

Procedimentos metodológicos para alcance da qualidade da pesquisa em ensino: empregando Rasch para validação de itens sobre corrente e circuitos elétricos

Methodological Procedures for Ensuring Quality in Educational Research: Using the Rasch Model to Validate Assessment Items on Electric Current and Electric Circuits

Elicarla de Jesus Souza¹
Amanda Amantes²

Resumo

Este artigo apresenta um estudo metodológico sobre a construção e a validação de um banco de itens relacionados à corrente elétrica e aos circuitos elétricos. Foram elaborados 40 itens, incluindo questões dicotômicas (de múltipla escolha e verdadeiro/falso) e discursivas, com base em instrumentos previamente disponíveis na literatura. A validação foi realizada com 205 estudantes da 3ª série do Ensino Médio de escolas públicas federais e estaduais que já haviam estudado o conteúdo. O Modelo Rasch foi utilizado para avaliar a dimensionalidade, a consistência interna, o ajuste dos itens e a adequação da amostra. Os resultados indicaram que o banco de itens é multidimensional e está bem ajustado. No entanto, os índices de confiabilidade e os valores de separação de pessoas itens sugerem que o instrumento pode apresentar limitações para discriminar estudantes com diferentes níveis de proficiência, especialmente em populações muito distintas. Apesar disso, a hierarquia dos itens apresentou elevada confiabilidade, sendo identificado apenas um item com ajuste insuficiente. Os procedimentos adotados contribuem para o avanço metodológico em relação à validação de instrumentos educacionais, oferecendo uma ferramenta consistente para a investigação do conhecimento sobre corrente elétrica e circuitos elétricos. Ressaltamos a importância de conduzir estudos futuros com amostras mais diversificadas para ampliar as evidências de validade.

Palavras-chave: Modelagem Rasch; Validação; Ensino de Física; Corrente elétrica; Circuitos elétricos.

¹ Universidade Federal da Bahia | ericarla-matos@hotmail.com

² Universidade Federal da Bahia | amanda.amantes@ufba.br

Abstract

This paper discusses a methodological study focused on developing and validating an item pool regarding electric current and electric circuits. A total of 40 items were created, including dichotomous questions (both multiple-choice and true/false) as well as constructed-response questions, based on existing instruments previously reported in the literature. The validation process involved 205 third-year high school students from federal and state public schools who had already covered the topic. The Rasch model was employed to assess dimensionality, internal consistency, item fit, and sample adequacy. The results indicated that the item pool is multidimensional and demonstrates good overall fit. However, the reliability indices and the separation values for both persons and items suggest that the instrument may struggle to distinguish among students with varying proficiency levels, particularly within heterogeneous populations. Despite this limitation, the item hierarchy showed high reliability, with only one item exhibiting inadequate fit. The procedures adopted contribute to methodological advances in the validation of educational instruments, providing a robust tool for investigating knowledge of electric current and electric circuits. We emphasize the importance of conducting future studies with more diverse samples to further strengthen the evidence of validity.

Keywords: Rasch model; Item validation; Physics Education; Electric current; Electric circuits.

Introdução

Avaliar atributos relacionados à aprendizagem tais como conhecimento, entendimento e concepção é uma tarefa complexa e repleta de desafios no campo educacional. Dada a natureza latente desses atributos e a falta de consenso teórico sobre suas próprias definições, é praticamente impossível estabelecer uma métrica absoluta que permita acessá-los e, conseqüentemente, avaliá-los de forma exata. Contudo, uma maneira de estabelecer interpretações de forma coerente sobre como as pessoas concebem ou entendem determinados conteúdos – objetivo fundamental para o estudo da aprendizagem – é valer-se de instrumentos construídos com base em um aporte teórico-metodológico consistente.

O processo de construção e validação de bancos de itens é uma prática comum no campo da avaliação (Xavier, 2018; Bayerl et al., 2021; Silva, 2021; Melo; Amantes, 2021; Oliveira; Wartha, 2024). Esse procedimento propicia o alcance de testes calibrados, favorecendo, dessa forma, o acesso mais fidedigno de um traço latente (Pasquali, 2007). Além disso, a partir de instrumentos validados realizamos comparações mais fidedignas, o que possibilita a avaliação de diferentes aspectos da aprendizagem. Segundo Shohamy (1994), o uso de testes pode contribuir para definir e identificar maneiras de mensurar habilidades, testar hipóteses e aprimorar a qualidade dos instrumentos de coleta de dados. No entanto, um dos principais problemas enfrentados no uso desses testes está relacionado à sua confiabilidade.

A confiabilidade é um atributo essencial na avaliação da qualidade de instrumentos de acesso e mensuração, sendo compreendida como o grau em que os resultados obtidos refletem, com precisão, o valor verdadeiro, minimizando os efeitos de erros aleatórios (Souza, 2017; Hayes, 1998). Nesse sentido, Bollen (1989) destaca que a confiabilidade está relacionada à consistência da medida. A ausência dessa propriedade pode comprometer significativamente a validade dos resultados, uma vez que dados inconsistentes prejudicam a interpretação e a comparação dos achados. De acordo com Terwee et al. (2007), a

confiabilidade diz respeito à capacidade do instrumento de produzir resultados estáveis e coerentes ao longo do tempo, em diferentes contextos e entre distintos avaliadores, abrangendo aspectos como precisão, estabilidade e homogeneidade, sendo, portanto, um dos critérios centrais para aferir a qualidade de um instrumento.

Nos manuais de Psicometria, um teste é considerado válido quando, de fato, mede aquilo que se propõe a medir (Pasquali, 2009). Raymundo (2009) destaca que a ausência de validação dos testes utilizados em uma pesquisa pode resultar em inadequações decorrentes da falta de critérios de qualidade, gerando interpretações equivocadas durante a análise dos dados obtidos. Do ponto de vista pedagógico, a validação dos instrumentos de avaliação é crucial para garantir que os resultados reflitam, com maior acurácia, o conhecimento dos estudantes. Uma avaliação nesses termos permite um diagnóstico mais preciso das dificuldades e dos conhecimentos dos alunos.

No âmbito da pesquisa acadêmica, a validação dos instrumentos garante que os dados coletados sejam compatíveis com o que é estabelecido teoricamente, apresentando, inclusive, um caráter de replicabilidade, requisito fundamental para a maioria dos estudos científicos. A partir de dados confiáveis, coletados por instrumentos validados, reduz-se a arbitrariedade das inferências, o que fornece parâmetros mais robustos tanto para desenvolver quanto para verificar teorias educacionais. Além disso, resultados provenientes de análises cujos dados foram coletados por instrumentos validados podem orientar a formulação de políticas educacionais mais eficazes e equitativas.

Nesse contexto, o presente estudo tem como objetivo apresentar a elaboração e validação de um banco de itens sobre corrente elétrica e circuitos elétricos. O banco abarca: i) itens previamente validados na literatura, como os de Coelho (2011), Mazur (2015), Silveira, Moreira e Axt (1989) e Silveira (2011); ii) itens adaptados de testes usuais de livros didáticos e vestibulares e iii) itens para acessar conhecimentos não contemplados pelos materiais existentes. Como resultado, temos um instrumento destinado a diagnosticar o conhecimento dos estudantes sobre esses conteúdos em diferentes situações educacionais, sendo relevante para pesquisas sobre aprendizagem, currículo e efeitos instrucionais. Para além do instrumento, os procedimentos metodológicos proporcionam um arcabouço de parâmetros a serem empregados em pesquisas educacionais e de ensino cujo foco se concentra na avaliação de traços latentes.

