Nota técnica de justificativa em favor do modelo de 1 parâmetro no ENAMED e na PND

Ricardo Primi,
Universidade São Francisco & EduLab21, Instituto Ayrton Senna.

O presente documento apresenta uma justificativa técnica para a adoção do modelo de um parâmetro da Teoria de Resposta ao Item (TRI), também conhecido como Modelo de Rasch ou 1PL nas avaliações do INEP, especialmente no ENAMED e na PND. A argumentação descreve a lógica clássica das abordagens estatísticas e psicométricas da TRI, situando suas diferenças epistemológicas e operacionais, discutindo a violação prática de pressupostos (especialmente unidimensionalidade) e examinando as consequências da escolha de modelos mais ou menos flexíveis.

Duas tradições da TRI: modelagem estatística e mensuração

Há duas principais vertentes no campo da Teoria de Resposta ao Item (TRI), aqui denominadas abordagem com ênfase em modelagem estatística e abordagem com ênfase psicométrica. Historicamente associada a Lord, Birnbaum e à tradição da *IRT* norte-americana (Lord & Novick, 1968; Hambleton, Swaminathan & Rogers, 1991) a abordagem de modelagem estatística privilegia modelos de dois e três parâmetros (2PL, 3PL). Nessa perspectiva, o objetivo central é modelar estatisticamente o padrão empírico entre a probabilidade de acerto e o traço latente. A flexibilidade dos parâmetros de discriminação (*a*) e chute (*c*) permite acomodar a variedade de formas observadas nas curvas características dos itens. De tradição distinta a abordagem psicométrica do Modelo de Rasch (Rasch, 1960; Wright & Stone, 1979) foca a medida do construto latente e a construção de uma régua aditiva com intervalos constantes e interpretáveis. A prioridade não é ajustar a forma empírica das curvas a qualquer custo, mas definir uma escala substantivamente coerente, com propriedades fortes de mensuração, como separabilidade de parâmetros e objetividade específica.

É evidente que os modelos 2PL e 3PL oferecem maior flexibilidade para reproduzir curvas empíricas. O modelo de Rasch, ao se restringir ao parâmetro de dificuldade, tende a gerar ajustes menos precisos no sentido estritamente descritivo, o que pode resultar em resíduos maiores. Na abordagem estatística, presume-se que essa menor precisão de ajuste produza erros nas estimativas de habilidade dos examinados. No entanto, esse texto questiona essa presunção de que modelos menos flexíveis implicam necessariamente perda de validade.

Unidimensionalidade: pressuposto central e suas limitações

Ambas as abordagens partem da suposição de unidimensionalidade: existe uma única dimensão responsável pelos acertos aos itens; assim, as covariâncias entre os itens são determinadas apenas pela associação de cada item à variável latente (que é capturada pelo parâmetro de discriminação e/ou pela carga fatorial e/ou pela correlação do item com o escore total do teste) e pelo seu nível de dificuldade. Os itens seriam organizados numa hierarquia ordenada única, em uma régua, dos mais fáceis aos mais difíceis.

Essa concepção implica que acertos inesperados em itens difíceis por pessoas de baixa habilidade seriam interpretados como exceções atribuíveis ao chute, uma vez que o domínio do conhecimento deveria seguir a mesma sequência para todos os indivíduos, determinada pela hierarquia (ordem) de dificuldade dos itens ao longo da dimensão latente. No contexto do conhecimento escolar, isso corresponderia a assumir que, em nível latente, as habilidades avaliadas no ENAMED e na PND, como matemática ou medicina, se organizam segundo uma única trilha de progressão, sem reconhecer caminhos alternativos de aprendizagem. Essa suposição impediria a possibilidade de que um estudante domine tarefas mais complexas em áreas específicas antes de dominar tarefas mais simples pertencentes a outras subáreas do conhecimento.

No entanto, o conhecimento humano não é unidimensional. Teorias contemporâneas de inteligência mostram distinções importantes entre inteligência cristalizada (Gc) e conhecimentos específicos (Gkn). Estudos sobre o desenvolvimento intelectual na vida adulta indicam crescente especialização por áreas e evidenciam que, mesmo dentro de uma única profissão, o conhecimento se estrutura em múltiplas competências, habilidades e saberes relacionados mas com relativa independência. Assim, a unidimensionalidade estrita é raramente observada (Ackerman, 1996; Primi, McGrew et al., 2023). Como consequência, torna-se muito menos plausível supor que os trajetos de aprendizagem sigam uma ordem rígida e uniforme para todos os estudantes, ditada exclusivamente pela sequência de dificuldade dos itens.

