ANÁLISE ESTATÍSTICA DE ITENS DE TESTES

(TCT E TRI)

Denilson Junio Marques Soares

Sumário

inep
MINISTÉRIO DA EDUCAÇÃO
GOVERNO FEDERAL
BRASIL
UNIÃO E RECONSTRUÇÃO

APRESENTAÇÃO

É com grande satisfação que apresentamos o curso "Análise Estatística de Itens de Testes (TCT e TRI)", parte do projeto "Café com Medidas: Introdução ao Desenvolvimento de Testes Educacionais em Larga Escala" promovido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) por meio da equipe técnica da Coordenação-Geral de Medidas da Educação Básica (CGMEB/DAEB).

A avaliação educacional desempenha um papel crucial no desenvolvimento de políticas públicas e na melhoria da qualidade da educação em nosso país. Ela permite analisar o desempenho dos estudantes, identificar lacunas no aprendizado, avaliar a eficácia das políticas educacionais e direcionar melhorias no sistema educacional. Um componente fundamental da avaliação educacional são os exames padronizados, que fornecem dados essenciais para essas análises.

Para garantir que esses exames sejam precisos, válidos e justos, é fundamental a aplicação de técnicas estatísticas de avaliação da qualidade, tanto no processo de construção dos testes, quanto na análise dos itens que o compõem. Neste curso, focaremos em duas teorias, oriundas da Psicometria, que oferecem subsídios para tal: a Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI). Nosso objetivo é oferecer uma base sólida para que você, participante, desenvolva competências essenciais para compreendê-las.

Agradecemos por sua participação e desejamos um curso produtivo e enriquecedor.

Prof. Dr. Denilson Junio Marques Soares

Comissão de Análises de Dados de Testes e Questionários da Educação Básica em Larga Escala

1. INTRODUÇÃO

No cenário educacional contemporâneo, os exames padronizados desempenham um papel fundamental no monitoramento da qualidade da educação e na formulação de políticas públicas. No Brasil, o Exame Nacional do Ensino Médio (Enem), o Sistema de Avaliação da Educação Básica (Saeb) e o Exame Nacional para Certificação de Competências de Jovens e Adultos (Encceja) são exemplos de iniciativas que utilizam a análise estatística de itens de testes para aferir o desempenho escolar dos estudantes em diferentes níveis de ensino.

A Psicometria é a base metodológica que sustenta essas avaliações. Trata-se de uma área que une Psicologia e Estatística com o objetivo de desenvolver instrumentos de medida para avaliar o conhecimento e o comportamento humano. Dentro deste campo, destacam-se duas teorias para a análise de itens de testes: a Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI). A TCT concentra-se na investigação das propriedades do conjunto de itens que compõem um teste, analisando o desempenho geral do instrumento. Em contrapartida, a TRI examina as propriedades de cada item individualmente, permitindo uma análise mais detalhada de cada pergunta.

Cada uma dessas teorias oferece um conjunto de ferramentas e abordagens para garantir a precisão e a validade das avaliações, contribuindo para a construção de instrumentos de medida confiáveis. A TCT é mais simples de aplicar e interpretar, contudo possui limitações em sua capacidade de fornecer informações detalhadas sobre itens específicos. A TRI, por sua vez, oferece análises mais precisas, porém é mais complexa e requer recursos computacionais específicos e amostras de tamanho adequado. Quando utilizadas em conjunto, a TCT e a TRI oferecem um amplo espectro de possibilidades de análises, enriquecendo a avaliação e a interpretação dos testes.

Este curso abordará essas teorias de forma introdutória, discutindo suas origens, conceitos básicos, pressupostos, modelos, procedimentos de análise de itens, aplicações práticas e limitações. Para ilustrar os conceitos e técnicas discutidos, utilizaremos gráficos construídos no software estatístico R (R CORE TEAM, 2024). Os códigos-fonte utilizados e as imagens geradas estão disponíveis em: https://github.com/denilsonjms/Curso_Inep.

2. TEORIA CLÁSSICA DOS TESTES (TCT)

A Teoria Clássica dos Testes (TCT) tem suas raízes no início do século XX, com contribuições fundamentais de Charles Spearman, que introduziu o conceito de fator geral na psicometria, e L.L. Thurstone, com suas análises fatoriais. A Teoria foi consolidada sobretudo com os trabalhos de Guilford (1936; 1954), Gulliksen (1950) e Nunnally (1967) e, desde então, é uma das abordagens mais utilizadas para a construção e análise de testes e questionários. Ela se baseia na ideia de que qualquer escore observado em um teste é composto por um escore verdadeiro e um erro de medição, ambos componentes que precisam ser compreendidos e controlados para que os testes sejam válidos e fidedignos.

Na TCT, o escore bruto (X) é o escore obtido pelo indivíduo em um teste. Ele é composto por dois componentes principais: o escore verdadeiro (T), que representa a habilidade real do indivíduo, e o erro de medição (E), que inclui todas as variações não sistemáticas que afetam a pontuação. Matematicamente, isso pode ser representado pela fórmula:

$$X = T + E$$

O escore verdadeiro (T) é a média dos escores que um indivíduo obteria se pudesse realizar o mesmo teste um número infinito de vezes, sob condições idênticas. Já o erro de medição (E) é a diferença entre o escore bruto e o escore verdadeiro, refletindo todas as influências que não são consistentes e que podem incluir fatores como cansaço, motivação, condições de teste, entre outros (PASQUALI, 2003).

A TCT assume que o erro de medição é aleatório e, portanto, sua média é zero. Isso implica que, em média, o erro não afeta a estimativa do escore verdadeiro. Essa relação entre os componentes é crucial para entender a precisão e a validade dos testes. O objetivo é minimizar o erro (E) para que o escore bruto (X) seja o mais próximo possível do escore verdadeiro (T).

2.1 ANÁLISE DE ITENS PELA TCT

Na TCT, diversas informações são consideradas para avaliar a adequação de um item e seu comportamento em relação aos outros itens do teste. Nesta seção, definiremos e exploraremos alguns dos parâmetros psicométricos essenciais que são utilizados para verificar a precisão, validade e confiabilidade de um teste. A análise desses parâmetros é fundamental para a melhoria contínua dos instrumentos de avaliação, assegurando que eles sejam eficazes e justos na medição das habilidades dos indivíduos.

2.1.1 Estatísticas Descritivas

É comum iniciar a análise de um conjunto de dados pelas estatísticas descritivas, e o mesmo se aplica aos itens de um teste. Geralmente, quanto maior a variabilidade do item e quanto mais a média do item estiver no ponto central da distribuição, melhor será o item (KLINE, 2005). Outras análises descritivas como amplitudes, percentis, quartis e decis, e medidas de distribuição (achatamento e assimetria de distribuição), bem como análises gráficas (histogramas, blox-plots, etc.) também podem ser realizadas como instrumentos na avaliação de itens.

2.1.2 Índice de Dificuldade

Conforme Borgatto e Andrade (2012), a dificuldade de um determinado item é definida em termos da proporção de respondentes que o acertam. Assim:

$$\text{Índice de Dificuldade} = \frac{(\text{Número de acertos no item})}{(\text{Número total de respondentes})}$$

Segundo Condé (2001), considera-se o item como fácil quando o seu índice de dificuldade for superior a $0{,}70$, moderado quando estiver entre $0{,}30$ e $0{,}70$, e difícil quando for inferior a $0{,}30$. Pasquali (2003), por sua vez, apresenta mais dois níveis para essa classificação e recomenda um percentual esperado para os índices de dificuldade, conforme indicado na Tabela 1.

Tabela 1 – Distribuição ideal de Itens com base no Índice de Dificuldade

Classificação do item	Índice de Dificuldade	Percentual esperado de itens na avaliação
Muito Fáceis	Superior à 0,9	10%
Fáceis	Entre 0,7 e 0,9	20%
Medianos	Entre 0,3 e 0,7	40%
Difíceis	Entre 0,1 e 0,3	20%
Muito Difíceis	Inferior à 0,1	10%

Fonte: Adaptado de Pasquali (2003).

2.1.3 Discriminação dos Itens

A discriminação de um item avalia a capacidade de diferenciar indivíduos com bom desempenho daqueles com baixo rendimento no mesmo teste. Esse parâmetro é crucial para garantir que um teste possa efetivamente separar os respondentes com diferentes níveis de habilidade ou conhecimento. Para avaliar a discriminação de um item, existem várias abordagens estatísticas que podem ser utilizadas:

2.1.3.1 Criação de Grupos-Critério:

Nesta abordagem, os respondentes são divididos em grupos com base em seus desempenhos gerais no teste. Kelley (1939) propôs que essa separação fosse feita utilizando os 27% superiores e os 27% inferiores dos participantes. Com base nessa divisão, Pasquali (2003) sugere que a discriminação do item seja calculada como a diferença na proporção de acertos entre o grupo de alto desempenho e o grupo de baixo desempenho. Segundo Arias, Lloreda e Lloreda (2006), uma boa referência para a classificação da qualidade discriminativa de um item é a descrita na Tabela 2.

