a estimativa da proficiência e calibração dos parâmetros dos itens pelo software BILOG MG 3.0. Para a construção da escala de proficiência, foram selecionados itens de discriminação a > 1, que, submetidos a agrupamento, formaram seis subconjuntos com níveis crescentes de proficiência. A interpretação pedagógica da escala foi dada pela relação desses níveis com os conteúdos testados. Resultados: A frequência de alunos que realizaram o teste foi de 94,2%. A proficiência variou de 470 a 580 pontos, com média de 500 pontos e aumentou progressivamente na graduação, um indicador de validade das medidas obtidas. A curva de ajuste indicou um teste difícil; 61,5% dos itens com discriminação moderada a alta. Do total de alunos, 22% obtiveram proficiência no teste

---

* Mestre em Saúde Brasileira pela Universidade Federal de Juiz de Fora. Docente da disciplina de Pediatria, da Faculdade de Ciências Médicas e da Saúde de Juiz de Fora (FCMS/JF).E-mail:monicasejanes@hotmail.com.
**Doutor em Gastroenterologia pela Universidade Federal de São Paulo. Professor Titular da Disciplina de Gastroenterologia da Faculdade de Medicina da Universidade Federal de Juiz de Fora. E-mail:julio.chebli@medicina.ufjf.br.
***Doutora em Saúde Brasileira pela Universidade Federal de Juiz de Fora. Docente efetiva do Estágio de Medicina Comunitária/ Atenção Primária em saúde na UFJF.E-mail: shctibi@gmail.com.
****Doutor em Educação pela Pontifícia Universidade Católica do Rio de Janeiro. Coordenador de medidas educacionais do Centro de Políticas Públicas e Avaliação da Educação - CAEd/UFJF. E-mail:wellington@caed.ufjf.br.

superior ao nível 2, 490 a 510 pontos. Conclusões: Apesar das limitações descritas, o Teste de Progresso mostrou-se um valioso instrumento de avaliação de desempenho quando analisado pela Teoria de Resposta ao Item. A relação do nível de proficiência necessário para a aprendizagem dos domínios cognitivos do teste pode ser explicada através das etapas de construção do raciocínio clínico, importantes no desenvolvimento da competência diagnóstica, fundamental na formação do estudante de Medicina.
Palavras-chave: Educação médica. avaliação educacional. Cognição. Competência profissional. Psicometria.

ABSTRACT

Objective: This cross-sectional study, with 1559 Brazilian medical students, emphasized the Progress Test as an assessment tool and described a methodology for the analysis of the results, by Item Response Theory. Materials and methods: Calculated confidence interval of 95% and applied Tukey test to analyse the difference between proficiency means. Estimation of proficiency and calibration of the parameters of the items by software BILOG MG 3.0. To construct the proficiency scale, items of discrimination a > 1 were selected, which, subjected to clustering, formed six subsets with increasing levels of proficiency. The pedagogical interpretation of the scale was given by the relationship of these levels with the tested contents. Results: The frequency of students who took the test was 94.2%. Proficiency ranged from 470 to 580 points, with a mean of 500 points and increased progressively at graduation, an indicator of validity of the measures obtained. The adjustment curve indicated a difficult test; 61.5% of items with moderate to high discrimination. Among the total number of students, 22% obtained a test proficiency higher than level 2, 490 to 510 points. Conclusion: Despite the limitations described, Progress Test proved to be a valuable tool for performance assessment when analysed by Item Response Theory. The relationship of the level of proficiency required for learning the cognitive domains of the test may be explained through the stages of construction of clinical reasoning, important in the development of diagnostic competence, fundamental in the training of medical students.
Keywords: Medical education, Educational assessment, Cognition, Professional competence, Psychometrics.

Submetido em 08/05/23

1 INTRODUÇÃO

O currículo da escola médica deve ter por missão a formação de profissionais competentes, com domínio dos recursos essenciais para um desempenho de êxito, em situações complexas de vigilância de saúde (TSUJI; AGUILARDA-SILVA, 2004), garantindo assim, a qualidade dos cuidados de saúde (BELAY; SENDEKIE; EYOWAS, 2022).É importante acompanhar o desempenho dos estudantes durante a graduação e inferir se as competências profissionais foram alcançadas (BAIRD *et al.*, 2017), utilizando para tal, avaliações de abordagem programática, por meio de diferentes ferramentas e metodologias, alinhadas com as competências esperadas (COLLARES; LOGULO; GREC, 2012).Este estudo enfatiza a validade do Teste de Progresso (TP) como instrumento de avaliação de desempenho e a importância da utilização de metodologias de análise dos resultados, que produzam medidas de qualidade e valor pedagógico.

O TP, introduzido inicialmente na década de 1970, é uma ferramenta de avaliação cognitiva (OLIVEIRA et al., 2022). utilizada em diversas escolas médicas no mundo. Entre as vantagens, as aplicações repetidas, estimulam a aprendizagem longitudinal, com base no desempenho no teste (VAN DER VLEUTEN; FREEMAN; COLLARES, 2018 ; BICUDO et al., 2019). Permite às escolas verificar a eficácia do currículo, possibilitando intervenções corretivas no processo de ensino e aprendizagem (VAN DER VLEUTEN; FREEMAN; COLLARES, 2018 ;BELAY; SENDEKIE; EYOWAS, 2022).

No Brasil, atualmente, o TP é aplicado por mais de 180 escolas médicas (52,7% do total), integrando dezoito núcleos regionais (OLIVEIRA et al., 2022). Os resultados são analisados, em sua maioria, segundo a Teoria Clássica dos Testes (TCT).

A literatura (PASQUALI, 2009 ; COLLARES; LOGULO; GREC, 2012), no entanto, apresenta vantagens da Teoria de Resposta ao Item (TRI) sobre a TCT. A TRI estima a probabilidade de um aluno com um determinado desempenho, denominado proficiência, acertar um item do teste (PASQUALI, LUIZ ;PRIMI, 2003); (COLLARES; LOGULO; GREC, 2012). Posiciona esses valores numa única métrica, possibilitando a comparação entre eles e construção de uma escala de proficiência, que permite acompanhar o desenvolvimento progressivo do desempenho do estudante (ANDRADE; LAROS; GOUVEIA, 2010). Para que esse processo seja pedagogicamente significativo, é necessário interpretar a escala, relacionando os

níveis de proficiência ao conteúdo do teste (KÁTIA MESQUITA DE OLIVEIRA; FRANCO; SOARES, 2007; ANDRADE; LAROS; GOUVEIA, 2010).

As escalas possibilitam comparar médias de proficiência entre os estudantes e entre os períodos de escolaridade; verificar a distribuição dos estudantes por nível de proficiência, permitindo avaliar se o resultado alcançado foi satisfatório ou insatisfatório e requer tratamento pedagógico; avaliar o estágio de desenvolvimento de cada desempenho cognitivo testado (início, pico, consolidação e desenvolvimento máximo) e comparar os resultados com o momento do curso no qual o estudante se encontra (KÁTIA MESQUITA DE OLIVEIRA; FRANCO; SOARES, 2007).

No Brasil, o número de consórcios de TP tem aumentado nos últimos anos, melhorando a qualidade da avaliação. Por outro lado, a análise dos resultados por TCT, subestima as informações fornecidas pelo teste, tanto sobre o desempenho dos estudantes , como sobre o currículo e a escola, importantes para a tomada de decisão dos gestores.

Este estudo apresentou uma análise do TP por meio da TRI, potencializando as informações fornecidas pelo teste. A precisão da mensuração do desempenho do estudante na avaliação, permite intervenções oportunas e contribui para a aquisição da competência necessária ao profissional de saúde ao final do curso.

2 MATERIAIS E MÉTODOS

Este estudo transversal, avaliou o desempenho cognitivo de estudantes de Medicina de Minas Gerais, no TP de 2014, por meio da estimativa de proficiência pela TRI.

2.1 CONTEXTO

Atualmente existem três consórcios de escolas médicas para aplicação do TP em Minas Gerais, nomeados TepMinas I, II e III. O TepMinas I, foi criado em 2013 e em 2014, seis das sete escolas que integravam o consórcio na época e totalizavam 4557 estudantes matriculados no curso de Medicina, participaram da aplicação do TP analisado no estudo. Os docentes de uma das escolas ainda estavam em treinamento, por isso a escola não participou.