Diante dessa realidade, surge naturalmente a questão sobre a legitimidade do uso da TRI, dado que ela se fundamenta no pressuposto da unidimensionalidade. A resposta é que, embora a unidimensionalidade estrita raramente se verifique, costuma existir uma unidimensionalidade essencial, na qual uma dimensão geral convive com fatores locais de grupo, aproximando o fenômeno de um modelo hierárquico mais realista (Reise, 2012). Mesmo com esses desvios, a TRI pode ser utilizada de forma pragmática, desde que suas limitações sejam reconhecidas e que se adote um critério essencialmente utilitário baseado na validade preditiva das medidas.

Se, apesar das condições não ideais, as estimativas produzirem escores empiricamente úteis — especialmente para finalidades como equalização, comparabilidade e construção de métricas comuns — então a utilidade prática desses benefícios justifica o emprego da TRI. Em outras palavras, a adequação do modelo decorre da capacidade de gerar medidas funcionalmente válidas, mesmo sob pressupostos apenas parcialmente atendidos.[1]

Consequências práticas da violação da unidimensionalidade

Quando há múltiplas dimensões (ainda que organizadas sob um fator geral principal): (a) padrões aparentemente incoerentes tornam-se possíveis: indivíduos podem acertar itens difíceis e errar itens mais fáceis e (b) Itens de boa qualidade conceitual podem apresentar baixa discriminação simplesmente porque não há outros itens daquela habilidade na prova para formar covariâncias suficientes, e (c) Itens que representam habilidades pouco desenvolvidas no sistema educacional podem parecer “ruins” estatisticamente, embora sejam essenciais para a validade de conteúdo.

Discriminação como multidimensionalidade devido a flutuações locais

Ao aplicar a modelagem estatística nessas condições, a estimativa da discriminação faz com que itens mais fortemente correlacionados à dimensão geral predominante na prova, que podem ou não coincidir com a dimensão mais válida, recebam maior peso (Primi, et. al. 2018). Em provas com distribuição desigual de conteúdo, isso pode gerar ponderações discutíveis.

Considere o seguinte exemplo: em um teste com 10 itens de matemática, sendo 7 de aritmética, 1 de álgebra, 1 de geometria e 1 de análise gráfica, os 7 itens de aritmética dominarão a covariância e, portanto, apresentarão maior discriminação. Um aluno A que acerte os 7 itens de

[1] Essa argumentação de natureza utilitarista foi precisamente empregada por Lord e Novick (1968) ao reconhecerem que as medidas psicológicas não possuem, em sentido estrito, propriedades intervalares. Ainda assim, no tratado seminal da psicometria, os autores adotam a suposição de medidas intervalares como um artifício pragmático, justificando-a não por fundamentos estritamente teóricos, mas por sua utilidade empírica. Para eles, a legitimidade de tratar escores como intervalares decorre do fato de que tal suposição frequentemente produz escalas mais úteis, interpretáveis e empiricamente preditivas. Assim, a validade dessa operação depende menos da fidelidade aos pressupostos formais e mais de quão bem a escala resultante cumpre sua função prática, especialmente em termos de predição e utilidade estatística. Em suas palavras: “Trataremos uma medida como possuindo propriedades de escala intervalar, embora seja claro que o procedimento de mensuração e a teoria que o fundamenta produzam apenas uma escala nominal ou, na melhor das hipóteses, ordinal. Ao tratar dados por meio de métodos de intervalo nesses casos, estamos, de fato, estipulando uma função de distância específica para nossa escala, mesmo quando o processo subjacente de mensuração e a teoria que o apoia não a fornecem. Isso poderia ser considerado um reforço arbitrário de nosso modelo. Entretanto, do ponto de vista pragmático, a única avaliação significativa desse procedimento é aquela baseada na utilidade da escala resultante. Se construímos um escore de teste contando respostas corretas (escore zero–um) e tratamos os escores resultantes como possuindo propriedades intervalares, o procedimento pode ou não produzir um bom preditor de algum critério. Na medida em que essa forma de escalonamento produz um bom preditor empírico, a estipulação da escala intervalar é justificada (Lord e Novick, 1968, p.22)”.

aritmética terá uma habilidade estimada maior do que um aluno B que acerte 4 itens de aritmética e os três itens especializados. Mas, nesse caso, quem realmente sabe mais? O aluno A, que demonstra domínio apenas sobre aproximadamente um terço do conteúdo avaliado? Ou o aluno B, que revela conhecimento nas diferentes áreas contempladas pelo teste?