Validação e Modelagem Rasch

A validação de instrumentos avaliativos pode ser conduzida de diversas maneiras, empregando-se tanto métodos clássicos, como a Análise Fatorial Exploratória e Confirmatória (Brown, 2015), quanto modelagens, como a Teoria de Resposta ao Item (Bond; Fox, 2007; Richardson, 1936) e o Modelo Rasch (Rasch, 1960). Cada procedimento apresenta características distintas, mas todos têm como objetivo averiguar se os pressupostos teóricos e empíricos são observados nos dados empíricos submetidos à análise, o que fornece evidências que dão suporte às inferências e interpretação sobre a adequação dos instrumentos. Neste trabalho empregamos o Modelo Rasch para averiguar aspectos sobre: i) a dimensionalidade do banco de itens; ii) a adequação dos itens e das pessoas para acessar o traço “conhecimento sobre corrente elétrica e circuitos elétricos”; iii) a consistência do teste e iv) a adequação da amostra.

Em que consiste o processo de validação?

Validar consiste em um processo de verificar se os instrumentos acessam, de forma confiável, aquilo que se propõem a acessar. Gonçalves Junior (2020) explica que esse processo fornece indícios sobre o grau de concordância das respostas fornecidas aos itens, testes e questionários, a um modelo esperado; esse modelo se fundamenta em uma proposição teórica cujo padrão de resposta é formulado a partir da hipótese de que os itens realmente medem um atributo psicológico único e coerente. Se a hipótese é verdadeira, a comparação entre o comportamento dos dados empíricos e o modelo esperado atesta a validade do instrumento.

Para acessar o traço latente almejado, faz-se necessário elaborar o modelo teórico que oriente a construção dos instrumentos de avaliação. Esse processo é essencial para garantir a qualidade dos dados coletados. Existem diversos métodos utilizados para validar instrumentos dessa natureza. Entre eles, os métodos pautados na psicometria são amplamente empregados para avaliar a validade e confiabilidade dos instrumentos avaliativos (Cunha; Almeida Neto; Stackfleth, 2016). Alguns procedimentos comuns incluem a validação de conteúdo, a validação de critério e a validação de constructo.

Xavier (2018) conceitua a validação amostral como uma das etapas em que os dados são oriundos das respostas de estudantes a testes de conhecimentos. Trata-se de um processo em que uma amostra representativa do público-alvo é utilizada para testar a consistência interna, dentre outros parâmetros. Esse processo envolve a aplicação do instrumento a uma amostra piloto e a análise dos dados para verificar a consistência interna do teste e a precisão das medições. A validação amostral contribui para identificar e corrigir problemas antes da aplicação em larga escala, tornando o instrumento mais adequado e confiável para o público-alvo.

Modelagem Rasch e Validação

A Modelagem Rasch é uma abordagem proposta pelo matemático dinamarquês Georg Rasch (1960) que atualmente é empregada, dentre outras finalidades, para avaliar a confiabilidade e a validade de instrumentos, especialmente daqueles destinados a acessar traços latentes, como habilidades, atitudes ou percepção. Ela oferece uma escala intervalar que possibilita a interpretação das distâncias entre as medidas de diferentes sujeitos, assim como a avaliação das características dos instrumentos aplicados (Amantes; Coelho; Marinho, 2015). Esse modelo assume que a probabilidade de um indivíduo acertar um item específico depende de sua habilidade e da dificuldade do item (Linacre, 2010).

A dificuldade do item indica o ponto da escala de habilidade em que ele apresenta 50% de probabilidade de ser respondido corretamente, enquanto a habilidade do indivíduo representa o seu traço latente. O ajuste do item está associado ao quanto ele se adéqua ao conjunto de itens (o teste em geral) para avaliar o traço almejado.

A medida de habilidade refere-se à mensuração de um atributo latente, como o entendimento, e deve seguir os mesmos princípios de uma medida objetiva. Nessa perspectiva, a habilidade de um indivíduo é estimada com base em suas respostas a itens específicos, que funcionam como ferramentas de acesso. Para que essa medida seja considerada válida e independente, as estimativas de proficiência não podem variar em função do conjunto de itens utilizados, assim como as estimativas de dificuldade dos itens não devem depender do grupo de respondentes. Esse princípio de independência entre

pessoa e item constitui um dos fundamentos do Modelo Rasch, conferindo ao processo de mensuração características semelhantes às das medidas nas ciências naturais.

A Modelagem Rasch oferece diversos benefícios para a avaliação da confiabilidade e da validade de instrumentos. Por meio da análise dos parâmetros de dificuldade dos itens e das habilidades dos indivíduos, é possível identificar itens inconsistentes ou inadequados, permitindo a revisão e o aprimoramento do instrumento. Além disso, no contexto da modelagem de testes, é possível verificar se o instrumento avalia um único construto (unidimensionalidade) ou múltiplos construtos relacionados (multidimensionalidade), aspecto fundamental para a validade das inferências sobre o desempenho dos participantes (Britto, 2021).

O Modelo Rasch apoia-se no princípio da objetividade específica. Segundo Verdú-Soriano e Torres (2024), uma medida só é considerada válida e generalizável quando não depende das condições específicas de obtenção. Se os dados se ajustam ao modelo, as comparações entre pessoas são independentes dos itens administrados, e os parâmetros dos itens não são influenciados pela amostra de calibração. Esse princípio assegura que a estimativa da habilidade de um indivíduo não depende do conjunto particular de itens aplicados, assim como a dificuldade de um item não é condicionada pela amostra de participantes utilizada (Britto, 2021).

Como implicação, o processo de validação resulta em uma escala única e comum, na qual itens e habilidades são expressos na mesma métrica. Essa característica permite calcular a probabilidade de sucesso a partir da diferença entre habilidade e dificuldade, garantindo robustez, comparabilidade e generalização das inferências. Esse último aspecto da modelagem possibilita análises que vão além da validação, pois uma vez satisfeitos os critérios de ajuste dos itens, das pessoas, do modelo e da dimensionalidade, a medida obtida pode ser empregada em análises diversas, como por exemplo em estudos preditivos (Cook; Beckman, 2006), de trajetória de aprendizagem (Amantes; Coelho; Marinho, 2015; Xavier; Amantes, 2021) e de perfil amostral (Calheiros; Melo; Amantes, 2023), dentre outras.

Método

Construção do banco de itens

Segundo Wright e Stone (1999) um banco de itens é constituído por um conjunto de itens elaborados e calibrados de forma metódica, que serve para definir e quantificar um tema comum. Desenvolvemos um instrumento composto por um conjunto de itens que abrange questões relacionadas ao conhecimento formal sobre corrente elétrica e circuitos elétricos no nível do Ensino Médio. Esses itens foram construídos ou adaptados da literatura (Coelho, 2011; Mazur, 2015; Silveira; Moreira; Axt, 1989; Silveira, 2011), sendo classificados como conceituais, procedimentais ou de aplicação de ferramentas matemáticas.