Tabela 2 – Classificação dos itens, conforme o índice de discriminação

Índice de Discriminação	Classificação do item
Inferior à 0,20	Item ineficiente (deve ser eliminado ou totalmente revisado)
Entre 0,20 e 0,30	Item necessita ser revisado
Entre 0,30 e 0,40	Item aceitável
Superior à 0,40	Item idealizado.

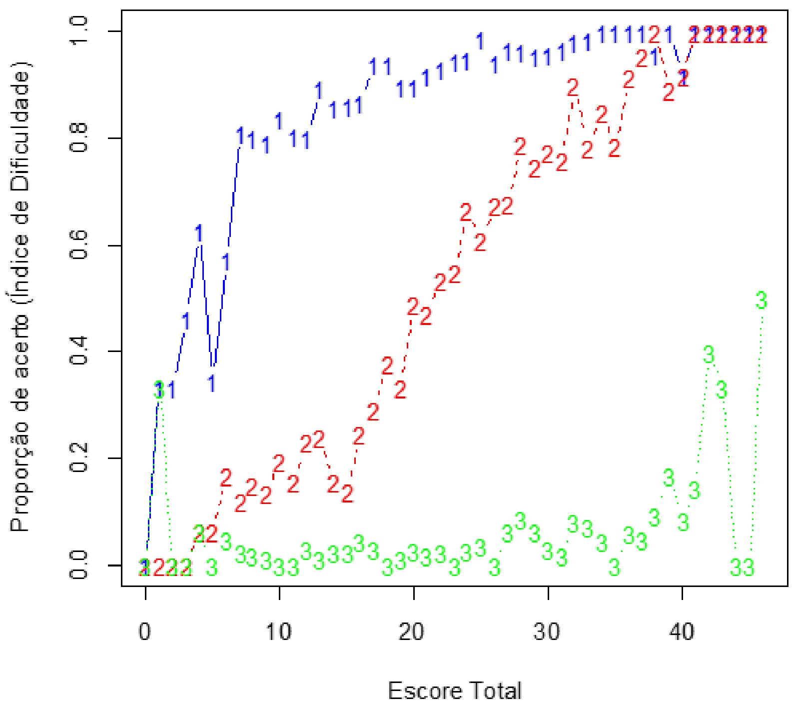
Fonte: Adaptado de Arias, Lloreda e Lloreda (2006).

Também se pode, mediante a aplicação do teste $t\ de\ Student$ para duas médias de amostras independentes (BUSSAB; MORETTIN, 2004), averiguar se um item é considerado não discriminativo ou inadequado, ao nível de significância α. Isto ocorrerá caso o valor da estatística de teste estiver na região de não rejeição da hipótese nula, indicando que as médias dos grupos inferior e superior são estatisticamente iguais.

2.1.3.2 Análise Gráfica do Item (AGI)

Outra maneira de medir o poder de discriminação de um item é por meio da análise gráfica da proporção de acertos dos itens do teste em relação ao escore total. Este método foi desenvolvido por Rasch (1960), que plotou os escores totais de um teste em função das taxas de acerto em itens cognitivos. A Figura 1 ilustra a análise de três itens do teste.

Figura 1 – Escore total versus proporção de acerto para três itens de um teste



Fonte: Elaborado pelo autor (2024).

Observe que o item 1 tem um grau de dificuldade muito baixo, pois, independentemente do escore do respondente, a proporção de acertos permanece próxima de 1. Em contrapartida, o item 3 apresenta um grau de dificuldade elevado, pois para obter uma boa proporção de acertos, é necessário que o respondente tenha obtido um escore alto no teste. Esses itens não possuem um bom poder discriminativo, pois a proporção de acertos não varia significativamente entre diferentes grupos-critério. Por outro lado, o item 2 se destaca em termos de discriminação, pois apresenta diferentes proporções de acertos para respondentes com diferentes escores.

2.1.3.3 Correlação com o Escore Total Observado

A correlação de cada item com o escore total do teste é outra maneira de avaliar a discriminação. Itens que têm alta correlação com o escore total são considerados bons discriminadores, pois indicam que o desempenho no item é consistente com o desempenho geral no teste. Esta correlação é frequentemente calculada utilizando a correlação ponto-bisserial (r_{pb}), que é adequada para dados binários (correto/incorreto). A fórmula é dada por:

$$r_{pb} = \frac{M_1 - M_0}{s} \sqrt{\frac{pq}{n}}$$

Em que:

- M_1 = média dos escores da variável contínua para o grupo que possui o valor 1 na variável dicotômica
- M_0= média dos escores da variável contínua para o grupo que possui o valor 0 na variável dicotômica
- s = desvio padrão total dos escores da variável contínua
- p = proporção de observações com o valor 1 na variável dicotômica
- q = proporção de observações com o valor 0 na variável dicotômica ($q = 1 - p$)
- n = número total de observações

De acordo com Tôrres (2015), de maneira geral, espera-se que o coeficiente de correlação ponto-bisserial assuma valores positivos e superiores a 0,30 para que sejam considerados de boa discriminação.

2.2 FIDEDIGNIDADE DOS TESTES

A fidedignidade de um teste refere-se à consistência/estabilidade dos resultados que ele produz. Em outras palavras, é a capacidade do teste de medir algo de forma consistente ao longo do tempo, entre diferentes itens do teste ou entre diferentes avaliadores. Existem várias maneiras de atestá-la, sendo as mais comuns: o método teste-reteste, a correlação entre duas metades e a consistência interna.

a) Teste-Reteste

O método teste-reteste avalia a estabilidade de um teste ao longo do tempo. Consiste em aplicar o mesmo teste a um grupo de indivíduos em dois momentos distintos e, em seguida, calcular a correlação entre os resultados das duas aplicações. Se o teste for fidedigno, os resultados devem ser semelhantes em ambas as ocasiões, indicando que a medida é estável ao longo do tempo. No entanto, este método pode ser influenciado pela memória dos participantes ou por mudanças reais nas características medidas entre os dois momentos.

b) Correlação entre Duas Metades

Este método divide o teste em duas metades equivalentes e calcula a correlação entre os escores obtidos em cada uma delas. A correlação é então ajustada por uma fórmula de Spearman-Brown para estimar a fidedignidade do teste completo. Ele é útil para verificar a consistência interna, mas a divisão em duas metades pode não ser sempre prática ou ideal, especialmente se o teste não for suficientemente longo.

c) Consistência Interna

A consistência interna mede a coerência entre os itens de um teste. O coeficiente Alfa de Cronbach é uma das medidas mais utilizadas para avaliá-la, considerando a variância dos itens individuais e a variância total do teste. Matematicamente, o coeficiente Alfa de Cronbach (α) é calculado pela fórmula:

$$\alpha = \frac{N}{N+1}\left(1 - \frac{\sum s_i^2}{\sum s_t^2}\right)$$

em que:

- N = número de itens no teste
- s_i^2 = variância dos escores de cada item
- s_t^2 = variância total dos escores do teste

Em síntese, quando os itens são mais homogêneos, o valor de α será maior, indicando maior consistência interna. O valor do coeficiente Alfa de Cronbach varia de 0 a 1, contudo, é importante considerar que um coeficiente muito alto ($\alpha \geq 0{,}95$) pode indicar redundância entre os itens, sugerindo que alguns deles podem ser removidos sem perda significativa de informação (SOARES, 2018).

A interpretação típica dos valores de α é a seguinte:

- $\alpha \geq 0{,}9$: Excelente
- $0{,}8 \leq \alpha < 0{,}9$: Bom
- $0{,}7 \leq \alpha < 0{,}8$: Aceitável
- $0{,}6 \leq \alpha < 0{,}7$: Questionável
- $0{,}5 \leq \alpha < 0{,}6$: Pobre
- $\alpha < 0{,}5$: Inaceitável

Apesar de ser a medida mais utilizada para avaliar a consistência interna de um teste, o coeficiente Alfa de Cronbach apresenta várias limitações que podem comprometer a precisão das estimativas de fidedignidade. Uma delas é o pressuposto de tau-equivalência, que implica que todos os itens de um teste têm a mesma importância ou peso para o fator subjacente que está sendo medido. Em outras palavras, cada item é considerado igualmente confiável e contribui de forma igual para a medida total.

No entanto, na prática, essa suposição raramente se sustenta, pois os itens podem variar significativamente em suas cargas fatoriais.

Outra limitação do coeficiente Alfa de Cronbach é o fato de ser um estimador *lower-bound* (limite inferior) para a fidedignidade. Isso significa que os valores de Alfa de Cronbach tendem a ser menores do que os valores obtidos por técnicas mais adequadas de estimativa de fidedignidade. Essa característica pode resultar em uma subestimação da verdadeira consistência interna do teste. Assim, técnicas mais avançadas, como a Fidedignidade Composta e o Ômega de McDonald, tem ganhado maior notoriedade por fornecer estimativas mais precisas da consistência interna de um teste.

