

UNIVERSIDADE JOSÉ DO ROSÁRIO VELLANO - UNIFENAS

Layara Maria Garcia

**COMPARAÇÃO DA CONFIABILIDADE ENTRE TRÊS ESCALAS PARA
AVALIAÇÃO DE HABILIDADES CLÍNICAS EM LABORATÓRIO DE
SIMULAÇÃO E TRÊS TIPOS DIFERENTES DE AVALIADORES. ESTUDO
EXPERIMENTAL UNICÊNTRICO, 2019**

Belo Horizonte

2019

Layara Maria Garcia

**COMPARAÇÃO DA CONFIABILIDADE DE TRÊS ESCALAS PARA AVALIAÇÃO
DE HABILIDADES CLÍNICAS EM LABORATÓRIO DE SIMULAÇÃO E TRÊS
TIPOS DIFERENTES DE AVALIADORES. ESTUDO EXPERIMENTAL
UNICÊNTRICO, 2019**

**Dissertação apresentada ao curso de Mestrado
Profissional em Ensino em Saúde da Universidade
José do Rosário Vellano para obtenção do título de
Mestre em Ensino em Saúde.**

**Orientador: Prof. Antonio Carlos de Castro Toledo
Jr.**

Belo Horizonte

2019

Ficha catalográfica elaborada pela Biblioteca Itapoã UNIFENAS
Conforme os padrões do Código de Catalogação Anglo Americano (AACR2)

61-057 (043.3)

G216c Garcia, Layara Maria.

Comparação da confiabilidade entre três escalas para avaliação de habilidades clínicas em laboratório de simulação e três tipos diferentes de avaliadores: estudo experimental unicêntrico, 2019. [manuscrito] / Layara Maria Garcia. -- Belo Horizonte, 2019.

54 f. : il.

Dissertação (Mestrado) - Universidade José do Rosário Vellano, Programa de Pós-Graduação Stricto Sensu em Ensino em Saúde, 2019.

Orientador : Prof. Dr. Antonio Carlos de Castro Toledo Jr.

1. Avaliação educacional. 2. Competência clínica. 3. Simulação. 4. Treinamento por simulação. I. Toledo Jr, Antonio Carlos de Castro. II. Título.

Bibliotecária responsável: Jéssica M. Queiroz CRB6/3254



Presidente da Fundação Mantenedora - FETA

Larissa Araújo Velano Dozza

Reitora

Maria do Rosário Velano

Vice-Reitora

Viviane Araújo Velano Cassis

Pró-Reitor Acadêmico

Mário Sérgio Oliveira Swerts

Pró-Reitora Administrativo-Financeira

Larissa Araújo Velano Dozza

Pró-Reitora de Planejamento e Desenvolvimento

Viviane Araújo Velano Cassis

Diretora de Pesquisa e Pós-graduação

Laura Helena Órfão

Coordenador do Curso de Mestrado Profissional em Ensino em Saúde

Antonio Carlos de Castro Toledo Jr.

Certificado de Aprovação

“COMPARAÇÃO DA CONFIABILIDADE ENTRE TRÊS ESCALAS PARA AVALIAÇÃO DE HABILIDADES CLÍNICAS EM LABORATÓRIO DE SIMULAÇÃO E TRÊS TIPOS DIFERENTES DE AVALIADORES. ESTUDO EXPERIMENTAL UNICÊNTRICO, 2019”

AUTOR: Layara Maria Garcia

ORIENTADOR: Prof. Dr. Antonio Carlos de Castro Toledo Jr

Aprovado como parte das exigências para obtenção do Título de **Mestre Profissional em Ensino em Saúde** pela Comissão Examinadora.



Prof. Dr. Antonio Carlos de Castro Toledo Jr



Profa. Dra. Maria Aparecida Turci



Profa. Dra. Juliana Silva Barra

Belo Horizonte, 06 de dezembro de 2019.



Prof. Dr. Antonio Carlos de Castro Toledo Jr.
Coordenador do Mestrado Profissional
Em Ensino em Saúde
UNIFENAS

Para meu orientador, Antonio Carlos de Castro Toledo Jr.,
pelo incentivo durante todo o projeto. Sua paciência e
apoio foram essenciais para a sua realização.

AGRADECIMENTOS

Primeiramente agradeço a Deus que me deu forças para concluir este projeto de forma satisfatória.

Aos meus pais, Eliane e Laércio (*in memorian*), que trabalharam muito e fizeram de tudo para que o estudo se tornasse um sonho possível. Aos meus irmãos, Lidiane, Laércio Jr. e Maria Julia, pelo apoio e carinho. Ao meu padrasto, Gaspar, que cuidou do meu filho durante horas de estudo.

Aos queridos professores que me guiaram neste projeto e aos meus colegas que tornaram tudo mais leve.

A todos os professores e alunos que foram voluntários, sem a colaboração de vocês este projeto não teria se tornado uma realidade.

A colega e professora Flávia Soares de Matos, que me ajudou e incentivou nesta jornada.

E, finalmente, agradeço ao meu filho Leandro, razão da minha vida, que me impulsiona a acordar todos os dias e buscar o meu melhor.

“Não há saber mais ou saber menos: há saberes diferentes”.

PAULO FREIRE.

RESUMO

A incorporação do treinamento de habilidades clínicas através da simulação nos currículos das escolas médicas foi uma inovação importante para educação, permitindo que os estudantes se capacitem de forma ética em ambiente seguro e sem oferecer riscos aos pacientes. A melhor forma de avaliar habilidades em ambiente simulado não está totalmente definida. Poucos estudos compararam escalas de avaliação e avaliadores em relação a sua confiabilidade. **Objetivo:** descrever como diferentes escalas de avaliação de habilidades clínicas em laboratório de simulação e diferentes perfis de avaliadores interferem na confiabilidade da avaliação. **Metodologia:** foram recrutados 28 alunos do terceiro e 20 alunos do sexto anos do curso de medicina da UNIFENAS-BH. Os estudantes realizaram uma prova prática de habilidades sobre antropometria do recém-nascido no laboratório de habilidades, onde foram filmados para posterior avaliação do desempenho através de três escalas: escala de avaliação do nível de treinamento (NT), escala de classificação global (EG) e lista de verificação ou *checklist* (CL). Foram selecionados nove avaliadores divididos em três perfis: três pediatras sem experiência em ensino de habilidades, três pediatras com experiência em ensino de habilidade e três professores de habilidades que não eram pediatras. Um avaliador de cada perfil utilizou apenas uma das três escalas propostas, de forma que as três escalas foram utilizadas em cada um dos três perfis de avaliadores. Foi calculado o escore total para cada uma das escalas, num total possível de 9 pontos. **Resultados:** ao comparar as diferentes escalas, independentemente do perfil do avaliador, observou-se apenas que a nota isolada do CL foi capaz de discriminar os alunos do terceiro e do sexto anos. A análise multivariada mostrou que a NT, isolada ($p=0,346$) ou total ($p=0,407$), foi à única escala que não foi influenciada pelo perfil do avaliador. A análise de confiabilidade realizada de acordo com as escalas isoladas demonstrou que apenas a NT apresenta boa confiabilidade. O acréscimo da nota das medidas antropométricas aumentou a confiabilidade entre as escalas e avaliadores. **Conclusão:** existe diferença na confiabilidade entre as escalas. A NT acrescida de dados antropométricos foi a de maior confiabilidade. O perfil do avaliador interferiu diretamente na confiabilidade das escalas, porém os avaliadores especialistas apresentaram confiabilidade aceitável quando a escala foi somada aos dados objetivos. Apenas a NT foi capaz de discriminar corretamente os dois grupos e não foi influenciada pelo perfil do avaliador.

Palavras-chave: Avaliação educacional. Competência clínica. Simulação. Treinamento por simulação.

ABSTRACT

The use of clinical skills training through simulation into medical school curricula has been a major innovation for education, enabling students to train ethically in a safe environment without risk to patients. The best way to evaluate skills in a simulated environment is not well defined. Few studies have compared teachers and rating scales reliability. **Objective:** To describe how different clinical skill assessment scales in simulation laboratory and different teachers profiles interfere with assessment reliability. **Methodology:** 28 students from the third year and 20 students from the sixth year of UNIFENAS-BH Medical School were recruited. Students performed a practical skills test on newborn anthropometry in the simulation laboratory, where they were filmed for further performance assessment using three scales: Training Level Rating Scale (TL), Global Rating Scale (EG); and checklist (CL). Nine evaluators were selected divided into three profiles: three pediatricians with no experience in teaching skills, three pediatricians with experience in teaching skills, and three skill teachers who were not pediatricians. One teacher from each profile used only one of the three proposed scales, so that the three scales were used in each of the three profiles. The total score for each of the scales was calculated, with a possible total of 9 points. **Results:** when comparing the different scales, regardless of the teacher profile, it was observed only that the isolated grade of the CL was able to discriminate the third and sixth year students. Multivariate analysis showed that TL, isolated score ($p = 0.346$) or total score ($p = 0.407$), was the only scale that was not influenced by the teacher profile. The reliability analysis performed according isolated scales score showed that only the TL has good reliability. The addition of anthropometric measurements increased the reliability of scales and teachers. **Conclusion:** there is a difference in reliability between the scales. TL plus anthropometric measurements was the most reliable. The teacher profile directly affected the reliability of the scales, but the expert teachers showed acceptable reliability when anthropometric measurements were added to the scales. Only TL was able to correctly discriminate the two groups and was not influenced by teacher profile.

Keywords: Educational Measurement. Clinical Competence. Simulation. Simulation Training.

LISTA DE ILUSTRAÇÕES

Figura 1 -	O modelo de aprendizagem experiencial de Kolb	19
Figura 2 -	Pirâmide de Miller	20
Figura 3-	Exemplo de lista de verificação	22
Figura 4-	Exemplo de escala de avaliação global	21
Figura 5-	Exemplo de escala de classificação por nível de treinamento	23
Figura 6 -	Escala de nível de treinamento do projeto	29
Figura 7 -	Escala de classificação global do projeto	29
Figura 8 -	Lista de verificação ou <i>checklist</i> do projeto	29
Figura 9 -	Fluxograma da pesquisa	30
Quadro 1 -	Caso Clínico do projeto - Avaliação de coleta de dados antropométricos do neonato	27
Quadro 2 -	Instruções enviadas aos avaliadores para a realização do estudo	30

LISTA DE TABELAS

Tabela 1 -	Distribuição de frequência de sexo de acordo com o ano do curso	33
Tabela 2 -	Nota isolada de cada escala, nota das medidas antropométricas e nota total por escala por ano do curso	34
Tabela 3 -	Comparação do percentual de acertos e erros das medidas antropométricas por ano do curso	34
Tabela 4 -	Notas médias de acordo por perfil de avaliador considerando-se as notas das escalas isoladas e nota total por ano	35
Tabela 5 -	Comparação entre a nota isolada e total da escala global de acordo com o perfil do avaliador por ano do curso	35
Tabela 6 -	Comparação entre a nota isolada e total do <i>checklist</i> de acordo com o perfil do avaliador por ano do curso	36
Tabela 7 -	Comparação entre as médias da nota da escala de nível de treinamento e as médias da nota total da escala de nível de treinamento de acordo com o perfil do avaliador, por ano do curso.....	37
Tabela 8 -	Comparação entre as notas do especialista experiente de acordo com as diferentes escalas isoladas e notas totais por ano do curso	37
Tabela 9 -	Comparação entre as notas do especialista sem experiência de acordo com as diferentes escalas isoladas e notas totais por ano do curso	38
Tabela 10 -	Comparação entre as notas do não especialista experiente de acordo com as diferentes escalas isoladas e notas totais por ano do curso	38
Tabela 11 -	Fatores que influenciaram na nota geral isolada e total e na nota isolada e total de cada escala	39
Tabela 12 -	Análise de confiabilidade (alfa de Cronbach) de acordo com escala isolada e nota total	40
Tabela 13 -	Análise de confiabilidade (alfa de Cronbach) de acordo com perfil do avaliador considerando a escala isolada e a nota total	40

LISTA DE ABREVIATURAS, SÍMBOLOS E SIGLAS

BH	Belo Horizonte
CL	Lista de Verificação ou <i>Checklist</i>
CONEP	Conselho Nacional de Ética em Pesquisa
CPE	<i>Clinical Performance Examination</i> ou exame do desempenho clínico
EG	Escala de Classificação Global
EN	Especialista sem experiência
EE	Especialista experiente
NEE	Não especialista experiente
NT	Escala de Avaliação do Nível de Treinamento
OSCE	<i>Objective structured clinical examination</i> ou exame clínico objetivo estruturado
TCLE	Termo de Consentimento Livre e Esclarecido
UNIFENAS	Universidade José do Rosário Vellano

SUMÁRIO

1	INTRODUÇÃO	13
1.1	Características da avaliação.....	16
1.2	Teorias de aprendizagem e simulação.....	18
1.3	Avaliação de habilidades e competências.....	20
1.4	Instrumentos de avaliação.....	22
2	JUSTIFICATIVA	24
3	OBJETIVOS.....	25
4	MATERIAIS E MÉTODOS.....	26
4.1	Desenho do estudo.....	26
4.2	População alvo e amostra.....	26
4.3	Métodos.....	27
4.4	Análise estatística.....	31
4.5	Aspectos éticos.....	32
5	RESULTADOS.....	33
5.1	Comparação entre as escalas.....	33
5.2	Comparação entre os perfis de avaliadores	34
5.3	Comparação entre os perfis do avaliador de acordo com escala utilizada	35
5.4	Comparação entre as escalas de acordo com o perfil do avaliador	37
5.5	Análise multivariada.....	39
5.6	Análise de confiabilidade.....	39
6	DISCUSSÃO.....	41
7	CONCLUSÃO.....	45
8	CONSIDERAÇÕES FINAIS.....	46
	REFERÊNCIAS.....	47
	APÊNDICE.....	50
	ANEXO.....	53

1 INTRODUÇÃO

Os avanços e descobertas tecnológicas na área de saúde acontecem de forma dinâmica e com proporção vultosa na atualidade, sendo necessário que a formação médica também evolua, visando melhorar os cuidados na saúde da população. Apesar desta explosão de avanços tecnológicos, as habilidades clínicas de entrevista médica, exame físico e aconselhamento continuam sendo vitais para o sucesso do tratamento dos pacientes (HOLMBOE, 2004).