O instrumento consiste em um banco composto por 40 itens, distribuídos da seguinte maneira: 14 itens de verdadeiro ou falso, 23 itens de múltipla escolha com quatro alternativas e 3 itens discursivos. Sua estrutura está organizada da seguinte forma: os itens i1 a i37 são dicotômicos, enquanto os itens i38 a i40 são discursivos. O Quadro 1 apresenta exemplos de itens de cada tipo.

Quadro 1: Exemplos de itens dicotômicos, múltipla escolha e discursivos

Item do tipo dicotômico verdadeiro ou falso
i5- Nesse tipo de associação de lâmpadas mostrada na figura, a resistência equivalente é dada pela soma das resistências de cada lâmpada.

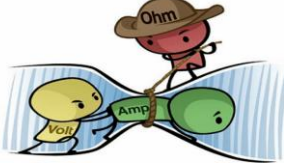
Fonte: http://www.10emtudo.com.br/aula/ensino/resistores_emserie_e_paralelo/
() Verdadeiro
() Falso
Item do tipo múltipla escolha
i17- A partir da imagem abaixo, marque a alternativa correta.

Fonte: https://embarcados.com.br/lei-de-ohm/
a) é correto afirmar que o fluxo de corrente elétrica não sofre interferência com a resistência do circuito.
b) a diferença de potencial não interfere no fluxo elétrico.
c) quanto maior for a resistência, menor vai ser o fluxo de corrente elétrica.
d) quanto maior a resistência, maior será o fluxo de corrente elétrica.
Item do tipo discursivo
i40 (Coelho, 2011) - Uma ação cotidiana e corriqueira é apertar um interruptor e acender uma lâmpada no teto ou no abajur. A figura mostra um modelo mais simples dessa situação: uma pilha comum está ligada a um interruptor e a uma lâmpada de lanterna. Ao pressionar o interruptor, a lâmpada se acende. Explique o que ocorre na pilha, nos fios, no interruptor e na lâmpada quando ela está acesa.

Figura: Representação de um circuito elétrico simples
Fonte: Coelho (2011, p. 60).

Fonte: Produção dos autores (2024).

Coleta e transformação dos dados

Validamos o banco de itens de maneira amostral, aplicando o teste a sujeitos que já estudaram os conteúdos em questão. O instrumento foi aplicado a 205 estudantes da 3ª série do Ensino Médio, provenientes da Rede Estadual do Espírito Santo e da Rede Federal de Minas Gerais. Na Rede Estadual do Espírito Santo, participaram alunos de uma escola regular e de uma escola em que o Ensino Médio é integrado ao ensino técnico (curso de Comércio). Já na Rede Federal de Minas Gerais, os participantes foram estudantes dos cursos técnicos de Agroecologia, Informática, Agrimensura e Meio Ambiente.

O banco de itens foi enviado aos professores de Física dessas instituições, que aceitaram contribuir com esta pesquisa mediante a aplicação do instrumento durante suas aulas. Os alunos responderam aos itens por meio do Google Forms, de forma presencial, ao longo de duas aulas geminadas de 50 minutos cada. Os estudantes não tiveram acesso a material de

consulta e responderam às questões apenas com base no conhecimento que já possuíam sobre os conteúdos.

Os resultados obtidos foram organizados em forma de tabela e codificados para transformar as respostas em dados numéricos. As respostas corretas dos itens de verdadeiro ou falso e de múltipla escolha foram substituídas pelo número 1 (um), enquanto as respostas incorretas foram representadas pelo número 0 (zero). Esse procedimento permitiu a construção de uma matriz de dados dicotômicos.

Método de análise

As respostas dos 205 estudantes constituíram os dados submetidos ao processo de validação, composto por duas etapas: a) validação dos itens dicotômicos; e b) validação dos itens discursivos.

Realizamos uma análise preliminar para identificar possíveis *outliers*, ou seja, dados discrepantes, que podem incluir: i) participantes que erraram ou acertaram todas as questões ou deixaram de responder à maioria delas; e ii) itens que foram acertados ou errados por todos os participantes. Não foi excluída nenhuma pessoa e nenhum item, pois nenhum deles se caracterizou como *outlier*.

a) Itens dicotômicos

O Modelo Rasch foi utilizado para a análise dos itens dicotômicos por meio do software Winsteps® 3.70 (Linacre, 2010). Neste trabalho, foi utilizada a versão gratuita de avaliação/estudante/demonstração do Winsteps, denominada MINISTEP. Foram avaliados os critérios de dimensionalidade, ajuste dos itens, confiabilidade e adequação ao modelo.

i) Dimensionalidade: a primeira análise consiste em identificar o número de dimensões do banco de itens a partir da variância residual padronizada, considerando a variância explicada pelas medidas empíricas e modeladas e a variância explicada pelo primeiro contraste. Avaliar a dimensionalidade significa verificar se o teste está acessando uma ou mais dimensões de conhecimento (Pasquali, 2009). O estudo da dimensionalidade é crucial para obter evidências de validade baseadas na estrutura interna do instrumento, favorecendo uma interpretação adequada dos resultados (Sisto, 2005).

ii) Ajuste dos itens: a análise estatística do ajuste dos itens é realizada a partir dos valores *Mean Square Statistic* (MNSQ), por meio dos índices *infit* e *outfit*. De acordo com Xavier (2018), tais índices “demonstram a relação entre o escore esperado pelo modelo e o escore bruto obtido empiricamente”. Esses índices permitem identificar itens que não se ajustam ao modelo, ou seja, que apresentam um padrão de resposta diferente do esperado. Um exemplo ocorre quando itens difíceis, que supostamente deveriam ser acertados por indivíduos com alta habilidade, são errados por eles e acertados por indivíduos com baixa habilidade.

iii) Confiabilidade: a confiabilidade é avaliada em termos da reprodutibilidade das medidas e da consistência interna do teste. Ela é reportada pelo coeficiente alfa de Cronbach, índice clássico da literatura, e pelos índices de separação de pessoas e de itens obtidos por meio da Modelagem Rasch.

iv) Adequação ao modelo: essa análise avalia se os dados observados se ajustam ao Modelo Rasch como um todo, verificando se a probabilidade de acerto em cada item depende apenas da habilidade do respondente e da dificuldade do item. Essa avaliação permite determinar se o comportamento dos itens e dos respondentes está em conformidade com as probabilidades previstas pelo modelo esperado, garantindo que as medidas obtidas

sejam válidas e interpretáveis, confirmando que o instrumento mede adequadamente o construto de interesse.

b) Itens discursivos

As respostas aos itens discursivos foram analisadas qualitativamente, tomando a perspectiva docente como referência. Adotamos como parâmetro a concepção acadêmica desses conceitos para avaliar o nível de entendimento dos estudantes.