2.3 LIMITAÇÕES DA TCT

Desde sua formulação inicial a TCT tem sido apreciada por sua simplicidade, facilidade de aplicação e por fornecer uma base sólida para a compreensão das propriedades dos testes. No entanto, essa teoria também apresenta algumas limitações que precisam ser consideradas e que limitam sua generalização. Assim, as conclusões tiradas de um teste específico podem não ser aplicáveis a outros contextos ou populações, restringindo a utilidade dos resultados para tomadas de decisão mais amplas. Dentre essas limitações, com base em Pasquali (2020), destacamos:

a) Dependência dos Itens do Teste (*Test-Dependent*)

Uma das principais limitações da TCT é que os resultados de um teste são altamente dependentes dos itens que o compõem. Isso significa que as características psicométricas de um teste podem mudar significativamente se os itens forem alterados. Portanto, a validade e a fidedignidade do teste podem ser comprometidas se houver mudanças nos itens, dificultando a comparação entre diferentes versões do teste.

b) Dependência dos Sujeitos (*Subject-Dependent*)

Os parâmetros dos itens de um teste (como dificuldade e discriminação) são calculados com base na amostra de sujeitos que realizou o teste. Nesse caso, esses parâmetros podem variar de uma amostra para outra. A dependência dos sujeitos implica que os resultados de um teste podem não ser generalizáveis para outras populações, limitando a aplicabilidade dos resultados a diferentes grupos de respondentes.

c) Problemas no Cálculo da Fidedignidade

A fidedignidade de um teste na TCT é frequentemente definida em termos de formas paralelas do teste. Para que duas formas sejam consideradas estritamente paralelas, elas devem produzir escores verdadeiros idênticos e variâncias iguais. Na prática, criar formas paralelas que atendam rigorosamente a esses critérios é extremamente difícil, o que pode comprometer a avaliação da fidedignidade do teste.

d) Suposição de Erro de Medida Constante

A TCT assume que a variância dos erros de medida é a mesma para todos os respondentes. Na realidade, essa suposição raramente se verifica, pois os erros de medida podem variar entre indivíduos com diferentes níveis de habilidade ou em diferentes condições de teste. Essa limitação pode levar a estimativas imprecisas da fidedignidade e validade dos testes.

e) Problemas com a Escala de Medida

A TCT trabalha com escores observados que são resultados de somas de itens. Esses escores assumem que cada item contribui igualmente para o escore total, o que nem sempre é verdadeiro. Além disso, a TCT não leva em consideração a relação não-linear entre os itens e a habilidade subjacente.

3 TEORIA DE RESPOSTA AO ITEM (TRI)

A Teoria de Resposta ao Item (TRI) surgiu na década de 1950 por pioneiros como Frederic Lord e Georg Rasch e se refere à uma família de modelos matemáticos que buscam explicar a relação entre o nível de habilidade ou traço latente (θ) e o padrão de resposta aos itens de um determinado teste ou escala. Ao contrário da TCT, cujo foco principal é o teste como um todo, a TRI concentra-se nos itens individuais.

A TRI trouxe uma abordagem complementar à TCT, permitindo superar suas limitações. Uma das suas principais vantagens é a propriedade de invariância da medida, o que significa que as estimativas da habilidade dos indivíduos são independentes dos itens específicos utilizados no teste. Da mesma forma, os parâmetros dos itens (como a dificuldade e a discriminação) são invariantes em relação às habilidades dos indivíduos que os respondem. Isso contrasta com a TCT, cujas estatísticas dos itens podem variar significativamente entre diferentes amostras.

Atualmente, a TRI tem sido utilizada em diversos campos, incluindo a Educação, a Psicologia e a Avaliação Profissional. Um dos exemplos mais notáveis de avaliação utilizando a TRI é o exame de proficiência em língua inglesa (Toefl). Outro exame de grande importância é o SAT (*Scholastic Aptitude Test ou Scholastic Assessment Test*), um exame educacional padronizado nos Estados Unidos, aplicado a estudantes do Ensino Médio e utilizado como critério de admissão nas universidades norte-americanas. No Brasil, sua aplicação em avaliações educacionais começou com o Saeb em 1995, e posteriormente no Encceja e Enem. Internacionalmente, a metodologia da TRI é adotada pelo Programa Internacional de Avaliação de Estudantes (Pisa) e em sistemas de avaliação educacional de diversos países como França, Holanda, Coreia do Sul e China.

3.1 PRESSUPOSTOS DA TRI

A TRI baseia-se em dois pressupostos fundamentais: a unidimensionalidade e a independência local. A unidimensionalidade refere-se à ideia de que um conjunto de itens em um teste mede um único traço latente ou habilidade subjacente. Esse traço latente é representado pela variável θ, e todos os itens do teste devem estar relacionados a essa única dimensão. A unidimensionalidade é importante para garantir que as estimativas de habilidade sejam significativas e interpretem corretamente a relação entre os itens e o traço latente (HAIR JR *et al.*, 2010).

O pressuposto de independência local, por sua vez, postula que as respostas aos itens de um teste são independentes umas das outras, quando controlamos o traço latente. Isso significa que, para indivíduos com o mesmo nível de habilidade (θ), a probabilidade de responder corretamente a um item não deve ser influenciada pelas respostas a outros itens do teste. A independência local garante que a correlação entre as respostas aos itens seja atribuída exclusivamente ao traço latente medido pelo teste, e não a fatores externos ou interações entre os itens (EMBRETSON; REISE, 2013).

Segundo Pasquali e Primi (2003), tais pressupostos não podem ser empiricamente demonstrados e nem possuem bases lógicas; eles são simplesmente aceitos ou não aceitos, ou seja, são axiomas. Além disso, conforme Hambleton, Swaminathan e Rogers (1991), a unidimensionalidade implica independência local e, de acordo com Lord e Novick (1968), a independência local também implica unidimensionalidade. Dessa forma, na prática, há apenas uma suposição a ser verificada. Isso simplifica a verificação dos pressupostos da TRI, mas não diminui a importância de garantir que eles sejam atendidos para assegurar a validade e a precisão das medidas obtidas por meio dessa teoria.

3.2 MODELOS UNIDIMENSIONAIS

Os modelos da TRI mais utilizados são os modelos logísticos, devido à sua capacidade de fornecer um tratamento estatístico detalhado e à sua prevalência na literatura da área. A escolha do modelo de TRI depende do tipo de item que está sendo avaliado e dos objetivos da avaliação. Para compreender e descrever esses modelos, é essencial entender os parâmetros que os regem:

- Parâmetro a (discriminação): Capacidade do item de diferenciar indivíduos com habilidades ou proficiências distintas. Valores próximos a zero indicam que indivíduos com baixa ou alta proficiência possuem aproximadamente a mesma probabilidade de responder corretamente ao item e, nesse caso, o item não está cumprindo um de seus objetivos fundamentais. Em geral, é esperado para um bom item valores de a superiores a $0{,}70$. Rabelo (2013) traz uma categorização de acordo com as seguintes faixas, descrita na Tabela 3.

Tabela 3 – Classificação dos itens quanto à Discriminação na TRI

Valores	Discriminação
$a = 0{,}0$	Nenhuma
$0{,}0 < a \leq 0{,}35$	Muito baixa
$0{,}35 < a \leq 0{,}65$	Baixa
$0{,}65 < a \leq 1{,}35$	Moderada
$1{,}35 < a \leq 1{,}70$	Alta
$a > 1{,}70$	Muito Alta

Fonte: Rabelo (2013, p. 138)

- Parâmetro b (dificuldade): Nível mínimo de proficiência necessário para que um respondente possua uma probabilidade $\frac{1+c}{2}$ de responder corretamente a um item igual, em que c é a probabilidade de acerto ao acaso, a ser definida abaixo. Caso o modelo não considere esse parâmetro, define-se essa probabilidade como 50%. Devido a sua representação em uma escala padronizada, os valores do parâmetro da dificuldade geralmente situam-se entre -3 e 3, representando cerca de $99{,}7\%$ dos casos, sendo que, quanto mais perto desses extremos mais fácil e difícil, respectivamente, é o item. A Tabela 4 representa a distribuição e a classificação do item de acordo com esse parâmetro, conforme indicado por Rabelo (2013).

Tabela 4 – Classificação e percentual esperado para os índices de dificuldade na TRI

Classificação	Valores	% esperado
Muito fáceis	Até -1,28	10%
Fáceis	De -1,27 a -0,52	20%
Medianos	De -0,51 a 0,51	40%
Difíceis	De 0,52 a 1,27	20%
Muito Difíceis	1,28 ou mais	10%

Fonte: Rabelo (2013, p. 134)

- Parâmetro c (acerto ao acaso): Probabilidade de um respondente acertar o item sem possuir a habilidade necessária para tal. Para um item com cinco opções de resposta, espera-se valores inferiores à $0{,}20$. Caso contrário, pode haver um indicativo de que a opção correta se difere das demais, atraindo respondentes com baixa proficiência.

Os modelos logísticos da TRI podem ser classificados em três tipos principais. O Modelo Logístico de Um Parâmetro (ML1P), conhecido como modelo de Rasch, considera apenas a dificuldade do item (parâmetro b). O Modelo Logístico de Dois Parâmetros (ML2P) considera a dificuldade (parâmetro b) e a discriminação do item (parâmetro a). Já o Modelo Logístico de Três Parâmetros (ML3P) inclui a dificuldade (parâmetro b), a discriminação (parâmetro a) e a probabilidade de acerto ao acaso do item (parâmetro c).