A proficiência em habilidades clínicas pelos estudantes de medicina requer tempo e prática. Os programas de ensino médico enfrentam limitações para cumprimento destes preceitos por vários motivos, como o crescente número de alunos e a falta de pacientes “apropriados”, como pacientes que apresentem os sinais e sintomas que preencham o conteúdo programático no momento do processo ensino-aprendizagem (HOLMBOE, 2004). A incorporação do treinamento de habilidades clínicas, através da simulação em ambientes controlados, nos currículos das escolas médicas foi inovação importante para educação, pois permite que os estudantes se capacitem de forma ética em ambiente seguro e tolerante a falhas, sem oferecer riscos aos pacientes (BUGAJ; NIKENDEI, 2016).

O uso da simulação no ensino em saúde não é recente. O primeiro registro sobre o ensino de habilidades por simulação ocorreu na França, no século XVIII, onde Madame Angélique du Coudray construiu uma manequim de pano, em tamanho real, para treinamento de médicos e parteiras com o intuito de diminuir os óbitos maternos e neonatais que ocorriam na época (MORAN, 2010).

A era moderna da simulação médica tem suas origens na segunda metade do século XX. Três acontecimentos distintos podem ser identificados como seus precursores. O primeiro deles foi quando Asmund Laerdal, publicitário e fabricante de brinquedos norueguês, produziu um manequim desenvolvido por dois médicos anestesistas, o americano Dr. Peter Safar e o norueguês Dr. Bjorn Lind, para o treinamento de habilidades em ressuscitação cardiorrespiratória. Até os dias de hoje, esses manequins desenvolvidos por equipes de anesthesiologistas têm sido fundamentais para o entendimento e aprendizagem baseados na simulação em diversas especialidades médicas e em outras profissões relacionadas à saúde. O segundo marco importante ocorreu entre o final da década de 1960 e 1980, quando foram desenvolvidos simuladores sofisticados e de alta fidedignidade para reproduzir aspectos, sons

e movimentos semelhantes aos de pacientes humanos. Já o terceiro movimento se iniciou no final do século XX com o reconhecimento mundial da necessidade de aprimorar o ensino das habilidades clínicas em graduandos e pós-graduandos, ocasionando a reforma curricular das escolas médicas. As faculdades começaram a construir instalações para o treinamento dos alunos simulando os cenários reais (BRADLEY, 2006; OKUDA *et al.*, 2009; KHAN; PATTISON; SHERWOOD, 2011). Antes de incluir a simulação médica nos programas de graduação, a aquisição de habilidades era limitada ao encontro de pacientes “apropriados” para o treinamento e aprendizagem dos alunos, dificultando assim a uniformização do ensino (BUGAJ; NIKENDEI, 2016).

O ensino de habilidades e competências clínicas através da simulação tem como objetivo facilitar o aprendizado por meio da imersão, reflexão, feedback e prática do aluno em cenário controlado, isento dos riscos inerentes a vida real (DATTA; UPADHYAY; JAIDEEP, 2012). Além disso, a psicologia cognitiva afirma que a retenção de informações e sua aplicação são melhores quando ensinados e treinados em ambientes simulados, do que quando ensinados em sala de aula (KHAN; PATTISON; SHERWOOD, 2011). Ziv *et al.* (2003) ressaltam que o uso de treinamento baseado em simulação na educação médica, similar ao seu uso nos campos da aviação ou das forças armadas, representa, por assim dizer, um imperativo ético e, portanto, a aquisição de habilidades básicas de atendimento antes da prática com o “paciente real”.

Estudantes de medicina, assim como os de outros cursos voltados às profissões da saúde, devem não só adquirir conjunto de conhecimentos fundamentais com base científica, mas também desenvolver competência clínica, que engloba raciocínio clínico, habilidades práticas, integridade, compaixão e respeito ao próximo (COOKE *et al.*, 2006; TRONCON, 1996).

Epstein (2007) definiu competência em medicina como “o habitual e sagaz uso da comunicação, conhecimento, aptidões técnicas, raciocínio clínico, emoções, valores e reflexão na prática diária para o benefício dos indivíduos e comunidades que são atendidas”. Do ponto de vista educacional, pode-se definir habilidade como sendo a capacidade que uma pessoa adquire para realizar uma tarefa específica com destreza e competência, como o conjunto de habilidades de natureza intelectual ou física, relevantes para um determinado fim, passível de ser ensinada, aprendida e avaliada (BLOOM, 1956). As habilidades podem ser descritas como pertencentes a três grandes domínios: o cognitivo (que se refere às habilidades de natureza

puramente intelectual, como a aquisição de conhecimento, compreensão, análise, síntese, entre outras), o psicomotor (são as habilidades que demandam os órgãos do sentido e efetuação neuromuscular para o desempenho de tarefas específicas) e o afetivo (que compreende as atitudes, as crenças, os valores, os juízos acerca das coisas). Na medicina, as competências envolvidas constituem, quase sempre, habilidades complexas e que compreendem inter-relações variadas entre esses três domínios (TRONCON, 1996).

A aprendizagem baseada em simulação é uma metodologia ativa de ensino, que oferece aos estudantes a possibilidade de treinar habilidades e competências clínicas de forma segura e controlada. Esta técnica de ensino tem conquistado grande popularidade a nível global e vem sendo considerada um componente essencial no currículo da formação médica (VANTINI; BENINI, 2008; KER; BRADLEY, 2010). A definição redutora, e mais conhecida, de simulação no ensino médico, consiste na utilização de técnicas para recriar habilidades clínicas que não se limitam ao desempenho técnico e psicomotor, incluindo também o domínio cognitivo; este conceito engloba qualquer ambiente simulado no qual os alunos treinam competências médicas, interagindo e comunicando com os seus pares, equipamentos ou pacientes reais ou simulados (ALINIER, 2007). Portanto, a aprendizagem em simulação não é realizada exclusivamente em laboratórios de simulação e nem se restringe à utilização de tecnologias, podendo necessitar apenas de uma folha de papel (KER; BRADLEY, 2010). É importante ressaltar que quanto mais o estudante acreditar que a situação e a experiência são reais, melhor será a sua utilidade educacional (MURPHY *et al.*, 2007).

No Brasil, o primeiro laboratório de simulação foi criado em 1998 no curso de medicina da Faculdade de Londrina, Paraná, quando de sua reestruturação curricular (DIAS; SCALABRINI NETO, 2012). Desde então, os laboratórios de simulação têm sido adotados em maior ou menor grau em vários cursos no Brasil. O primeiro laboratório de habilidades clínicas da Universidade José do Rosário Vellano (UNIFENAS), Minas Gerais, campus de Belo Horizonte, foi criado em 2003 para atender os cursos de medicina e enfermagem. Em 2012, foi inaugurado um novo laboratório de habilidades clínicas e simulação realística do Curso de Medicina, na unidade Itapoã (MATOS, 2017). Já o campus da UNIFENAS em Alfenas iniciou o uso da simulação para treinamento de seus alunos no ano de 2011, com a inauguração do laboratório de habilidades clínicas (comunicação pessoal).

1.1 Características da avaliação

A avaliação do estudante deve ser parte integrante de todo processo educacional, do início ao fim, pois permite a obtenção de informações sobre os domínios trabalhados para cumprir diferentes finalidades do currículo programático e pode ser considerada parte motivadora no processo ensino/aprendizagem, contribuindo assim na formação profissional (DIAS; SCALABRINI NETO, 2012; TRONCON, 1996; TRONCON *et al.*, 2016).

Uma boa avaliação deve cumprir diferentes funções, sendo as mais importantes: embasar a tomada de decisões (função somativa), estimular o aprendizado (função formativa) e prover elementos para o controle da qualidade do currículo (função informativa) (MUSICK, 2006; TRONCON, 1996). A função somativa é a mais tradicional e implica em utilizar os resultados dos procedimentos avaliativos para a tomada de decisões sobre o destino do estudante em relação à sua progressão ao longo do curso. A mais nobre das funções é a formativa, que utiliza a avaliação para orientar a correção de deficiências e reforçar o aprendizado do aluno, bem como para estimular as capacidades de reflexão e de autoavaliação. Já a função informativa é cumprida quando os seus resultados fornecem informações importantes sobre a qualidade dos processos educacionais, permitindo que as escolas atuem no aperfeiçoamento do currículo (TRONCON *et al.*, 2016).

A história do desenvolvimento das avaliações é acompanhada dos conceitos de validade, confiabilidade (fidedignidade), viabilidade, aceitabilidade, equivalência, impacto educacional e efeito catalítico, pois estes constituem atributos necessários para que os métodos avaliativos sejam considerados adequados ao uso (PACICO; HUTZ, 2015; ZANON; HAUCK FILHO, 2015; NORCINI *et al.*, 2011; TRONCON, 1996).

Durante décadas vários autores utilizavam o termo validade com diferentes significados. Apenas em 1921, nos Estados Unidos, a *National Association of Directors of Educational* apresentou a definição clássica de que validade se refere ao grau em que um teste mede aquilo que se propõe a medir (BUCKINGHAM, 1921; MARKUS; BORSBOOM, 2013). Em 1954, a *American Psychological Association*, a *American Educational Research Association* e o *National Council on Measurement in Education* publicaram a primeira versão dos padrões norte-americanos para testes e classificaram a validade em três tipos: validade de conteúdo (refere-se ao quanto o teste pode ser uma amostra representativa do conjunto de itens ou

questões do construto), validade de critério (refere-se ao quanto o teste pode prever o desempenho do sujeito em tarefas específicas) e validade de construto ou conceito teórico (referem-se ao quanto os construtos hipotéticos podem explicar os escores de um teste). As duas primeiras definições de validade baseiam-se na teoria para o teste e terceira avalia se os itens (ou questões) podem ou não confirmar determinado constructo. No decorrer dos anos, alguns autores propuseram um movimento que buscava a unificação das diferentes categorias da validade resultando em diversas publicações neste campo. Atualmente, muitos pesquisadores concordam que validade e validação são tópicos fundamentais quando se fala em instrumentos de mensuração (PACICO; HUTZ, 2015). Validade refere-se à qualidade das inferências, conclusões e decisões tomadas com base nos escores obtidos pelo uso de um instrumento. Validação é o processo em que se busca dar evidências de validade que apoiem a adequação, o significado e a utilidade das decisões tomadas com base nas inferências feitas a partir dos escores obtidos do teste (ZUMBO; CHAN, 2014).

A confiabilidade (termo que muitas vezes pode ser substituído por fidedignidade, consistência interna ou precisão) de um teste refere-se à estabilidade com que os escores dos estudantes conservam-se em aplicações repetidas de um mesmo teste ou em formas equivalentes de testes distintos, ou seja, é relacionada à objetividade e acurácia, que indicam a reprodutibilidade dos resultados de um teste. É uma propriedade essencial para a validade de um teste, de modo que um teste com baixa confiabilidade não será válido, pois não mede apropriadamente o constructo de interesse. Apesar de ser essencial, a confiabilidade não é condição suficiente para o uso de um teste, pois este deve apresentar evidências de validade conjuntamente; sendo assim se não houver evidências de confiabilidade também não haverá de validade. A capacidade de um teste conseguir produzir escores capazes de distinguir os estudantes com diferentes graus de habilidades, personalidades, entre outros e ao mesmo tempo garantir condições necessárias para uma adequada replicação dos resultados faz com que este teste seja considerado de alta confiabilidade (ZANON; HAUCK FILHO, 2015; TRONCON, 1996).

