UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
Instituto de Física
Programa de Pós-Graduação em Ensino de Física
Mestrado Profissional em Ensino de Física

EXPLORANDO O MUNDO DAS AVALIAÇÕES DE LARGA ESCALA

Wanderley Paulo Gonçalves Jr

&

Marta Feijó Barroso

Material instrucional associado à dissertação de mestrado de Wanderley Paulo Gonçalves Jr, apresentada ao Programa de Pós-Graduação em Ensino de Física da Universidade Federal do Rio de Janeiro.

Rio de Janeiro
2012

Explorando o Mundo das Avaliações de Larga Escala
(Versão Preliminar)

Wanderley P. Gonçalves Jr e Marta F. Barroso

Cara colega, caro colega

Avaliar é uma das tarefas mais executadas por professores, de todos os níveis: no ensino fundamental, no ensino médio e no ensino superior. Na maior parte das vezes, essa tarefa é considerada pesada por todos nós.

Hoje, cada vez mais, também somos, de diferentes formas, avaliados externamente – ou melhor, nossos alunos o são. E os resultados são divulgados, na forma de rankings nos jornais, de premiações e outros.

Este trabalho que você tem em suas mãos foi elaborado pensando em compartilhar algumas das coisas que aprendemos, ao analisar avaliações em educação e ao pensar o que fazemos em sala de aula com os resultados dessas avaliações em mãos.

Esperamos que você chegue aonde chegamos: a pensar que a avaliação precisa ser objeto de reflexão por nós, professoras e professores, para que seus resultados possam ter consequências nos trabalhos que executamos em nosso dia a dia. E a principal consequência deveria ser nos permitir melhorar o nosso trabalho!

Esse texto propõe-se a discutir a avaliação da aprendizagem, e como analisamos esta avaliação. Começamos com uma breve revisão e comentários a respeito de processos e técnicas de medida. Inicia-se pela ideia de medição nas ciências físicas e a dificuldade de transpor esse conceito para as ciências sociais. Essa discussão é importante porque... exatamente o que queremos dizer com "avaliar a aprendizagem"?.

Tudo isso vem recheado com alguns exemplos da Física.

Em seguida, uma questão que sempre parece árida: lidar com as ideias numéricas e as técnicas envolvidas para pensar os resultados de avaliações. Em outras palavras, discutir um pouco sobre estatística: as ideias de média, de desvio padrão, de análise de dados, de fazer modelos para os dados com o "ajuste" de dados aos modelos. São apresentadas algumas formas de fazer essas discussões de forma quase automática, com auxílio de alguns softwares.

Finalmente, chega-se ao ponto: as técnicas de avaliação de aprendizagem! A medida de aprendizagem através da aplicação de testes ("provas"), e as teorias da psicometria para obter "notas", ou escores, a partir desses testes, com a apresentação da Teoria Clássica dos Testes (TCT) e as ideias básicas da Teoria da Resposta ao Item (TRI), utilizada no ENEM.

Sumário

Fazer medidas 2

Exemplos de Medida: um exemplo de avaliação de aprendizagem em Física .. 5

Como fazer medidas: mais um exemplo da Física 7

Como apresentar resultados das medidas 9

Probabilidade: conceitos básicos 10

Melhor valor: medidas de tendência central 14

Medidas de dispersão 17

Função de distribuição, ou densidade de probabilidade 20

Distribuição de Gauss ou distribuição normal 21

Ajuste de dados – Regressão Linear 24

Testes: a Teoria Clássica de Testes 25

Testes: a Teoria da Resposta ao Item 27

Construir uma ICC (curva característica do item) e o modelo de Rasch 30

Modelos para a Teoria da Resposta ao Item 36

Fazer medidas

Ao trabalhar com ideias e grandezas em Física, ou em alguma das ciências ditas "exatas", parece natural imaginar que tudo pode ser expresso em termos de grandezas mensuráveis que se relacionam por meio de equações.

As medidas são feitas, em geral, com dois objetivos em mente: descrever comportamentos (função descritiva), ou seja, estudar como os resultados experimentais se comportam, ou obter relações entre as variáveis medidas (função explicativa).