O Enem e o Saeb utilizam o ML3P, proposto por Birnbaum (1968). Conforme Andrade, Tavares e Valle (2000), a expressão matemática desse modelo é dada por:

$$P\left(X_{ij} = 1 \,\middle|\, \theta_i, a_j, b_j, c_j\right) = c_j + (1 - c_j)\frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$$

em que $P(X_{ij} = 1)$ é a probabilidade de o indivíduo i com habilidade θ_i acertar o item j e a_j, b_j e c_j são, respectivamente, os parâmetros de discriminação, dificuldade e acerto ao acaso do item. A fórmula, desenvolvida por Birnbaum (1968), para o modelo logístico de dois parâmetros pode ser encontrada substituindo $c_j = 0$ na equação acima. Analogamente, para encontrarmos a fórmula do modelo logístico de um parâmetro, criado por Rasch (1960), basta substituirmos $c_j = 0$ e $a_j = 1$.

Além desses modelos, existe o Modelo Logístico de Quatro Parâmetros (ML4P) que incorpora um parâmetro d para representar a assimetria da curva característica do item. Esse quarto parâmetro permite capturar desvios na probabilidade de acerto em função de fatores como erros sistemáticos ou desatenção, proporcionando ainda mais precisão na modelagem das respostas dos indivíduos. Embora menos comum, o ML4P pode ser útil em situações onde uma modelagem mais complexa dos dados é necessária.

Cabe destacar, ainda, que a TRI não só permite a análise das propriedades dos itens de teste, como também oferece diversos métodos, cada um com suas particularidades e aplicações específicas, para estimar as proficiências dos indivíduos. Os mais comuns incluem a Estimação pela Máxima Verossimilhança (MLE), a Estimação Bayesiana a Posteriori (EAP) e a Estimação pelo Escore Máximo a Posteriori (MAP) (BAKER, 1992; 2001).

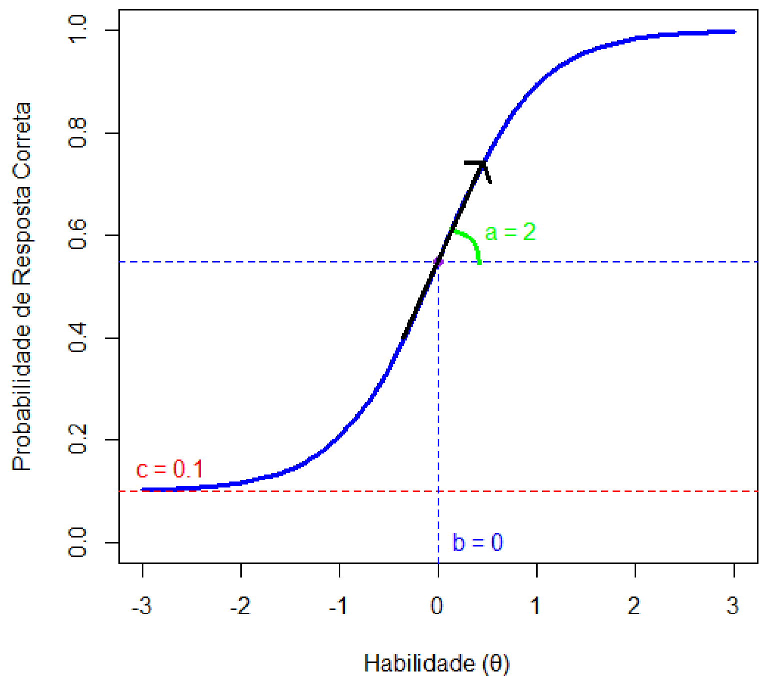
A MLE busca estimar a habilidade que maximiza a probabilidade das respostas observadas, sendo amplamente utilizada por sua simplicidade e eficiência em grandes amostras. A EAP baseia-se na integração das distribuições a priori e das respostas observadas para estimar a proficiência, sendo útil quando se dispõe de informações prévias sobre a habilidade dos indivíduos. A MAP combina a informação das respostas com uma distribuição a priori, buscando o valor de habilidade que maximiza a função posterior, sendo especialmente valiosa em contextos com amostras menores ou quando se deseja incorporar informações adicionais sobre os respondentes (BAKER, 1992; 2001). A análise desses métodos de estimação não é o foco deste material. Contudo, aos interessados na temática, recomenda-se a leitura de Baker (1992) que oferece uma exploração aprofundada dessas técnicas e suas aplicações.

3.3 CURVA CARACTERÍSTICA DO ITEM

A Curva Característica do Item (CCI) é uma ferramenta fundamental na TRI utilizada para representar graficamente a probabilidade de um participante responder corretamente a um item de teste em função do seu nível de habilidade na área avaliada. Esta curva, construída a partir das respostas de uma ampla amostra de participantes com diferentes níveis de habilidade, permite analisar, geometricamente, os parâmetros dos itens do teste, considerando sua qualidade e eficácia.

A CCI é uma curva sigmoide que se assemelha à função logística, começando com uma probabilidade muito baixa de resposta correta para participantes com níveis de habilidade muito baixos. À medida que o nível de habilidade aumenta, a probabilidade de resposta correta cresce rapidamente até se estabilizar em um nível máximo para participantes com níveis muito altos de habilidade. Esta forma sigmoide ilustra como a probabilidade de acerto de um item varia de acordo com a habilidade do respondente. Na Figura 2 tem-se a representação de uma CCI.

Figura 2 – Representação de uma Curva Característica do Item

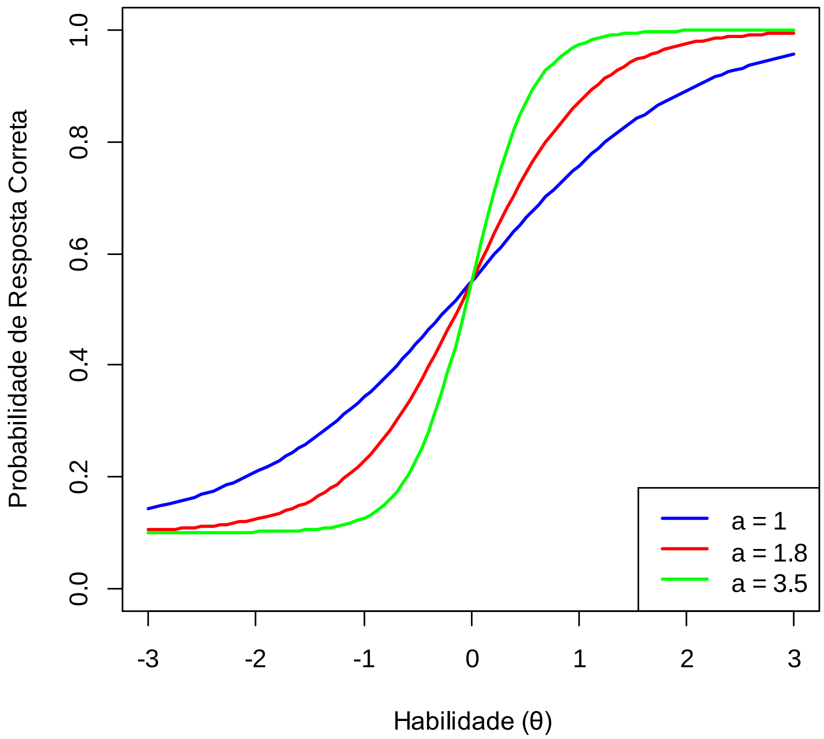


Fonte: Elaborado pelo autor (2024).

Observe, na Figura 2, que o eixo horizontal (θ) representa a habilidade do indivíduo e o eixo vertical representa a probabilidade de resposta correta ao item. A curva começa no valor correspondente ao parâmetro c no eixo vertical, se inclina de acordo com o parâmetro a, e atinge a probabilidade de 55% $(\frac{1+c}{2})$ no ponto correspondente ao parâmetro b no eixo horizontal. A inclinação da curva no ponto b é controlada por a, indicando como a probabilidade de acerto aumenta com a habilidade.

Desse modo, a inclinação da curva reflete a capacidade do item de discriminar participantes com diferentes níveis de habilidade. Itens altamente discriminativos têm curvas com uma inclinação mais íngreme, indicando que a probabilidade de resposta correta aumenta rapidamente em um intervalo estreito de habilidades. Itens com baixa discriminação apresentam uma curva mais suave, sugerindo que a probabilidade de resposta correta muda mais gradualmente com o aumento da habilidade. A Figura 3 ilustra como a variação do parâmetro a influencia a CCI.