O conceito de viabilidade em avaliação faz referência à disponibilidade e a boa utilização dos recursos materiais, humanos e de tempo existentes, necessários para que o emprego de um dado método seja bem sucedido. A aceitabilidade é o reconhecimento da avaliação por todos os partícipes como sendo pertinente, adequada e justa. Equivalência de uma avaliação é a alta semelhança dos resultados de exames, mesmo quando o procedimento é aplicado para grupos

diversos de estudantes, simultaneamente ou em um curto intervalo de tempo correspondente (NORCINI *et al.*, 2011; TRONCON *et al.*, 2016).

O impacto educacional é relacionado ao potencial de um procedimento avaliativo, pela sua própria existência ou pelo modo como é empregado, influenciando positivamente ou negativamente os processos de ensino e aprendizagem. Já o efeito catalítico é a capacidade do método avaliativo produzir transformações positivas nas instituições, como a introdução de sistemas inovadores que exijam novos recursos e desenvolvimento docente (TRONCON *et al.*, 2016).

Outro atributo essencial para que o método avaliativo em simulação seja considerado adequado é a fidelidade (é o quanto a simulação se aproxima da realidade), pois quanto maior o grau de realismo utilizado na avaliação maior será sua validade e confiabilidade (MCGAGHIE *et al.*, 2010).

Todo processo avaliativo, se realizado de forma planejada e sistemática, deve ser capaz de fornecer inúmeros aspectos da aprendizagem do estudante, além de permitir um olhar para o ensino e para o professor na sua prática pedagógica, implicando em reflexões sobre o processo ensino/aprendizagem (PINTO; TRONCON, 2014; TRONCON, 1996).

1.2 Teorias de aprendizagem e simulação

Algumas teorias de aprendizagem embasam o ensino de habilidades em ambiente simulado e também são a base para a avaliação de habilidades e competências, como o modelo de Dreyfus, o ciclo de Kolb e a pirâmide de Miller. O modelo de Dreyfus de aquisição de habilidades foi desenvolvido pelos irmãos Stuarty e Hubert Dreyfus na década de 1980, para a Força Aérea Americana, como referência para treinamento de pilotos de aviões de caça. O modelo define cinco níveis de competência pelos quais o estudante passa durante sua formação (DREYFUS; DREYFUS, 1980):

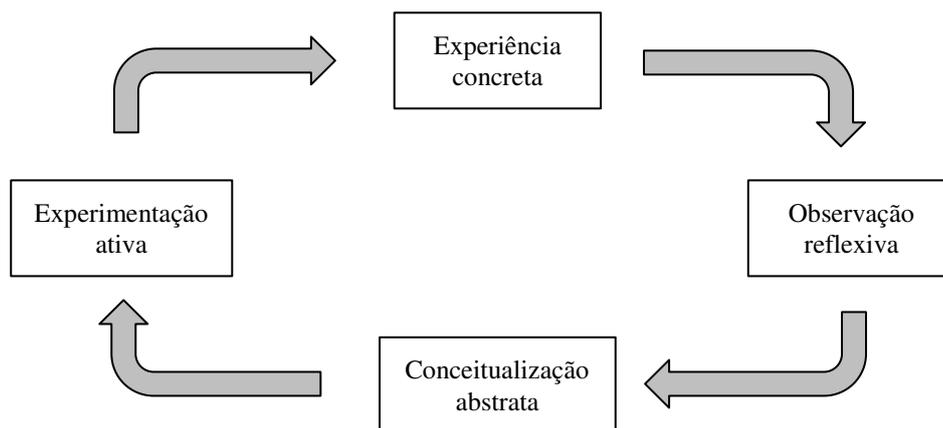
- novato - segue as regras independentemente o contexto;
- iniciante avançado - segue as regras adaptando-se ao contexto;
- competente - questiona e percebe consequências em longo prazo;

- profissional - quase não utiliza as regras, usa a intuição; e
- especialista - toma decisões e resolve problemas sem esforço.

O ciclo de Kolb (Figura 1) foi criado em 1984 pelo teórico de educação David A. Kolb. É um modelo de aprendizagem experiencial que apresenta a aprendizagem como um ciclo contínuo de quatro estágios (LEUNG; MARTIN; BATTY, 2009):

- experiência concreta - agir;
- observação reflexiva - refletir;
- conceitualização abstrata - conceituar; e
- experimentação ativa - aplicar.

Figura 1 - O modelo de aprendizagem experiencial de Kolb



Fonte: LEUNG; MARTIN; BATTY, 2009.

Em 1990, o médico norte-americano George Miller propôs um modelo representado por uma pirâmide, no qual os alunos seriam avaliados em quatro níveis distintos. A pirâmide é muito útil para demonstrar que o ensino médico não pode ser restrito à teoria. Para a formação profissional é necessário que o aluno saia do conhecimento meramente factual (saber), mas que também saiba aplicar esse conhecimento (saber como fazer) e executá-lo de maneira prática em laboratórios e/ou ambientes simulados (mostrar como faz) e, finalmente, aplicá-lo na vida real (fazer). Os dois primeiros níveis da pirâmide estão relacionados apenas com o domínio cognitivo, enquanto os dois últimos exigem a aplicação de habilidades psicomotoras (MILLER, 1990).

Figura 2 - Pirâmide de Miller



Fonte: PINTO; TRONCON, 2014.

1.3 Avaliação de habilidades e competências

A avaliação do aluno à “beira leito” é tão antiga quanto o ensino da medicina, mas somente nas últimas décadas vem ganhando atenção adequada para a necessidade de padronização. No início do século XX, Flexner já afirmava que a melhor maneira para avaliar um aluno de medicina seria através de casos concretos, nos quais o aluno pudesse realizar todas as habilidades que envolvem a prática diária da profissão: entrevista e coleta de dados, exame físico e conduta (COOKE *et al.*, 2006).

Até a década de 1970, a avaliação de habilidades e de competências clínicas era realizada por métodos tradicionais, como provas de múltipla escolha, pequenos ensaios e provas com questões abertas, que avaliavam o conhecimento teórico. O conhecimento clínico era avaliado através de observações assistemáticas e exames de desempenho do tipo casos longos e casos curtos (HARDEN *et al.*, 1975; TRONCON, 1996).

Nos exames do tipo casos longos e casos curtos o estudante permanecia com o paciente cerca de 50 minutos nos casos longos e 20 minutos nos casos curtos, para realizar a entrevista, exame físico e então apresentar ao examinador suas conclusões e condutas. A confiabilidade deste método avaliativo é questionável. Por outro lado, estas modalidades avaliativas

apresentavam diversas limitações. A baixa validade, pois a avaliação dependia exclusivamente dos pacientes hospitalizados naquele momento, o que dificultava medir o conhecimento sobre doenças e sintomas estabelecidos no currículo. Realizar a consulta completa em um paciente demanda tempo, levando a restrição da amostragem em relação ao conteúdo programático (um estudante que tenha baixo desempenho numa matéria específica pode ter excelente desempenho em outras). A confiabilidade destes testes é muito baixa, pela falta de padronização e objetividade, tornando difícil uma reprodutibilidade confiável. Existe ainda nestes exames a possibilidade do efeito halo, em que o professor pode julgar, analisar e tirar conclusões de uma estudante a partir de uma única característica, formulando um estereótipo global do indivíduo com único fator (como aparência, vestimenta, postura, fala, entre outros), interferindo no resultado (SOUZA, 2012; KHAN *et al.*, 2013). Essas modalidades de avaliação carecem de maior objetividade e padronização, pois os resultados obtidos são fortemente influenciados pelo tipo de paciente disponível para determinado exame e toda a avaliação depende muito do avaliador (HARDEN *et al.*, 1975; TRONCON, 1996; TRONCON *et al.*, 2016).

Procurando diminuir as deficiências inerentes ao uso das modalidades tradicionais para a avaliação de habilidades e competências clínicas foram desenvolvidos novos métodos, sendo importante destacar dois que são amplamente utilizados atualmente, o *objective structured clinical examination* (OSCE) ou exame clínico objetivo estruturado e o *Clinical Performance Examination* (CPE) ou exame do desempenho clínico (TRONCON, 1996; TRONCON *et al.*, 2016; HARDEN, 1988; KHAN *et al.*, 2013).

O OSCE foi descrito pela primeira vez por Harden (1975) e tinha como objetivo melhorar a validade e confiabilidade dos métodos de avaliação de habilidades e competências clínicas em ambientes simulados. Trata-se de abordagem para avaliação em habilidades ou competências clínicas em que os componentes são analisados de forma planejada e estruturada em concordância com o objetivo principal ao que se propôs o exame (HARDEN, 1988; KHAN *et al.*, 2013). É uma prova flexível, onde os alunos são avaliados em 6 a 25 estações com duração de 4 a 20 minutos cada, sendo as competências clínicas divididas e distribuídas pelas estações. Os alunos rodam por elas e cada uma possuiu um professor fixo, que avaliará todos os alunos que passarem pela estação com base em um instrumento objetivo. Ao final da prova, é realizado feedback, o que faz com que o OSCE tenha também papel formativo (KHAN *et al.*, 2013; SMEE, 2003). Atualmente, ele é considerado o padrão ouro para

avaliação de competências clínicas, tendo como principais vantagens a flexibilidade e alta validade e confiabilidade. Suas limitações se concentram no alto custo e complexidade, pois envolvendo grande número de professores, materiais, área física adequada e necessidade de planejamento prévio (HARDEN, 1988; KHAN *et al.*, 2013).

O CPE foi desenvolvido em escolas americanas como alternativa e solução de possíveis inconveniências do OSCE, como a artificialidade e a fragmentação na abordagem dos casos. Em geral, o estudante encontra um paciente simulado que deve ser entrevistado, submetido ao exame físico e em seguida informado e orientado sobre as medidas diagnósticas e terapêuticas necessárias para a solução do seu problema. A principal vantagem do CPE é a abordagem integral do caso, tornado a avaliação mais próxima da realidade profissional. Já a maior desvantagem desta modalidade é de ordem operacional, pois o CPE deve conter de 15 a 18 estações (cada estação dura aproximadamente 40 minutos, portanto o tempo do exame de um aluno dura no mínimo 10 horas), fazendo com que a avaliação completa possa durar dias (TRONCON, 1996; SOUZA, 2012).

Independentemente do formato da avaliação a ser utilizada, a adoção de instrumentos avaliativos específicos, como *checklists*, aumentam a validade e a confiabilidade da avaliação prática (TRONCON *et al.*, 2016).

1.4 Instrumentos de avaliação

Poucos estudos compararam instrumentos diferentes na avaliação de habilidades. Pugh *et al.* (2014) realizaram estudo sobre avaliação de habilidades que comparou três instrumentos: a lista de verificação ou *checklist* (CL), a escala de classificação global (EG) e escala de avaliação do nível de treinamento (NT). OCL (Figura 3) é desenvolvido para que o avaliador constate se os estudantes realizaram os itens considerados indispensáveis para determinada habilidade ou competência clínica. Geralmente utiliza-se uma escala dicotômica tipo sim ou não (PUGH *et al.*, 2014; SCHUWIRTH; ASH, 2013).

Figura 3 - Exemplo de lista de verificação ou *checklist*

Teste do Coraçãozinho	SIM	NÃO
1. Foi realizado após 24 horas do nascimento?		
2. O sensor do oxímetro foi colocado na mão direita e no tornozelo direito?		

3. Foi realizada a leitura da saturação após 3 a 5 minutos da instalação do sensor?		
---	--	--

Fonte: elaborado pela autora.

Já a EG (Figura4) permite que o examinador determine o desempenho geral dos alunos em cada habilidade ou competência clínica, como sendo suficiente ou insuficiente e ainda subdividindo estes resultados para determinação mais específica de cada desempenho. Os alunos avaliados com desempenho suficiente podem variar entre limite satisfatório, bom e excelente, em contrapartida os de desempenho insuficiente podem variar entre limite insatisfatório, ruim e péssimo. ANT (Figura5) visa classificar o desempenho do candidato correspondente ao desempenho esperado para o aluno ou residente de medicina de acordo com o ano do curso (PUGH *et al.*, 2014).

Figura 4 - Exemplo de escala de avaliação global

Insuficiente			Suficiente		
Péssimo	Ruim	Limítrofe	Limítrofe	Bom	Excelente

Fonte: elaborado pela autora.

Figura 5 - Exemplo de escala de classificação por nível de treinamento: em sua opinião o desempenho do candidato é equivalente a qual ano do curso de medicina?

Estudante de Medicina	Primeiro ano	Segundo ano	Terceiro ano	Quarto ano	Quinto ano	Sexto ano

Fonte: elaborado pela autora.