Os itens foram elaborados seguindo uma matriz construída pelo consórcio. O teste foi composto por 120 questões de múltipla escolha, de resposta única, com

quatro alternativas, nas áreas de Ciências Básicas, Clínica Médica, Clínica Cirúrgica, Pediatria, Saúde Coletiva, Ética, Ginecologia e Obstetrícia.

A metodologia utilizada pelo consórcio para análise dos resultados, foi a TCT O resultado individual por escola e global não identificado, foi fornecido aos respectivos coordenadores.

O acesso ao banco de dados da prova, para a realização das análises tratadas no estudo, foi solicitado oficialmente à coordenação do curso de medicina das seis escolas do TepMinas I. Apenas duas escolas, a Universidade Federal de Juiz de Fora - UFJF (instituição pública) e a Faculdade de Ciências Médicas e da Saúde de Juiz de Fora (FCMS/JF) - SUPREMA (instituição privada), concordaram em ceder seus resultados.

2.2 PARTICIPANTES

O estudo envolveu 1655 estudantes de Medicina, 973 matriculados na UFJF e 682 na FCMS/JF. Foram excluídos 42 estudantes da UFJF (4,32%) e 54 da FCMS/JF (7,92%) que estavam ausentes. Os 1559 estudantes incluídos corresponderam a 46% dos estudantes que realizaram o teste. O cálculo do tamanho amostral não foi realizado devido à ausência de dados da população total de estudantes.

2.3 ÉTICA

Aprovado pelo Comité de Ética e Investigação da FCMS/JF, parecer 2.076.924.

2.4 ANÁLISE ESTATÍSTICA DOS RESULTADOS

A frequência de participação dos estudantes foi avaliada por meio de análise descritiva.

As proficiências médias de cada período foram comparadas, a fim de, verificar o aumento ao longo da graduação. O intervalo de confiança contendo o valor verdadeiro da média do período, foi calculado e comparado com as médias dos outros intervalos, para verificar se houve mudança significativa entre eles. O cálculo do intervalo de confiança de 95% para as médias de proficiência dos estudantes, utilizou a seguinte fórmula (OGLIARI; ANDRADE, 2005)

$$IC = \bar{x} \pm Z\alpha/2 * s/\sqrt{n} \quad (1)$$

Onde,

$\bar{x}$ = Média da amostra, nesse caso PRF_TEP.

Zα/2 = Valor associado ao nível de significância α/2, nesse caso 1,96.

s = Desvio padrão da amostra, nesse caso PRF_TEP_sd.

n = Tamanho da amostra, nesse caso ESTUDANTES.

Variação = $Z\alpha/2 * s/\sqrt{n}$

Foi utilizado o teste de Tukey para verificar se as diferenças nos valores de proficiência entre os períodos de ensino foram significativas. A distribuição da amplitude estudentizada, o quadrado médio dos resíduos da ANOVA e o tamanho amostral dos grupos, foram calculados para determinar a diferença mínima significativa (D.M.S.), considerando os percentis dos grupos.

2.5 ANÁLISE PSICOMÉTRICA DO TESTE

Foram realizadas análises para verificar a possibilidade de utilização do modelo unidimensional da TRI e, em seguida, as análises de ajuste do teste aos estudantes e de ajuste dos itens ao modelo da TRI.

2.5.1 Análise de dimensionalidade

A análise de dimensionalidade foi realizada por meio de técnicas fatoriais, para verificar se os dados do TepMinas I possuíam característica unidimensional, condição essencial para a utilização do modelo unidimensional da TRI para itens dicotômicos.

A análise fatorial exploratória (AFE) constatou que dois fatores contribuíram significativamente para a variância dos itens e podendo ser extraídos. Essa análise está ancorada na literatura (MATOS; RODRIGUES, 2019) e nos arquivos gerados (ScreePlot e Variância acumulada), considerando a base de dados do teste. Em seguida, se os fatores 1 e 2 explicam a variância dos itens, pela análise fatorial confirmatória (AFC) verificou-se que os dois fatores extraídos eram autocorrelacionados, o que permitiu inferir que as respostas aos itens poderiam ser ajustadas por um modelo unidimensional da TRI. Esta análise está ancorada na literatura (BROWN, 2007; HAIR JUNIOR *et al.*, 2005) e nos ficheiros (Item Factor, Correlation Factor e Goodness fit) gerados tendo em conta a base de dados do teste.

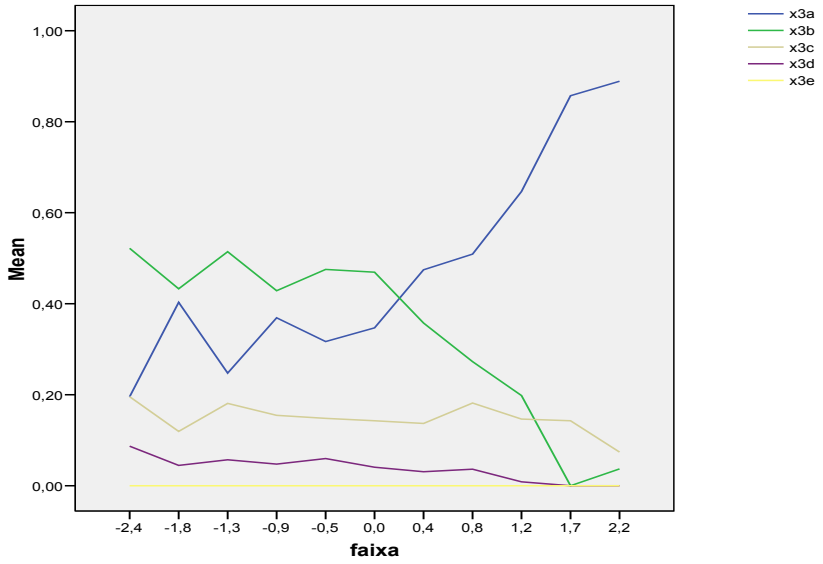
2.5.2 Análise de ajuste do teste aos estudantes

Para avaliar a qualidade do teste, a escala foi dividida em intervalos de meio desvio-padrão e verificado os percentuais de estudantes (proficiência) e itens (parâmetro de dificuldade) em cada um dos intervalos construídos.

2.5.3 Ajuste dos itens ao modelo da TRI

Os ajustes gráficos de cada item (Figura 1) permitiram avaliar o comportamento de cada alternativa do item, possibilitando a identificação de problemas de elaboração e de conteúdo dos itens. Três itens foram excluídos por inadequação ao modelo.

Figura 1 – Representação gráfica de ajuste de um item do teste.

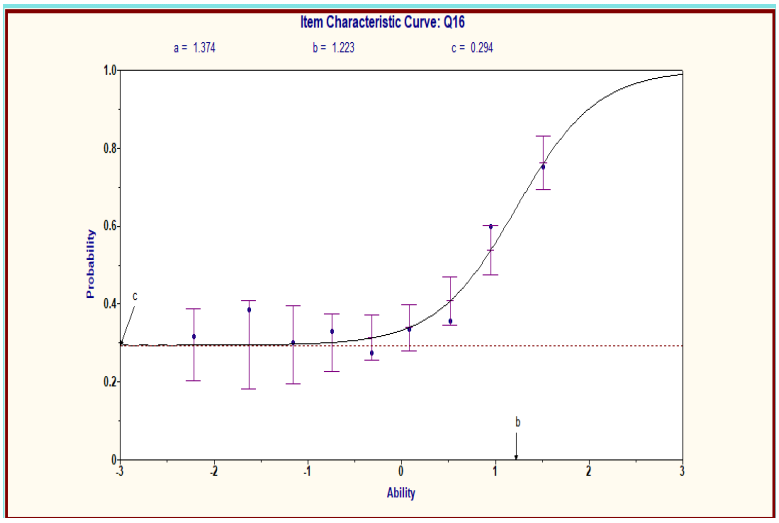


Fonte: Software BILOG MG 3.0.