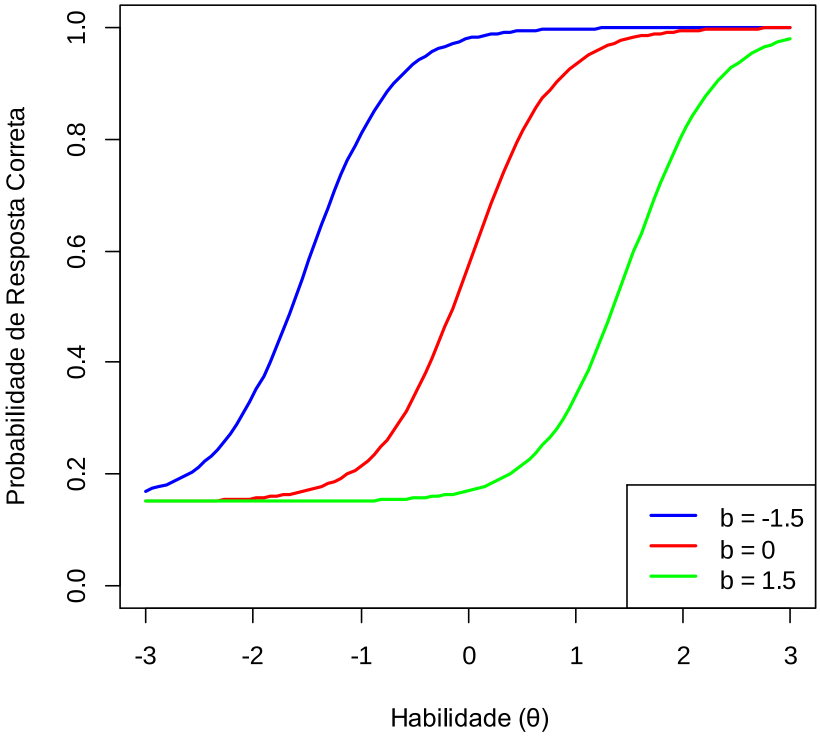
Figura 3 – CCIs com diferentes valores de a (discriminação)



Fonte: Elaborado pelo autor (2024).

A localização da curva na escala de habilidade, por sua vez, indica a dificuldade do item. Itens mais difíceis têm curvas deslocadas para a direita, exigindo um nível mais alto de habilidade para que a probabilidade de resposta correta aumente significativamente. A Figura 4 ilustra como a variação do parâmetro b influencia a CCI.

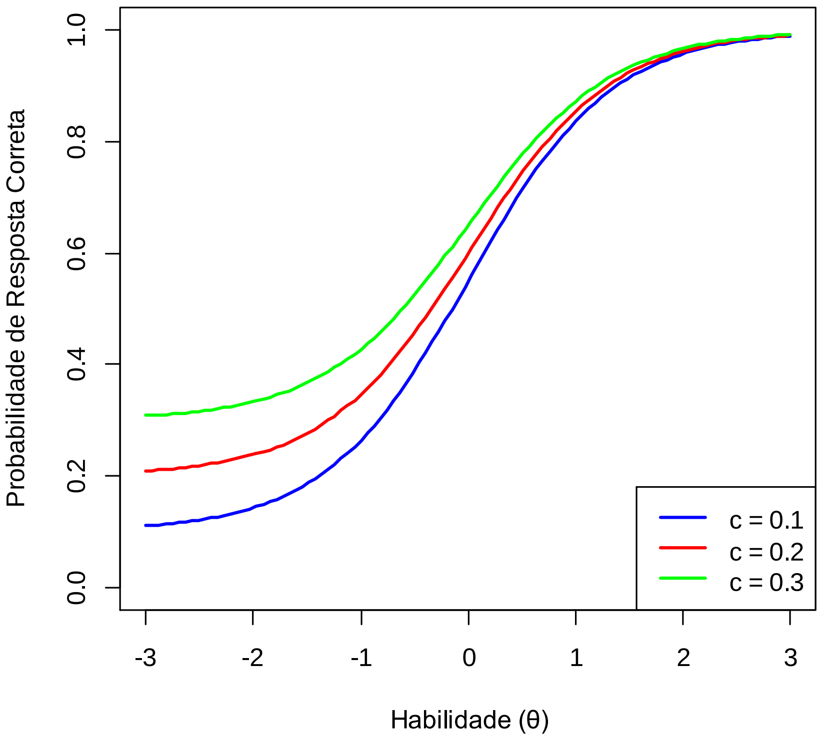
Figura 4 – CCIs com diferentes valores de b (dificuldade)



Fonte: Elaborado pelo autor (2024).

A variação do parâmetro c em uma CCI afeta a altura mínima da curva no eixo vertical. Em outras palavras, o parâmetro c determina o ponto onde a CCI intercepta o eixo y, ou seja, a probabilidade mínima de responder corretamente ao item, independentemente da habilidade do indivíduo. Portanto, à medida que o valor de c aumenta, a curva se desloca para cima no eixo y, elevando a probabilidade de acerto ao acaso. Isso significa que, mesmo para indivíduos com baixa habilidade, a probabilidade de acertar o item nunca será inferior ao valor de c. A Figura 5 ilustra como a variação do parâmetro c influencia a CCI.

Figura 5 – CCIs com diferentes valores de c (acerto ao acaso)



Fonte: Elaborado pelo autor (2024).

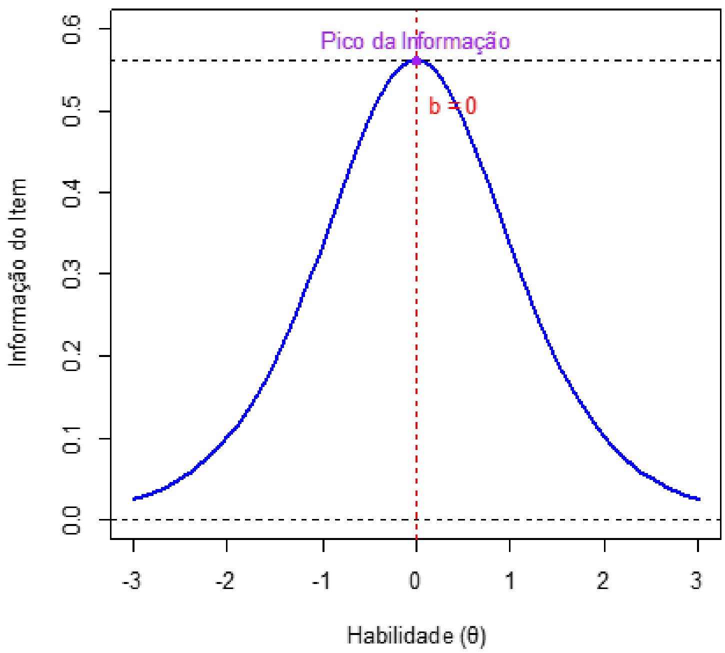
3.4 CURVA DE INFORMAÇÃO

A Curva de Informação de um item representa a quantidade de informação que ele fornece sobre a habilidade do respondente em diferentes níveis dessa habilidade. Em termos gerais, a Curva de Informação mostra como a precisão das estimativas de habilidade varia ao longo do continuum de habilidade (θ), indicando quanto um item (ou teste) contém de informação para a medida geral. Desse modo, ela permite identificar quais itens discriminam eficazmente o traço latente, possibilitando decisões informadas sobre sua adequação e permanência na escala de avaliação. (EMBRETSON; REISE, 2013).

Geometricamente, a altura da Curva de Informação em um ponto específico do continuum de habilidade indica a quantidade de informação fornecida pelo item nesse nível de habilidade. Curvas mais altas significam maior informação e, portanto, maior precisão na estimativa da habilidade nesse ponto. A largura da curva mostra a faixa de habilidade onde o item fornece informações úteis; itens com alta discriminação (a elevado) tendem a ter curvas de informação mais estreitas e altas, concentrando a informação em uma faixa mais específica de habilidades. O pico da Curva de Informação geralmente ocorre próximo ao ponto de dificuldade do item (b), indicando que o item é mais informativo para indivíduos cuja habilidade está próxima desse valor.

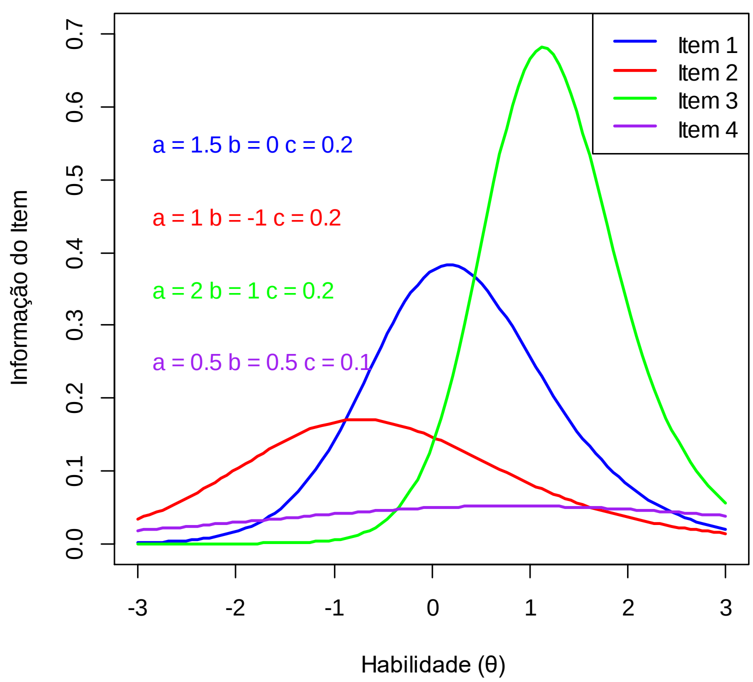
Para fins de ilustração, a Figura 6 representa uma Curva de Informação para um item em um ML3P e a Figura 7 traz exemplos dessas curvas modeladas para itens de um mesmo teste que possuem diferentes parâmetros psicométricos, permitindo a comparação da informação trazida em cada caso.