No campo das ciências naturais, e para ser mais específico, na Física (Alonso e Finn 1972)

> "*medir é um processo que nos permite atribuir um número a uma propriedade física como resultado de comparações entre quantidades semelhantes, sendo uma delas padronizada e adotada como unidade*." (pág. 13)

Mas é muito comum imaginar que não se pode aplicar os raciocínios da física às demais ciências, como a psicologia, sociologia, economia, antropologia, ensino e aprendizagem em física, etc. Isso parece sugerir que as ciências como a Física constroem afirmações verdadeiras, enquanto as demais ciências não (Babbie 2005), como se houvesse um questionamento do status "científico" dessas outras áreas. Em outras palavras, será que o comportamento humano pode ser sujeito a medidas? Será que o "método científico" pode ser aplicado ao comportamento humano?

Vejamos: em qualquer área do conhecimento, as teorias existentes quase sempre resultam de uma combinação de processos dedutivos e indutivos, isso é, uma explicação inicial para uma observação é testada, reformulando a ideia inicial até a construção de um corpo teórico plausível e que explique as observações conhecidas. Neste processo, conceitos são utilizados, e muitas vezes para defini-los melhor é necessário propor uma forma de medi-lo, fazendo o que é chamado de definição operacional do conceito.

A generalização da ideia de medir pode ser feita (Allen e Yen 2002):

> "*A medida é a atribuição sistemática de números para indivíduos de um conjunto, tendo como objetivo a representação das propriedades desses indivíduos. Os números atribuídos a esses indivíduos devem ser obtidos através de procedimentos cuidadosamente prescritos e reprodutíveis. Por exemplo, testes de personalidade geram suas pontuações a partir do uso das mesmas instruções, perguntas, e procedimentos de pontuação para cada examinando. Essas pontuações não podem ser comparadas de forma significativa, se a cada um dos examinandos forem dadas instruções diferentes ou se diferentes procedimentos de pontuação forem utilizados para estabelecê-los. Na mensuração, os números são atribuídos de forma sistemática e podem possuir formatos variados. Por exemplo, pode-se atribuir-se às pessoas com cabelos vermelhos o número "1" e às pessoas com cabelos castanhos "2" como forma de medida. Neste caso, os números são atribuídos aos indivíduos de uma forma sistemática de maneira que as diferenças entre os escores representam as diferenças na propriedade que está sendo medida (cor do cabelo). Da mesma forma, dando-se a um examinando uma nota 98 na prova de matemática ou a nota 54 num teste de personalidade, estará se realizando uma medição, desde de que os números sejam sistematicamente atribuídos para representar as diferenças de desempenho nos testes de matemática ou de personalidade*". (pág. 2)

Nas áreas da ciência em que o objeto de estudo é o conhecimento humano, é possível e muitas vezes desejável a realização de medidas sistemáticas. Por exemplo, as características sociais "cor de pele", "idade", "cidade natal", "sexo", "renda familiar" são facilmente mensuráveis.

Mas o que seria medir "aprendizagem em física"? Como seria medir a "religiosidade" de uma pessoa? Isto é, conceitos abstratos, atitudes, são passíveis de medida? Uma reflexão sobre isso é feita por Babbie (Babbie 2005)

> *"Deve-se reconhecer que todas estas medidas (todas medidas, aliás) são basicamente arbitrárias. O cientista social não pode descrever uma pessoa inequivocamente como "alienada" e outra como "não alienada". Pessoas serão, ao invés, descritas como relativamente mais ou menos alienadas – ou seja, comparando uma com a outra. Esta característica, entretanto, não é prerrogativa das ciências sociais, como demonstram a "escala de dureza" usada nas ciências físicas, a "escala Ritchter" para terremotos, etc. Ninguém pode dizer que um metal é "duro" ou que um terremoto é "severo", apenas que é mais "duro" ou mais "severo" que o outro."* (p. 59)

Muitas vezes, por motivos diversos, temos necessidade ou desejo de medir conceitos abstratos como classe social, racismo ou a aprendizagem em física.