O CL foi projetado e incorporado nas avaliações de habilidades e competências clínicas com o objetivo de aumentar a objetividade e confiabilidade. No entanto, alguns autores criticam a sua validade, pois relatam tendência em induzir o avaliador ao trivial. Alguns estudos indicam que a EG é mais confiável e que a subjetividade dela pode ser favorável à avaliação (VAN DER VLEUTEN *et al.*, 1991; CUSHING, 2002).O trabalho de Pugh *et al.* (2014) indicou que a NT foi o instrumento mais adequado para avaliar os residentes de clínica médica em um OSCE. Na revisão da literatura, não foram identificados estudos comparativos que indiquem qual o melhor instrumento na avaliação de habilidades e competências em estudantes de medicina.

2 JUSTIFICATIVA

Vários cursos de medicina no Brasil e no mundo têm adotado a aprendizagem baseada em simulação nos últimos anos. Ela representa grande avanço da educação médica, possibilitando o desenvolvimento de habilidades e competências antes de iniciar o atendimento aos pacientes, o que aumenta a segurança do paciente e do aluno. Apesar de todos esses avanços, a avaliação de habilidades em ambientes simulados ainda enfrenta desafios.

A utilização de instrumentos, como *checklists*, teoricamente aumenta a confiabilidade e a validade da avaliação de habilidades, mas poucos estudos compararam diferentes instrumentos em relação a sua confiabilidade e validade. Diferentes instrumentos podem ter aplicações diferentes. Determinado instrumento pode ser mais adequado para o início ou para final do curso, ou para avaliador novíço ou experiente etc. Trata-se de tema pouco estudado até o momento e pode ter impacto na avaliação e formação de estudantes de medicina.

3 OBJETIVOS

O objetivo geral desse estudo foi descrever como diferentes escalas de avaliação de habilidades clínicas em laboratório de simulação e diferentes perfis de avaliadores interferem na confiabilidade da avaliação dessas habilidades.

Os objetivos específicos foram:

- avaliar se existe diferença na confiabilidade das escalas utilizadas para avaliação de habilidades clínicas em laboratório de simulação;
- identificar qual a melhor escala para avaliar as habilidades clínicas do aluno de medicina em laboratórios de simulação;
- avaliar se o perfil do avaliador interfere na confiabilidade da escala de avaliação;
- identificar se as escalas são capazes de discriminar os alunos entre os anos do curso.

4 MATERIAIS E MÉTODOS

4.1 Desenho do estudo

Trata-se de estudo experimental para comparação de três instrumentos de avaliação de provas de habilidades clínicas em ambiente simulado e três tipos diferentes de avaliadores (avaliador não especialista experiente, especialista sem experiência e especialista experiente).

4.2 População e amostra

A população desse estudo foi constituída por estudantes do terceiro e sexto anos do curso medicina da UNIFENAS-BH. O curso possui aprendizado em ambiente simulado do primeiro ao oitavo períodos. No terceiro período, eles aprendem a realizar a antropometria do recém-nascido em ambiente simulado no bloco Nascimento, Crescimento e Desenvolvimento, utilizando manequim. Do quinto ao oitavo períodos eles têm contato com pacientes pediátricos durante atendimento em Unidades Básicas de Saúde do sistema público de saúde. No nono período eles realizam o internato em Clínica Médica e Cirurgia e no décimo período o internato em Pediatria e Ginecologia e Obstetrícia, onde tem contato com neonatologia. Com base nisso, optou-se por incluir no estudo alunos do quinto e sexto períodos (terceiro ano) e do 11º e 12º períodos (sexto ano), com objetivo de aumentar a chance de discriminação entre os dois grupos, considerando que o terceiro ano teve apenas treinamento básico em ambiente simulado e que o sexto ano já realizou o internato de pediatria e neonatologia.

Foram incluídos no estudo os alunos:

- regularmente matriculados no Curso de Medicina da UNIFENAS-BH;
- aprovados no módulo Nascimento, Crescimento e Desenvolvimento, no Internato de Pediatria e no Internato de Saúde da Mulher até o 2º semestre de 2018;
- que estavam cursando o terceiro ou sexto ano no primeiro semestre de 2019;
- que assinaram o Termo de Consentimento Livre e Esclarecido (TCLE - Apêndice).

Foram excluídos do estudo os alunos:

- com situação acadêmica irregular (cursando diferentes períodos); ou
- que foram reprovados no módulo Nascimento, Crescimento e Desenvolvimento, no Internato de Pediatria ou no Internato de Saúde da Mulher; ou
- que não assinaramo TCLE.

Foi utilizada uma amostra não probabilística de conveniência. Inicialmente, os alunos potenciais foram convidados por meio de e-mail, WhatsApp e outros canais de comunicação institucionais, como Diretório Acadêmico, Atlética e Ligas. Depois de aproximadamente uma semana, foi realizado convite presencial nas salas de aulas para esclarecimento das dúvidas sobre o estudo e agendada a prova no laboratório de habilidades do Campus Itapõa.

4.3 Métodos

Todos os voluntários assinaram o TCLE antes de sua inclusão no estudo. Eles realizaram uma prova de habilidades sobre antropometria do recém-nascido (esta habilidade foi selecionada pela possibilidade de ser avaliada adequadamente através de um vídeo) em manequins no laboratório de simulação. Eles foram avaliados quanto à habilidade de realizar parte do exame antropométrico que constitui a consulta de puericultura do recém-nascido. Os alunos receberam um caso clínico, solicitando que realizassem a pesagem, a medida do comprimento e do perímetro cefálico (Quadro 1).

Quadro 1 – Avaliação de coleta de dados antropométricos do neonato

Você está passando pelo estágio de pediatria, no ambulatório de puericultura, chega um lactente com 15 dias de vida para uma consulta de rotina. A mãe relata que o bebê está bem e não apresenta queixas no momento. **Considere que a balança já está tarada. Apesar das luvas não serem necessárias no exame habitual do recém-nascido, calce as luvas para preservação do manequim.**

Realize a avaliação antropométrica da criança (peso, comprimento e perímetro cefálico). **Registre o peso, comprimento e perímetro cefálico na folha própria e coloque na caixa ao lado da maca.**

Fonte: elaborado pela autora.

Em outubro de 2018, foi realizado pré-teste com cinco alunos que cursavam o terceiro ano no momento (estes estariam no quarto ano durante a execução do projeto, evitando a contaminação da amostra) e com dois avaliadores (que não participaram da pesquisa) com intuito de orientar e apontar possíveis falhas metodológicas.

Os voluntários do terceiro ano realizaram a prova em uma única data e os do sexto ano em outra data, com o intuito de evitar a comunicação do conteúdo da prova entre os participantes de períodos diferentes e contaminar o resultado da avaliação. Os alunos dos dois grupos têm atividades curriculares em locais diferentes, o que dificulta o contato entre eles, reduzindo assim o risco de contaminação. Todos os voluntários foram confinados no laboratório de simulação e realizaram a prova um a um, de acordo com a ordem de chegada. Após a realização desta, não tiveram contato com os próximos participantes, de modo a evitar contaminação intragrupo.

A prova foi filmada para posterior avaliação do desempenho dos voluntários que anotaram em formulário próprio os valores numéricos dos dados antropométricos. A realização da prova foi supervisionada por um dos pesquisadores. Os vídeos foram editados individualmente de modo a conter exclusivamente os procedimentos relacionados à avaliação e foram misturados entre os dois grupos em blocos de cinco. Cada avaliação individual tinha em média três minutos e a duração total do vídeo contendo a performance dos 48 alunos que participaram foi de 2 horas e 26 minutos. O uso do vídeo permitiu aos pesquisadores a utilização de número maior de avaliadores).

Foram utilizados três instrumentos para avaliação do desempenho dos alunos: escala de avaliação do nível de treinamento (NT), escala de classificação global (EG) e lista de verificação ou *checklist* (CL). A NT foi formulada com o intuito de comparar o desempenho dos alunos iniciantes (terceiro ano) com outros alunos e especialistas (Figura 6). A EG foi dividida em desempenho insuficiente e suficiente. Cada uma dessas divisões tinha três subdivisões para determinação mais específica de cada desempenho (Figura 7). Já o CL foi constituído por seis microhabilidades, com respostas dicotômicas: realizou ou não realizou (Figura 8). Tanto a EG quanto a NT são menos objetivas e, a princípio, dependem do julgamento e experiência do avaliador. Todas as escalas continham seis itens, cada item correspondia a um ponto em ordem crescente, totalizando a pontuação máxima de seis em cada uma das escalas. A medida objetiva de cada um dos três dados antropométricos foi

somada a nota de cada uma das escalas, considerando-se um ponto para o acerto de cada um deles. As medidas dos dados antropométricos foram consideradas corretas utilizando-se uma margem de erro de: ± 50 gramas para o peso, ± 1 cm para o comprimento e $\pm 0,5$ cm para o perímetro cefálico. Para fins de análise estatística, utilizou-se a nota isolada de cada escala, num máximo de seis pontos, e a nota total (nota isolada de escala somada a nota dos dados antropométricos), num total máximo de nove pontos.

Figura 6 - Escala de nível de treinamento

Não realizou	Aluno 3º ano	Aluno 4º ano	Aluno 5º ano	Aluno do 6º ano	Médico residente	Pediatra
Zero	1 ponto	2 pontos	3 pontos	4 pontos	5 pontos	6 pontos

Fonte: elaborado pela autora.

Figura 7 - Escala de classificação global

Não realizou	Insatisfatório			Satisfatório		
	Péssimo	Ruim	Limítrofe	Limítrofe	Bom	Excelente
Zero	1 ponto	2 pontos	3 pontos	4 pontos	5 pontos	6 pontos

Fonte: elaborado pela autora.

Figura 8 - Lista de verificação ou *checklist*

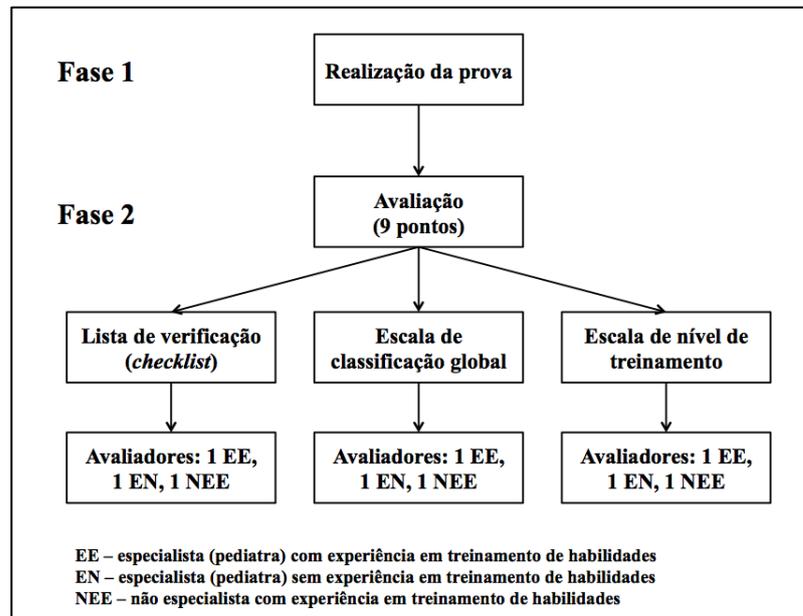
Pesagem do recém-nascido	Realizou	Não realizou
1) Despiu completamente o manequim?	1 ponto	Zero
2) Pesou corretamente?	1 ponto	Zero
Medida do comprimento e perímetro cefálico do recém-nascido		
3) Colocou o bebê na posição supina?	1 ponto	Zero
4) Posicionou a parte fixa da régua na extremidade da cabeça do bebê?	1 ponto	Zero
5) Estendeu corretamente os membros inferiores do bebê e apoiou os calcanhares na superfície móvel da régua?	1 ponto	Zero
6) Passou a fita métrica na altura da região occipital, imediatamente acima das orelhas e das sobrancelhas do bebê?	1 ponto	Zero

Fonte: elaborado pela autora.

A aplicação das escalas foi feita posteriormente, com base nas filmagens da prova. Os avaliadores desconheciam o grupo dos alunos participantes. A avaliação foi realizada por nove avaliadores divididos em três perfis: três pediatras sem experiência em ensino de habilidades (especialista sem experiência - EN), três pediatras com experiência em ensino de habilidades (especialista experiente - EE) e três professores de habilidades que não eram pediatras (não especialista experiente - NEE). Cada avaliador de cada perfil utilizou apenas uma das três escalas propostas, de forma que as três escalas foram utilizadas em cada um dos

três grupos de avaliadores. Cada aluno foi avaliado três vezes pela mesma escala, mas por avaliadores de perfis de diferentes, num total de nove avaliações por aluno (Figura 9).

Figura 9 - Fluxograma da pesquisa



Fonte: elaborado pela autora.

Os avaliadores receberam as instruções para realização da prova, conforme o Quadro 2, e um *pendrive* com os vídeos. Não houve nenhum treinamento sobre o uso das escalas.