A resposta depende do critério substantivo adotado, e não da estrutura estatística da prova. A modelagem da discriminação pode capturar, em parte, características locais do teste, que não correspondem necessariamente a propriedades "verdadeiras" do conhecimento avaliado. Como argumenta Wright (1992, 1995), as inclinações empíricas das curvas características dos itens (CCIs) sempre variam; a questão inferencial central é decidir se essas variações devem ser tratadas como características estáveis e substantivas dos itens — como pressupõem os modelos 2PL e 3PL — ou como flutuações locais e instáveis, dependentes da amostra, ou uma combinação desses dois fatores.

Ao estimar o parâmetro a, corre-se o risco de perpetuar idiossincrasias amostrais, marcadas por instabilidade e forte dependência do conjunto específico de itens e respondentes, reificando padrões locais como se fossem propriedades essenciais do item. Em alguns casos, a discriminação pode estar associada ao funcionamento diferencial do item, como demonstrado por Masters (1988). Em síntese, embora a discriminação possa sinalizar itens mais informativos e validos sob certas condições, ela também pode refletir uma idiossincrasia do conjunto de itens que compõe a prova, não devendo ser interpretada automaticamente como evidência de maior validade.

Considere ainda um segundo exemplo: duas provas de matemática a serem equalizadas, cada uma composta por 10 itens. A Prova A contém 7 itens de probabilidade e 3 de geometria; a Prova B contém 7 itens de geometria e 3 de probabilidade. Desde o início, essas provas não são plenamente equivalentes em termos de conteúdo, o que já dificulta o processo de equalização. Entretanto, ao modelarmos a discriminação, um problema adicional emerge: na Prova A, os itens de probabilidade tenderão a apresentar maior discriminação devido à homogeneidade interna desse bloco, adquirindo peso desproporcional; na Prova B ocorrerá o inverso, com os itens de geometria tornando-se mais discriminativos e, portanto, mais influentes na estimativa de habilidade.

A discriminação, nesse caso, exacerba a discrepância entre as provas. Isso ocorre porque a discriminação não representa exclusivamente uma característica absoluta do item, refletindo uma suposta capacidade intrínseca de medir um traço latente verdadeiro. Ao contrário, ela também é influenciada pelas condições locais e contextuais dos demais itens que compõem a prova. Assim, quando se utiliza esse parâmetro, amplifica-se artificialmente uma diferença que já existia no conteúdo, agravando a desigualdade entre as provas e dificultando ainda mais a equalização.

Implicações da modelagem da discriminação para a equalização no ENAMED e na PND

Ao examinar as propriedades fundamentais de uma escala de medida, Wright (1992) argumenta que a definição do construto exige que a hierarquia de dificuldade dos itens permaneça estável para todos os níveis de habilidade. Isso significa que as curvas características dos itens (CCIs) devem ser paralelas e não se cruzar. Se se cruzarem, em razão de diferentes inclinações, a ordem relativa de dificuldade dos itens muda conforme o nível de habilidade da pessoa. Esse fenômeno implica que o construto subjacente não é definido por uma régua única, mas por múltiplas réguas — uma para cada faixa de habilidade. Assim, um item pode ser "mais fácil" para indivíduos de baixa habilidade, mas "mais difícil" para indivíduos de alta habilidade, o que inviabiliza a noção de um traço único e estável.

No contexto do ENAMED e na PND, essa discussão torna-se particularmente relevante porque se busca uma equalização de formas distintas de prova, frequentemente sem itens comuns entre elas. Quando não há itens comuns, recorre-se a métodos de vinculação indireta, onde identificam-se itens de conteúdo e dificuldade similares entre as formas, constroem-se hierarquias de itens para cada prova, e então essas hierarquias são alinhadas para produzir um

sistema comum de medidas. (ver *virtual equating of test forms* em https://www.winsteps.com/winman/equating.htm).