2.5.4 Calibração de itens e estimativa de proficiência

Os itens foram calibrados utilizando-se o método MMAP (Maximum Marginal A Posteriori), obtendo-se os parâmetros de discriminação (a), dificuldade (b) e acerto ao acaso (c) para cada item (FRANCISCO DE ANDRADE; TAVARES; DA CUNHA VALLE, 2000). As proficiências foram estimadas pelo método EAP (Expected A Posteriori), com base no padrão de acerto do examinando. A relação entre a proficiência do estudante (variável latente medida) e a probabilidade de responder corretamente ao item (FRANCISCO DE ANDRADE; TAVARES; DA CUNHA VALLE, 2000) foi representada graficamente pelas CCIs (Figura 2). A inclinação da curva do item, informou os pontos de início, auge e consolidação da aprendizagem (i,b_esc e s) de cada domínio testado. O software BILOG MG 3.0 também forneceu os gráficos das curvas de informação de cada item e do teste.

Figura 2 – Representação gráfica da curva característica de um item



Fonte: Software BILOG MG 3.0.

2.6 CONSTRUÇÃO DA ESCALA DE PROFICIÊNCIA

Os valores obtidos como parâmetros dos itens e como pontos de início, auge e consolidação da aprendizagem de cada desempenho testado, permitiram a construção de uma escala, com média 0 e desvio padrão 1, representativa do construto proficiência para o teste aplicado. Foram selecionados como representativos da escala de proficiência, itens com parâmetro alto de discriminação a > 1 (TREVISAN *et al.*, 2019) e baixa possibilidade de acerto ao acaso, c < 0,5. Por estarem em uma mesma escala métrica, os valores do parâmetro b são considerados representativos da proficiência na escala criada e para evitar valores pequenos ou negativos na escala, estes foram multiplicados e somados a valores arbitrários, criando uma escala arbitrária , com 28 intervalos de distância de 5 pontos entre si.

2.6.1 Determinação dos níveis de proficiência

A interpretação pedagógica da escala, no entanto, depende da relação entre a proficiência e os desempenhos cognitivos abordados no teste. Para estabelecer essa relação, os itens da escala foram agrupados por proximidade em subconjuntos (clusters), de acordo com a medida da distância entre os pontos b e s da sua CCI, por meio da metodologia K-means. O nível de proficiência associado a cada subconjunto de itens, foi definido pela média dos pontos relativos ao auge (b) e à consolidação da aprendizagem (s) dos itens desse subconjunto.

2.7 INTERPRETAÇÃO PEDAGÓGICA DA ESCALA

Em cada subconjunto de itens, o desempenho cognitivo foi avaliado por meio de quatro intervalos de aprendizagem, considerados significativos: abaixo do ponto que permite acertar os itens; o crescimento máximo da aprendizagem, necessário ao acerto dos itens; início da consolidação da aprendizagem e aprendizagem consolidada (KÁTIA MESQUITA DE OLIVEIRA; FRANCO; SOARES, 2007).

A discussão sobre a similaridade dos itens em cada subconjunto e sua relação com os níveis de proficiência, foi realizada com base nas teorias de construção do raciocínio clínico (MAMEDE *et al.*, 2014; PEIXOTO; SANTOS; FARIA, 2018).

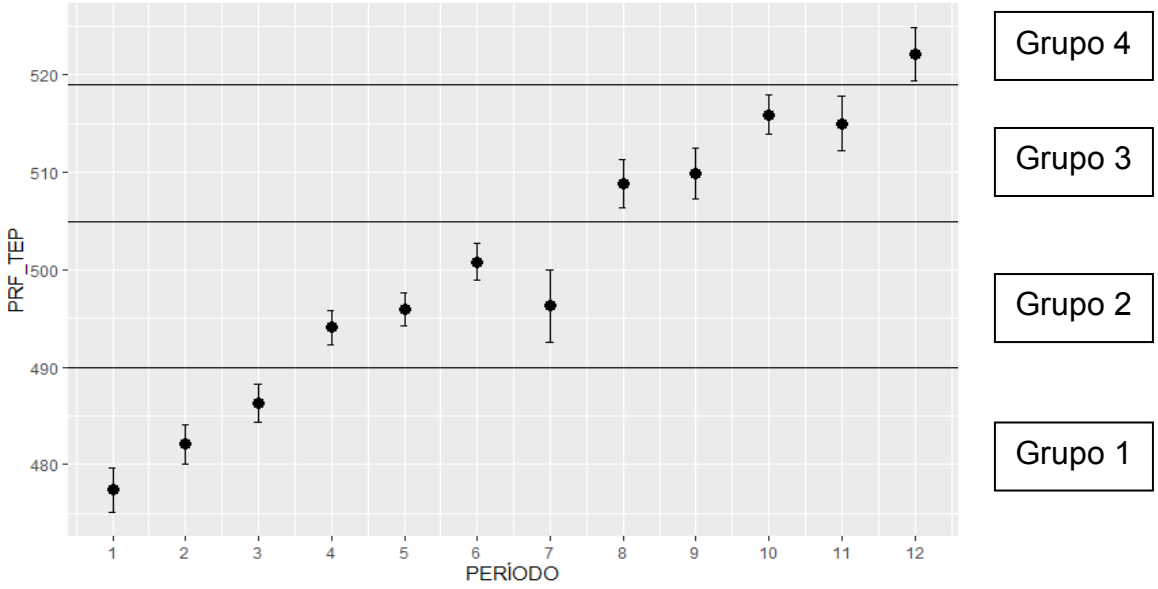
2 RESULTADOS

A frequência de alunos que realizaram o teste foi de 94,2%, sendo 59,7% da UFJF e 40,3% da Suprema.

Apresentamos na figura 3, as médias por período, os respectivos ICs e os resultados obtidos pelo teste de Tukey. Verificamos quatro grupos com proficiência crescente ao longo da graduação.

Os períodos 1 a 3 (grupo 1), se posicionaram na mesma faixa de proficiência, o que também ocorreu para os períodos 4 a 7 (grupo 2) e 8 a 11 (grupo 3). Os valores de proficiência aumentaram progressivamente na ordem crescente dos grupos, atingindo o auge no período 12 (grupo 4), ou seja, os estudantes estão no ápice de seus respectivos níveis de conhecimento neste último período. Os resultados foram sensíveis à evolução da escolaridade, um indicador de validade das medidas obtidas.

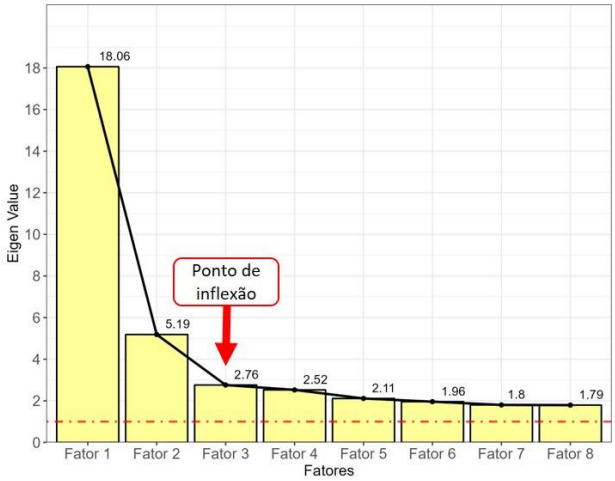
Figura 3. Teste de Tukey para as médias dos períodos avaliados.



Fonte: Software BILOG MG 3.0.

A análise da dimensionalidade verificou que dois fatores contribuíram significativamente para a variância dos itens. A figura 4 apresenta o diagrama de inclinação (= Scree Test), método utilizado para definir o número de fatores a extrair.