“Consideramos que o estudante demonstra um entendimento mais avançado quando é capaz de mobilizar um conteúdo de forma coerente, tanto em termos procedimentais quanto de conceitualização” (Amantes, 2009). Para examinar a compreensão desses estudantes do Ensino Médio sobre corrente elétrica e circuitos elétricos, utilizamos as Categorias de Explicitação, sistema desenvolvido por Amantes (2009), cuja concepção também é adotada em outros estudos (Saglam-Arslan, 2010; Oliveira; Amantes, 2021; Gadéa; Amantes, 2019).

De acordo com esse sistema, as respostas dos alunos podem ser classificadas em três categorias: Explícito, Parcialmente Explícito e Não Explícito. Esse sistema incorpora níveis de complexidade empregados em outras taxonomias, como o Sistema de Habilidades Dinâmicas de Fischer (1980) e a Taxonomia SOLO de Biggs e Collis (1982). Entretanto, em vez de se basear em pressupostos teóricos do desenvolvimento, o Sistema de Explicitação atribui os níveis de acordo com a proximidade em relação à concepção acadêmica, ou seja, avalia a exatidão das respostas com base em um parâmetro formal previamente estabelecido.

Resultados

Itens dicotômicos

Dimensionalidade e validade

Para verificar a dimensionalidade, analisam-se a variância e a correspondência entre as previsões do modelo e os dados observados, expressa pelos resíduos. A hipótese inicial é a de que é possível diferenciar pelo menos três tipos de conhecimento: procedimental, conceitual e aplicação de ferramentas matemáticas. É importante destacar que, embora o banco de itens construído aborde os conteúdos de corrente elétrica e circuitos elétricos, contemplando itens de natureza conceitual, procedimental e de aplicação de ferramentas matemáticas, é possível que esses itens mensurem outras dimensões ou diferentes composições de múltiplas habilidades (Xavier, 2018).

Tabela 1: Variância residual padronizada e variância explicada pelo primeiro contraste

Banco de itens	Empírico	Modelado
Variância bruta explicada pelas medidas	24,0%	24,0%
Variância bruta explicada pelas pessoas	7,5%	7,6%
Variância bruta explicada pelos itens	16,4%	16,5%
Variância não explicada pelo 1º contraste	2,9	5,9%

Fonte: Produção dos autores (2024).

A Tabela 1 apresenta a variância residual padronizada e a variância bruta explicada pelas pessoas, pelas medidas e pelos itens, constituindo uma síntese dos dados gerados pela Modelagem Rasch a respeito da dimensionalidade do banco de itens aplicado. Segundo Linacre (2009), quando os valores empíricos e modelados da variância bruta explicada pelas

medidas são próximos, isso constitui um indício de que o Modelo Rasch é uma estratégia adequada para avaliar os dados obtidos pelo teste. É possível observar, na Tabela1, a igualdade entre as porcentagens dos valores empíricos e modelados das variâncias brutas explicadas pelas medidas, pelas pessoas e pelos itens.

Uma das propriedades para que uma escala seja considerada unidimensional é apresentar autovalor do primeiro contraste menor ou igual a 2,0 (Almiro e Simões, 2015). Analisando o resultado da variância não explicada no primeiro contraste, apresentado na Tabela 1, observa-se um autovalor de 2,9, acima do limite estabelecido (2,0). Esse valor constitui um indicativo de que o teste possivelmente não seja unidimensional, ou seja, de que não esteja acessando apenas uma única dimensão correspondente ao conhecimento geral sobre corrente elétrica e circuitos elétricos.

Outro aspecto relacionado à dimensionalidade é a identificação da dependência entre os itens. Valores de correlação inferiores a 0,25 indicam possível unidimensionalidade (Linacre, 2009). A Tabela 2 apresenta as correlações residuais padronizadas utilizadas para identificar dependência entre os itens. Observa-se que todos os valores, em módulo, são superiores a 0,25, indicando elevada correlação entre os itens e sugerindo, portanto, a existência de outra dimensão além da principal.

Tabela 2: Correlações residuais padronizadas

Itens		Correlação
4	5	.30
3	4	.30
25	31	.27
15	18	.27
27	34	.25
5	24	-.28
30	34	-.27
5	27	-.26
23	25	-.26
15	26	-.26

Fonte: Produção dos autores (2024).

Interpretamos que, em relação ao aspecto da dimensionalidade, os resultados da modelagem indicam a presença de mais de uma dimensão acessada, as quais podem corresponder às facetas previamente estabelecidas ou a outras facetas que distingam os conhecimentos sobre corrente elétrica e circuitos elétricos. Essa hipótese pode ser confirmada por meio de uma Análise Fatorial Exploratória, que constitui uma etapa complementar do estudo da estrutura do teste.

Confiabilidade e adequação do modelo

O objetivo desta análise é examinar a reprodutibilidade das respostas e a consistência interna do teste. Na Modelagem Rasch, a confiabilidade é analisada a partir do coeficiente alfa de Cronbach e dos índices de separação e confiabilidade das pessoas e dos itens. O alfa de Cronbach é frequentemente referido como o principal estimador de confiabilidade, embora não seja o único (Almeida; Santos; Costa, 2010). Fox e Jones (1998) destacam que a confiabilidade das pessoas depende não apenas de um conjunto de itens adequadamente calibrados para estimar a habilidade, mas também de uma distribuição suficientemente ampla

dessa habilidade entre os respondentes, condição essencial para que as medidas Rasch evidenciem uma hierarquia de capacidade/desenvolvimento (separação de pessoas) no construto avaliado. Assim, a análise combina o alfa de Cronbach com os índices Rasch de separação para oferecer um diagnóstico mais robusto da qualidade do instrumento.

O coeficiente alfa de Cronbach é uma das estatísticas mais amplamente empregadas em pesquisas que envolvem a construção, a validação e a aplicação de instrumentos de medida (Cortina, 1993). Para Streiner (2003), esse coeficiente não é uma propriedade fixa da escala, mas uma característica relacionada ao padrão de respostas da população analisada. Seu valor varia entre 0 e 1 (ou entre 0% e 100%), sendo que coeficientes superiores a 0,7 indicam níveis aceitáveis de confiabilidade das medidas (Martins, 2006; Maroco; Garcia-Marques, 2006).

O índice de confiabilidade das pessoas indica a replicabilidade da ordenação dos participantes que seria esperada caso essa amostra respondesse a outro conjunto paralelo de itens destinados a medir o mesmo construto (Wright; Masters, 1982). O coeficiente de separação entre as pessoas fornece uma estimativa mais precisa, uma vez que o alfa de Cronbach tende a ultrapassar o valor máximo esperado para a confiabilidade das medidas obtidas por meio da Modelagem Rasch (Linacre, 2009). Os índices relativos às pessoas estão apresentados na Tabela 3. O alfa de Cronbach foi de 0,76, indicando bom ajuste quanto à consistência interna.