Esse processo depende criticamente da suposição de que cada forma de prova mede o mesmo construto e que suas hierarquias de dificuldade são comparáveis. No entanto, quando modelos 2PL ou 3PL são utilizados, a introdução dos parâmetros de discriminação e chute pode potencialmente alterar a ordenação dos itens, tornando a hierarquia dependente do contexto local da prova — isto é, dos demais itens que compõem aquela forma específica. Como a discriminação não expressa estritamente uma propriedade absoluta do item, mas uma característica também influenciada pela homogeneidade ou heterogeneidade dos itens ao seu redor, duas provas diferentes do ENAMED e na PND podem gerar ordenações distintas para itens conceitualmente equivalentes criando uma dificuldade para se achar itens ancoras. Se a hierarquia muda de uma forma para outra não porque o construto é diferente, mas porque a discriminação responde ao contexto local dos itens, torna-se desafiador identificar com segurança quais itens são realmente equivalentes.

O modelo de Rasch, ao contrário, porque se restringe a abstrair uma relação média geral entre probabilidade de acerto e o traço latente, garante que todas as CCIs permaneçam paralelas e que a hierarquia de dificuldade dos itens seja estável. Assim, Prova A e Prova B podem diferir no conteúdo, mas a ordenação de seus itens não dependerá da heterogeneidade ou homogeneidade relativa de cada forma. Isso preserva a condição fundamental para o virtual equating: uma definição única do construto, com itens colocados em posições comparáveis na escala.

Para o ENAMED e na PND, nos quais a equalização entre formas é essencial e muitas vezes ocorre sem itens comuns, esse ponto é decisivo: a modelagem da discriminação tem o potencial de tornar a equalização mais difícil, menos estável e menos interpretável, enquanto o modelo de Rasch preserva exatamente as propriedades necessárias para viabilizar uma escala comum coerente entre formas distintas.

O parâmetro de chute (c) e suas limitações

O parâmetro *c* do 3PL modela a probabilidade de acerto por pessoas com baixa habilidade, isto é, assintota inferior. Essa modelagem produz um efeito ainda mais radical nas pontuações dos sujeitos. Fundamentada na suposição de unidimensionalidade estrita os acertos em itens com *c* maior que zero, cuja dificuldade excede substancialmente o nível estimado de habilidade do respondente são interpretados como “chute”. Com base nessa inferência, o modelo reduz o peso desses acertos, podendo inclusive anulá-los quando a distância entre habilidade e dificuldade é grande. Contudo, essa suposição é problemática, pois pressupõe uma unidimensionalidade perfeita que raramente se verifica em dados reais de avaliação educacional (Primi & Cicchetto, 2018).

Como já advertido por Lord e Novick (1968) e posteriormente enfatizado por Lord (1980), a estimação do parâmetro de chute é estatisticamente instável e fortemente dependente da amostra, além de carecer de base empírica sólida quando múltiplas habilidades influenciam as respostas. Wright (1995) acrescenta que, na maioria das aplicações práticas, há pouca evidência de chute sistemático nos dados; assim, a introdução do parâmetro *c* tende a penalizar indiscriminadamente os respondentes — especialmente aqueles com perfis de conhecimento específicos — em vez de corrigir efetivos comportamentos de chute. Nesses contextos, o parâmetro *c* não apenas falha em melhorar a validade da medida, como pode introduzir viés adicional nos escores.

Wright (1995) argumenta que o fenômeno do chute não deve ser tratado como uma propriedade estável dos itens, mas como um comportamento ocasional de determinados respondentes. Segundo essa perspectiva, não são os itens que “chutam”, mas as pessoas que, em circunstâncias específicas, recorrem ao chute — e nem todos os indivíduos o fazem, tampouco de forma consistente nos mesmos itens. Dessa forma, a identificação do chute torna-se mais precisa quando se analisa o padrão de respostas dos indivíduos, via *outfit*, permitindo que esse comportamento seja diagnosticado, rotulado e tratado de maneira explícita, em vez de ser

incorporado de forma indiscriminada ao modelo por meio de um parâmetro estrutural do item. Por exemplo, o *outfit* elevado pode informar o erro de medida da pontuação de um sujeito que contenha padrões inesperados de resposta, reduzindo a confiança na estimativa pontual de sua habilidade. Essa abordagem preserva a integridade da hierarquia de dificuldade dos itens e evita penalizações generalizadas, reforçando a ideia de que o tratamento do chute deve ocorrer no nível do respondente.