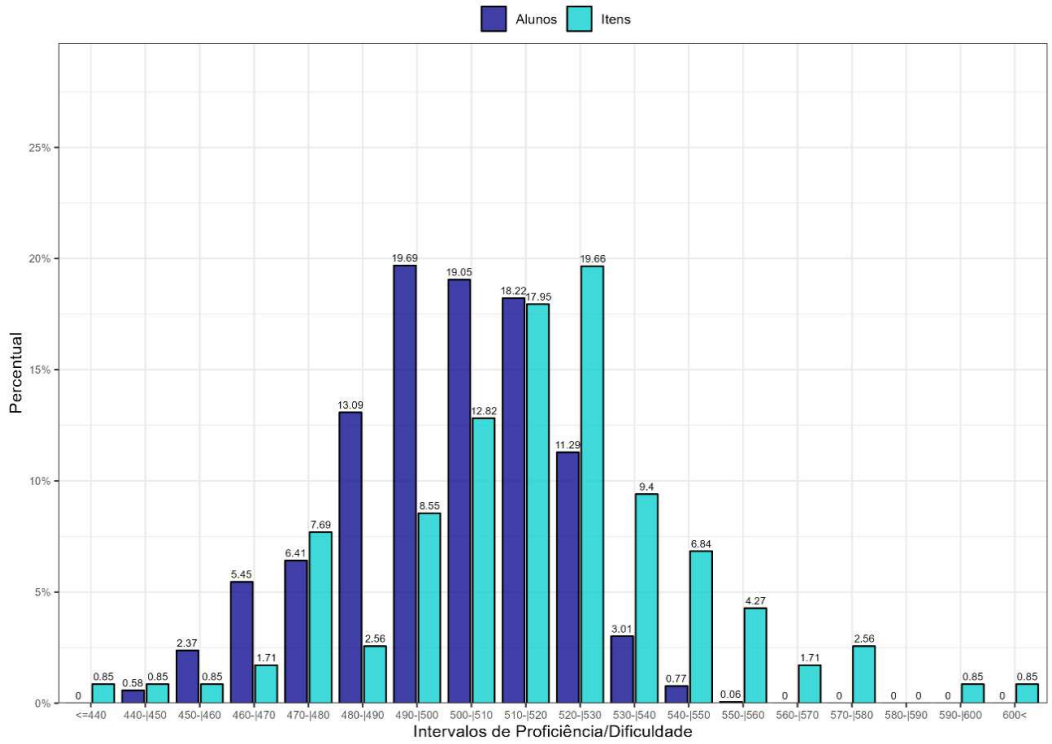
Figura 4. Diagrama de inclinação.



Fonte: Elaboração Própria.

Na curva de ajuste do teste aos estudantes, observou-se alta concentração de itens em torno da proficiência média, 500 pontos e nos intervalos de 460 a 530 pontos. Verificou-se também que o teste está desalinhado à direita da escala, ou seja, o teste foi difícil para os alunos (Figura 5).

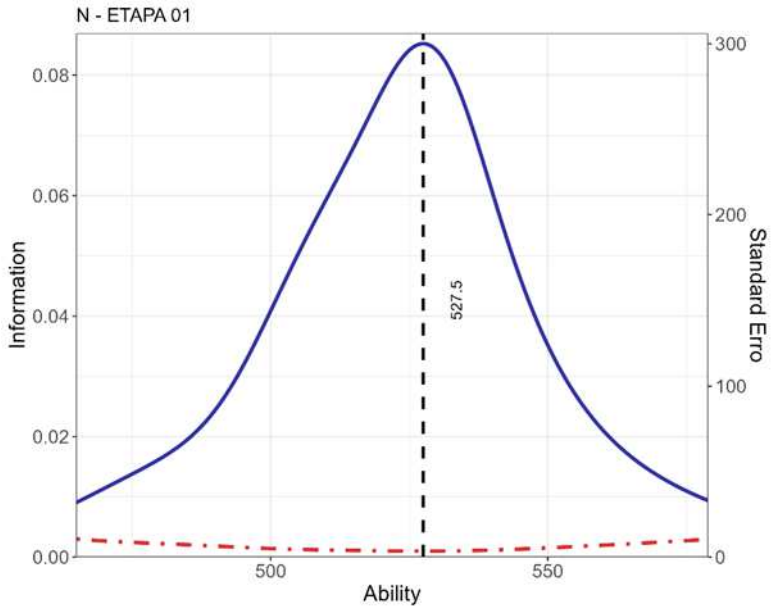
Figura 5. Curva de ajuste do teste aos estudantes.



Fonte: Software BILOG MG 3.0.

O teste apresentou uma boa estimativa de proficiência, como visualizado na curva de informação do teste, figura 6. O máximo de informação ocorreu no valor de proficiência de 527,5 (linha contínua) e coincidiu com o mínimo erro de medida (linha pontilhada).

Figura 6. Curva de informação do teste.



Fonte: Software BILOG MG 3.0.

A calibração dos itens revelou uma discriminação moderada a alta (a > 0,65) (RABELO, 2013) em 61,5% dos itens. Relativamente ao parâmetro dificuldade, 15,4% dos itens eram fáceis (b < - 0,52), 23% moderados (- 0,52 > b < 0,52) e 61,5% difíceis (b > 0,52) (RABELO, 2013). Por se tratar de itens de múltipla escolha com quatro alternativas, a probabilidade de acerto ao acaso foi considerada para valores de c > 0,25 e foi de 46% (RABELO, 2013).

Foram selecionados 48 itens como representativos da escala de proficiência. Os valores de proficiência (parâmetro b) desses itens foram arbitrariamente multiplicados por 20 e somados a 500, obtendo-se uma escala de proficiência que compreende um intervalo entre 455 e 590 pontos (Figura 7).

Figura 7.Representação dos parâmetros que geraram a escala de proficiência do TP de 2014.

n	item	selecao	a	b	c	i	b_erc	r
112,00	Q115	1,00	1,07209	-1,43495	0,13654	456,85	471,30	485,75
100,00	Q103	1,00	1,06162	-1,36239	0,12219	458,16	472,75	487,35
111,00	Q114	1,00	1,06872	-1,19840	0,09940	461,53	476,03	490,53
93,00	Q96	1,00	1,11879	-0,78365	0,26997	470,48	484,33	498,18
97,00	Q100	1,00	1,11529	-0,27975	0,25936	480,51	494,41	508,30
26,00	Q27	1,00	1,80258	-0,09574	0,50000	489,49	498,09	506,68
17,00	Q18	1,00	1,03110	-0,07780	0,32135	483,42	498,44	513,47
52,00	Q54	1,00	1,46529	-0,01134	0,35077	489,20	499,77	510,35
51,00	Q53	1,00	1,16576	0,01812	0,32113	487,07	500,36	513,65
73,00	Q75	1,00	1,19739	0,06026	0,26102	488,27	501,21	514,14
71,00	Q73	1,00	1,52577	0,11258	0,28005	492,10	502,25	512,41
31,00	Q33	1,00	1,69440	0,21393	0,20263	495,13	504,28	513,42
61,00	Q63	1,00	1,01225	0,23933	0,33084	489,48	504,79	520,09
23,00	Q24	1,00	1,46910	0,44985	0,14996	498,45	509,00	519,54
36,00	Q38	1,00	1,31334	0,48877	0,21166	497,98	509,78	521,57
91,00	Q94	1,00	1,13653	0,51320	0,26786	496,63	510,26	523,90
83,00	Q86	1,00	1,73023	0,53597	0,37242	501,76	510,72	519,67
14,00	Q14	1,00	1,51076	0,54121	0,40268	500,57	510,82	521,08
114,00	Q117	1,00	1,20042	0,62394	0,18096	499,57	512,48	525,39
53,00	Q55	1,00	1,40138	0,67989	0,17081	502,54	513,60	524,65
39,00	Q41	1,00	1,41256	0,72658	0,31553	503,56	514,53	525,50
42,00	Q44	1,00	1,31500	0,73734	0,40376	502,96	514,75	526,53
44,00	Q46	1,00	1,36209	0,74099	0,18598	503,44	514,82	526,19
86,00	Q89	1,00	1,59515	1,01257	0,32710	510,54	520,25	529,96
64,00	Q66	1,00	1,65806	1,14987	0,21133	513,65	523,00	532,34
60,00	Q62	1,00	1,79923	1,15705	0,23997	514,53	523,14	531,75
16,00	Q16	1,00	1,37425	1,22345	0,29418	513,19	524,47	535,74
11,00	Q11	1,00	1,03364	1,23069	0,43102	509,62	524,61	539,60
62,00	Q64	1,00	1,86579	1,23916	0,20141	516,48	524,78	533,09
102,00	Q105	1,00	1,90783	1,25494	0,17894	516,98	525,10	533,22
67,00	Q69	1,00	1,14419	1,30108	0,14756	512,48	526,02	539,56
66,00	Q68	1,00	1,58283	1,31845	0,12608	516,58	526,37	536,16
78,00	Q81	1,00	1,35283	1,34447	0,16056	515,44	526,89	538,34
25,00	Q26	1,00	1,61368	1,40929	0,28010	518,58	528,19	537,79
54,00	Q56	1,00	1,17053	1,42120	0,31523	515,19	528,42	541,66
89,00	Q92	1,00	1,19401	1,42599	0,21907	515,54	528,52	541,50
63,00	Q65	1,00	1,30729	1,49644	0,26265	518,08	529,93	541,78
80,00	Q83	1,00	2,10533	1,50162	0,21210	522,67	530,03	537,39
65,00	Q67	1,00	1,62431	1,52592	0,44034	520,98	530,52	540,06
20,00	Q21	1,00	1,02020	1,56806	0,23598	516,17	531,36	546,55
33,00	Q35	1,00	1,49300	1,64395	0,13314	522,50	532,88	543,26
68,00	Q70	1,00	2,16397	1,68605	0,19179	526,56	533,72	540,88
56,00	Q58	1,00	1,08439	1,82834	0,11867	522,28	536,57	550,85
50,00	Q52	1,00	1,14789	1,96935	0,17857	525,89	539,39	552,88
94,00	Q97	1,00	1,58347	2,06324	0,24183	531,48	541,26	551,05
104,00	Q107	1,00	1,35150	2,40284	0,24251	536,59	548,06	559,52
43,00	Q45	1,00	1,15788	3,17093	0,12083	550,04	563,42	576,80
24,00	Q25	1,00	1,07154	3,36558	0,33125	552,85	567,31	581,77

Fonte: Software BILOG MG 3.0.