Por serem abstratos, esses conceitos, em geral, não passam de ideias gerais na mente do pesquisador. Para que se possa medi-los, é necessário melhorar essas ideias, fornecer uma definição operacional, ou um conceito diferente. Começa-se levantando as características desses conceitos, reduzindo-se as características a indicadores empíricos específicos. Essa não será certamente a definição completa e indiscutível do conceito, mas representarão as características dele que julgamos úteis para a necessidade.

Por exemplo, ao se perguntar a professores de ensino médio o que entendem por aprendizagem em física, alguns afirmariam que saber física é saber resolver problemas numéricos de física; outros diriam que é solucionar problemas que necessitam de conhecimento dos conceitos físicos estudados, ou ainda, que a aprendizagem se revelaria se o aluno conseguisse aplicar os conceitos estudados em seu cotidiano. Cada uma dessas posições apresenta uma característica que pode ser utilizada para uma definição operacional do que seria aprendizagem em física.

Podemos olhar para definições acadêmicas mais conhecidas. Como exemplo, o PISA (Programa Internacional de Avaliação de Estudantes) avalia, em ciências, o quanto o aluno adquiriu de "letramento científico", definido como "a capacidade de usar o conhecimento científico para identificar questões e chegar a conclusões baseadas em evidências para entender e ajudar a tomar decisões a respeito do mundo e as mudanças causadas a ele pela atividade humana".

Já nos Parâmetros Curriculares Nacionais para o Ensino Médio, na área de Física, definem como aprendizagem em física "o resgate do espírito questionador do aluno e seu desejo de explorar o mundo, reconhecendo a física como cultura e como possibilidade de conhecer este mundo que o cerca". Neste processo, o aluno deve desenvolver uma atitude reflexiva e autocrítica diante dos erros cometidos, gerenciar os conhecimentos adquiridos e compreender a predominância de aspectos técnicos e científicos na tomada de decisões sociais significativas e os conflitos gerados, nestes, pela negociação política.

O ENEM (Exame Nacional de Ensino Médio) entende que aprender física significa que o aluno adquiriu ou desenvolveu um conjunto de habilidades que, em conjunto, representam a aquisição e o desenvolvimento de algumas competências (aptidões) elencadas na Matriz de Referência desse exame.

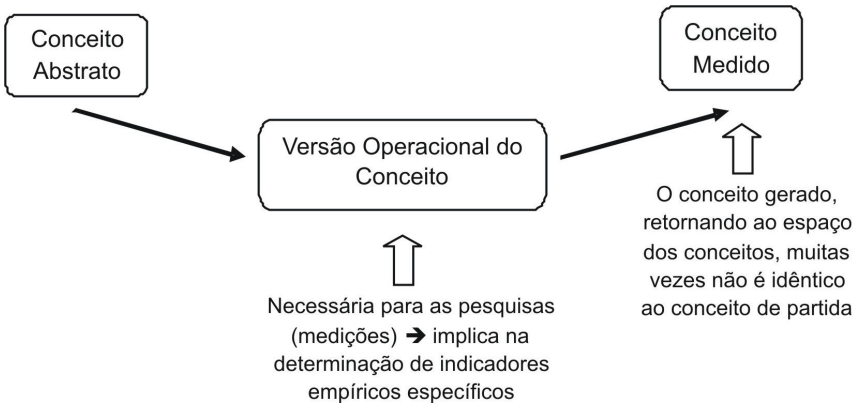
Em resumo, conceitos abstratos podem ser medidos se houver uma definição operacional plausível para eles elaborada pelas pessoas envolvidas; quase sempre não é possível construir-se um significado universal para esses conceitos. Por isso, não faz sentido dizer que o conceito operacional de um conceito abstrato está certo ou errado, mas sim dizer se ele é mais ou menos útil para responder às questões propostas.

É necessário refletir sobre as consequências desta discussão. A partir do momento que o pesquisador define o conceito operacional e quais aspectos desse conceito serão estudados, pode-se dizer que ele não coleta dados para o seu estudo, mas sim que cria esses dados.