Justificativa do uso do Modelo de Rasch (1PL)

A abordagem de modelagem estatística sustenta que, sem modelar as nuances (discriminação e chute), incorremos em erros de estimação. Mas, na presença de violação estrutural da unidimensionalidade, ocorre o oposto: a flexibilidade pode introduzir vieses sistemáticos, como superponderar habilidades super-representadas e desvalorizar acertos de habilidades incomuns. Esse conjunto de características reforça a posição clássica segundo a qual, quando pressupostos estruturais são violados, a modelagem de parâmetros adicionais não aumenta a validade—e pode, inclusive, reduzi-la.

Dado esse cenário, o modelo de Rasch oferece vantagens importantes: embora também pressuponha unidimensionalidade, o Rasch não modela variações potencialmente idiossincráticas de discriminação e chute. Assim, ele não ira re-ponderar itens em função de padrões locais de covariância, evitando vieses derivados da distribuição desigual de conteúdo. Os acertos são ponderados igualmente, preservando melhor a diversidade de habilidades presentes no teste. Acertos inesperados são tratados por meio de estatísticas de ajuste (*person fit, infit outfit*), que servem para qualificar a precisão da medida sem alterar a estrutura do escore.

Do ponto de vista da validade de critério, a simplicidade do Rasch pode ser vantajosa: testes muito homogêneos (alta precisão interna) frequentemente exibem menor poder preditivo, pois sacrificam diversidade de conteúdo. O Rasch, ao não concentrar peso excessivo em itens redundantes, pode preservar melhor a amplitude do construto.

Em resumo, a justificativa para o uso do modelo de 1 parâmetro decorre de: (a) violação inevitável da unidimensionalidade estrita em avaliações educacionais reais; (b) vieses introduzidos pelos parâmetros de discriminação e chute quando a estrutura de conteúdo é desigual; (c) instabilidade estatística dos parâmetros adicionais reconhecida por Lord; (d) maior preservação da validade de conteúdo ao evitar reponderações artificiais; (e) maior robustez interpretativa: uma régua única, linear e teoricamente coerente; (f) pragmatismo científico: quando pressupostos estão apenas parcialmente atendidos, modelos mais simples produzem medidas mais estáveis, transparentes e úteis — exatamente o tipo de posição defendida por Lord ao enfatizar que modelos estatísticos devem servir como ferramentas pragmáticas, não como descrições ontológicas perfeitas.

Referências

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*(2), 227–257. https://doi.org/10.1016/S0160-2896(96)90016-1

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), 7–16.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25*(1), 15–29.

Primi, R., & Cicchetto, A. A. (2018). Como os escores do ENEM são atribuídos pela TRI? In *Anais do VI CONBRATRI: Métodos para detecção de fraudes em testes* (pp. 1–1). ABAVE/CONBRATRI.

Primi, R., Nakano, T. C., McGrew, K., & Schneider, J. (2023). *Educação no século XXI: Inteligência, pensamento crítico e criatividade* (Vol. 1). Hografe & Instituto Ayrton Senna.

Primi, R., McGrew, K., Schneider, J., Nakano, T. C., & Dias, N. M. (2023). Inteligência: Como é concebida com base nos modelos psicométricos? In R. Primi, T. C. Nakano, K. McGrew, & J. Schneider (Eds.), *Educação no século XXI: Inteligência, pensamento crítico e criatividade* (Vol. 1, pp. 13–77). Hografe & Instituto Ayrton Senna.

Primi, R., Nakano, T. C., & Wechsler, S. M. (2018). Using four-parameter item response theory to model human figure drawings. *Avaliação Psicológica, 17*(4), 473–483. https://doi.org/10.15689/ap.2018.1704.7.07

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Wright, B. D. (1992). IRT in the 1990s: Which models work best? 3PL or Rasch? *Rasch Measurement Transactions, 6*(1), 196–200. https://www.rasch.org/rmt/rmt61a.htm

Wright, B. D. (1995). 3PL IRT or Rasch? *Rasch Measurement Transactions, 9*(1), 408. https://www.rasch.org/rmt/rmt91b.htm

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.