

Verbetes Metodológico

Teoria da Resposta ao Item

Ednaldo Ribeiro¹ 

¹Professor da Universidade Estadual de Maringá, UEM, Maringá, PR, Brasil.

E-mail: ednaldorip@gmail.com

DOI: 10.17666/ebcs.2026.041

PREÂMBULO

A Teoria da Resposta ao Item (TRI) é um conjunto de modelos de mensuração que relacionam a probabilidade de uma resposta a um item (uma questão de prova, uma pergunta de questionário, um voto nominal) a um traço latente não observado do respondente, como uma habilidade, uma atitude ou uma posição ideológica, e a características do próprio item, como sua dificuldade e seu poder de discriminação. Desenvolvida na psicometria para superar a dependência dos escores em relação ao instrumento usado, própria da Teoria Clássica dos Testes, a TRI situa itens e respondentes em uma mesma escala e permite comparar grupos avaliados com instrumentos distintos ou em momentos diferentes. Nas ciências sociais, responde a perguntas de mensuração: como estimar, com incerteza explícita, grandezas que não podem ser observadas diretamente, como o grau de democracia de um país ou a posição ideológica de um parlamentar. O verbete apresenta a origem e os fundamentos da TRI, seus modelos e pressupostos, suas vantagens, limites e críticas, aplicações na ciência política e os debates atuais sobre estimação, identificação e equivalência de mensuração.

SUMÁRIO

1. Histórico e fundamentos epistemológicos
2. Procedimentos principais
3. Vantagens, limites de validade e críticas conhecidas
4. Exemplos de aplicação em ciências sociais
5. Debates contemporâneos e variações
6. Referências bibliográficas

Aprovado em: 28/09/2026

Editor(a) responsável: Adriano Codato



Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution (<https://creativecommons.org/licenses/by/4.0/>), que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

Enciclopédia Brasileira de Ciências Sociais (EBCS). ANPOCS, 041 - 2026

1. Histórico e fundamentos epistemológicos

A Teoria da Resposta ao Item (TRI) surgiu na psicometria como resposta a uma limitação estrutural da Teoria Clássica dos Testes, cujas bases remontam ao artigo “*The Proof and Measurement of Association between Two Things*”, publicado por Charles Spearman em 1904: os escores produzidos por qualquer instrumento dependem do conjunto específico de itens que o compõe. Na teoria clássica, a dificuldade de um item é a proporção de acertos observada em uma amostra e, por isso, muda conforme a amostra é mais ou menos habilidosa. O escore de um respondente, por sua vez, muda conforme o teste é mais fácil ou mais difícil. Nesse quadro, medida e instrumento são indissociáveis. O mesmo estudante obtém escores diferentes em provas diferentes sobre o mesmo conteúdo, e o mesmo item muda de dificuldade quando aplicado a outra amostra. Comparações entre grupos testados com instrumentos distintos, ou entre aplicações em momentos diferentes, ficam metodologicamente indefensáveis.

O conceito de traço latente, introduzido por Paul Lazarsfeld em “*The Logical and Mathematical Foundation of Latent Structure Analysis*”, de 1950, no contexto da medição de atitudes, forneceu a base para uma solução. O postulado central é que comportamentos observáveis (as respostas aos itens) são efeitos de processos hipotéticos subjacentes que não podem ser observados diretamente. O postulado tem uma consequência direta para a interpretação dos resultados: o pesquisador ou pesquisadora não está descrevendo uma distribuição de acertos, está estimando um traço que existe no respondente e que gera as respostas observadas. O interesse passou a ser a estimativa desse traço, não o escore bruto.

A primeira formulação sistemática veio de Frederic Lord, no “*Educational Testing Service*”, em “*A Theory of Test Scores*”, de 1952. Lord desenvolveu um modelo de dois parâmetros que resolve o problema de dependência do instrumento, isto é, os parâmetros dos itens passaram a ser estimados de forma invariante em relação à amostra e a habilidade passou a ser estimada de forma invariante em relação ao teste. Na prática, isso significa que dois pesquisadores usando conjuntos diferentes de itens para medir a mesma habilidade chegam a resultados comparáveis, o que é impossível dentro da teoria clássica. A constatação de que respondentes com habilidades muito baixas às vezes acertavam itens difíceis por acaso levou à incorporação de um terceiro parâmetro para representar essa probabilidade, formalizada por Allan Birnbaum nos capítulos que escreveu para o volume de Lord e Novick (1968/2008) e consolidada por Lord (1980).