Tabela 3: Índices para pessoas

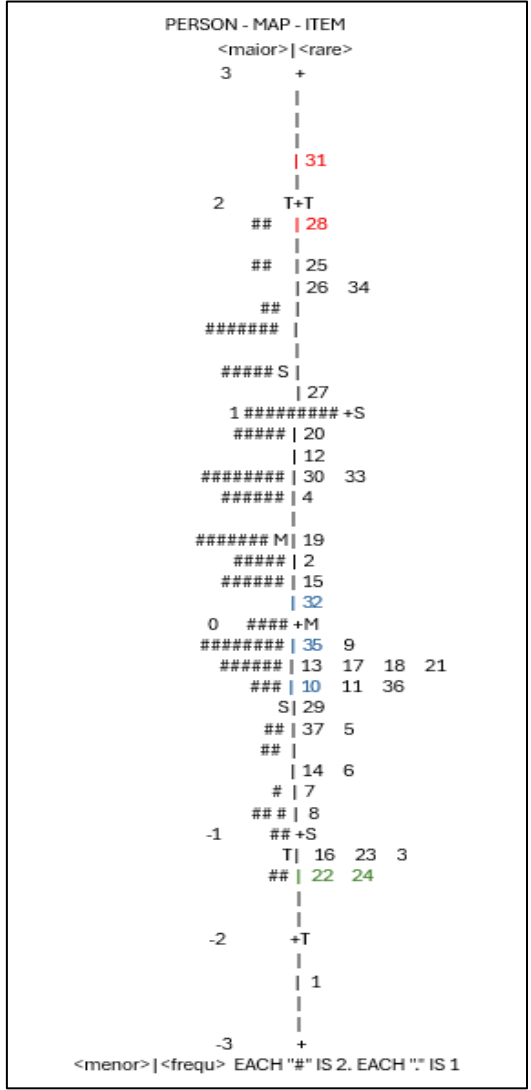
	Alfa de Cronbach		Média	Desvio Padrão	MNSQ infit	T infit	MNSQ outfit	T outfit	Separação	Confiabilidade
Banco de itens	0,76	Item	.00	.99	1.00	.0	.99	.1	5.72	.97
		Pessoa	.43	.78	.99	.0	.99	.1	1.69	.74

Fonte: Produção dos autores (2024).

Na mesma tabela, observa-se o coeficiente de separação dos itens. Segundo Linacre (2017), valores considerados adequados para esses índices incluem separação de itens e de pessoas superior a 2,0, associada a confiabilidade acima de 0,8. Conforme evidenciado na Tabela 3, o índice de separação dos itens apresentou valor de 5,72, excedendo significativamente o parâmetro recomendado pela literatura, com confiabilidade correspondente de 0,97.

Já em relação ao coeficiente de separação das pessoas, que indica a precisão com que a variabilidade dos sujeitos é capturada pelos itens, o valor obtido foi de 1,69, com índice de confiabilidade de 0,74. Valores de separação de pessoas inferiores a 2, geralmente associados a uma confiabilidade das pessoas inferior a 0,8, sugerem que o teste pode não ser adequado para distinguir indivíduos com alta e baixa proficiência (Linacre, 2017). Isso significa que o instrumento pode apresentar limitações para discriminar participantes com diferentes níveis de desempenho. Por outro lado, a hierarquia dos itens mostrou-se confiável, resultado importante para atestar a estrutura do instrumento. Outro parâmetro empregado para avaliar a adequação dos itens e do teste, de modo geral, é a distribuição das medidas dos itens e das pessoas no mapa de distribuição, que fornece uma escala em *logits* e está representado na Figura 1.

Figura 1 – Índice alfa de Cronbach para pessoas



Fonte: Produção dos autores (2024).

No mapa de itens apresentado na Figura 1, os estudantes são organizados de acordo com sua medida de habilidade em relação ao traço latente, enquanto os itens são dispostos segundo seu nível de dificuldade, formando uma hierarquia vertical. Itens de maior dificuldade, como os de números 31 e 28 (em vermelho), bem como participantes com maior habilidade, localizam-se na parte superior do mapa. Já os itens de dificuldade intermediária, como os de números 32, 35 e 10 (em azul), situam-se na faixa central, enquanto os itens de menor dificuldade, como os de números 1, 22 e 24 (em verde), aparecem na parte inferior.

Espera-se, nesse tipo de análise, uma distribuição aproximadamente normal que cubra toda a faixa de habilidades dos participantes e de dificuldades dos itens. Isso indica, por um lado, que a amostra é heterogênea e, por outro, que o instrumento é capaz de captar essa heterogeneidade, dimensionando tanto os níveis mais elevados quanto os mais baixos de habilidade relacionados ao traço investigado. No presente estudo, a média de proficiência dos participantes mostrou-se superior à média de dificuldade do teste (+M > M), o que sugere que o banco de itens era relativamente “fácil” para essa amostra específica. Ainda assim, a distribuição geral de itens e participantes revelou-se adequada, corroborando a consistência e a aplicabilidade do instrumento.

Confiabilidade e adequação do modelo

Tabela 4: Índice de ajustes dos itens

Banco de itens		
Item	Infit (MNSQ)	Outfit (MNSQ)
1	1,01	1
2	1,01	1
3	1,09	1,02
4	1,13	1,15
5	1,18	1,21
6	1,1	1,16
7	0,85	0,75
8	0,92	0,83
9	0,99	0,96
10	1,11	1,13
11	1,01	0,97
12	1,16	1,22
13	0,95	0,95
14	0,99	0,89
15	1,03	1,01
16	0,98	0,96
17	1,01	1,01
18	0,9	0,88
19	0,93	0,91
20	1,02	1,07
21	0,85	0,8
22	0,85	0,74
23	0,82	0,66
24	0,81	0,6
25	1,18	1,38
26	1,02	0,97
27	1	0,97
28	1,25	1,51
29	0,93	0,98
30	1,02	1,03
31	1,1	1,22
32	0,87	0,84
33	0,9	0,9
34	1,08	1,06
35	1,06	1,07
36	1,02	1,02
37	0,89	0,84
MNSQ	Infit	Outfit
Média	1,00	0,99
(SD)	0,11	0,18

Fonte: Produção dos autores (2024).

Espera-se que os itens do banco discriminem adequadamente diferentes níveis de conhecimento. No entanto, a análise da confiabilidade sugeriu uma possível limitação nessa

distinção entre indivíduos com altas e baixas proficiências. Essa evidência pode estar relacionada ao tamanho da amostra ou à presença de itens específicos que interferem nas estimativas de desempenho. Para aprofundar essa análise, recorre-se às estatísticas de ajuste dos itens.

A análise estatística do ajuste dos itens considera os valores de *Mean Square Statistic* (MNSQ), por meio dos índices *infit* e *outfit*. A MNSQ é uma estatística de qui-quadrado dividida por seus graus de liberdade, e seus valores expressam a relação entre o escore esperado pelo modelo e o escore bruto obtido empiricamente (Xavier, 2018, p. 67). De acordo com Wright e Stone (2004), a análise dos índices *infit* e *outfit* permite identificar itens que apresentam padrões de resposta inesperados. Wright e Linacre (1994) ressaltam que os valores das medidas de ajuste, MNSQ *infit* e *outfit*, devem situar-se no intervalo considerado adequado, entre 0,5 e 1,5.

A Tabela 4 apresenta os índices de ajuste dos itens com base nos valores de MNSQ *infit* e *outfit*. Dos 37 itens, apenas o item 28 apresentou índice fora do intervalo de ajuste. Para o banco de itens analisado, as médias dos índices de *infit* e *outfit* foram de 1,0 ($\pm 0,11$) e 0,99 ($\pm 0,18$), respectivamente, indicando que os itens, em seu conjunto, estão bem ajustados.