Figura 6 – Curva de Informação do Item (ML3P)



Fonte: Elaborado pelo autor (2024).

Figura 7 – Curvas de Informação de Itens com diferentes parâmetros psicométricos (ML3P)



Fonte: Elaborado pelo autor (2024).

A Figura 7 apresenta quatro Curvas de Informação para itens com diferentes combinações dos parâmetros a, b e c. A curva azul, correspondente ao item com $a = 1{,}5,\ b = 0$ e $c = 0{,}2$, mostra um pico de informação em torno de $\theta = 0$, indicando que é mais informativa para indivíduos com habilidade média. A curva vermelha, com $a = 1{,}0,\ b = -1$ e $c = 0{,}2$ desloca o pico de informação para a esquerda, em torno de $\theta = -1,$ sendo mais informativa para indivíduos com habilidades mais baixas. A curva verde, com $a = 2{,}0,\ b = 1{,}0$ e $c = 0{,}2$, é mais alta e estreita, com o pico de informação em $\theta = 1$, indicando que discrimina melhor entre indivíduos com habilidades próximas a esse valor. A curva roxa, com $a = 0{,}5,\ b = 0{,}5$ e $c = 0{,}1$, é mais baixa e larga, oferecendo informação em uma faixa mais ampla de habilidades, mas com menos precisão.

Entre as curvas, a curva verde possui o maior pico de informação, indicando que o item com $a = 2{,}0,\ b = 1{,}0$ e $c = 0{,}2$, fornece mais informação do que os outros itens em seu ponto de maior discriminação, em torno de $\theta = 1$. Isso significa que esse item é particularmente eficaz em discriminar entre indivíduos com habilidades próximas a 1, fornecendo estimativas de habilidade mais precisas nesse intervalo. Em contraste, a curva roxa, embora mais ampla, possui o menor pico de informação, indicando que fornece menos informação e, portanto, menos precisão na estimativa de habilidade em qualquer ponto específico.

Cabe destacar que a Curva de Informação tem diversas aplicações práticas. Na construção de testes adaptativos computadorizados (CAT), itens são escolhidos com base na quantidade de informação que fornecem em relação ao nível de habilidade estimado do respondente, com itens mais informativos sendo preferidos para melhorar a precisão das estimativas de habilidade. No desenho de testes, é importante escolher itens que, coletivamente, forneçam informação suficiente em toda a faixa de habilidades de interesse, garantindo que o teste seja preciso para todos os níveis de habilidade dos respondentes. Além disso, a análise das curvas de informação pode ajudar a identificar itens que não estão contribuindo significativamente para a precisão do teste ou que são mais informativos em faixas de habilidade específicas.

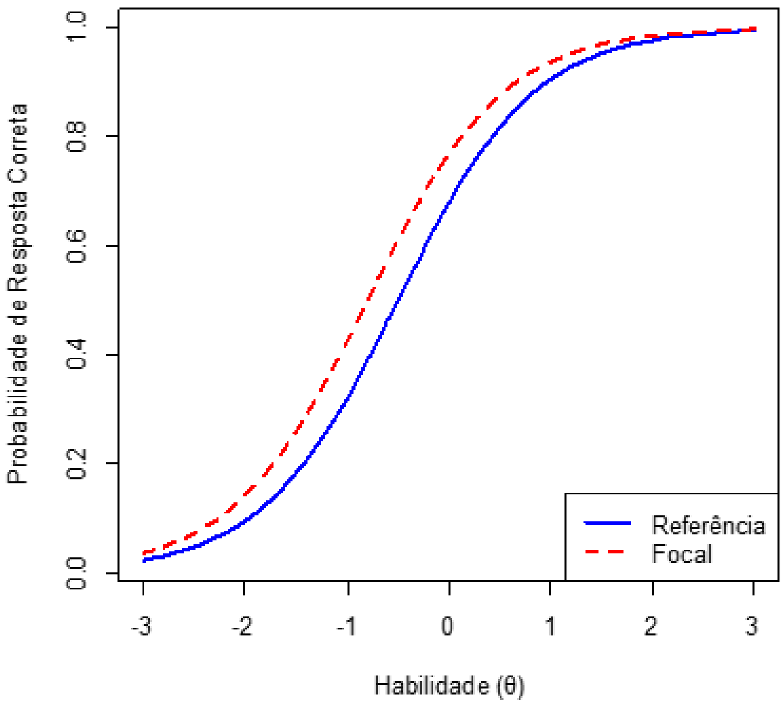
3.5 FUNCIONAMENTO DIFERENCIAL DOS ITENS

O Funcionamento Diferencial dos Itens (DIF) é usado para identificar itens que se comportam de maneira diferente para diferentes grupos de indivíduos, mesmo quando esses indivíduos possuem o mesmo nível de habilidade. Em outras palavras, um item apresenta DIF quando as probabilidades de resposta correta variam entre grupos, sugerindo um potencial viés ou injustiça no item. Cabe destacar que o DIF pode ser classificado em duas categorias principais: DIF uniforme e DIF não uniforme, que podem ser compreendidos analisando as diferenças no parâmetro de dificuldade (b) e no parâmetro de discriminação (a).

O DIF uniforme ocorre quando a diferença na probabilidade de acerto de um item é constante ao longo de todos os níveis de habilidade (θ) entre os dois grupos. Essa diferença é refletida no parâmetro de dificuldade (b). No caso de DIF uniforme, o parâmetro de dificuldade (b) difere entre os grupos, enquanto o parâmetro de discriminação (a) permanece o mesmo.

Exemplo: Imagine um item de um teste de matemática que é mais fácil para um grupo de alunos que estudaram em uma escola que enfatiza técnicas específicas de resolução de problemas (Grupo Focal), enquanto é consistentemente mais difícil para outro grupo que não teve essa preparação (Grupo Referência. Nesse caso, o parâmetro de dificuldade (b) seria menor (indicando que a questão é mais fácil) para o grupo que recebeu a preparação específica. A Figura 8 ilustra essa situação.

Figura 8 – Exemplo de DIF Uniforme



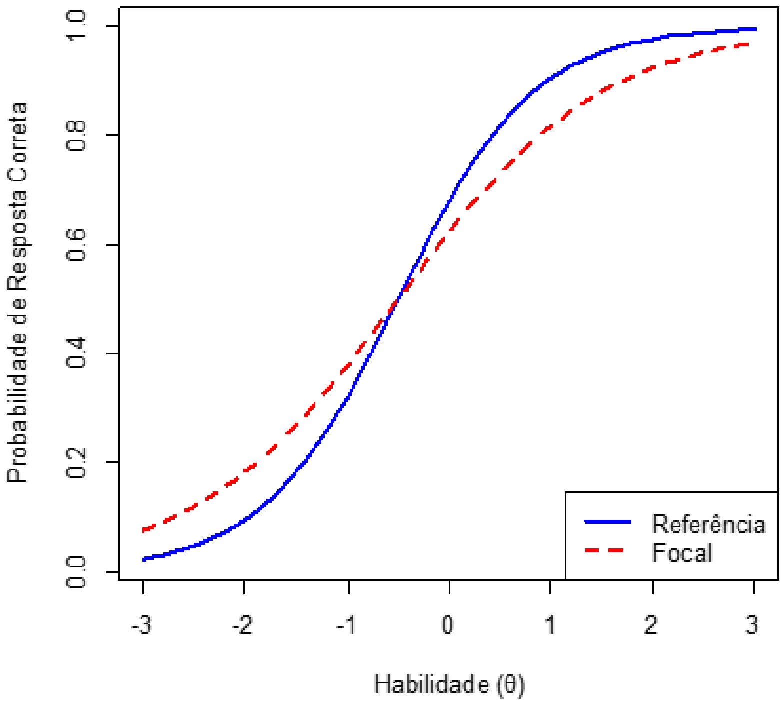
Fonte: Elaborado pelo autor (2024).

Na Figura acima, a CCI para o grupo focal está deslocada em relação à curva do grupo de referência. Se o parâmetro de dificuldade (b) é $-0{,}5$ para o grupo de referência e $-0{,}8$ para o grupo focal, isso indica que a questão é mais fácil para o grupo focal, resultando em uma curva deslocada para a direita. Ambas as curvas têm a mesma inclinação, refletindo a mesma discriminação (a) entre os grupos.

O DIF não uniforme ocorre quando a diferença na probabilidade de acerto de um item varia ao longo dos níveis de habilidade (θ). Isso significa que tanto o parâmetro de dificuldade (b) quanto o parâmetro de discriminação (a) podem diferir entre os grupos. No DIF não uniforme, o parâmetro de discriminação (a) difere, o que causa uma mudança na inclinação das CCIs entre os grupos.

Exemplo: Considere uma questão de um teste de linguagem que avalia habilidades de vocabulário avançado. Esta questão pode ser mais fácil para falantes nativos de um idioma específico em níveis mais baixos de habilidade, mas à medida que a habilidade aumenta, essa vantagem pode diminuir ou até se inverter. Nesse caso, o parâmetro de discriminação (a) pode ser maior para o grupo de referência do que para o grupo focal, refletindo uma diferença na sensibilidade do item em detectar habilidades. A Figura 9 ilustra essa situação.

Figura 9 – Exemplo de DIF Não Uniforme



Fonte: Elaborado pelo autor (2024).