A análise de conglomerados deu origem a seis subconjuntos de itens, que de acordo com a média dos pontos b e s de seus itens, foram associados a seis níveis de proficiência.

Os níveis foram representados por intervalos de proficiência com limiares semelhantes aos valores determinados na escala. Nível 1: até 490; Nível 2: 491 a 510; Nível 3: 511 a 525; Nível 4: 526 a 540; Nível 5: 541 a 560; Nível 6: 561 a 580. Como o último nível era representado por apenas dois itens, optou-se por não o considerar (Quadro 1).

A relação entre o período do curso e o nível de proficiência definido em cada subconjunto de itens foi avaliada, observando-se um aumento da proficiência de acordo com o avanço na graduação, como mostra o Quadro 1.

Quadro 1: Relação entre o período do curso e o nível de proficiência dos estudantes.

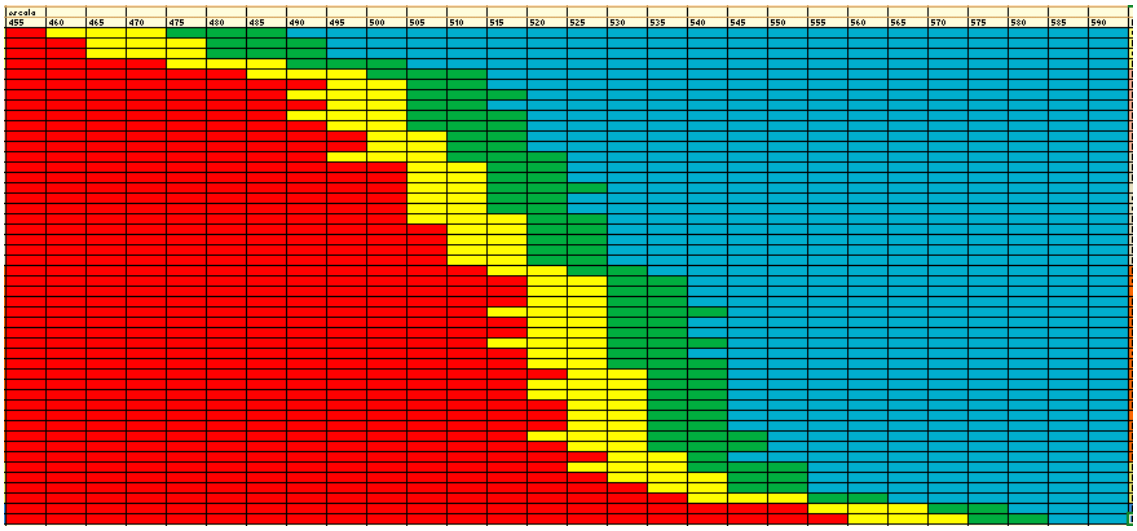
Período do Curso		Faixas de Proficiência por nível					Total de estudantes
		Nível 1 até 490	Nível 2 491 - 510	Nível 3 511 - 525	Nível 4 526 - 540	Nível 5 541- 560	
1	Count	102	28	0	0	0	130
	%	78,5%	21,5%	,0%	,0%	,0%	100,0%
2	Count	85	47	0	0	0	132
	%	64,4%	35,6%	,0%	,0%	,0%	100,0%
3	Count	61	70	0	0	0	131
	%	46,6%	53,4%	,0%	,0%	,0%	100,0%
4	Count	28	114	0	0	0	142
	%	19,7%	80,3%	,0%	,0%	,0%	100,0%
5	Count	20	121	1	0	0	142
	%	14,1%	85,2%	,7%	,0%	,0%	100,0%
6	Count	16	117	6	0	0	139
	%	11,5%	84,2%	4,3%	,0%	,0%	100,0%
7	Count	39	69	10	7	0	125
	%	31,2%	55,2%	8,0%	5,6%	,0%	100,0%
8	Count	15	78	29	19	0	141
	%	10,6%	55,3%	20,6%	13,5%	,0%	100,0%
9	Count	10	64	24	29	0	127
	%	7,9%	50,4%	18,9%	22,8%	,0%	100,0%
10	Count	2	43	22	41	2	110
	%	1,8 %	39,1%	20,0%	37,3%	1,8%	100,0%
11	Count	7	44	15	38	3	107
	%	6,5%	41,1%	14,0%	35,5%	2,8%	100,0%
12	Count	5	28	8	72	20	133
	%	3,8%	21,1%	6,0%	54,1%	15,0%	100,0%
Total	Count	390	823	115	206	25	1559
%		25,0%	52,8%	7,4%	13,2%	1,6%	100,0%

Fonte: Software BILOG MG 3.0.

Entre o total de estudantes, 22% obtiveram uma proficiência no teste superior ao Nível 2, de 490 a 510 pontos (Figura 8). Desses estudantes, 57,5% estavam no estágio (9º ao 10º período) e 13% no ciclo clínico (5º ao 8º período).

A proficiência também foi relacionada com a aprendizagem dos desempenhos cognitivos testados em cada subconjunto de itens, por meio de quatro intervalos considerados significativos: abaixo do ponto que permitiria acertar os itens; auge da aprendizagem necessária para acertar os itens; início da consolidação da aprendizagem e aprendizagem já consolidada (Figura 9).

Figura 9 - Representação da relação da proficiência com a aprendizagem de cada domínio cognitivo testado.



Fonte: Software BILOG MG 3.0.

Legenda:

- Aquém
- Auge do desenvolvimento
- Início de consolidação
- Consolidação

4 DISCUSSÃO

No TP Minas 2014, a frequência de 94,2% dos alunos das duas instituições envolvidas foi semelhante à de 92% (SAKAI; FERREIRA FILHO; MATSUO, 2011; TOMIC, Eliane R. *et al.*, 2005) descrita pela Universidade Estadual de Londrina (UEL) de 2004 a 2006, assim como o percentual de participação por período do curso (SAKAI; FERREIRA FILHO; MATSUO, 2011; TOMIC et al., 2005).

Alamro et al. (2023) relata participação em torno de 100% na aplicação do TP por um consórcio saudita de 20 escolas nos anos de 2020 e 2021.O aumento da participação foi atribuído à percepção adequada dos dirigentes universitários sobre os benefícios do teste. Nouns e Georg (2010) relatam uma participação de 85-95% no TP realizado por 13 escolas na Alemanha e na Áustria e atribuem-na principalmente ao conceito formativo sem decisão de aprovação/reprovação.

A proficiência média no teste foi de 500 pontos e apresentou o aumento progressivo esperado ao longo da graduação (477,3 para 522,8), contribuindo como medida de validade do teste (DIJKSTERHUIS *et al.*, 2009; GREEN; HEALES, 2023; TRONCON, 1996). Em estudos nacionais também podemos observar esse resultado, mas com base na pontuação total de respostas corretas na avaliação, calculada pela TCT (BICUDO *et al.*, 2019; SAKAI; FERREIRA FILHO; MATSUO, 2011; TOMIC, Eliane R. *et al.*, 2005) e como os itens de um instrumento de medida não são tratados individualmente, não é possível comparar. Holanda e a Alemanha, com experiência em TP multicêntrico utilizando a TRI, apresentam resultados de progressão semelhantes aos deste estudo e, para além da importância formativa, destacam o TP como fonte de informação para avaliação e monitorização de alterações curriculares e formação de professores (ALAMRO *et al.*, 2023; TIO *et al.*, 2016).