Em paralelo, Georg Rasch chegou de forma independente a um modelo de um parâmetro ao analisar dados de testes de leitura em Copenhague entre 1951 e 1953 (Rasch, 1960/1980). Rasch buscava uma propriedade mais restrita: a separabilidade completa entre os parâmetros de item e de pessoa, que chamou de objetividade específica. Um modelo com essa propriedade tem o mesmo estatuto epistemológico da medição em física: a estimativa da habilidade independe totalmente dos itens usados; e a estimativa da dificuldade independe totalmente das pessoas testadas. Lord e Novick (1968/2008) formalizaram as condições matemáticas para que esse tipo de inferência seja válida. No mesmo volume, Birnbaum reformulou os modelos de Lord substituindo a função de ogiva normal pela função logística. Os resultados são praticamente idênticos, mas a nova formulação tornou os cálculos computacionalmente

tratáveis. Sem essa mudança, os modelos teriam permanecido restritos a laboratórios com acesso a computadores de grande porte.

Fumiko Samejima estendeu a teoria a itens politômicos e propôs o modelo de resposta gradual, para respostas em categorias ordenadas, e não apenas no par certo/errado (Samejima, 1969). Depois, em “*A General Model for Free-Response Data*”, de 1972, estendeu-a a respostas livres. Essa extensão abriu a TRI para instrumentos de atitude e personalidade, ampliando seu alcance para além dos testes de desempenho.

A aplicabilidade em larga escala dependeu do desenvolvimento de programas específicos entre o fim dos anos 1970 e meados dos anos 1980, como BICAL, em 1979, LOGIST, em 1982, e BILOG, em 1984. No Brasil, a adoção da TRI respondeu a uma necessidade específica dos sistemas de avaliação educacional: comparar o desempenho de estudantes em diferentes anos e regiões usando provas distintas, o que só é possível com parâmetros invariantes em relação à amostra. A primeira aplicação em larga escala ocorreu em 1995, no SAEB. Andrade et al. (2000) documentaram esse processo e produziram o primeiro manual de referência sobre TRI para o público brasileiro.

2. Procedimentos principais

A TRI parte de uma premissa direta de que para qualquer item existe uma relação entre o nível de habilidade do respondente e a probabilidade de resposta correta. Quanto maior a habilidade, maior a probabilidade de acerto. Essa relação tem a forma de uma curva em S, crescente de zero a um, chamada de Curva Característica do Item (CCI). Cada item tem a sua CCI e o conjunto das curvas de um instrumento descreve como cada questão se comporta ao longo de toda a amplitude da habilidade medida.

Os modelos da TRI variam conforme a natureza dos itens (dicotômicos ou politômicos), o número de populações envolvidas e a quantidade de traços latentes medidos, como sistematiza o “*Handbook of Modern Item Response Theory*”, organizado por Wim van der Linden e Ronald Hambleton em 1997. Os modelos unidimensionais para um único grupo são os mais utilizados nas ciências sociais. Dentro desse conjunto, os itens dicotômicos admitem três variantes que diferem no número de parâmetros usados para descrever cada item.

O modelo logístico de 3 parâmetros (ML3) é o mais completo entre os modelos dicotômicos. Cada item é descrito por três características. A discriminação (a) indica quanto o item diferencia respondentes com habilidades ligeiramente distintas na região de dificuldade do item: um item muito discriminativo reage a pequenas diferenças de habilidade, enquanto um item pouco discriminativo produz taxas de acerto semelhantes para respondentes de níveis muito diferentes. A dificuldade (b) indica o nível de habilidade no qual o respondente tem 50% de probabilidade de acertar, descontado o efeito do acaso: itens com b alto exigem mais habilidade, itens com b baixo são acessíveis a respondentes de menor habilidade. O acerto casual (c) é a probabilidade de um respondente com habilidade muito baixa acertar o item, relevante em testes de múltipla escolha, em que a eliminação de alternativas tem peso real (Hambleton & Swaminathan, 1985).

A escolha entre os três modelos tem consequências metodológicas concretas. O modelo de 1 parâmetro (modelo de Rasch) impõe a restrição de que todos os itens discriminam igualmente. Quando satisfeita, essa condição garante que a estimativa de habilidade independe dos itens usados e que a dificuldade do item independe da amostra, o que torna possível comparar grupos testados com instrumentos distintos. O modelo de 2 parâmetros relaxa essa restrição e permite que itens tenham poderes discriminativos diferentes, ao custo de exigir amostras maiores. O modelo de 3 parâmetros é o padrão nas avaliações educacionais de larga escala com itens de múltipla escolha. A escolha entre eles reflete o que o pesquisador ou pesquisadora quer poder afirmar com os dados, e não apenas uma preferência técnica.