Medida de tempo

O que é tempo? Essa pergunta é título de livros, de teses filosóficas, de discussões ardorosas... Mas existe alguma dúvida quanto ao fato que sabemos medir o tempo? Isto é, que temos uma definição operacional simples para ele? Contar o ângulo girado por um motor (relógio de pulso), as batidas de um metrônomo, entre outros!

A nossa conclusão então é que medir um conceito abstrato, muitas vezes, implica em "recriar" este conceito através de parâmetros concretos que sejam possíveis de ser medidos, gerando assim um novo conceito (o conceito operacional), que apesar de útil, não é a medida real do conceito abstrato em questão. De forma esquemática,



Define-se como a "operacionalização" de um conceito (Babbie 2005) o processo pelo qual são especificadas observações empíricas que podem ser tomadas como indicadores dos atributos contidos deste conceito.

Exemplos de Medida: um exemplo de avaliação de aprendizagem em Física

Pode-se agora exemplificar como um mesmo conceito pode ser tratado ou medido de diferentes formas, fornecendo diferentes visões do conceito.

No ensino médio, a discussão de movimentos de objetos próximos à superfície da Terra faz parte do conteúdo abordado na maior parte das escolas. Este tópico também é abordado no primeiro ano dos cursos universitários da área de ciência e tecnologia.

Tradicionalmente (a forma de trabalho da maior parte das escolas) este tópico é ensinado com auxílio de exercícios que envolvem trabalho exaustivo com equações e gráficos na resolução de problemas numéricos. Coerentemente, a avaliação da aprendizagem desse conteúdo é feita através de provas e trabalhos com exercícios numéricos desse mesmo tipo.

Vejamos um exemplo.

> Um corpo é lançado verticalmente para cima, em um local onde o efeito do atrito com o ar é desprezível. Se a velocidade de lançamento foi de 30 m/s qual a altura máxima atingida pelo corpo? (Considere g = 10 m/s^2.)

O processo de resolução é tradicional e trivial. Utilizam-se as equações que descrevem movimentos com aceleração constante. Em particular, a chamada "equação de Torricelli", $\mathbf{v^2 = v_0^2 + 2a\Delta s}$, onde $\mathbf{v}$ é a componente da velocidade no eixo vertical, $\mathbf{v_0}$ é a velocidade inicial, $\mathbf{a}$ é a aceleração e $\mathbf{\Delta s}$ o deslocamento realizado.

Sabendo que ao atingir a altura máxima a componente vertical da velocidade do corpo se anula, o procedimento habitual de resolução dos estudantes é:

$$\mathbf{v^2 = v_0^2 + 2a\Delta s}$$
$$\mathbf{0^2 = 30^2 + 2(-10)\Delta s}$$
$$\mathbf{20\Delta s = 900}$$
$$\mathbf{\Delta s = 45\,m}$$

Ao corrigir esse problema nos testes e provas, observa-se que o aluno, para o ponto de altura máxima e consequentemente de velocidade instantânea zero, substitui o valor da aceleração da gravidade por 10 m/s^2.

Para um professor que utiliza este tipo de problema, medir a aprendizagem em física nesse conteúdo consiste na verificação da memorização e aplicação de equações em situações numéricas. Dentro dos parâmetros estabelecidos pelo professor para o conceito de aprendizagem deste conteúdo, há total coerência em suas ações e ele pode afirmar que o aluno que acerta o exercício aprendeu o conteúdo.

No entanto, se a concepção de aprendizagem dos conceitos físicos deste conteúdo implicar na exigência de uma compreensão mais conceitual das grandezas físicas envolvidas no fenômeno, pode-se propor outro tipo de exercício.

Observe as questões a seguir. Elas foram aplicadas a alunos calouros de um curso da área de ciências exatas da UFRJ, que na sua maioria são ingressam vindos de escolas tradicionais (comunicação privada, M.F.Barroso).