O teste apresentou um grau de dificuldade elevado, com 61,5% de itens difíceis (RABELO, 2013). A discriminação foi adequada, pois 61,5% dos itens apresentaram discriminação moderada a alta (a > 0,65) (RABELO, 2013). Ware e Vik (2009) delinearam alguns critérios para itens de qualidade, como a presença de pelo menos 50% dos itens em níveis cognitivos superiores (aplicação e raciocínio), com discriminação maior ou igual a 60%, o que ocorreu no teste estudado. A probabilidade de acerto ao acaso de 46% foi considerada alta (RABELO, 2013). Quanto maior a discriminação e menor o acerto ao acaso, maior a informação do item (RABELO, 2013).

O conjunto de informações do item gera as informações do teste (FRANCISCO DE ANDRADE; TAVARES; DA CUNHA VALLE, 2000). Em uma avaliação em larga escala, espera-se que muitos itens se localizem em torno da média, onde o erro de medida é mínimo, e em menor número nos extremos da escala (FRANCISCO DE ANDRADE; TAVARES; DA CUNHA VALLE, 2000), como observado no TP 2014, com uma elevada concentração de itens em torno da proficiência média da prova (500) e nos intervalos 460 a 530 pontos.

Quanto maior o número e a qualidade dos itens em cada nível da escala, maior o grau de representatividade em relação às proficiências avaliadas naquele nível (KÁTIA MESQUITA DE OLIVEIRA; FRANCO; SOARES, 2007). Entre outros critérios, um item representativo da escala, deve ser discriminativo e ter um > 0,7 (MOREIRA JUNIOR, 2014). Na escala construída no estudo, os itens possuem discriminação maior ou igual a 1.

As avaliações em educação requerem testes com itens que abordem os desempenhos cognitivos contidos nas matrizes de referência e que possuam diferentes graus de dificuldade, proporcionando uma análise mais precisa do desempenho dos estudantes (KÁTIA MESQUITA DE OLIVEIRA; FRANCO; SOARES, 2007). Normalmente, os consórcios internacionais aplicam o TP com um número maior de itens e mais vezes ao ano (ALAMRO *et al.*, 2023; NOUNS; GEORG, 2010; TIO *et al.*, 2016). Além disso, para garantir a qualidade da informação de itens não testados, como os utilizados na TP de 2014 no consórcio holandês (TIO *et al.*, 2016) os especialistas desenvolveram cinco critérios de relevância para melhorar a consistência e a precisão da construção e revisão dos itens: testar conhecimentos específicos da especialidade médica; testar conhecimentos prontos (necessários como pré-requisito para funcionar numa situação prática); ser um conhecimento importante para a prática médica; ter relevância prática para o tratamento de situações de alta prevalência ou de alto risco; e o conhecimento deve formar a base de um ou mais conceitos importantes no currículo (SCHUWIRTH *et al.*, 2010; WRIGLEY *et al.*, 2012). No presente estudo, após o ajuste inicial aos modelos da TRI e, posteriormente, aos critérios de bons itens, representativos da escala de proficiência , apenas 48 itens puderam ser utilizados na escala, comprometendo a sua interpretação pedagógica, devido aos fatores acima descritos

A proficiência média do teste, 500 pontos, foi posicionada no nível 2, referente ao segundo subgrupo de itens. Como no Brasil não temos outras análises do TP por TRI e o currículo das escolas médicas internacionais não é semelhante ao nosso, não foram encontrados dados na literatura para comparação quanto ao valor encontrado para a proficiência média e para a relação dos valores de proficiência dos alunos com os períodos do curso e desempenhos testados.

Como pudemos observar, houve um aumento da proficiência durante a graduação. Até o quarto período (ciclo básico), as proficiências dos estudantes no teste ocuparam o primeiro e segundo níveis da escala, em grande parte relacionadas

a conceitos e conteúdos ministrados no ciclo básico, principalmente em farmacologia e epidemiologia. Como sabemos, o processo de desenvolvimento do raciocínio clínico ao longo da formação médica ocorre sob a forma de etapas sucessivas (PEIXOTO; SANTOS; FARIA, 2018). Durante os anos iniciais, os conceitos biológicos e fisiopatológicos e as informações semiológicas sobre sinais e sintomas são aprendidos de forma isolada, permitindo aos estudantes relacionar os dados semiológicos aos conhecimentos biomédicos previamente aprendidos, mas sem relacionar as manifestações clínicas a um determinado grupo de doenças (PEIXOTO; SANTOS; FARIA, 2018).

Os estudantes do quinto e sexto período (ciclo clínico), obtiveram médias de proficiência que se posicionaram até o terceiro nível. Nesse nível, os dois itens iniciais ainda mantinham as características do nível anterior, mas os itens seguintes, independentemente da área de conteúdo testada, exigiam o conhecimento de achados clínicos e laboratoriais referentes a doenças prevalentes na rotina diária de atendimento dos estudantes. Nessa etapa do curso, os estudantes têm maior exposição a casos reais e iniciam uma fase de raciocínio clínico em que a inter-relação das manifestações clínicas com a fisiopatologia favorece o encapsulamento de conceitos e a formação de roteiros de doenças (PEIXOTO; SANTOS; FARIA, 2018). O desempenho diagnóstico depende da existência de uma base de conhecimento extensa e bem organizada na memória, com um rico acervo de representações mentais ou scripts de doença (MAMEDE *et al.*, 2014).

A partir do sétimo, oitavo e nono períodos (ciclo clínico e estágio), os estudantes passam a dominar conhecimentos agrupados até ao quarto nível; os alunos do décimo ao décimo segundo períodos (estágio), até ao quinto nível. Nestes níveis, é bem marcada a presença de itens com casos clínicos diversos, onde é necessário organizar conhecimentos prévios e relacioná-los com sinais e sintomas, mecanismos causais e condições em que a doença é susceptível de ocorrer. Alguns itens localizados no final do quarto e quinto nível abordam conteúdos mais complexos relacionados às especialidades.

Este estudo deixa clara a possibilidade de avaliar o desempenho do estudante com uma medida de qualidade, através do TP, quando analisado pela TRI. Para uma interpretação pedagógica mais precisa, como exige o currículo médico, é necessária a aplicação de avaliações sucessivas com um número maior de itens de boa qualidade psicométrica. Tal medida, tanto do ponto de vista financeiro como pedagógico,

beneficiaria da cooperação entre as escolas médicas, partilhando experiências, pontos fortes e fracos.

5 CONSIDERAÇÕES FINAIS

Os resultados obtidos no estudo, apesar das limitações descritas, mostram que é possível monitorar o desempenho cognitivo dos estudantes de medicina pelo TP, além de serem bons indicadores para a avaliação do currículo e do curso.

A análise pela TRI qualifica o TP como um instrumento de medida educacional da qualidade do desempenho e de seus desdobramentos. É necessário, no entanto, garantir a validade de conteúdo do teste, utilizando uma matriz que assegure a elaboração de itens que atendam aos objetivos educacionais do projeto pedagógico do curso.

Os resultados constituem uma importante fonte de material a ser relacionado ao projeto pedagógico da escola médica, à estrutura curricular e ao trabalho educacional desenvolvido, embora sejam necessárias outras análises semelhantes para a construção de uma escala de proficiência que seja representativa do conhecimento cognitivo esperado de um graduando do curso de Medicina.

Parte final do artigo, na qual se apresentam as considerações correspondentes aos objetivos e/ou hipóteses.

Potencial conflito de interesse:
Os autores relatam nenhum conflito de interesses

Fontes de financiamento:
O presente estudo não teve fontes de financiamento externas.

Vinculação acadêmica:
Este artigo é parte da tese de doutorado de Monica Couto Guedes Sejanes da Rocha pela Universidade Federal de Juiz de Fora (UFJF).

REFERÊNCIAS

Alamro AS, Alghasham AA, Al-Shobaili HA, Alhomaidan HT, Salem TA, Wadi MM, Saleh MN. 2023. 10 anos de experiência na adoção, implementação e avaliação de testes de progresso para estudantes de medicina sauditas. J Taibah Univ Med Sci [Internet]. 18(1):175–185. https://doi.org/10.1016/j.jtumed.2022.07.008.