Dois postulados condicionam a aplicação de qualquer modelo TRI. O primeiro é a unidimensionalidade: o modelo pressupõe que um único traço latente dominante explica o desempenho no conjunto de itens. Se o instrumento mede dois construtos distintos com pesos equivalentes, a habilidade estimada será uma mistura dos dois e a interpretação substantiva fica comprometida. O segundo postulado é a independência local: dado o nível de habilidade do respondente, as respostas a quaisquer dois itens são estatisticamente independentes entre si. Esse postulado falha quando itens são contextualizados em torno de um mesmo texto, ou quando a sequência do instrumento induz efeitos de fadiga ou aprendizado. A violação produz erros padrão subestimados e índices de ajuste inflados (Hambleton & Swaminathan, 1985; Lord & Novick, 1968/2008).

A estimação dos parâmetros envolve um problema circular de que estimar a habilidade dos respondentes requer conhecer os parâmetros dos itens, e estimar os parâmetros dos itens requer conhecer as habilidades. R. Darrell Bock e Marcus Lieberman, em *"Fitting a Response Model for n Dichotomously Scored Items"*, de 1970, e Bock e Murray Aitkin, em *"Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm"*, de 1981, resolveram esse problema com métodos iterativos que tornaram viável a estimação em amostras de grande porte. Respondentes que acertaram ou erraram todos os itens não têm estimativa de habilidade possível por máxima verossimilhança. Métodos bayesianos produzem estimativas utilizáveis nesses casos, mas com maior incerteza.

A escala de habilidade é convencionalmente fixada com média 0 e desvio padrão 1. O que distingue a TRI da teoria clássica não é a escala em si, mas o fato de que itens e respondentes ficam expressos na mesma métrica. Isso possibilita a equalização: dois grupos testados com instrumentos diferentes podem ter suas habilidades comparadas, desde que os instrumentos compartilhem alguns itens comuns. Permite também acompanhar o desenvolvimento individual ao longo do tempo com provas progressivamente mais difíceis. São essas propriedades que tornaram a TRI o padrão metodológico dos grandes sistemas de avaliação educacional, como o *Programme for International Student Assessment* ou Programa Internacional de Avaliação de Estudantes (PISA) e o Sistema de Avaliação da Educação Básica (SAEB).

3. Vantagens, limites de validade e críticas conhecidas

Em relação à Teoria Clássica dos Testes, a TRI tem três vantagens mensuráveis. A primeira é a invariância dos parâmetros dos itens, isto é, a dificuldade e a discriminação

estimadas para um item não variam conforme a amostra usada para estimá-las, desde que o modelo se ajuste aos dados. A segunda é a invariância da estimativa de habilidade, pois o escore de um respondente não depende do conjunto específico de itens administrado. A terceira é a expressão de itens e respondentes na mesma escala, o que possibilita comparar grupos testados com instrumentos distintos, acompanhar o desenvolvimento individual ao longo do tempo e construir instrumentos adaptativos que concentram itens na região de habilidade do respondente.

A invariância dos parâmetros depende de condições que precisam ser verificadas empiricamente antes que as inferências sejam válidas. Identificar quando essas condições falham requer distinguir dois tipos de problema: limitações de escopo, que são restrições ao tipo de pergunta que o método consegue responder; e vieses sistemáticos, que são distorções nas estimativas produzidas quando os postulados do modelo não se sustentam nos dados.

As limitações de escopo da TRI decorrem de seus próprios pressupostos. O modelo de 3 parâmetros exige amostras de 500 a 1.000 respondentes para produzir estimativas estáveis; com amostras menores, os erros padrão crescem a ponto de tornar comparações entre grupos metodologicamente indefensáveis. A unidimensionalidade, pelas razões expostas na seção 2, é outra restrição de escopo, visto que quando o instrumento mede dois construtos com pesos equivalentes, o parâmetro estimado fica sem ancoragem empírica.