Itens discursivos

Quadro 2: Categorização dos itens discursivos

i39- O que será menos perigoso: ligar um aparelho de 110V a uma tomada de 220V ou ligar um aparelho de 220V a uma tomada de 110V? Justifique.		
Resposta do estudante	Nível de complexidade	Descrição do nível
Ligar um aparelho de 110V a uma tomada de 220 V.	0	Não Explícito: resposta equivocada, com erro de conceitos científicos, em que o estudante responde que ligar um aparelho de 220V a uma tomada de 110V é mais perigoso.
Ligar um aparelho de 220V em uma tomada de 110V	1	Parcialmente Explícito: o estudante reconhece que ligar um aparelho de 220V em uma tomada 110V é menos perigoso. Ou que ligar um aparelho de 110V em uma tomada de 220V é mais perigoso.
É menos perigoso ligar um aparelho de 220V em 110V, pois só causa mau funcionamento, apenas fica mais fraco. Quando um aparelho de 110V é ligado a uma tomada de 220V, é mais perigoso, pois, quando é conectado a uma voltagem maior do que a sua classificação, pode resultar em danos ao aparelho (como queimar) e até representar riscos de incêndio.	2	Explícito: o estudante reconhece que ligar um aparelho de 220V em uma tomada de 110V é menos perigoso. Ou que ligar um aparelho de 110V em uma tomada de 220V é mais perigoso. Justifica como funcionará esse aparelho e explica sobre a possibilidade de queimar ou danificar o aparelho.

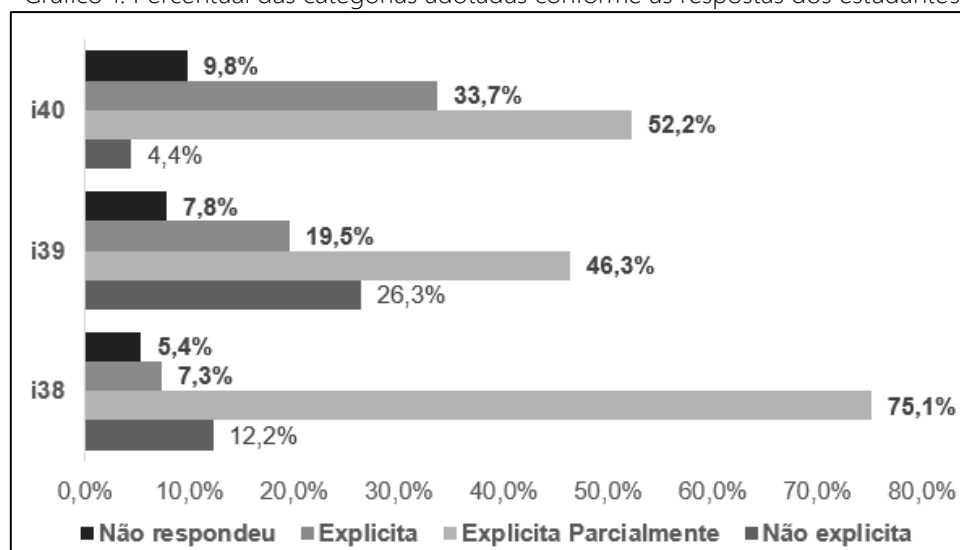
Fonte: Produção dos autores (2024).

As respostas aos itens i38 a i40 foram analisadas utilizando-se três categorias de avaliação: Explícita, Parcialmente Explícita e Não Explícita (Amantes, 2009). A categoria Explícita refere-se a um entendimento articulado, no qual as ideias são relacionadas de maneira coerente e adequada aos parâmetros acadêmicos. A categoria Parcialmente Explícita indica que o estudante possui algum entendimento do que foi perguntado, mas ainda apresenta equívocos ou uma compreensão limitada em termos das relações estabelecidas e das concepções generalizadas; ou seja, um entendimento menos articulado. Já a categoria Não Explícita contempla respostas das quais não é possível inferir algum entendimento ou que apresentam compreensão totalmente equivocada.

As respostas dos 205 estudantes foram classificadas de acordo com as três categorias de avaliação supracitadas. O Quadro 2 apresenta um exemplo da categorização da questão i39.

As Categorias de Explicitação constituem um sistema de classificação que pode ser facilmente incorporado às práticas docentes, uma vez que sua lógica se assemelha à empregada pelos professores na correção das atividades dos alunos (Amantes; Oliveira, 2012). Na abordagem adotada neste trabalho, considera-se que o uso apropriado da linguagem acadêmica para explicar corretamente os conteúdos indica um entendimento mais articulado. Foi realizada uma análise da frequência dessas categorias com o objetivo de avaliar sua distribuição em relação a cada uma das questões, considerando que, entre indivíduos que já estudaram o tema, é esperada uma distribuição aproximadamente normal.

Gráfico 1: Percentual das categorias adotadas conforme as respostas dos estudantes



Fonte: Produção dos autores (2024).

Os resultados apresentados no Gráfico 1 demonstram que, de maneira geral, as respostas aos itens i38, i39 e i40 concentram-se na categoria Parcialmente Explícita, o que era esperado. Do ponto de vista da avaliação dos itens, espera-se que, em uma população, o nível de conhecimento se concentre em torno da média.

Observa-se que o item 40 foi o que apresentou maior porcentagem de respostas na categoria de maior complexidade, enquanto o item 38 foi o que apresentou a menor porcentagem nessa categoria. Esse resultado fornece indícios de que o item 38 necessita de ajuste, pois pode conter algum elemento que introduza viés nas respostas dos participantes. Uma possível explicação é que estudantes com maior nível de entendimento tenham

produzido respostas classificadas como parcialmente explícitas por não terem compreendido plenamente o enunciado.

De modo geral, as respostas aos itens apresentaram um padrão consistente, tanto em relação à exatidão quanto aos equívocos observados, o que indica um bom design dos itens. Caso contrário, seria esperada uma maior dispersão das respostas, sugerindo possível falta de compreensão do que está sendo perguntado.

Discussão

As análises realizadas por meio da Modelagem Rasch revelaram que, de maneira geral, o banco de itens, composto por questões dicotômicas e discursivas, apresenta bom ajuste. O item 28 foi o único que apresentou índice de *outfit* MNSQ igual a 1,51, valor que não está em conformidade com o intervalo considerado ideal pela literatura. Segundo Linacre (2009), valores de *infit* e *outfit* entre 1,50 e 2,00 são considerados desajustes moderados, que não comprometem significativamente a qualidade da medida. Nesse contexto, optou-se por manter esse item de múltipla escolha no banco, uma vez que seu impacto sobre a confiabilidade geral do instrumento é limitado.

A análise da dimensionalidade levantou a hipótese de que o conjunto de itens pode estar acessando mais de uma dimensão do conhecimento sobre circuitos elétricos. Essa questão deverá ser investigada em estudos futuros, nos quais os testes de conhecimento serão formulados com base nos resultados obtidos, considerando as dificuldades e as correlações observadas. É fundamental que esses futuros testes contemplem um equilíbrio entre questões fáceis, médias e difíceis, de modo a garantir uma avaliação abrangente das habilidades dos estudantes.