12) Uma bola é lançada verticalmente para cima. No ponto mais alto da trajetória da bola,

Resposta	Média	Total
sua aceleração é nula, e sua velocidade é não nula.	8.5%	8
sua velocidade e aceleração são nulas.	57.4%	54
sua velocidade é nula, mas a aceleração não é nula.	35.1%	33

Dos resultados, pode-se verificar que mais da metade dos estudantes responde que a aceleração no ponto mais alto da trajetória é nula. Em outras palavras, a repetição de exercícios numéricos durante o ensino médio não garantiu a compreensão

do conceito da aceleração da gravidade. E essa interpretação dos resultados é reforçada com a questão seguinte, aplicada ao mesmo grupo.

13) Considere duas situações:

Situação I - Uma bola é lançada para cima, verticalmente.

Situação II - Uma bola é largada do alto de uma torre.

Podemos afirmar que

Resposta	Média	Total
a aceleração depende da velocidade com que a bola é lançada na situação I, e da altura que é largada na situação II.	24.5%	23
na primeira situação, a aceleração é vertical e para cima, e na segunda a aceleração é vertical e para baixo.	51.1%	48
nada podemos afirmar sobre a aceleração, pois não temos nenhuma informação sobre as velocidades.	1.1%	1
nas duas situações, a bola tem a mesma aceleração.	11.7%	11
Nenhuma das respostas anteriores está correta.	12.8%	12

Esses resultados apontam para o fato que resolver problemas numéricos não garante a aprendizagem do conceito aceleração da gravidade. Em outras palavras, um professor que entende a compreensão do conceito como aprendizagem do mesmo não terá nos problemas numéricos bons parâmetros para sua avaliação.

Pesquisas em Ensino de Física indicam (Arons 1997) que existe uma dificuldade tremenda dos estudantes em situações como a contemplada nas duas questões respondidas pelos calouros da UFRJ. Logo, o conceito de aprendizagem em física adotado pelo professor exemplificado no primeiro problema (numérico) não é útil quando se deseja que o estudante adquira uma compreensão maior sobre o conceito de aceleração num da gravidade, pois, apesar de utilizar matematicamente o valor desta aceleração para o cálculo da altura atingida pelo corpo no primeiro problema apresentado, na maior parte das vezes ele não consegue relacionar a operação matemática com o conceito solicitado nas duas questões.

As duas formas de avaliar a aprendizagem, pela apresentação de um problema numérico e pela apresentação de questões conceituais, revelam resultados diferentes. Isto é, propusemos duas formas diferentes de operacionalizar a medida "aprendizagem de queda livre". E obtemos resultados discrepantes. Em outras palavras, a operacionalização forneceu medidas diferentes, nenhuma certa ou errada, apenas mais ou menos útil para o objetivo do proponente da medida.

Como fazer medidas: mais um exemplo da Física

Vamos retornar aos exemplos de Física, para entender com exemplos mais simples a discussão de como tratar os dados.

Ao realizar uma medida para descrever um fenômeno – e vamos tomar como exemplo a determinação do período de um pêndulo simples (um pequeno objeto pendurado na extremidade de um fio) – é necessário entender que a medida a ser

realizada deve expressar, de forma clara, um resultado que seja compreensível (e muitas vezes reprodutível) por outra pessoa. Na prática, precisamos de uma unidade e, mais do que isso, precisamos apresentar um número e a exatidão deste número.

Sobre a "**exatidão**" de uma medida

Mesmo nas ciências exatas como a física, nem sempre conhecemos qual é o "valor verdadeiro" de uma medida. A medida que fazemos deve sempre se aproximar deste valor, mas não sabemos quão próxima deste valor está a medida feita. E para comparar resultados, avaliar teorias, é muito importante conhecer esta "incerteza" na medida – isso é, qual é a nossa avaliação da proximidade possível entre o valor verdadeiro e a medida que fazemos.