Os vieses sistemáticos surgem quando os postulados do modelo são violados sem que o pesquisador ou a pesquisadora o detecte. O mais frequente decorre da violação da independência local: quando itens compartilham um mesmo texto ou cenário, as respostas ficam correlacionadas além do traço latente, os erros padrão ficam subestimados e os índices de ajuste ficam inflados, induzindo à aceitação de modelos que não descrevem bem os dados, como mostrou Wendy Yen em *“Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model”*, de 1984. Um segundo viés surge da má especificação do modelo: aplicar o modelo de 1 parâmetro a dados com discriminação heterogênea produz estimativas distorcidas nas extremidades da escala, onde respondentes com habilidade alta ficam subvalorizados e os de habilidade baixa, superavaliados. Um terceiro viés é o funcionamento diferencial dos itens: quando um item funciona de forma diferente para subgrupos com a mesma habilidade latente, por razões de gênero, classe ou região, as estimativas favorecem sistematicamente um grupo sem que isso apareça nos índices globais de ajuste.

A crítica metodológica mais substantiva ao programa da TRI veio de Goldstein (1979). Segundo o autor, os testes utilizados para verificar se o modelo de Rasch se ajusta aos dados têm poder estatístico insuficiente para detectar as violações que invalidam a objetividade específica. O modelo é aceito não porque os dados confirmam suas restrições, mas porque os procedimentos de verificação não são sensíveis o suficiente para rejeitá-lo. Michell (1997) formula uma objeção de outra ordem, retomada em *“Normal Science, Pathological Science and Psychometrics”*, de 2000, de que a psicometria assume que os traços latentes são quantidades mensuráveis sem demonstrar que satisfazem as condições formais da mensuração conjunta, condições que raramente são testadas na prática. Borsboom et al. (2004) acrescentam que a interpretação causal das variáveis latentes, a ideia de que a habilidade gera as respostas, é

uma afirmação sobre a estrutura da realidade, não uma consequência do ajuste estatístico do modelo.

No contexto brasileiro, a aplicação da TRI no SAEB e no ENEM gerou controvérsia sobre a interpretação dos escores de proficiência. Klein (2019) mostrou que os pontos de corte usados para definir níveis de desempenho são decisões analíticas, não propriedades objetivas da escala, e que sua divulgação como parâmetros técnicos absolutos induz interpretações que os dados não sustentam.

4. Exemplos de aplicação em ciências sociais

Treier e Jackman (2008) partiram de uma questão de mensuração: a democracia, tratada nos estudos comparados como variável independente observável, é, na verdade, um traço latente medido com erro. O objeto eram os indicadores do *Polity IV*, usados em centenas de estudos para atribuir pontuações anuais de democracia a 184 países. A decisão metodológica central foi substituir a regra aditiva do *Polity* por um modelo ordinal bayesiano de resposta ao item, que estima os parâmetros de discriminação de cada indicador e produz, para cada país-ano, não um escore pontual, mas uma distribuição posterior com intervalo de credibilidade. O motivo para essa escolha se justificou pelo fato de que os indicadores do *Polity* são variáveis categóricas ordinais, não intervalares, e a agregação aditiva pressupõe pesos iguais e ausência de erro de medida, dois pressupostos que os dados não sustentam. O principal achado metodológico foi que o erro de medida no *Polity* é substancial. No ano 2000, 70 dos 153 países analisados tinham escores de democracia estatisticamente indistinguíveis dos Estados Unidos a 5% de significância. Um país-ano situado entre os percentis 60 e 70 da escala latente pode ter um escore *Polity* que varia de menos 2 a 6, uma amplitude de 8 pontos em uma escala de 21.

Coppedge et al. (2017) abordaram o mesmo problema pelo ângulo da coleta de dados. No projeto *Varieties of Democracy* (V-Dem), cada indicador em cada país é codificado, em regra, por cinco ou mais especialistas independentes. A decisão de usar múltiplos codificadores por célula, e não um único codificador ou um processo de consenso, foi tomada para que o modelo de mensuração pudesse estimar tanto o traço latente quanto o viés sistemático e a variância aleatória de cada codificador. O modelo aplicado é um modelo bayesiano ordinal de resposta ao item, desenvolvido pela equipe do V-Dem a partir da abordagem proposta por Daniel Pemstein, Stephen Meserve e James Melton em “*Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type*”, de 2010. O resultado metodológico foi que o V-Dem se tornou o único dos principais projetos de medição da democracia a fornecer estimativas de confiabilidade no nível do indicador, com intervalos de incerteza para cada estimativa de país-ano.