Quadro 2 - Instruções enviadas aos avaliadores para a realização do estudo. (Continua)

1. Você está recebendo um *pendrive* com dois arquivos de vídeo. Cada arquivo tem a duração aproximada de 1:10 horas e os formulários de avaliação. **Comece pelo mestradoP1.mp4 e depois o mestradoP2.mp4.**
2. Cada vídeo corresponde a uma prova de habilidade de avaliação antropométrica do recém-nascido (medida de peso, comprimento e perímetro cefálico).
3. Os alunos são de diferentes períodos do curso de medicina da UNIFENAS-BH, a partir do 5º período (3º ano). Todos eles já tiveram aula sobre o tema e foram aprovados na respectiva disciplina ou módulo.
4. Você deve assistir ao vídeo de cada avaliação (aluno) sem pausar e/ou repetir.
5. Durante a avaliação ou ao final dela, você deve preencher o formulário correspondente ao aluno. Os formulários estão numerados na mesma ordem dos vídeos.
6. Caso necessário, pode pausar entre um aluno e o outro para preencher o formulário.
7. Caso tenha que interromper a avaliação no meio do vídeo, anote o tempo transcorrido para que você possa recomeçar do mesmo ponto.
8. Caso erre o preenchimento do questionário, faça um traço transversal na marcação errada, de forma que fique legível, escreva nulo ao lado e faça a marcação correta com a mesma caneta. **Por favor, não use corretivo.**
9. Considere que a mãe da criança estava presente no momento do exame, apesar de ela não aparecer em cena. Todos alunos estarão usando luva por uma questão de conservação no manequim (boneco) e não pela necessidade do exame. Desconsidere o uso das luvas em sua avaliação.
10. A **escala de avaliação global** permite que o examinador determine o desempenho geral dos alunos em cada habilidade ou competência clínica, como sendo satisfatório ou insatisfatório e ainda subdividindo estes resultados para determinação mais específica de cada desempenho. Os alunos avaliados com desempenho

satisfatório podem variar entre limítrofe, bom e excelente; em contrapartida os de desempenho insatisfatório podem variar entre limítrofe, ruim e péssimo”.; OU

Quadro 2 - Instruções enviadas aos avaliadores para a realização do estudo. (Conclusão)

10. A lista de verificação ou *checklist* é desenvolvida para que o avaliador constate se os estudantes realizaram os itens considerados indispensáveis para determinada habilidade ou competência clínica. Geralmente utiliza-se uma escala dicotômica tipo sim ou não. OU

10.A **escala de avaliação do nível de treinamento** visa classificar o desempenho do candidato correspondente ao desempenho esperado para o aluno de medicina de acordo com o ano do curso, tendo o pediatra com referência do desempenho

Fonte: elaborado pela autora.

4.4 Análise estatística

O banco de dados foi formatado no Microsoft® Excel® (Microsoft, USA) e a análise estatística realizada com o software IBM® SPSS® Statistics v. 19 (IBM, USA).

Na análise descritiva foram utilizadas as medidas de tendência central para variáveis contínuas, como idade e notas, e distribuição de frequência para variáveis categóricas, como sexo e ano do curso. Foi calculado o escore total para cada uma das escalas por meio soma das notas relativas as medidas antropométricas e a nota de cada escala, num total possível de 9 pontos.

A análise comparativa entre os grupos foi realizada com teste de t de *Student* para variáveis contínuas e χ^2 ou teste de Fischer para variáveis categóricas. A comparação do desempenho dos alunos entre os três instrumentos e pelos três tipos de avaliadores foi feita por meio do teste ANOVA e teste *post hoc* de Turkey HSD. A análise multivariada foi feita por regressão linear. Foi considerado como nível de significância estatística o valor de 0,05.

A análise de confiabilidade e consistência interna foi feita pelo teste alfa de Cronbach. Considerou-se os seguintes valores na interpretação da confiabilidade:

- ruim - $\leq 0,70$;
- aceitável - $> 0,7$ e $\leq 0,8$;
- boa - $> 0,8$ e $\leq 0,9$; e
- excelente - $> 0,90$.

4.5 Aspectos éticos

Este estudo está em concordância com a versão atual da Declaração de Helsinque e a Resolução 466/2012 do CONEP e suas atualizações e foi aprovado pelo Comitê de Ética em Pesquisa da UINFENAS (Parecer nº 3.085.011 - Anexo).

Todos os voluntários receberam esclarecimento sobre os objetivos do estudo e, se concordaram em participar, assinaram o TCLE antes de sua inclusão. Os alunos que não concordaram em participar do estudo não sofreram nenhuma penalidade.

Nenhum dos voluntários recebeu pagamento ou benefício pessoal por sua participação no estudo. As avaliações foram aplicadas em dias de aula habitual dos alunos, sua participação também não implicou em nenhuma despesa adicional. A participação do aluno neste estudo não implicou em nenhum risco físico, pois não foi realizado nenhum procedimento invasivo. Pode ter havido algum constrangimento relacionado ao desempenho individual na avaliação. Todo esforço foi feito para manutenção do sigilo da identidade dos voluntários e de suas informações. O TCLE foi arquivado fisicamente separado dos questionários. Todas as informações relativas aos voluntários foram identificadas apenas pelo número de registro e apenas a equipe de pesquisadores teve acesso às informações e aos bancos de dados.

5 RESULTADOS

Quarenta e oito alunos foram recrutados, sendo 28 (58,3%) do terceiro ano e 20 (41,7%) do sexto ano do curso de Medicina da UNIFENAS-BH. Nenhum aluno foi excluído do estudo. A prova foi realizada no dia 14/03/2019 com os alunos do terceiro ano e no dia 15/03/2019 com os do sexto ano. A duração de cada prova variou de 2 a 4 minutos. No terceiro ano houve mais alunos do sexo feminino (57,1%) e no sexto ano a distribuição entre os sexos foi igual (Tabela 1). A idade média foi de $23,6 \pm 2,6$ anos, sendo que a idade média dos alunos do terceiro ano foi menor que do sexto ano ($22,4 \pm 2,1$ versus $25,5 \pm 2,1$; $p < 0,001$).

Tabela 1 - Distribuição de frequência de sexo de acordo com o ano do curso

Sexo	Terceiro ano		Sexto ano		Total	
	n	%	n	%	n	%
Feminino	16	57,1	10	50,0	26	54,2
Masculino	12	42,9	10	50,0	22	45,8
Total	28	100,0	20	100,0	48	100,0

$p = 0,422$

Participaram do estudo nove avaliadores, sendo três pediatras sem experiência em ensino de habilidades (EN), um deles pediatra geral e os outros dois pediatras neonatologistas. Entre os três pediatras com experiência em ensino de habilidades (EE), um era pediatra geral, um pediatra infectologista e o outro pediatra nefrologista. Todos tinham mais de 2 anos de experiência no ensino de habilidades. Já o terceiro grupo foi constituído por três professores de habilidades que não eram pediatras (NEE), sendo um endocrinologista, um gastroenterologista e o outro clínico geral. Todos os três também tinham mais de dois anos de experiência no ensino de habilidades.

5.1 Comparação entre as escalas

A Tabela 2 apresenta a média da nota isolada de cada escala, a nota das medidas antropométricas e a nota total por escala (nota isolada da escala + nota das medidas antropométricas) de acordo com o ano do curso. A nota isolada do CL foi maior entre os alunos do terceiro ano em comparação com o sexto ano ($5,1 \pm 0,54$ versus $4,7 \pm 0,61$, respectivamente; $p = 0,025$). Não se observou diferença entre as notas isoladas das EG e de NT. A nota das medidas antropométricas foi maior no sexto ano em comparação com o

terceiro ano ($2,6\pm 0,50$ versus $2,1\pm 0,81$, respectivamente; $p = 0,013$). Não houve diferença entre os dois anos em relação à nota total nas diferentes escalas.

Tabela 2 - Nota isolada de cada escala, nota das medidas antropométricas e nota total por escala por ano do curso

Instrumento	Terceiro ano		Sexto ano		Total		p*
	Média	D.P.	Média	D.P.	Média	D.P.	
Nota da escala global isolada	4,6	0,49	4,4	0,47	4,5	0,49	0,079
Nota do <i>checklist</i> isolado	5,1	0,54	4,7	0,61	4,9	0,60	0,025
Nota do nível de treinamento isolado	2,7	0,83	2,8	0,84	2,7	0,83	0,771
Nota das medidas antropométricas	2,1	0,81	2,6	0,50	2,3	0,74	0,013
Nota total escala global	6,7	0,83	7,0	0,66	6,8	0,77	0,221
Nota total <i>checklist</i>	7,2	1,03	7,3	0,73	7,2	0,91	0,610
Nota total nível de treinamento	4,8	1,07	5,4	0,95	5,0	1,05	0,051

D.P. - desvio padrão; *teste t de *Student*

Com objetivo de avaliar a diferença observada entre as notas das medidas antropométricas, comparou-se o percentual de acerto de cada medida de acordo com o ano do curso. A Tabela 3 mostra que houve diferença apenas na medida de peso, com percentual de acerto significativamente mais elevado entre os alunos do sexto ano ($95,0\%$ versus $53,6\%$; $p = 0,003$).

Tabela 3 - Comparação do percentual de acertos e erros das medidas antropométricas por ano do curso

Variável		Terceiro ano		Sexto ano		Total		p*
		n	%	n	%	n	%	
Peso	Errou	13	46,4	1	5,0	14	29,2	0,003
	Acertou	15	53,6	19	95,0	34	70,8	
Perímetro cefálico	Errou	10	35,7	3	15,0	13	27,1	0,102
	Acertou	18	64,3	17	85,0	35	72,9	
Comprimento	Errou	3	10,7	4	20,0	7	14,6	0,311
	Acertou	25	89,3	16	80,0	41	85,4	

*teste de χ^2

5.2 Comparação entre os perfis de avaliadores

A Tabela 4 representa as notas de acordo com o perfil do avaliador por ano considerando a nota isolada e total de cada escala. Não se observou diferença estatística em nenhuma das situações.

Tabela 4 - Notas médias de acordo por perfil de avaliador considerando-se as notas das escalas isoladas e nota total por ano

Escalas isoladas	Terceiro ano		Sexto ano		Total		p*
	Média	D.P.	Média	D.P.	Média	D.P.	
Especialista experiente	3,8	0,54	3,6	0,53	3,7	0,53	0,271
Especialista sem experiência	4,1	0,61	3,9	0,79	4,0	0,69	0,288
Não especialista experiente	4,5	0,51	4,4	0,52	4,5	0,51	0,233
Nota total	Média	D.P.	Média	D.P.	Média	D.P.	p
Especialista experiente	5,8	0,93	6,2	0,69	6,0	0,85	0,154
Especialista sem experiência	6,2	1,03	6,5	0,89	6,3	0,98	0,282
Não especialista experiente	6,6	0,82	7,0	0,67	6,8	0,77	0,124

D.P. - desvio padrão; *teste t de Student

5.3 Comparação entre os perfis do avaliador de acordo com escala utilizada

A Tabela 5 demonstra a comparação entre a nota isolada e total da EG de acordo com o perfil do avaliador por ano do curso. Houve diferença estatisticamente significativa na nota isolada do EE, sendo que a notado terceiro ano foi maior ($3,5 \pm 0,79$ versus $2,8 \pm 0,89$; $p = 0,004$). Também houve diferença na nota total do NEE, neste caso o sexto ano foi melhor ($8,3 \pm 0,57$ versus $7,6 \pm 0,83$; $p = 0,002$). Considerando a comparação entre os diferentes perfis de avaliador, o teste *post hoc* de Turkey HSD mostrou que as notas foram estatisticamente diferentes entre os três tipos de avaliador, sendo que a nota do EE foi sempre menor, seguida da nota do EN e do NEE.

Tabela 5 - Comparação entre a nota isolada e total da escala global de acordo com o perfil do avaliador por ano do curso

Escala global isolada	Terceiro ano		Sexto ano		Total		p*
	Média	D.P.	Média	D.P.	Média	D.P.	
Especialista experiente	3,5	0,79	2,8	0,89	3,2	0,90	0,004
Especialista sem experiência	4,8	0,59	4,6	0,68	4,7	0,62	0,418
Não especialista experiente	5,6	0,57	5,7	0,47	5,6	0,53	0,414
ANOVA**	<0,001						
Escala global total	Média	D.P.	Média	D.P.	Média	D.P.	p*
Especialista experiente	5,6	0,99	5,4	1,05	5,5	1,01	0,490
Especialista sem experiência	6,8	0,98	7,2	0,83	7,0	0,93	0,169

Não especialista experiente	7,6	0,83	8,3	0,57	7,9	0,79	0,002
ANOVA**						<0,001	

D.P. - desvio padrão; *teste de t de Student; **teste Post Hoc (Turkey HSD) p < 0,001 entre todos os grupos nas duas escalas

A Tabela 6 demonstra a comparação entre nota isolada e total do CL de acordo com o perfil do avaliador por ano do curso. Não houve diferença estatística entre os anos, mas a ANOVA mostrou diferença significativa entre as notas dos três tipos de avaliadores, sendo a nota do EN a mais baixa, seguida do NEE e do EE.