Andrade DF De, Valle C. 2000. Teoria de Resposta ao Item: Conceitos e Aplicações. SINAPE.

Andrade JM, Laros JA, Gouveia V V. 2010. O uso da Teoria de Resposta ao Item em avaliações educacionais: orientações para pesquisadores. Avaliação Psicológica. 9(3):421-435.

Baird JA, Andrich D, Hopfenbeck TN, Stobart G. 2017. Avaliação e aprendizagem: campos separados? Assess Educ Princ Policy Pract. 24(3):317–350. https://doi.org/10.1080/0969594X.2017.1319337.

Belay LM, Sendekie TY, Eyowas FA. 2022. Qualidade das questões de múltipla escolha no exame de qualificação do internato médico determinada pela Teoria da Resposta ao Item na Universidade Debre Tabor, Etiópia. BMC Med Educ. 22(1). https://doi.org/10.1186/s12909-022-03687-y.

Bicudo AM, Hamamoto Filho PT, Abbade JF, Hafner M de LMB, Maffei CML. 2019. Teste de Progresso em Consórcios para todas as Escolas Médicas do Brasil. Rev Bras Educ Med. 43(4):151–156. https://doi.org/10.1590/1981-52712015v43n4rb20190018.

Brown TA. 2007. Análise fatorial confirmatória para pesquisa aplicada. Choice Rev Online. 44(05):44-2769-44–2769. https://doi.org/10.5860/choice.44-2769.

Collares CF, Logulo W, Grec P. 2012. Psicometria e Qualidade do Ensino Médico : Conceitos e Aplicação na Teoria de Resposta ao Item. Sci Heal [Internet]. 3(1):33-49. https://www.researchgate.net/publication/230996347.

Dijksterhuis MGK, Scheele F, Schuwirth LWT, Essed GGM, Nijhuis JG, Braat DDM. 2009. Testes de progresso no ensino médico pós-graduado. Med Teach. 31(10):e464-8. https://doi.org/10.3109/01421590902849545.

Green DJ, Heales CJ. 2023. Teste de progresso: Uma perspectiva educativa que explora os fundamentos dos testes de progresso e a sua introdução num currículo de Radiografia de Diagnóstico. J Med Imaging Radiat Sci. 54(1):35-42. https://doi.org/10.1016/j.jmir.2022.12.009.

Kátia Mesquita De Oliveira L, Franco C, Soares TM. 2007. Revista Electrónica Iberoamericana de Calidad, Eficacia y Cambio en Educación Vol. 5, No. 2e.

Mamede S, Van Gog T, Sampaio AM, De Faria RMD, Maria JP, Schmidt HG. 2014. Como é que a competência diagnóstica dos alunos pode beneficiar mais da prática com casos clínicos? os efeitos da reflexão estruturada no diagnóstico futuro das mesmas e novas doenças. Acad Med. 89(1):121–127. https://doi.org/10.1097/ACM.0000000000000076.

Matos DAS, Rodrigues EC. 2019. Metodologias de Análise Fatorial. Escola Nacional de Administração Pública. Brasília: Enap, 2019.

Moreira Junior FDJ. 2014. Contribuições da Teoria de Resposta ao Item nas avaliações educacionais. Ciência e Nat. 36(3). https://doi.org/10.5902/2179460x13120.

Nouns ZM, Georg W. 2010. Testes de progresso em países de língua alemã. Med Teach. 32(6):467–470. https://doi.org/10.3109/0142159X.2010.485656.

Ogliari PJ, Andrade DF De. 2005. Estatística básica para ciências agrárias e biológicas. Universidade Federal de Santa Catarina, Centro Tecnológico Departamento de Informática e Estatística. Florianópolis, Santa Catarina - Brasil.

Oliveira SS de, Postal EA, Afonso DH, Merss CE, Cyrino EG, Abreu Junior AF de, Batista NA. 2022. Teste de Progresso da Abem: consolidando uma estratégia de avaliação da educação médica. Rev Bras Educ Med. 46(1). https://doi.org/10.1590/1981-5271v46.1-editorial.

Pasquali, Luiz ;Primi R. 2003. Fundamentos da Teoria de Resposta ao Item. 2(2):99-110.

Pasquali L. 2009. Psicometria. Rev da Esc Enferm da USP. 43(spe):992–999. https://doi.org/10.1590/s0080-62342009000500002.

Peixoto JM, Santos SME, Faria RMD de. 2018. Processos de desenvolvimento do raciocínio clínico em estudantes de medicina. Rev Bras Educ Med. 42(1):75–83. https://doi.org/10.1590/1981-52712015v41n4rb20160079.

Rabelo M. 2013. Avaliação educacional: fundamentos, metodologia. SBM.

Sakai MH, Ferreira Filho OF, Matsuo T. 2011. Avaliação do crescimento cognitivo de estudantes de medicina: aplicação do teste de equalização no Teste de Progresso. Rev Bras Educ Med. 35(4):493–501. https://doi.org/10.1590/s0100-55022011000400008.

Schuwirth L, Bosman G, Henning RH, Rinkel R, Wenink ACG. 2010. Colaboração em testes de progresso em escolas médicas na Holanda. Med Teach. 32(6):476–479. https://doi.org/10.3109/0142159X.2010.485658.

Tio RA, Schutte B, Meiboom AA, Greidanus J, Dubois EA, Bremers AJA. 2016. O teste de progresso da medicina: a experiência holandesa. Perspect Med Educ. 5(1):51-55. https://doi.org/10.1007/s40037-015-0237-1.

Tomic ER, Martins MA, Lotufo PA, Benseñor IM. 2005. Teste de progresso: avaliação de quatro anos de aplicação na Faculdade de Medicina da Universidade de São Paulo. Clínica Médica 60(5):389-396

Trevisan LMV, Barbetta PA, Andrade DF de, Rocha GT, Azevedo TCA de M. 2019. Dimensionalidade e escala de proficiência em uma prova interdisciplinar. Estud em Avaliação Educ. 30(74):392. https://doi.org/10.18222/eae.v30i74.5324.

Troncon LE de A. 1996. Avaliação do estudante de medicina. Med (Ribeirão Preto Online). 29(4):429–439. https://doi.org/10.11606/issn.2176-7262.v29i4p429-439.

Tsuji H, Aguilar-da-Silva RH. 2004. Experiência com um currículo baseado em problemas implementado na unidade de sistema endócrino do 2º ano do curso médico da Faculdade de Medicina de Marília-FAMEMA. Arq Bras Endocrinol Metabol. 48(4):535–543. https://doi.org/10.1590/s0004-27302004000400015.

van der Vleuten C, Freeman A, Collares CF. 2018. Utopia do Teste do Progresso. Perspect Med Educ. 7(2):136-138. https://doi.org/10.1007/s40037-018-0413-1.

Ware J, Vik T. 2009. Garantia de qualidade da redacção de itens: Durante a introdução de perguntas de múltipla escolha em medicina para exames de alto risco. Med Teach. 31(3):238–243. https://doi.org/10.1080/01421590802155597.

Wrigley W, Van Der Vleuten CP, Freeman A, Muijtjens A. 2012. Um quadro sistémico para o Teste de Progresso: Strengths, constraints and issues: Guia AMEE n.º 71. Med Teach. 34(9):683–697. https://doi.org/10.3109/0142159X.2012.704437.

,

ANEXO A – Confirmação de submissão do artigo

08/05/23, 23:31 ScholarOne Manuscripts

Medical Teacher

Home

Author

Submission Confirmation

Print

Thank you for your submission

Submitted to	Medical Teacher
Manuscript ID	CMTE-2023-0521
Title	PROGRESS TEST IN MEDICAL STUDENT ASSESSMENT: WHAT TO GLIMPSE FROM THE PERSPECTIVE OF ITEM RESPONSE THEORY?
Authors	da Rocha, Monica Tibiriça, Sandra Silva, Wellington Chebli, Julio
Date Submitted	08-May-2023

Author Dashboard

ANEXO B – Exemplo da matriz de competências do TepMinas I

PROBLEMA	NÍVEL DE ATENÇÃO	IDADE	GÊNERO	HABILIDADE
Artrite reumatoide	Primária	Idoso	Feminino	Anamnse/exame clínico
Hepatites	Primária	Adulto	Masculino	Interpretação de exames complementares
Ação de quimioterápicos no ciclo celular em neoplasias.	Terciária	Adulto	Feminino	Compreensão de aspectos morfofuncionais.