Além da estrutura dimensional do instrumento, os resultados referentes à separação das pessoas merecem consideração. O índice de separação obtido (1,69), associado à confiabilidade das pessoas (0,74), sugere que o banco de itens apresenta capacidade limitada para discriminar estudantes com diferentes níveis de proficiência. Entretanto, esse resultado deve ser interpretado à luz das características da amostra, uma vez que os índices de separação dependem da variabilidade das habilidades dos participantes (Streiner, 2003). No presente estudo, a média de proficiência dos estudantes foi superior à dificuldade média dos itens, indicando que o banco se mostrou relativamente fácil para essa população específica, o que pode ter reduzido o poder discriminativo observado.

Embora a amostra tenha sido composta por estudantes de escolas públicas federais e estaduais, o objetivo deste estudo foi a validação psicométrica do banco de itens, e não a comparação do desempenho entre diferentes contextos escolares. Dessa forma, não foram realizadas análises específicas sobre a influência do tipo de escola na proficiência dos estudantes. Considerando que diferentes contextos educacionais podem influenciar a distribuição das habilidades medidas, recomenda-se que estudos futuros investiguem essa questão utilizando amostras mais amplas e diversificadas, ampliando as evidências de validade do instrumento.

Em relação aos itens discursivos, a análise dos dados revela que o item i39 é o mais difícil, apresentando a maior porcentagem de respostas classificadas como Não Explícita e uma elevada frequência de respostas Parcialmente Explícita. Esse item pode se beneficiar de uma revisão para aprimorar sua clareza ou ajustar seu nível de complexidade. Em contraste, o item i40 é o mais acessível, com a maior porcentagem de respostas classificadas como Explícita,

embora a elevada taxa de não resposta sugira que ele ainda apresente desafios para alguns estudantes. O item i38, por sua vez, concentrou a maior parte das respostas na categoria Parcialmente Explícita, indicando que, embora os estudantes compreendam o item até certo ponto, muitos ainda não conseguem formular uma resposta completamente correta. Essas análises apontam áreas específicas em que os itens podem ser aperfeiçoados ou nas quais os estudantes podem necessitar de maior suporte pedagógico para aprimorar seu entendimento dos conceitos avaliados.

Considerações

Desenvolver instrumentos para acessar traços latentes é uma tarefa complexa e essencial para a realização de análises educacionais de qualidade. Este estudo apresentou a construção e a validação de um banco de 40 itens sobre corrente elétrica e circuitos elétricos, aplicado a 205 estudantes do Ensino Médio. A análise utilizou o Modelo Rasch para os itens dicotômicos e as Categorias de Explicitação para os itens discursivos, avaliando dimensionalidade, ajuste, consistência e hierarquia de dificuldades.

Os resultados apontam para uma possível multidimensionalidade, que deverá ser investigada em estudos futuros. Também foi observada confiabilidade interna adequada ($\alpha > 0,7$), boa separação dos itens e limitações na separação das pessoas. A maioria dos itens apresentou bom ajuste, com exceção do item 28, e as análises forneceram subsídios para o aprimoramento dos itens discursivos. O mapa de itens indicou predominância de questões fáceis, mas também forneceu subsídios para a construção de testes compostos por questões fáceis, médias e difíceis.

Tivemos como resultado a limitação do teste em discriminar pessoas com habilidades diferentes. No contexto do estudo, isso pode ter ocorrido devido à aplicação em escolas muito distintas que possivelmente tiveram instruções muito diferenciadas em relação ao conteúdo. Esse fato leva o teste a ficar muito fácil para um grupo e/ou muito difícil para outro, gerando estimativas de habilidades que não correspondem exatamente a um espectro amplo de uma amostra homogênea. Tal resultado, contudo, não é tão determinante para atestar a coerência interna do teste; ele nos indica que a aplicação do mesmo deve ser melhor direcionada para o público-alvo almejado.

O presente trabalho estabelece uma metodologia de validação que pode ser empregada em diferentes tipos de testes. A discussão e a condução de procedimentos dessa natureza são importantes para dimensionar a confiabilidade dos instrumentos e elevar a qualidade das pesquisas na área. O processo de validação fornece evidências sobre a adequação das ferramentas empregadas no estudo, sejam elas de coleta ou de análise, reduzindo a possibilidade de resultados e interpretações enviesados. Além disso, os resultados obtidos evidenciam a importância de compreender as características da população avaliada durante o processo de validação, uma vez que aspectos como a heterogeneidade/homogeneidade da amostra podem influenciar os indicadores psicométricos, especialmente aqueles relacionados à separação das pessoas.

Referências

- ALMIRO, P. A. A.; SIMÕES, M. R. Estudo da versão portuguesa do EPQ-R: uma aplicação do Modelo Rasch. In: GOLINO, H. G. et al. *Psicometria contemporânea: compreendendo os modelos Rasch*. Vol. I. 1. ed. São Paulo: Casa do Psicólogo, 2015. p. 242-281.
- AMANTES, A. *Contextualização no ensino de Física: efeitos sobre a evolução do entendimento dos estudantes*. 2009. 275f. Tese (Doutorado) – Universidade Federal de Minas Gerais, UFMG/FaE, 2009.
- AMANTES, A.; OLIVEIRA, E. A construção e o uso de sistemas de categorias para avaliar o entendimento dos estudantes. *Ensaio*, Belo Horizonte, v. 14, n. 2, p. 61-79, 2012.
- AMANTES, A.; COELHO, G. R.; MARINHO, R. A medida nas Pesquisas em Educação: empregando o modelo Rasch para acessar e avaliar traços latentes. *Ensaio Pesquisa em Educação em Ciências*, v. 17, n. 3, p. 657-684, 2015.
- BAYERL, Jonathan et al. Principles of signaling pathway modulation for enhancing human naive pluripotency induction. *Cell Stem Cell*, v. 28, n. 9, p. 1549-1565.e12, 2021.
- BIGGS, J.; COLLIS, K. *Evaluating the quality of learning: the SOLO Taxonomy*. New York: Academic Press, 1982. v. 296.
- BOND, T. G.; FOX, C. M. Important Principles of Measurements Made Explicit. In: BOND, T. G.; FOX, C. M. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 2. ed. London: LEA, 2007. p. 15-25.
- BOLLEN, K. *Equações estruturais com variáveis latentes*. Nova Iorque: John Wiley & Sons, 1989. p. 179-225.
- BRITTO, D. S. Práticas de modelagem matemática e dimensões da aprendizagem geométrica. *Revista de Educação em Ciência e Tecnologia*, v. 12, n. 1, p. 169-196, 2021.
- BROWN, T. A. *Confirmatory Factor Analysis for Applied Research*. 2. ed. New York, NY: The Guilford Press, 2015.
- CALHEIROS, F.; MELO, V. F. de; AMANTES, A. O autoconceito em Química de estudantes do Ensino Médio de escolas de Salvador: uma análise exploratória em busca de fatores preditivos. *Revista Exitus*, v. 13, n. 1, p. e023058, 2023. DOI: 10.24065/re.v13i1.2176.
- COELHO, Geide Rosa. *A evolução do entendimento dos estudantes em eletricidade: um estudo longitudinal*. 2011. Tese (Doutorado em Educação) — Universidade Federal de Minas Gerais, Belo Horizonte, 2011.
- COOK, D. A.; BECKMAN, T. J. Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine*, v. 119, n. 2, p. 166.e7-166.e16, 2006.
- CORTINA, J. M. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, v. 78, p. 98-104, 1993.
- CRONBACH, L. J.; GLESER, G. C.; NANDA, H.; RAJARATNAM, N. *The dependability of behavioral measurements: Theory for generalizability of scores and profiles*. New York: John Wiley, 1972.