Tabela 6 - Comparação entre a nota isolada e total do *checklist* de acordo com o perfil do avaliador por ano do curso

<i>Checklist isolado</i>	Terceiro ano		Sexto ano		Total		p*
	Média	D.P.	Média	D.P.	Média	D.P.	
Especialista experiente	5,8	0,52	5,6	0,51	5,7	0,52	0,191
Especialista sem experiência	4,4	0,92	4,0	1,10	4,2	1,02	0,108
Não especialista experiente	5,1	0,89	4,7	1,09	4,9	1,00	0,092
ANOVA**						<0,001	
<i>Checklist total</i>	Média	D.P.	Média	D.P.	Média	D.P.	p*
Especialista experiente	7,8	1,10	8,2	0,67	8,0	0,94	0,204
Especialista sem experiência	6,5	1,23	6,6	1,10	6,5	1,17	0,885
Não especialista experiente	7,2	1,23	7,3	1,21	7,2	1,21	0,921
ANOVA***						<0,001	

D.P. - desvio padrão; *teste de t de Student; **teste Post Hoc (Turkey HSD) p < 0,001 entre todos os grupos; ***teste Post Hoc (Turkey HSD) p < 0,01 entre todos os grupos

A Tabela 7 compara a nota da NT isolada e total de acordo com o perfil do avaliador por ano do curso. Observa-se que na nota total do EE do sexto ano foi significativamente maior que a nota do terceiro ano ($5,1 \pm 0,93$ versus $4,1 \pm 1,23$; $p = 0,006$). Já a ANOVA mostra que houve diferença entre a nota isolada do EE em comparação como EN ($2,2 \pm 0,99$ versus $3,1 \pm 1,16$; $p < 0,001$) e o NEE ($2,2 \pm 0,99$ versus $2,9 \pm 0,90$; $p = 0,007$). Também houve diferença significativa entre nota total do EE em comparação com o EN ($4,5 \pm 1,20$ versus $5,4 \pm 1,41$; $p = 0,002$) e o NEE ($4,5 \pm 1,20$ versus $5,1 \pm 0,99$; $p = 0,027$). Não se observou diferença entre a nota isolada e total do EN em comparação com o NEE.

Tabela 7 - Comparação entre as médias da nota da escala de nível de treinamento e as médias da nota total da escala de nível de treinamento de acordo com o perfil do avaliador, por ano do curso.

Nível de treinamento isolado	Terceiro ano		Sexto ano		Total		p*
	Média	D.P.	Média	D.P.	Média	D.P.	
Especialista experiente	2,0	1,04	2,5	0,89	2,2	0,99	0,154
Especialista sem experiência	3,1	1,22	3,1	1,10	3,1	1,16	0,950
Não especialista experiente	3,0	0,86	2,8	0,97	2,9	0,90	0,503
ANOVA**						<0,001	
Nível de treinamento total	Média	D.P.	Média	D.P.	Média	D.P.	p*
Especialista experiente	4,1	1,23	5,1	0,94	4,5	1,20	0,006
Especialista sem experiência	5,1	1,51	5,7	1,23	5,4	1,41	0,222
Não especialista experiente	5,0	0,94	5,4	1,04	5,1	0,99	0,231
ANOVA***						0,002	

D.P. - desvio padrão; *teste de t de Student; **teste *Post Hoc* (Turkey HSD) $p < 0,001$ entre especialista com experiência e sem experiência e $p = 0,007$ entre especialista e não especialista com experiência; ***teste *Post Hoc* (Turkey HSD) $p = 0,002$ entre especialista com experiência e sem experiência e $p = 0,027$ entre especialista e não especialista experiente

5.4 Comparação entre as escalas de acordo com o perfil do avaliador

A Tabela 8 compara a nota isolada e total do EE de acordo com a escala utilizada por ano do curso. Observa-se que houve diferença estatística na nota isolada da EG onde a nota do terceiro ano foi maior ($3,5 \pm 0,79$ versus $2,8 \pm 0,89$; $p = 0,004$) e na nota total da NT, em que a nota do sexto ano foi maior ($5,1 \pm 0,94$ versus $4,1 \pm 1,23$; $p = 0,006$). Já a ANOVA mostrou diferença significativa entre as notas das três diferentes escalas, sendo a nota da NT a mais baixa, seguida da EG e do CL.

Tabela 8 - Comparação entre as notas do especialista experiente de acordo com as diferentes escalas isoladas e notas totais por ano do curso

Escalas isoladas	Terceiro ano		Sexto ano		Total		p
	Média	D.P.	Média	D.P.	Média	D.P.	
Escala global isolada	3,5	0,79	2,8	0,89	3,2	0,90	0,004
Checklist isolado	5,8	0,52	5,6	0,51	5,7	0,52	0,191
Nível de treinamento isolado	2,0	1,04	2,5	0,89	2,2	0,99	0,154
ANOVA*						<0,001	
Nota total	Média	D.P.	Média	D.P.	Média	D.P.	
Escala global total	5,6	0,99	5,4	1,05	5,5	1,01	0,490
Checklist total	7,8	1,10	8,2	0,67	8,0	0,94	0,204
Nível de treinamento total	4,1	1,23	5,1	0,94	4,5	1,20	0,006
ANOVA*						<0,001	

*teste *Post Hoc* (Turkey HSD) $p < 0,05$ entre todas as escalas

A Tabela 9 compara a nota isolada e total do EN de acordo com a escala utilizada por ano do curso. Não houve diferença estatística entre os anos, mas a ANOVA mostrou diferença significativa entre as notas das três diferentes escalas, sendo a nota da NT a mais baixa, seguida do CL e EG.

Tabela 9 - Comparação entre as notas do especialista sem experiência de acordo com as diferentes escalas isoladas e notas totais por ano do curso

Escalas isoladas	Terceiro ano		Sexto ano		Total		p
	Média	D.P.	Média	D.P.	Média	D.P.	
Escala global isolada	4,8	0,59	4,6	0,68	4,7	0,62	0,418
Checklist isolado	4,4	0,92	4,0	1,10	4,2	1,02	0,108
Nível de treinamento isolado	3,1	1,22	3,1	1,10	3,1	1,16	0,950
ANOVA*	<0,001						
Nota total	Média	D.P.	Média	D.P.	Média	D.P.	
Escala global total	6,8	0,98	7,2	0,83	7,0	0,93	0,169
Checklist total	6,5	1,23	6,6	1,10	6,5	1,17	0,885
Nível de treinamento total	5,1	1,51	5,7	1,23	5,4	1,41	0,222
ANOVA**	<0,001						

*teste *Post Hoc* (Turkey HSD) $p < 0,05$ entre todas as escalas; **teste *Post Hoc* (Turkey HSD) $p < 0,05$ entre nota total do nível de treinamento e a nota total das outras duas escalas.

A Tabela 10 compara a nota isolada e total do NEE de acordo com a escala utilizada por ano do curso. Observa-se que na nota total da EG do sexto ano foi significativamente maior que a nota do terceiro ano ($8,3 \pm 0,57$ versus $7,6 \pm 0,83$; $p=0,002$). Já a ANOVA mostrou diferença significativa entre as notas das três diferentes escalas, sendo a nota da NT novamente a mais baixa, seguida do CL e da EG.

Tabela 10 - Comparação entre as notas do não especialista experiente de acordo com as diferentes escalas isoladas e notas totais por ano do curso

Escalas isoladas	Terceiro ano		Sexto ano		Total		p
	Média	D.P.	Média	D.P.	Média	D.P.	
Escala global isolada	5,6	0,57	5,7	0,47	5,6	0,53	0,414
Checklist isolado	5,1	0,89	4,7	1,09	4,9	1,00	0,092
Nível de treinamento isolado	3,0	0,86	2,8	0,97	2,9	0,90	0,503
ANOVA*	<0,001						
Nota total	Média	D.P.	Média	D.P.	Média	D.P.	
Escala global total	7,6	0,83	8,3	0,57	7,9	0,79	0,002
Checklist total	7,2	1,23	7,3	1,21	7,2	1,21	0,921
Nível de treinamento total	5,0	0,94	5,4	1,04	5,1	0,99	0,231
ANOVA*	<0,001						

*teste *Post Hoc* (Turkey HSD) $p < 0,05$ entre todas as escalas

5.5 Análise multivariada

Com objetivo de avaliar melhor qual das duas variáveis, perfil do avaliador ou ano do curso, apresenta maior influência da definição da nota realizou-se regressão linear considerando a nota isolada e nota total de cada uma das escalas (Tabela 11). O perfil do avaliador influenciou significativa a nota isolada e total da EG e a nota total do CL. A nota do CL isolado foi influenciada tanto pelo perfil do avaliador como pelo ano do curso. Deve-se destacar que nesse caso a nota do terceiro ano foi maior que a média do sexto ano, conforme já apresentado na Tabela 2. A nota isolada da NT não foi influenciada por nenhuma das duas variáveis. Já a nota total da NT foi influenciada apenas pelo ano do curso.

Tabela 11 - Fatores que influenciaram na nota geral isolada e total e na nota isolada e total de cada escala

Nota isolada escala global	P ajustado
Ano do curso	0,196
Perfil do avaliador	<0,001
Nota total escala global	P ajustado
Ano do curso	0,208
Perfil do avaliador	0,001
Nota isolada checklist	P ajustado
Ano do curso	0,022
Perfil do avaliador	0,001
Nota total checklist	P ajustado
Ano do curso	0,506
Perfil do avaliador	0,005
Nota isolada nível de treinamento	P ajustado
Ano do curso	0,696
Perfil do avaliador	0,346
Nota total nível de treinamento	P ajustado
Ano do curso	0,004
Perfil do avaliador	0,407

5.6 Análise de confiabilidade

A Tabela 12 apresenta os níveis de confiabilidade e consistência interna da nota isolada e total de cada escala de acordo com alfa de Cronbach. Em relação as notas isoladas, apenas a NT

apresenta boa confiabilidade. O acréscimo da nota das medidas antropométricas aumentou a confiabilidade de todas as escalas, sendo que a NT manteve o maior nível de confiabilidade.

Tabela 12 - Análise de confiabilidade (alfa de Cronbach) de acordo com escala isolada e nota total

Escala isolada	Alfa de Cronbach
Escala global	0,469
<i>Checklist</i>	0,439
Nível de treinamento	0,739
Nota total	Alfa de Cronbach
Escala global	0,782
<i>Checklist</i>	0,757
Nível de treinamento	0,839

Já a Tabela 13 avalia a confiabilidade de acordo com o perfil do avaliador. Na nota isolada das escalas, a confiabilidade foi muito baixa. Assim como na confiabilidade das escalas, o acréscimo da nota das medidas antropométricas aumentou a confiabilidade dos avaliadores. O NEE teve a menor confiabilidade nas duas análises.

Tabela 13 - Análise de confiabilidade (alfa de Cronbach) de acordo com perfil do avaliador considerando a escala isolada e a nota total

Avaliador com escala isolada	Alfa de Cronbach
Especialista sem experiência	0,533
Especialista experiente	0,292
Não especialista experiente	0,183
Avaliador com nota total	Alfa de Cronbach
Especialista sem experiência	0,769
Especialista experiente	0,718
Não especialista experiente	0,635

6 DISCUSSÃO

O objetivo do estudo foi avaliar se existe diferença na capacidade de discriminação e na confiabilidade entre três escalas utilizadas para avaliação de habilidades e entre três perfis diferentes de avaliadores. A principal limitação desse estudo é o tamanho da amostra, tanto de alunos, como de avaliadores, o que pode ter impactado negativamente a análise de confiabilidade. Outro ponto que pode ter influenciado no desempenho dos alunos é o fato da avaliação ter sido filmada, o que pode ter comprometido a performance de alguns deles. Por outro lado, a filmagem facilitou a parte operacional do estudo, facilitando o recrutamento dos avaliadores e a realização das avaliações, que foram feitas de forma independente por cada um deles. Como os avaliadores preencheram os questionários sem nenhum tipo de supervisão dos pesquisadores e apesar de terem sido orientados a não reverem as avaliações, pode ter ocorrido algum tipo de contaminação ou viés durante as avaliações. Outras limitações são a utilização de avaliadores sem treinamento para o uso das escalas e, talvez, as habilidades avaliadas sejam muito simples e não tenham a complexidade necessária para permitir a discriminação de um aluno menos experiente de um aluno mais experiente.