ANEXO C - Panorama da utilização internacional dos Testes de Progresso

School(s)	Description	Notes
Netherlands Group Five medical faculties in the Netherlands (Groningen, Leiden, Maastricht, Nijmegen and VU Amsterdam) and additionally, the Ghent University in Belgium use the test	This test began in 1977 in Maastricht. Collaboration with other schools began in 1999. The test is delivered, paper based, four times a year to a total of about 10,000 students in all 6 years of the undergraduate schools. There are 200 questions on each paper in single best answer format with a correction for guessing. Students are allowed 4 hours for the test. Questions are selected according to a blueprint from the whole of medical knowledge. The question bank has about 5000 items. The tests are summative and there are additional block tests. Scores are aggregated over the tests, scoring classification is based on mean and standard deviation of year cohort performance	Percentage scores are converted to qualifications (Good, Satisfactory Unsatisfactory). Tables have been defined with the final outcome for all possible combinations. Tables are based on two concepts: one exceptional qualification cannot save or ruin the general pattern, and the most recent test carries the largest weight There is a form of progress testing at Utrecht using a short answer format[4] Several postgraduate training programmes use progress testing
McMaster, including undergraduate programme, physician assistant programme, Canada Limerick University	This test began in 1992. The test is delivered three times a year to all 3 years of the programme. The test is delivered online to about 540 students (McMaster). There are 180 questions in single best answer format with a correction for guessing. Students are allowed 3 hours for the test. Questions are selected according to a blueprint from the whole of medical knowledge with a major focus on basic science. The question bank has about 3000 items. The tests are both summative and formative. There are additional Concept Application Exercises as summative tests. Scores are not aggregated over the tests but students are 'flagged' if their score is more than 1.5 standard deviations below the cohort mean score	Limerick Medical School, Ireland, use this McMaster PPI test, delivered on line for the graduate entry programme. The test is delivered twice a year to students in all 4 years of the course, (approximately 90 students/year in this new and expanding course) it is used formatively only
Charite, Germany (Germany – Berlin Regel, Berlin reform, Witten, Aachen, Bochum, LMU Munich, Koln, Munster, Hannover, Mannheim, Regensburg, Austria – Graz, Innsbruck)	This test began in Charité, Berlin in 1999 with an initiative led by the students of the school. This formative test is now delivered to all the listed institutions in Germany and Austria. The test is delivered twice a year, paper based. It is delivered to about 8500 students each time. It is purely formative and each school has different regulations about the test. In Berlin, it is given throughout the course for students on the PBL course, for those students on the traditional course, it is delivered in the clinical years of the course. There are 200 questions in single best answer format with a correction for guessing. The test lasts 3 hours. The question bank has 5300 items. There is a specified blueprint for each test representing the whole of medical knowledge. The student report has number of correct answers, wrong answers, don't-know answers and 'total score' correct–wrong each for the whole test and for each subscale (medical organs × medical disciplines)	There is another postgraduate progress test for 50 trainees in dermatology specialization
NBME 1 (Barts, St. George's London, Leeds and Queens University, Belfast)	This recent venture (2008) is a collaboration between these four UK medical schools and the National Board of Medical Examiners of the USA. Each school uses the test in different ways. NBME delivers two tests per year, web based. Each test is 120 items in single best answer format with no correction for guessing. The test lasts 3 hours. Generally, the tests are given in the last 3 years of the 5-year courses. The blueprint represents graduation objectives of the whole of medical knowledge. The bank provides an initial eight non overlapping test forms (question papers)	
NBME 2 (University of South Florida and Case Western Reserve University)	This is another initiative for two schools in the USA. As above, NBME provides test forms to the schools according to a specified blueprint that is generally drawn from the whole of medical knowledge. In south Florida, the test is used summatively. Initially (2008), it was given four times a year to third year students with 230 questions in each test, delivered online over 5 hours. They are currently reviewing their aims and objectives for this test	Currently, there is no detail from Case Western but I understand they use the test in their clinical years
Southern Illinois University, Vanderbilt, University of New Mexico, Penn State, Texas Tech, Medical College of Georgia, University of Minnesota	Beginning in 2004, these six schools deliver a single formative test each year. The tests have 70 questions currently delivered as paper test but planning for online delivery in 2010. The questions are single best answer and given to students in all 4 years of the course, a total of 3200 students. There is no correction for guessing, there is no aggregation of test scores. Questions are in two broad areas – clinical data interpretation and diagnostic pattern matching. The question bank has several hundred items of each type	

Manchester, UK	Manchester began using progress tests in 1997. The test is delivered twice a year to all 5 years of the course, about 2000 students. It is a paper-based test. There are 125 single best answer questions with no correction for guessing. The students have 2½ hours for each test. Test scores are simply totalled with no aggregation system. The tests are summative but have little progress decisions in the first 2 years of the course. In the final year of the course, there are additional summative knowledge tests. Questions are all selected from the UMAP (Universities Medical Assessment Partnership, a consortium of 14 UK medical schools) question bank. The blueprint is from the whole of medical knowledge	
Peninsula College of Medicine and Dentistry, UK	The school began using progress tests with its first student entry in 2002. The test is delivered four times a year to students in all 5 years of the course, 900 students. Paper based but delivered as a pilot online in 2009. There are 125 items in each test in single best answer format with a correction for guessing. Students have 3 hours for the test. The school's question bank has 2500 items and questions are also used from the UMAP question bank. The tests are summative. Questions are aggregated over the tests and scores are based on percentiles of each year cohort with a criterion referenced standard for the tests in the final year. There are no other knowledge-based tests in the school except for an additional end of year test at the end of year one. The blueprint is on the whole of medical knowledge and is organized in line with a General Medical Council developed blueprint for the knowledge of junior doctors	Progress tests are also used in the Dental School
Tampere, Finland	This school delivers formative progress tests three times a year to all the students on the course, about 600 students. There are a total of 18 tests over the length of the course and a student is expected to attend 80% of those tests. Each test has 224 items in True/False format. The students have 3 hours for this paper-based test. There is a correction for guessing. Students are scored for the total test and sub scores are discipline based. The blueprint is from the whole of medical knowledge with a discipline based structure. The bank has 10,000 items	
Otago Medical School, New Zealand	This school has a formative progress test delivered on line twice a year. There are 100 items in each test. The items are MCQ	
Sao Paulo Medical School, Brazil	Starting in 2001, a formative progress test was delivered twice a year to all students on the course, approximately 1000 students. The test is 100 questions in single best answer format with no correction for guessing. There is strict blueprint with 33 questions on basic sciences, 33 questions on clinical sciences and 34 on clerkship rotation questions. There is a bank of 2000 questions	See notes[b]
University of Indonesia, Medical School	The University of Indonesia began using progress testing in 2008. The tests are formative delivered to all undergraduate years, approximately 1200 students. The paper-based tests have 120 questions in Single Best Answer format, no correction for guessing, delivered twice a year. Blueprinting is from the whole of medical knowledge. The bank has 1200 items	
Mozambique	The Catholic University of Mozambique has been using progress testing since 2001. The tests are used both summatively and formatively. It is used in all 6 years of the programme and delivered to a total of 200 students. There are four tests each year delivered on paper. 200 questions in each test as single best answer. The students are allowed 4 hours for the test. There are corrections for guessing. The blueprint is for the whole of medical knowledge. There are additional block tests. There is a bank of several thousand questions	
Pretoria, South Africa	No direct information but publication suggests progress test may be taking place	See notes[c]

Notes: Where a test is used summatively, it is also (almost by definition for progress testing) used formatively but different schools will provide student feedback in different ways. Where a test is used formatively, there are no summative judgements made on the basis of progress test scores. [a]Rademakers et al. (2005). [b]The information about Brazil is from Tomic et al. (2005) suggesting that other schools in Brazil use progress testing. [c]Verhoeven et al. (2005). PPI, Personal Progress Index; NBME, The National Board of Medical Examiners; MCQ, multiple-choice question.

AGRADECIMENTOS

Agradecemos a Universidade Federal de Juiz de Fora e a Faculdade de Ciências Médicas e da Saúde de Juiz de Fora por cederem seus bancos de dados, tornando possível o estudo..