Desposato (2006) usou um modelo de pontos ideais para investigar por que mais de um terço dos deputados da Câmara brasileira troca de partido durante uma legislatura. A questão de pesquisa era identificar quais fatores (acesso a recursos do Estado, vantagem eleitoral ou compatibilidade ideológica) determinam as decisões de migração partidária. A decisão metodológica foi estimar os pontos ideais dos legisladores por meio do W-NOMINATE,

procedimento de escalonamento de votações nominais desenvolvido por Keith Poole e Howard Rosenthal e apresentado em *“Congress: A Political-Economic History of Roll Call Voting”*, de 1997. Trata-se de um modelo espacial de votação formalmente aparentado à TRI (Clinton et al., 2004), que expressa legisladores e centroides partidários na mesma escala latente e torna possível calcular a distância ideológica entre cada deputado e os partidos para os quais poderia migrar, o que não seria possível com escores de concordância brutos, que dependem do conjunto específico de votações analisadas em cada período. A pesquisa concluiu que a distância ideológica prevê a migração partidária especialmente em regiões com eleitorados mais desenvolvidos, onde incentivos programáticos pesam mais do que o acesso a recursos distributivos.

5. Debates contemporâneos e variações

A estimação bayesiana por métodos de Monte Carlo em cadeias de Markov tornou-se a variante dominante da TRI nas ciências sociais, principalmente na ciência política (Clinton et al., 2004). A abordagem trata parâmetros de itens e de pessoas como quantidades incertas com distribuições de probabilidade, o que torna possíveis extensões que a máxima verossimilhança não comporta. Andrew Martin e Kevin Quinn, em *“Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999”*, de 2002, estimaram pontos ideais de juízes da Suprema Corte dos Estados Unidos que variam ao longo do tempo. Devin Caughey e Christopher Warshaw, em *“Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model”*, de 2015, desenvolveram um modelo hierárquico dinâmico que estima a opinião agregada de grupos a partir de *surveys* esparsos e heterogêneos. O custo computacional desses modelos criou uma agenda própria, visto que Kosuke Imai, James Lo e Jonathan Olmsted, em *“Fast Estimation of Ideal Points with Massive Data”*, de 2016, propuseram algoritmos de aproximação que estimam pontos ideais em bases com milhões de observações em minutos, onde o MCMC exigiria semanas.

Um segundo debate trata da identificação e da interpretação substantiva da escala latente. A escala de um modelo TRI só fica definida após restrições de localização, de escala e, em modelos multidimensionais, de rotação. Essas restrições são escolhas do pesquisador ou pesquisadora e afetam as conclusões, como mostram Joseph Bafumi, Andrew Gelman, David Park e Noah Kaplan em *“Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation”*, de 2005. Fariss (2014) mostrou a dimensão empírica do problema, isto é, que as medidas internacionais de respeito a direitos humanos pareciam estagnadas porque o padrão de codificação dos relatórios de monitoramento ficou mais exigente ao longo das décadas, e um modelo que assume parâmetros de itens constantes no tempo confunde mudança no instrumento com mudança no fenômeno. A discussão sobre como validar modelos de variáveis latentes, retomada pelo próprio Fariss em *“Are Things Really Getting Better? How to Validate Latent Variable Models of Human Rights”*, de 2018, e por Kevin Reuning, Michael Kenwick e Fariss em *“Exploring the Dynamics of Latent Variable Models”*, de 2019, e a proposta de identificação orientada por teoria apresentada por Marco Morucci, Margaret Foster,

Kaitlyn Webster, So Jin Lee e David Siegel em “*Measurement That Matches Theory: Theory-Driven Identification in Item Response Theory Models*”, de 2025, respondem a essa dificuldade.

Na pesquisa comparada, a questão central é a equivalência de mensuração. Itens aplicados em países ou grupos culturais distintos podem funcionar de modo diferente para respondentes com o mesmo nível do traço latente, o que reproduz em escala internacional o problema do funcionamento diferencial dos itens, como discutem Eldad Davidov, Bart Meuleman, Jan Cieciuch, Peter Schmidt e Jaak Billiet em “*Measurement Equivalence in Cross-National Research*”, de 2014. Gary King, Christopher J. L. Murray, Joshua A. Salomon e Ajay Tandon, em “*Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research*”, de 2004, propuseram as vinhetas de ancoragem como correção: os respondentes avaliam cenários hipotéticos fixos e essas avaliações calibram as diferenças de uso da escala entre grupos.