Quando se comparou as diferentes escalas, independentemente do perfil do avaliador, observou-se apenas que a nota isolada do *checklist* foi capaz de discriminar os alunos do terceiro e do sexto anos. No entanto, o resultado foi o contrário do esperado, a nota do terceiro ano foi superior a nota do sexto ano. Esse resultado pode ser explicado pelo fato dos alunos do terceiro ano terem passado pelo módulo de treinamento em ambiente simulado da habilidade solicitada neste projeto mais recentemente, ocasionando um viés devido ao condicionamento prévio destes alunos. Além disto, os *checklists* são construídos a partir de padrões utilizados nos programas educacionais, sendo geralmente muito explícitos e detalhados podendo induzir o avaliador ao trivial (VAN DER VLEUTEN *et al.*, 1991; CUSHING, 2002). Apesar da EG isolada não ter apresentado diferença estatisticamente significativa para discriminar os alunos, os do terceiro ano também tiveram nota média superior aos do sexto ano, fato este que também poderia ser explicado pelo condicionamento mais recente desses alunos.

Em relação aos resultados das medidas antropométricas isoladas, observou-se que os alunos do sexto ano apresentaram taxa de acerto do peso significativamente maior que os alunos do terceiro ano. A observação dos vídeos permitiu identificar que um percentual importante dos

alunos do terceiro ano não tirou a roupa do manequim antes de pesá-lo (dados não apresentados), o que pode explicar essa diferença observada, já que a presença da roupa não interfere diretamente nas outras medidas antropométricas avaliadas (perímetro cefálico e comprimento). Esse erro pode ser explicado pela ansiedade dos alunos mais novos e/ou pela falta de experiência prática, uma vez que eles realizaram esse procedimento poucas vezes e sempre em ambiente simulado.

Já na análise da nota total (nota da escala + nota dos dados antropométricos) das três escalas, evidenciou-se melhor desempenho dos alunos do sexto ano. O modelo da pirâmide de Miller pode explicar estes resultados. Os alunos do terceiro ano aprenderam ao final do segundo ano a habilidade do projeto, no laboratório de simulação seguindo um passo a passo semelhante ao *checklist* (mostrar como faz- “desempenho condicionado”), ao contrário, os do sexto ano estariam em um nível avançado, aplicando a habilidade na vida real (fazer) com técnicas mais consolidadas, explicando o melhor desempenho em dados precisos (MILLER, 1990).

Quando se comparou os diferentes perfis de avaliadores em cada escala, observou-se que a EG isolada, quando utilizada pelo especialista experiente, permitiu discriminar os alunos entre os anos, mas a exemplo do *checklist*, o resultado foi o inverso do esperado. Há um corpo de pesquisa sobre o uso de CL, que descrevem com precisão a ocorrência de resultados mais objetivos e EG que descrevem a qualidade de um desempenho, permitindo mais interpretação pelo avaliador. Certamente um examinador experiente na habilidade avaliada é capaz de observar e pontuar com mais precisão os aspectos positivos e negativos da prova, pois sabe o que é relevante para um desempenho de sucesso (REGEHR *et al.*, 1999; HODGES *et al.*, 2002). Já o fato dos alunos do terceiro ano terem médias maiores pode ser explicado pelo condicionamento recente deles, como já discutido. A escala global somada às medidas objetivas permitiu ao não especialista discriminar entre o terceiro e o sexto anos. A escala de nível de treinamento somada às objetivas permitiu ao especialista experiente diferenciar corretamente os dois grupos. De acordo com esses resultados, talvez a escala mais indicada para o avaliador especialista experiente seja a NT total. Já para o avaliador não especialista experiente seria a EG associada a dados objetivos.

A análise multivariada mostrou que a NT, isolada ($p=0,346$) ou total ($p=0,407$), foi a única escala que não foi influenciada pelo perfil do avaliador. Além disso, a NT total foi capaz de discriminar corretamente os dois grupos de alunos ($p=0,004$). O CL isolado também permitiu

discriminar entre os anos ($p=0,022$), mas o resultado foi o inverso esperado, como já discutido, e foi influenciado pelo perfil do avaliador ($p=0,001$). As demais escalas não permitiram discriminar entre os anos. Esses resultados sugerem que talvez a NT seja a melhor escala quando os avaliadores tiverem perfis diferentes. Pugh *et al.* (2014) em seu estudo evidenciou resultados semelhantes, que indicaram que a escala de avaliação do nível de treinamento apresentou maior confiabilidade para avaliar as habilidades clínicas em laboratórios de simulação. Importante destacar, no entanto, que ela trabalhou com médicos residentes.

As escalas utilizadas apresentam características diferentes. O CL é uma escala mais objetiva, de fácil aplicação (respostas dicotômicas de sim ou não) e mais fácil de ser utilizada pelo avaliador não especialista com pouco treinamento. No entanto, tende a subestimar a avaliação, mantendo os resultados no mínimo desejável e pode ser falseada pelo condicionamento do aluno avaliado, como sugerido pelos resultados desse estudo. Já a EG é uma escala subjetiva, que depende do conhecimento prévio do avaliador sobre a habilidade a ser analisada ou que ele construa um modelo mental sobre ela, que pode se modificar ao longo da avaliação, ao ver um aluno que desempenha a habilidade melhor do que o seu modelo inicial. Além disso, apresentará baixa confiabilidade entre avaliadores diferentes, principalmente em relação a classificação dos conceitos limítrofes. Também pode ser influenciada pelo condicionamento prévio, como sugerido por esse estudo. A NT também é uma escala subjetiva, mas ela possui um parâmetro de comparação, que são os níveis mais altos de pontuação, nesse estudo o Médico residente e o Pediatra. Essa referência auxilia o avaliador na criação do modelo mental de comparação e pode tornar a escala menos subjetiva quando comparada com a EG. Isso pode explicar a maior confiabilidade e a capacidade de discriminação observadas na NT nesse estudo.

A análise de confiabilidade realizada de acordo com os tipos de escalas demonstrou que apenas a escala de nível de treinamento apresentou boa confiabilidade e que o acréscimo da nota das medidas antropométricas aumentou ainda mais a sua confiabilidade. A EG e o CL isolados apresentaram baixa confiabilidade, que melhorou com a utilização das medidas objetivas. Em relação à confiabilidade de acordo com o perfil do avaliador evidenciou que as escalas isoladas apresentaram confiabilidade muito baixa e que neste caso o acréscimo da nota das medidas antropométricas também melhorou a confiabilidade dos avaliadores. Considerando que não houve nenhum treinamento prévio dos avaliadores para cada um dos

instrumentos, era de se esperar que a confiabilidade não fosse alta. Chesser *et al.* (2009) realizaram estudo em quatro escolas de medicina do Reino Unido através da incorporação de uma estação de OSCE nas avaliações tradicionais e envio de um examinador treinado para cada uma destas escolas, depois compararam os resultados dos examinadores treinados previamente com os sem treinamento. Eles concluíram que o treinamento dos avaliadores melhora a confiabilidade entre eles, garantindo que todos interpretem as questões de forma semelhante, facilitando uma avaliação mais uniforme para o desempenho os alunos. Inquestionavelmente, o treinamento dos avaliadores em qualquer uma das escalas, aumentaria a confiabilidade delas, independentemente do perfil do avaliador (CHESSER *et al.*, 2009), mas um dos objetivos do estudo era justamente verificar a adequação das diferentes escalas aos perfis dos avaliadores.

Apesar das limitações do estudo, trata-se de um estudo experimental, que comparou três escalas de avaliação de habilidades e três perfis de avaliadores, onde cada estudante foi avaliado nove vezes, três vezes pela mesma escala e outras três pelos diferentes avaliadores, aumentando o número de registros analisados. As avaliações foram cegas, o que reduziu o viés por parte dos avaliadores. Até onde foi possível identificar na revisão de literatura, trata-se do primeiro estudo que comparou três escalas em estudantes da graduação e que utilizou a NT. Os resultados sugerem a possibilidade da utilização da escala de nível de treinamento para avaliar as habilidades de alunos da graduação em laboratório de simulação.

7 CONCLUSÃO

Os resultados indicaram que existe diferença na confiabilidade entre as escalas utilizadas para avaliação de habilidades clínicas em laboratório de simulação. A escala de nível de treinamento acrescida de dados objetivos foi a de maior confiabilidade. O perfil do avaliador interferiu diretamente na confiabilidade das escalas, porém os avaliadores especialistas nas habilidades avaliadas apresentaram confiabilidade aceitável quando a escala foi somada aos dados objetivos. Apenas a escala de nível de treinamento foi capaz de discriminar corretamente os dois grupos e não foi influenciada pelo perfil do avaliador.

8 CONSIDERAÇÕES FINAIS

Trata-se de um tema relevante e pouco estudado, sendo necessários novos estudos com estas escalas, principalmente sobre a escala de nível de treinamento na graduação, uma vez que essa escala apresentou bons resultados e, de acordo com a revisão da literatura, nunca havia sido usada nessa população.

Novos estudos experimentais, com maior número de voluntários e de avaliadores, poderão avaliar de forma mais precisa a confiabilidade das diferentes escalas. Sugere-se ainda a utilização de habilidades mais complexas, com o intuito de aumentar a discriminação entre alunos de diferentes anos.

Os resultados apontam para a possibilidade de utilização da escala de nível de treinamento acrescida de dados objetivos na avaliação de alunos em laboratórios de habilidade. Não foi identificado uma escala mais adequada aos diferentes perfis de avaliadores analisados, devendo-se considerar o treinamento deles antes da aplicação da escala; pois é evidente na literatura que o treinamento dos examinadores contribui para um aumento na confiabilidade do instrumento utilizado.

REFERÊNCIAS

- ALINIER, G. A typology of educationally focused medical simulation tools. **Medical Teacher**, [S. l.], v. 29, n. 8, p. 243-250, Oct. 2007.
- BLOMM, B. S. **Taxonomy of education objectives**. New York: David McKay Company, 1956. 216p.
- BRADLEY, P. The history of simulation in medical education and possible future directions. **Medical Education**, [S. l.], v. 40, p. 254-262, 2006.
- BRANNICK, M. T.; EROL-KORKMAZ, H. T.; PREWETT, M. A systematic review of the reliability of objective structured clinical examination scores. **Medical Education**, [S. l.], v. 45, n. 12, p. 1181-1189, Oct. 2011.
- BUCKINGHAM, B. R. Intelligence and its measurement: A symposium. **Journal of Educational Psychology**, [S. l.], v. 12, n. 3, p. 123-147, 1921.
- BUGAJ, T. J.; NIKENDEI, C. Practical Clinical Training in Skills Labs: Theory and Practice. **GMS Journal for Medical Education**, [S. l.], v. 33, n. 4, Aug. 2016.
- CHESSER, A. *et al.* Sources of variation in performance on a shared OSCE station across four UK medical schools. **Medical Education**, [S. l.], v. 43, n. 6, p. 526-532, May. 2009.
- COOKE, M. *et al.* American Medical Education 100 Years after the Flexner Report. **The New England Journal of Medicine**, [S. l.], v. 355, p. 1339-1344, Sept. 2006.
- CUSHING, A. Assessment of Non-Cognitive Factors. *In*: Norman, G.R. *et al.* **International Handbook of Research in Medical Education**. 1. ed. Dordrecht: Kluwer Academic Publishers, 2002. p. 711-755.
- DATTA, R.; UPADHYAY, K.; JAIDEEP, C. Simulation and its role in medical education. **Medical Journal, Armed Forces India**, [S. l.], v. 68, n. 2, p. 167-172, Apr. 2012.
- DIAS, R. D.; SCALABRINI NETO, A. Importância do laboratório de habilidades na capacitação e avaliação prática de alunos e profissionais da área da saúde. *In*: TIBÉRIO, I. F. L. C. *et al.* **Avaliação prática de habilidades clínicas em medicina**. Rio de Janeiro: Atheneu, 2012.
- DREYFUS, S. E.; DREYFUS, H. L. **A Five-stage model of the mental activities involved in directed skill acquisition**. Washington, DC: Storming Media, 1980. 22 p.
- EPSTEIN, R. M. Assessment in Medical Education. **The New England Journal of Medicine**, [S. l.], v. 356, p. 387-396, Jan. 2007.
- HARDEN, R. M. *et al.* Assessment of clinical competence using an objective structured clinical examination. **British Medical Journal**, [S. l.], v. 1, p. 447-451, 1975.
- HARDEN, R. M. What is an OSCE? **Medical Teacher**, [S. l.], v. 10, p. 10-23, 1988.

HODGES *et al.* The challenge of creating new OSCE measures to capture the characteristics of expertise. **Medical Education**, [S. l.], v. 36, n. 8, p. 742-748, Aug. 2002.

HOLMBOE, E. S. Faculty and the observation of trainees' clinical skills: problems and opportunities. **Academic Medicine**, [S. l.], v. 79, n. 5, p. 16-22, Jan. 2004.