No gráfico de DIF não uniforme, as curvas características dos itens para os grupos de referência e focal não são paralelas. Se o parâmetro de discriminação (a) é $1{,}5$ para o grupo de referência e $1{,}0$ para o grupo focal, as curvas terão diferentes inclinações, indicando que a dificuldade relativa do item muda ao longo dos níveis de habilidade. Por exemplo, a curva do grupo focal pode começar mais alta em níveis baixos de habilidade e cruzar a curva do grupo de referência em níveis mais altos, indicando uma mudança na vantagem entre os grupos.

Existem diversos métodos estatísticos para identificar e corrigir itens com DIF, garantindo assim a integridade do teste. Entre os mais comuns está o método de Mantel-Haenszel, sobretudo para detectar DIF uniforme. Outros métodos incluem modelos logísticos e análises de regressão logística multinível, que podem considerar variáveis adicionais e estruturas hierárquicas nos dados. Modelos de TRI multigrupo são também amplamente utilizados, pois ajustam simultaneamente os parâmetros dos itens para diferentes grupos, permitindo uma comparação direta e precisa dos parâmetros dos itens.

REFERÊNCIAS

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. **Teoria da resposta ao item**: conceitos e aplicações. In: Simpósio Nacional de Probabilidade e Estatística (SINAPE), 14., 2000, Caxambu. Anais... Caxambu: Associação Brasileira de Estatística, 2000. 164 p.

ARIAS, M. R. M.; LLOREDA, M. V. H.; LLOREDA, M. J. H. **Psicometría**. [S.1.]: Alianza Editorial, 2006. 488 p

BAKER, F. B. **Item Response Theory**: Parameter Estimation Techniques. New York: University of Wisconsin-Madison, 1992. 528 p.

BAKER, F. B. **The basics of Item Response Theory**. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD, 2001. 185 p.

BIRNBAUM, A. **Some latent trait models and their use in inferring an examinee's ability**. In F. M. Lord & M. R. Novick. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, Boston, p. 397-479, 1968.

BORGATTO, A. F.; ANDRADE, D. F. Análise clássica de testes com diferentes graus de dificuldade. **Estudos em Avaliação Educacional**, São Paulo, v. 23, n. 52, p. 146-156, 2012.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística básica**. 5 ed. São Paulo: Saraiva, 2004. 526 p.

CONDÉ, F. N. **Análise empírica de itens**. Technical report, Instituto Nacional de Estudos e Pesquisas Educacionais-DAEB/INEP/MEC, Brasília, 2001. 193 p.

EMBRETSON, S. E.; REISE, S. P. **Item Response Theory**. Psychology Press, Hove, United Kingdom, 2013. 384 p.

HAIR JÚNIOR, J.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E. **Multivariate Data Analysis**. 7th ed. Upper Saddle River: Prentice Hall, 2010. 785 p.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. **Fundamentals of Item Response Theory**. Newbury Park: Sage Publications, 1991. 184 p.

KELLEY, T. L. The selection of upper and lower groups for the validation of test items. **Journal of Educational Psychology**, Warwick & york, v. 30, n. 1, p. 17-24, 1939.

KLINE, T. **Psychological Testing**: A practical approach to design and evaluation. Thousand Oaks, CA: Sage, 2005. 363 p.

LORD, F. M.; NOVICK, M. R. **Statistical theories of mental test scores**. Reading, MA: Addison-Wesley, 1968. 568 p.

PASQUALI, L. **Psicometria**: teoria dos testes na psicologia e na educação. Petrópolis: Vozes, 2003. 397 p.

PASQUALI, L.; PRIMI, R. Fundamentos da teoria da resposta ao item TRI. **Avaliação Psicológica**, Campinas, v. 2, n. 2, p. 99-110, 2003.

PASQUALI, L. **TRI–Teoria de resposta ao item**: Teoria, procedimentos e aplicações. Editora Appris, 2020.

R DEVELOPMENT CORE TEAM. **R**: A Language and Environment for Statistical Computing. Vienna: R Foundation on Statistical Computing, 2024. Disponível em: https://www.r-project.org.

RABELO, M. L. **Avaliação Educacional**: Fundamentos. Metodologia e Aplicaçoes no Contexto Brasileiro, Rio de Janeiro, 1a ediçao, Coleçao Profmat, Sociedade Brasileira de Matemática, 2013.

RASCH, G. **Probabilistic models for some intelligence and achievement tests**. Copenhagen: Danish Institute for Education Research, 1960. 184 p.

SOARES, D. J. M. **Teoria Clássica dos Testes e Teoria de Resposta ao Item aplicadas em uma avaliação de matemática básica**. 2018. 121 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) - Universidade Federal de Viçosa, Viçosa, 2018.

TÔRRES, F. C. **Uma aplicação da teoria de resposta ao item em um simulado de matemática no modelo Enem**. 2015. 116 p. Dissertação (Mestrado em Matemática) - Programa de Mestrado Profissional em Matemática em Rede Nacional, Universidade de Brasília, Brasília, 2015.