Uma maneira mais ou menos óbvia de estimar a exatidão de nossa medida é repetí-la, em iguais condições. É intuitivo que quanto mais vezes repetirmos a medida, melhor vai ser o resultado final (tratando os dados obtidos). Mas... a repetição de uma medida em idênticas condições não fornece resultados idênticos! Experimente medir algumas vezes com um cronômetro o período de oscilação de um pêndulo... Essas pequenas diferenças são flutuações estatísticas em nossos resultados – e as incertezas associadas são chamadas de erros aleatórios.

Há outro tipo de inexatidão da medida, que não podem ser corrigidos ou minimizados por repetição. O exemplo é um instrumento cuja calibração se altera com o tempo. Outro exemplo é você medir a altura da pessoa em diferentes momentos do dia (sabe-se que em geral ao longo do dia a pessoa "encolhe"). Esse tipo de inexatidão é denominada erro sistemático, e é difícil de ser evitado.

Os conceitos de precisão e acurácia estão associados a esses dois tipos de inexatidões experimentais. Veja o exemplo:

"Um jogador de futebol está treinando cobranças de penâlti. Ele chuta a bola 20 vezes, e 20 vezes acerta na trave do lado direito do goleiro. Ele é extremamente preciso, pois seus resultados não apresentam nenhuma variação em nenhuma das 20 vezes. Em compensação, sua acurácia é nula – ele nunca consegue acertar o "valor verdadeiro", o gol.

Isso também tem efeitos em outros tipos de medida. Por exemplo, numa pesquisa eleitoral pode acontecer que o pesquisador entre numa sala com 20 pessoas que decidem informar ao pesquisador que votarão em um determinado candidato – o que não é necessariamente verdadeiro. Então, o levantamento de dados feito por este pesquisador é totalmente inacurado...

Em um laboratório de física, propusemos um experimento: cada um dos alunos mede um certo número de vezes o período de oscilação de um pêndulo simples com um cronômetro. Este cronômetro determinava o intervalo de tempo entre o ligar e desligar com medidas até centésimo de segundo.

Observou-se que era muito pouco provável a repetição dos valores encontrados na medida realizada por cada aluno. Essas medidas diferentes são resultado de muitos pequenos fatores, não controláveis pelo observador: a dificuldade de definir exatamente o final da oscilação, a demora ou a rapidez em apertar o botão de ligar e desligar, entre outros. E aí, o que devemos fazer? Qual das medidas é o período do pêndulo?

Como apresentar resultados de medidas

Para se determinar a cor preferida de um grupo de 30 pessoas, pode-se fazer a pergunta a cada uma delas. Para pesquisar a intenção de voto para Presidente da República, é praticamente impossível consultar todos os eleitores (a *população* ou o *universo estatístico*).

população
amostra

Recorre-se neste caso ao que se denomina de *amostra*, isto é, um grupo de indivíduos ou objetos pertencente ao universo pesquisado; a pergunta é feita a este subgrupo, e espera-se chegar ao resultado que reflita o todo. Cada pessoa da amostra é denominado *indivíduo* ou *objeto*. No caso da intenção de voto, as pessoas são os indivíduos.

Imagine que uma construtora pretenda lançar um empreendimento imobiliário num bairro da cidade. Para isso, a empresa faz uma pesquisa para sondar a preferência dos possíveis compradores em relação ao tamanho dos apartamentos, número de vagas de garagem, cor de fachada, área de lazer, sistema de segurança, etc. Cada uma dessas características é uma variável da pesquisa.

Na variável número de vagas de garagem, as opções podem ser "nenhuma", "uma", "duas" ou "mais de duas" vagas. Diz-se que esses são os *valores* ou *realizações da variável* "numero de vagas de garagem".