KER, J.; BRADLEY, P. Simulation in medical education. *In*: SWANWICK, T. **Understanding Medical Education: Evidence, Theory and Practice**. 1. ed. London: Wiley-Blackwell, 2010.

KHAN, K. Z. *et al.* The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoretical perspective. **Medical Teacher**, [S. l.], v. 35, n. 9, p. 1437-46, Sep. 2013.

KHAN, K. Z.; PATTISON, T.; SHERWOOD, M. Simulation in medical education. **Medical Teacher**, [S. l.], v. 33, p. 1-3, 2011.

LEUNG, F. H.; MARTIN, D.; BATTY, H. A theory-based curriculum design for remediation of residents' communication skills. **Medical Teacher**, [S. l.], v. 31, n. 12, p. 555-559, dez. 2009.

MARKUS, K. A.; BORSBOOM, D. Frontiers of test validity theory: Measurement, causation, an meaning. **Journal of Educational Measurement**, New York, v. 51, n. 5, p. 463-469, 2013.

MATOS, F. S. **A prova prática-oral estruturada é comparável a uma estação do exame clínico objetivo estruturado, na avaliação de habilidades clínicas em estudantes de medicina?** Estudo experimental. 2017, 45 f. Dissertação (Mestrado Profissional em Ensino em Saúde) - Universidade José do Rosário Vellano, Belo Horizonte, 2017.

MCGAGHIE, W. C. *et al.* A critical review of simulation-based medical education research: 2003–2009. **Medical Education**, [S. l.], v. 44, n. 1, p. 50-63, 2010.

MILLER, G. E. The assessment of clinical skills/competence/performance. **Academic Medicine**, [S. l.], v. 65, n. 9, p. 63-67, Sep. 1990.

MORAN, M. D. Enlightenment via Simulation: “Crone-ology’s” First Woman. **Journal of Endourology**, [S. l.], v. 24, n. 1, p. 5-8, Jan. 2010.

MURPHY, J. G. *et al.* Is simulation based medicine training the future of clinical medicine? **European Review for Medical and Pharmacological Science**, [S. l.], v. 11, n. 1, p. 1-8, Jan. 2007.

MUSICK, D. A conceptual model for program evaluation in graduate medical education. **Academic Medicine**, [S. l.], v. 81, n. 8, p. 759-765, Aug. 2006.

NORCINI, J. *et al.* Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. **Medical Teacher**, [S. l.], v. 33, n. 3, p. 206-214, 2011.

OKUDA, Y. *et al.* The utility of simulation in Medical Education: What is the evidence? **Mount Sinai Journal of Medicine**, [S. l.], v. 76, p. 330-343, 2009.

PACICO, J. C.; HUTZ, C. S. **Validade**. In: HUTZ, C. S.; BANDEIRA, D. R.; TRENTINI, C. M. **Psicometria**. 1. ed. Porto Alegre: Artmed, 2015. p. 71-84.

PINTO, M. P. P.; TRONCON, L. E. A. Avaliação do estudante – aspectos gerais. **Medicina**, Ribeirão Preto, v. 47, n. 3, p. 314-323, 2014.

PUGH, D. *et al.* Progress testing: is there a role for the OSCE? **Medical Education**, [S. l.], v. 48, n. 6, p. 623-31, 2014.

REGEHR, G. *et al.* Assessing the generalisability of OSCE measures across content domains. **Academic Medicine**, [S. l.], v. 74, n. 12, p. 1320-1322, Dec. 1999,

SCHUWIRTH, L.; ASH, J. Assessing tomorrow's learners: in competency-based education only a radically different holistic method of assessment will work. Sixthings we could forget. **Medical Teacher**, [S. l.], v. 35, n. 7, p. 555-559, Jul. 2013.

SMEE, S. ABC of learning and teaching in medicine. Skill based assessment. **British Medical Journal**, [S. l.], v. 37, p. 703-706, 2003.

SOUZA, R. G. S. Atributos Fundamentais dos Procedimentos de Avaliação. In: TIBÉRIO, I. F. L. C. *et al.* **Avaliação prática de habilidades clínicas em medicina**. Rio de Janeiro: Atheneu, 2012.

TRONCON, L. E. A. Avaliação do Estudante de Medicina. **Medicina**, Ribeirão Preto, v. 29, p. 429-439, out./dez. 1996.

TRONCON, L. E. A. *et al.* Estruturação de Sistemas para Avaliação Programática do Estudante de Medicina. **Revista Brasileira de Educação Médica**, [S. l.], v. 40, n. 1, p. 30-42, 2016.

VANDER VLEUTEN, C. P. M. *et al.* Pitfalls in the pursuit of objectivity: Issues of reliability. **Medical Education**, [S. l.], v. 25, n. 2, p. 110-118, Mar. 1991,

VANTINI, I.; BENINI, L. Models of learning, training and progress evaluation of medical students. **ClinicaChimica Acta**, [S. l.], v. 393, n. 1, p. 13-16, 2008.

ZANON, C.; HAUCK FILHO, N. Fidedignidade. In: HUTZ, C. S.; BANDEIRA, D. R.; TRENTINI, C. M. **Psicometria**. 1. ed. Porto Alegre: Artmed, 2015. p. 85-95.

ZIV, A. *et al.* Simulation-based medical education: na ethical imperative. **Academic Medicine**, [S. l.], v. 78, n. 8, p. 783-788, 2003.

ZUMBO, B. D.; CHAN, E. K. H. **Validity and Validation in Social, Behavioral, and Health Sciences**. 1. ed. New York: Springer, 2014. 344 p.

6. Descrição dos desconfortos e riscos da pesquisa

(X) Risco Mínimo () Risco Baixo () Risco Médio () Risco Alto

Sua participação neste estudo não implica em nenhum risco físico, pois não será realizado nenhum procedimento invasivo. Pode haver algum constrangimento relacionado ao confinamento antes da prova e à aplicação da prova. Por tratar-se de uma prova de habilidades em que o aluno será filmado para posterior avaliação do desempenho. Todo esforço será feito para manutenção das gravações somente entre a equipe de pesquisadores. O TCLE será arquivado em espaço fisicamente separado do espaço onde serão arquivadas as gravações das provas. Todas as informações relativas aos voluntários serão identificadas apenas pelo número de registro e apenas a equipe de pesquisadores terá acesso às informações e aos bancos de dados.

7. Descrição dos benefícios da pesquisa

Sua participação não trará nenhum benefício direto para você, mas melhorar os critérios de confiabilidade para avaliação nos laboratórios de simulação pode indicar mudanças no sistema de avaliação que podem beneficiar muitos estudantes de vários cursos de medicina.

8. Despesas, compensações e indenizações

Nenhum dos voluntários receberá pagamento ou benefício pessoal por sua participação no projeto. As avaliações serão aplicadas em dias de aula habitual, logo você não terá nenhuma despesa pessoal nessa pesquisa, incluindo transporte, exames e consultas.

9. Direito de confidencialidade

Você tem assegurado que todas as suas informações pessoais obtidas durante a pesquisa serão consideradas estritamente confidenciais e os registros estarão disponíveis apenas para os pesquisadores envolvidos no estudo.

Os resultados obtidos nessa pesquisa poderão ser publicados com fins científicos, mas sua identidade será mantida em sigilo.

Imagens ou fotografias que possam ser realizadas se forem publicadas, não permitirão sua identificação.

10. Acesso aos resultados da pesquisa

Você tem direito de acesso atualizado aos resultados da pesquisa, ainda que os mesmos possam afetar sua vontade em continuar participando da mesma.

11. Liberdade de retirada do consentimento

Você tem direito de retirar seu consentimento, a qualquer momento, deixando de participar da pesquisa, sem qualquer prejuízo à continuidade de seu cuidado e tratamento na instituição.

12. Acesso aos profissionais responsáveis pela pesquisa

Você tem garantido o acesso, em qualquer etapa da pesquisa, aos profissionais responsáveis pela mesma, para esclarecimento de eventuais dúvidas acerca de procedimentos, riscos, benefícios, etc., através dos contatos abaixo:

Prof. Orientador: Antonio Carlos de Castro Toledo Jr. - **Mestranda:** Layara Maria Garcia Brasileiro

Tel. (31) 2536-5678 - **E-mail:** antonio.toledo@unifenas.br

13. Acesso à instituição responsável pela pesquisa

Você tem garantido o acesso, em qualquer etapa da pesquisa, à instituição responsável pela mesma, para esclarecimento de eventuais dúvidas acerca dos procedimentos éticos, através do contato abaixo:

Comitê de Ética - UNIFENAS

Rodovia MG 179, Km 0, Alfenas – MG

Tel: (35) 3299-3137

E-mail: comitedeetica@unifenas.br

segunda à sexta-feira das 14:00h às 16:00h

14. Consentimento

Fui informado verbalmente e por escrito sobre os dados dessa pesquisa e minhas dúvidas com relação a minha participação foram satisfatoriamente respondidas.

Ficaram claros para mim quais são os propósitos do estudo, os procedimentos a serem realizados, os desconfortos e riscos, as garantias de confidencialidade e de esclarecimentos permanentes. Ficou claro também que a minha participação é isenta de despesas e que tenho garantia do acesso aos pesquisadores e à instituição de ensino.

Tive tempo suficiente para decidir sobre minha participação e concordo voluntariamente em participar desta pesquisa e poderei retirar o meu consentimento a qualquer hora, antes ou durante a mesma, sem penalidades, prejuízo ou perda de qualquer benefício que eu possa ter adquirido.

A minha assinatura neste Termo de Consentimento Livre e Esclarecido dará autorização aos pesquisadores e ao Comitê de Ética em Pesquisa da Universidade José do Rosário Vellano, de utilizarem os dados obtidos quando se fizer necessário, incluindo a divulgação dos mesmos, sempre preservando minha identidade.

Assino o presente documento em duas vias de igual teor e forma, ficando uma em minha posse.

Belo Horizonte, _____ de _____ de 2019

Assinatura do aluno que quer participar da pesquisa

Nome pesquisador: Layara Maria Garcia Brasileiro

Assinatura pesquisador

ANEXO A - Parecer consubstanciado do CEP

UNIVERSIDADE JOSÉ
ROSÁRIO VELLANO/UNIFENAS



PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Comparação da confiabilidade entre três escalas para avaliação de habilidades clínicas em Laboratório de Simulação e três tipos diferentes de avaliadores. Estudo experimental unicêntrico, 2019

Pesquisador: Antonio Carlos de Castro Toledo Junior

Área Temática:

Versão: 3

CAAE: 02657318.4.0000.5143

Instituição Proponente: Universidade José Rosário Vellano/UNIFENAS

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 3.085.011

Apresentação do Projeto:

Adequada.

Objetivo da Pesquisa:

Adequado.

Avaliação dos Riscos e Benefícios:

Adequados.

Comentários e Considerações sobre a Pesquisa:

Pesquisa relevante.

Considerações sobre os Termos de apresentação obrigatória:

Adequados.

Conclusões ou Pendências e Lista de Inadequações:

Nada digno de nota.

Considerações Finais a critério do CEP:

Este parecer foi elaborado baseado nos documentos abaixo relacionados:

Endereço: Rodovia MG 179 km 0

Bairro: Campus Universitário

CEP: 37.130-000

UF: MG

Município: ALFENAS

Telefone: (35)3299-3137

Fax: (35)3299-3137

E-mail: comitedeetica@unifenas.br

**UNIVERSIDADE JOSÉ
ROSÁRIO VELLANO/UNIFENAS**



Continuação do Parecer: 3.085.011

Tipo Documento	Arquivo	Postagem	Autor	Situação
Informações Básicas do Projeto	PB_INFORMAÇÕES_BÁSICAS_DO_PROJETO_1250506.pdf	10/12/2018 10:29:06		Aceito
TCLE / Termos de Assentimento / Justificativa de Ausência	TCLE_layara.pdf	26/11/2018 10:14:02	Antonio Carlos de Castro Toledo Junior	Aceito
Outros	questionario_layara.docx	06/11/2018 18:17:52	Antonio Carlos de Castro Toledo Junior	Aceito
Declaração de Instituição e Infraestrutura	declaracao_layara.pdf	06/11/2018 17:59:43	Antonio Carlos de Castro Toledo Junior	Aceito
Projeto Detalhado / Brochura Investigador	ilovepdf_merged.pdf	03/11/2018 21:47:47	Antonio Carlos de Castro Toledo Junior	Aceito
Folha de Rosto	folharostolayara.pdf	03/11/2018 21:46:43	Antonio Carlos de Castro Toledo Junior	Aceito

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

ALFENAS, 14 de Dezembro de 2018

Assinado por:
MARCELO REIS DA COSTA
(Coordenador(a))

Endereço: Rodovia MG 179 km 0
Bairro: Campus Universitário **CEP:** 37.130-000
UF: MG **Município:** ALFENAS
Telefone: (35)3299-3137 **Fax:** (35)3299-3137 **E-mail:** comitedeetica@unifenas.br