ANEXO I – R SCRIPTS

```
################################################################################
###################### Instalando e Carregando Pacotes ##########################
################################################################################

if(!require(ltm)) install.packages("ltm", dependencies=TRUE)
if(!require(mirt)) install.packages("mirt", dependencies=TRUE)
if(!require(psych)) install.packages("psych", dependencies=TRUE)
if(!require(ShinyItemAnalysis)) install.packages("ShinyItemAnalysis", dependencies=TRUE)
library(ltm)
library(mirt)
library(psych)
library(ShinyItemAnalysis)

################################################################################
################################## Figura 1 ##################################
################################################################################

#Usando o conjunto de dados "dataMedical" do pacote "ShinyItemAnalysis"
dados=dataMedical[,1:50]
plot(descript(dados),items=c(49,30,48),type="b",includeFirstLast=TRUE,
xlab = "Escore Total", ylab = "Proporção de acerto (Índice de Dificuldade)",
col = c("blue", "red", "green"))
```

```
###############################################################################
################################### Figura 2 ##################################
###############################################################################

# Definindo os parâmetros a, b e c para um item específico
a <- 2  # Discriminação
b <- 0    # Dificuldade
c <- 0.1  # Acerto ao acaso

# Criando uma função para calcular a CCI
calculate_cci <- function(theta, a, b, c) {
  return(c + (1 - c) / (1 + exp(-a * (theta - b))))
}

# Definindo a faixa de habilidades (theta)
theta <- seq(-3, 3, length.out = 100)

# Calculando a probabilidade de resposta correta para a faixa de habilidades
prob <- calculate_cci(theta, a, b, c)

# Plotando a CCI
plot(theta, prob, type = "l", lwd = 3, col = "blue", ylim = c(0, 1),
     xlab = "Habilidade (θ)", ylab = "Probabilidade de Resposta Correta",
     main = "Curva Característica do Item (CCI)")

# Adicionando uma linha horizontal para o parâmetro c (pseudo-chance level)
abline(h = c, col = "red", lty = 2)
text(x = -2.7, y = c, labels = paste("c =", round(c, 2)), pos = 3, col = "red")
```

```
# Adicionando uma linha vertical para o parâmetro b (dificuldade do item)
abline(v = 0:0.55, h=0.55, col = "blue", lty = 2)
text(x = b, y = 0, labels = paste("b =", round(b, 2)), pos = 4, col = "blue")

# Adicionando um ponto e texto no ponto (b, 0.5) para ilustrar b no eixo x
points(x = b, y = 0.55, col = "purple", pch = 19)
text(x = 0.3, y = 0.6, labels = paste("a =", round(a, 2)), pos = 4, col = "green")


##############################################################################
################################## Figura 3 ##################################
##############################################################################

# Definindo os parâmetros constantes b e c
b <- 0  # Dificuldade
c <- 0.1  # Acerto ao acaso

# Definindo diferentes valores para o parâmetro a (discriminação)
a_values <- c(1, 1.8, 3.5)

# Criando uma função para calcular a CCI
calculate_cci <- function(theta, a, b, c) {
  return(c + (1 - c) / (1 + exp(-a * (theta - b))))
}

# Definindo a faixa de habilidades (theta)
theta <- seq(-3, 3, length.out = 100)
```

```
# Calculando as probabilidades de resposta correta para cada valor de a
cci_curves <- sapply(a_values, function(a) calculate_cci(theta, a, b, c))

# Plotando as CCIs
plot(theta, cci_curves[,1], type = "l", lwd = 2, col = "blue", ylim = c(0, 1),
     xlab = "Habilidade (θ)", ylab = "Probabilidade de Resposta Correta",
     main = "CCIs com diferentes valores de a (discriminação)")
lines(theta, cci_curves[,2], lwd = 2, col = "red")
lines(theta, cci_curves[,3], lwd = 2, col = "green")

# Adicionando uma legenda
legend("bottomright", legend = paste("a =", a_values), col = c("blue", "red", "green"), lwd = 2)


################################################################################
#################################### Figura 4 ####################################
################################################################################

# Definindo os parâmetros constantes a e c
a <- 2.5  # Discriminação
c <- 0.15  # Acerto ao acaso

# Definindo diferentes valores para o parâmetro b (dificuldade)
b_values <- c(-1.5, 0, 1.5)

# Criando uma função para calcular a CCI
calculate_cci <- function(theta, a, b, c) {
```

```
  return(c + (1 - c) / (1 + exp(-a * (theta - b))))
}

# Definindo a faixa de habilidades (theta)
theta <- seq(-3, 3, length.out = 100)

# Calculando as probabilidades de resposta correta para cada valor de b
cci_curves <- sapply(b_values, function(b) calculate_cci(theta, a, b, c))

# Plotando as CCIs
plot(theta, cci_curves[,1], type = "l", lwd = 2, col = "blue", ylim = c(0, 1),
     xlab = "Habilidade (θ)", ylab = "Probabilidade de Resposta Correta",
     main = "CCIs com diferentes valores de b (dificuldade)")
lines(theta, cci_curves[,2], lwd = 2, col = "red")
lines(theta, cci_curves[,3], lwd = 2, col = "green")

# Adicionando uma legenda
legend("bottomright", legend = paste("b =", b_values), col = c("blue", "red", "green"), lwd = 2)


################################################################################
################################### Figura 5 ###################################
################################################################################

# Definindo os parâmetros constantes a e b
a <- 1.5  # Discriminação
b <- 0  # Dificuldade
```

```
# Definindo diferentes valores para o parâmetro c (acerto ao acaso)
c_values <- c(0.1, 0.2, 0.3)

# Criando uma função para calcular a CCI
calculate_cci <- function(theta, a, b, c) {
  return(c + (1 - c) / (1 + exp(-a * (theta - b))))
}

# Definindo a faixa de habilidades (theta)
theta <- seq(-3, 3, length.out = 100)

# Calculando as probabilidades de resposta correta para cada valor de c
cci_curves <- sapply(c_values, function(c) calculate_cci(theta, a, b, c))

# Plotando as CCIs
plot(theta, cci_curves[,1], type = "l", lwd = 2, col = "blue", ylim = c(0, 1),
     xlab = "Habilidade (θ)", ylab = "Probabilidade de Resposta Correta",
     main = "CCIs com diferentes valores de c (acerto a acaso)")
lines(theta, cci_curves[,2], lwd = 2, col = "red")
lines(theta, cci_curves[,3], lwd = 2, col = "green")

# Adicionando uma legenda
legend("bottomright", legend = paste("c =", c_values), col = c("blue", "red", "green"), lwd = 2)
```

```
##############################################################################
################################### Figura 6 ###################################
##############################################################################

# Definindo os parâmetros a, b e c para um item específico
a <- 1.5  # Discriminação
b <- 0    # Dificuldade
c <- 0  # Acerto ao acaso

# Criando uma função para calcular a probabilidade de resposta correta no modelo 3PL
calculate_prob_3pl <- function(theta, a, b, c) {
  return(c + (1 - c) / (1 + exp(-a * (theta - b))))
}

# Criando uma função para calcular a informação do item no modelo 3PL
calculate_item_information <- function(theta, a, b, c) {
  P_theta <- calculate_prob_3pl(theta, a, b, c)
  return((a^2 * (1 - P_theta) * (P_theta - c)^2) / ((1 - c)^2 * P_theta))
}

# Definindo a faixa de habilidades (theta)
theta <- seq(-3, 3, length.out = 100)

# Calculando a informação do item para a faixa de habilidades
information <- sapply(theta, function(t) calculate_item_information(t, a, b, c))

# Plotando a Curva de Informação do Item
plot(theta, information, type = "l", lwd = 2, col = "blue", ylim = c(0, 0.6),
```

```
  xlab = "Habilidade (θ)", ylab = "Informação do Item",
  main = "Curva de Informação do Item (3PL)")

# Adicionando linhas e textos para destacar os parâmetros
abline(v = b, col = "red", lty = 2) # Linha vertical no ponto de dificuldade b
text(x = b, y = max(information) * 0.9, labels = paste("b =", round(b, 2)), pos = 4, col = "red")

abline(h = 0, col = "black", lty = 2) # Linha horizontal no eixo y
abline(h = max(information), col = "black", lty = 2) # Linha horizontal no pico da informação

# Adicionando pontos de destaque para os parâmetros
points(x = b, y = max(information), col = "purple", pch = 19)
text(x = b, y = max(information), labels = paste("Pico da Informação"), pos = 3, col = "purple")


################################################################################
################################### Figura 7 ###################################
################################################################################

# Definindo os parâmetros a, b e c para vários itens
parameters <- list(
  list(a = 1.5, b = 0, c = 0.2),
  list(a = 1.0, b = -1, c = 0.2),
  list(a = 2.0, b = 1, c = 0.2),
  list(a = 0.5, b = 0.5, c = 0.1)
)
```

```
# Criando uma função para calcular a probabilidade de resposta correta no modelo 3PL
calculate_prob_3pl <- function(theta, a, b, c) {
  return(c + (1 - c) / (1 + exp(-a * (theta - b))))
}

# Criando uma função para calcular a informação do item no modelo 3PL
calculate_item_information <- function(theta, a, b, c) {
  P_theta <- calculate_prob_3pl(theta, a, b, c)
  return((a^2 * (1 - P_theta) * (P_theta - c)^2) / ((1 - c)^2 * P_theta))
}

# Definindo a faixa de habilidades (theta)
theta <- seq(-3, 3, length.out = 100)

# Plotando as Curvas de Informação dos Itens
plot(NULL, xlim = c(-3, 3), ylim = c(0, 0.7), xlab = "Habilidade (θ)", ylab = "Informação do Item",
     main = "Curvas de Informação dos Itens (3PL)")

# Definindo cores para as diferentes curvas
colors <- c("blue", "red", "green", "purple")

# Iterando sobre os parâmetros e plote cada curva de informação
for (i in 1:length(parameters)) {
  a <- parameters[[i]]$a
  b <- parameters[[i]]$b
  c <- parameters[[i]]$c

  information <- sapply(theta, function(t) calculate_item_information(t, a, b, c))
```

```
  lines(theta, information, lwd = 2, col = colors[i])
  text(x = -3, y = 0.65 - 0.1 * i, labels = paste("a =", a, "b =", b, "c =", c), col = colors[i], pos = 4)
}

# Adicionando uma legenda
legend("topright", legend = paste("Item", 1:length(parameters)), col = colors, lwd = 2)


##############################################################################
################################### Figura 8 #################################
##############################################################################

# Definindo os parâmetros
set.seed(123)
n <- 1000 # Número de indivíduos em cada grupo
theta <- rnorm(n) # Habilidade latente
alpha <- 1.5 # Parâmetro de discriminação

# Parâmetros de dificuldade para os dois grupos
beta_ref <- -0.5 # Grupo de referência
beta_focal <- -0.8 # Grupo focal (DIF presente)

# Função para calcular a probabilidade de acerto
p_function <- function(alpha, beta, theta) {
  exp(alpha * (theta - beta)) / (1 + exp(alpha * (theta - beta)))
}
```

```
# Valores de theta para as CCIs
theta_values <- seq(-3, 3, by = 0.1)

# Probabilidades para cada grupo ao longo de theta
prob_ref <- p_function(alpha, beta_ref, theta_values)
prob_focal <- p_function(alpha, beta_focal, theta_values)

# Plotando CCIs usando a função plot
plot(theta_values, prob_ref, type = "l", col = "blue", lwd = 2,
    xlab = "Habilidade (θ)", ylab = "Probabilidade de Resposta Correta",
    main = "DIF Uniforme")
lines(theta_values, prob_focal, col = "red", lwd = 2, lty = 2)
legend("bottomright", legend = c("Referência", "Focal"), col = c("blue", "red"), lwd = 2, lty = c(1, 2))


##################################################################################
#################################### Figura 9 ####################################
##################################################################################

# Definindo os parâmetros
set.seed(123)
n <- 1000 # Número de indivíduos em cada grupo
theta <- rnorm(n) # Habilidade latente

# Parâmetros de discriminação e dificuldade para os dois grupos
alpha_ref <- 1.5 # Parâmetro de discriminação para o grupo de referência
beta_ref <- -0.5 # Parâmetro de dificuldade para o grupo de referência
```

```
alpha_focal <- 1.0 # Parâmetro de discriminação para o grupo focal (DIF não uniforme)
beta_focal <- -0.5 # Parâmetro de dificuldade para o grupo focal

# Função para calcular a probabilidade de acerto
p_function <- function(alpha, beta, theta) {
  exp(alpha * (theta - beta)) / (1 + exp(alpha * (theta - beta)))
}

# Valores de theta para as CCIs
theta_values <- seq(-3, 3, by = 0.1)

# Probabilidades para cada grupo ao longo de theta
prob_ref <- p_function(alpha_ref, beta_ref, theta_values)
prob_focal <- p_function(alpha_focal, beta_focal, theta_values)

# Plotando CCIs usando a função plot
plot(theta_values, prob_ref, type = "l", col = "blue", lwd = 2,
      xlab = "Habilidade (θ)", ylab = "Probabilidade de Resposta Correta",
     main = "DIF Não Uniforme")
lines(theta_values, prob_focal, col = "red", lwd = 2, lty = 2)
legend("bottomright", legend = c("Referência", "Focal"), col = c("blue", "red"), lwd = 2, lty = c(1, 2))
```