CUNHA, C. M.; ALMEIDA NETO, O. P. de; STACKFLETH, R. Principais métodos de avaliação psicométrica da validade de instrumentos de medida. *Revista de Atenção à Saúde*, v. 14, n. 47, p. 75-83, 2016. DOI: <https://doi.org/10.13037/ras.vol14n47.3391>. ISSN 2359-4330.

FISCHER, K. W. A theory of cognitive development: the control and construction of hierarchies of skills. *Psychological Review*, v. 87, p. 477-531, 1980. DOI: 10.1037/0033-295X.87.6.477.

FOX, Christine M.; JONES, James A. Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology*, v. 45, n. 1, p. 30-45, 1998. DOI: 10.1037/0022-0167-45.1.30

GADÉA, S. J. S.; AMANTES, A. Avaliando a aprendizagem de estudantes dos anos iniciais sobre flutuação em uma atividade investigativa. *Revista EDUCAmazônia*, Ano 12, v. XXIII, n. 2, p. 21-46, jul./dez. 2019.

GONÇALVES JUNIOR, Wanderley Paulo. *A programação como ferramenta para o ensino de Física: aprendizagem sobre força por meio do Scratch*. 2021. Tese (Doutorado em Ensino, Filosofia e História das Ciências) — Universidade Federal da Bahia, Salvador, 2021.

HAYES, B. E. *Measuring Customer Satisfaction: Survey design, use, and statistical analysis methods*. Milwaukee, Wisconsin: ASQC Quality Press, 1998.

LINACRE, J. M. *A User's Guide to WINSTEPS MINISTEPS Rasch Model Computer Programs*. Chicago, 2006.

MAROCO, J.; GARCIA-MARQUES, T. Qual a fiabilidade do alfa de Conbrach? Questões antigas e soluções modernas? *Laboratório de Psicologia*, Lisboa, v. IV, n. 1, p. 65-90, 2006.

MARTINS, G. A. Sobre confiabilidade e validade. *Revista Brasileira de Gestão de Negócios*, São Paulo, v. VIII, n. 1, p. 1-12, jan. 2006.

MAZUR, E. *Peer Instruction: a revolução da aprendizagem ativa*. Porto Alegre: Penso Editora, 2015.

MELO, Viviane Florentino; AMANTES, Amanda. Mapeando elementos do perfil epistemológico de densidade. *Amazônia: Revista de Educação em Ciências e Matemáticas*, Belém, v. 17, n. 38, p. 153-172, 2021.

OLIVEIRA, Isabela Torres; WARTHA, Edson José. Construção e validação de um instrumento de coleta de dados: percepção de professores de possíveis impactos da pandemia da covid-19 na educação básica. *Amazônia: Revista de Educação em Ciências e Matemáticas*, Belém, v. 20, n. 44, p. 155-169, 2024.

OLIVEIRA, E. A. G.; AMANTES, A. Astronomy teaching in the elementary education in accordance with the new guidelines of the Base Nacional Comum Curricular. *REVES - Revista Relações Sociais*, v. 4, n. 4, p. 12825-12801, 2021.

PASQUALI, L. Validade dos testes psicológicos: será possível reencontrar o caminho?. *Psicologia: teoria e pesquisa*, v. 23, p. 99-107, 2007.

PASQUALI, L. Psicometria. *Revista da Escola de Enfermagem da USP*, v. 43, p. 992-999, 2009.

RAYMUNDO, Valéria Pinheiro. Construção e validação de instrumentos: um desafio para a psicolinguística. *Letras de Hoje*, Porto Alegre, v. 44, n. 3, p. 86-93, set. 2009.

RICHARDSON, M. W. Notes on the rationale of item analysis. *Psychometrika*, v. 1, p. 69-76.

- SAGLAM-ARSLAN, A. Cross-Grade Comparison of Students' Understanding of Energy Concepts. *Journal of Science Educational and Technology*, Netherlands, n. 19, p. 303-313, 2010.
- SHOHAMY, E. The role of language tests in the construction and validation of second-language acquisition theories. In: TARONE, E. E.; GASS, S. M.; COHEN, A. D. (Orgs.). *Research methodology in second language acquisition*. New Jersey: Lawrence Erlbaum Associates publishers, 1994.
- SILVEIRA, F. L.; MOREIRA, M. A.; AXT, R. Validação de um teste para verificar se o aluno possui concepções científicas sobre corrente elétrica em circuitos simples. *Ciência e Cultura*, São Paulo, v. 41, n. 11, p. 1129-1133, nov. 1989.
- SILVEIRA, F. L. Um teste para verificar se o respondente possui concepções científicas sobre corrente elétrica em circuitos simples. In: ROCHA FILHO, J. B. *Física no ensino médio: falhas e soluções*. Porto Alegre: Edipucrs, 2011. p. 61-67.
- SISTO, F. F. Um estudo sobre a dimensionalidade do teste do desenho da figura humana. *Interação em Psicologia*, v. 9, p. 11-19, 2005.
- SOUZA, A. C. de; ALEXANDRE, N. M. C.; GUIRARDELLO, E. de B. Propriedades psicométricas na avaliação de instrumentos: avaliação da confiabilidade e da validade. *Epidemiol. Serv. Saúde*, Brasília, v. 26, n. 3, p. 649-659, 2017.
- STREINER, D. L. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. *Journal of Personality Assessment*, v. 80, p. 217-222, 2003.
- TERWEE, C. B.; BOT, S. D.; BOER, M. R. de et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.*, v. 60, n. 1, p. 34-42, jan. 2007.
- VERDÚ-SORIANO, J.; TORRES, H. G. de la. Aplicación del análisis Rasch en la investigación enfermera. *Enfermería Clínica*, 35, 2024.
- WRIGHT, B.; MASTERS, G. N. *Análise de escala de classificação*. MESA Press, 1982.
- WRIGHT, B.; STONE, M. Item banking. In: WRIGHT, B.; STONE, M. *Measurement Essentials*. 2. ed. Wilmington: Wide Range, 1999. p. 107-117.
- XAVIER, A. P. *Laboratório Virtual Versus Laboratório Material: A Aprendizagem de Física com Intervenções Tradicionais e Investigativas*. Tese (Doutorado em Ensino, Filosofia e História das Ciências) – Universidade Federal da Bahia, Salvador, 2018.
- XAVIER, A. P.; AMANTES, A. Validação de roteiro estruturado para sua utilização como instrumento de intervenção no laboratório didático de física. *Brazilian Journal of Development*, v. 7, n. 9, p. 91917-91927, 2021.