Dois frentes mais recentes ampliam o alcance da técnica. A primeira é a aplicação a dados que não são respostas a questionários, como votações nominais em massa, textos políticos e, mais recentemente, estimativas de pontos ideais produzidas com modelos de linguagem de grande porte, como em “*Semantic Scaling: Bayesian Ideal Point Estimates with Large Language Models*”, de Michael Burnham, de 2024. A segunda é a consolidação de implementações de código aberto, como os pacotes mirt para R, apresentado por R. Philip Chalmers em “*mirt: A Multidimensional Item Response Theory Package for the R Environment*”, de 2012, e py-irt para Python, apresentado por John Lalor e Pedro Rodriguez em “*py-irt: A Scalable Item Response Theory Library for Python*”, de 2023, que reduziram a dependência de programas comerciais e facilitaram a verificação independente das análises, em linha com a agenda da ciência aberta. No Brasil, Jorge Luis Bazán, Márcia D'Elia Branco e Heleno Bolfarine propuseram, em “*A Skew Item Response Model*”, de 2006, um modelo com função de ligação assimétrica, que relaxa a suposição de simetria da curva característica do item.

6. Referências bibliográficas

Textos fundadores e genealogia

- Lord, Frederic M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum.
- Lord, Frederic M., & Novick, Melvin Robert. (2008). *Statistical theories of mental test scores*. Information Age Publishing. (Obra original publicada em 1968).
- Rasch, Georg. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). University of Chicago Press. (Obra original publicada em 1960).
- Samejima, Fumiko. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34 (Suppl. 1), 1-97. <https://doi.org/10.1007/BF03372160>.

Manuais e guias de referência

- Andrade, Dalton Francisco de, Tavares, Heliton Ribeiro, & Valle, Raquel da Cunha. (2000). *Teoria da resposta ao item: Conceitos e aplicações*. Associação Brasileira de Estatística.
- Hambleton, Ronald Kenneth, & Swaminathan, Hariharan. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff. <https://doi.org/10.1007/978-94-017-1988-9>.

Crítica metodológica

- Borsboom, Denny, Mellenbergh, Gideon Jan, & van Heerden, Jaap. (2004). The concept of validity. *Psychological Review*, 111 (4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>.
- Goldstein, Harvey. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal*, 5 (2), 211–220. <https://doi.org/10.1080/0141192790050207>.
- Michell, Joel. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88 (3), 355–383. <https://doi.org/10.1111/j.2044-8295.1997.tb02641.x>.

Estimação e aplicações nas ciências sociais

- Clinton, Joshua, Jackman, Simon, & Rivers, Douglas. (2004). The statistical analysis of roll call data. *American Political Science Review*, 98 (2), 355–370. <https://doi.org/10.1017/S0003055404001194>.
- Coppedge, Michael, Gerring, John, Lindberg, Staffan Ingemar, Skaaning, Svend-Erik, & Teorell, Jan. (2017). *V-Dem comparisons and contrasts with other measurement projects* (V-Dem Working Paper No. 2017:45). Varieties of Democracy Institute. <https://doi.org/10.2139/ssrn.2951014>.
- Desposato, Scott William. (2006). Parties for rent? Ambition, ideology, and party switching in Brazil's Chamber of Deputies. *American Journal of Political Science*, 50 (1), 62–80. <https://doi.org/10.1111/j.1540-5907.2006.00170.x>.
- Fariss, Christopher J. (2014). Respect for human rights has improved over time: Modeling the changing standard of accountability. *American Political Science Review*, 108 (2), 297–318. <https://doi.org/10.1017/S0003055414000070>.
- Klein, Ruben. (2019). Uma solução para a divergência de diferentes padrões no SAEB. *Ensaio: Avaliação e Políticas Públicas em Educação*, 27 (103), 229–249. <https://doi.org/10.1590/S0104-40362019002702060>.
- Treier, Shawn, & Jackman, Simon. (2008). Democracy as a latent variable. *American Journal of Political Science*, 52 (1), 201–217. <https://doi.org/10.1111/j.1540-5907.2007.00308.x>.

VER TAMBÉM:

Mensuração | Validade | Confiabilidade | Análise fatorial | Survey | Comparação

COMO CITAR ESTE VERBETE:

Ribeiro, Ednaldo. (2026). Teoria da Resposta ao Item. In *Enciclopédia Brasileira de Ciências Sociais (EBCS)*. ANPOCS. 10.17666/ebcs.2026.041.