Essas variáveis, por sua vez, podem ser classificadas em dois grupos.

variáveis qualitativas
ordinais e nominais

O primeiro deles, das variáveis qualitativas, é composto pelas variáveis que apresentam como possíveis valores uma qualidade ou atributo dos indivíduos; como exemplo, a cor da pele, o esporte favorito, o grau de instrução, o gênero. Se essas variáveis qualitativas possuem uma ordem em seus valores, como ocorre com grau de instrução (fundamental, médio, superior), essa é uma variável qualitativa *ordinal*. No caso do esporte favorito, essa variável é uma variável qualitativa *nominal*.

variáveis quantitativas
discretas e contínuas

No segundo grupo estão as variáveis quantitativas, que possuem valores numéricos como possíveis realizações. Os exemplos são a idade do indivíduo, sua altura, seu peso, o número de irmãos ou dependentes. Essas variáveis também podem ser dividas em dois subgrupos: o das variáveis quantitativas *discretas*, que constituem as variáveis numéricas representadas por números inteiros como o número de irmãos ou dependentes, e o das variáveis quantitativas *contínuas*, quando a medida pode ser representada por números reais, como a altura ou o peso do indivíduo.

frequência
absoluta e relativa

Após a realização de medidas, costuma-se apresentar o resultado de diversas formas. Uma das formas corresponde à apresentação do número de vezes que um valor da variável é obtido, a denominada *freqüência absoluta* do valor. Quando se registra a frequência absoluta de um valor em relação ao total de valores obtidos, tem-se a *frequência relativa*.

Por exemplo, suponha que em uma pequena sala de aula tenha sido feito um levantamento sobre a descendência étnica por parte de mãe de 10 alunos e que o resultado seja o seguinte: três alunos descendentes de mãe negra, um de japonesa, quatro de italiana e dois de portuguesa. A frequência absoluta da variável "descendência" de cada um dos seus valores é: negros, 3; japoneses, 1; italianos, 4; portugueses, 2. Já a freqüência relativa da descendência italiana é: 4 em 10 ou 0,4 ou 4/10 ou 2/5 ou 40% (os dados estão apresentados na Tabela 1).

Tabela 1. Exemplo de frequência de uma medida qualitativa nominal.

Descendência étnica por parte de mãe			
	frequência	frequência relativa	frequência percentual
afrodescendente	3	0,3 (3/10)	30%
japonesa	1	0,1 (1/10)	10%
italiana	4	0,4 (4/10)	40%
portuguesa	2	0,2 (2/10)	20%
Total	10	1 (10/10)	100%

Probabilidade: conceitos básicos

Ao lançar uma moeda para o alto, sabe-se intuitivamente que a "chance" de cada uma das suas faces cair para cima é de 1 para 2, ou de 50%. Se rolarmos um dado, a "chance" de o número dois ficar virado para cima é de uma em seis ou 1/6. O termo "chance", entre aspas, refere-se ao que se denomina de *probabilidade*. Afirmar que a probabilidade de se obter "cara", no lançamento da moeda é de 50% significa dizer que se ela for lançada um grande número de vezes, a quantidade de vezes que se obterá a face com "cara" voltada para cima será aproximadamente a metade dos lançamentos. Da mesma forma, ao se jogar o dado, também um grande número de vezes, ter-se-á que o número dois ficará na face superior do dado cerca de 1/6 das vezes. À medida que o número de eventos realizados para um determinado fenômeno aumenta, aumenta também a proximidade entre os resultados obtidos estatística e experimentalmente. De acordo com Young (Young 1962):

> *"No problema do lançamento da moeda, deve-se ressaltar que a razão entre o número de vezes que se obtém cara e o número total de eventos se aproxima de 0,5 à medida que o número de eventos se torna muito grande. Isso não é a mesma coisa que dizer que o número de caras obtidas se aproxima do número de coroas. Por exemplo, em 100 lançamentos, um resultado razoável a ser obtido seria obter cara 52 duas vezes. Para 10.000 lançamentos um provável resultado seria obter 5020 caras. Nesse segundo caso, a razão se aproxima muito mais de 0,5 do que no primeiro caso, mas a diferença entre o número de caras e coroas obtidas é maior. Na realidade, pode-se motrar que a diferença entre o número de caras e o número de coroas obtidas tende a se tornar cada vez maior, independente do fato que a razão de cada uma delas em relação ao total de eventos se aproxima de 0,5. Portanto, se você está tirando cara e coroa com alguém e está perdendo, você não necessariamente recuperará suas perdas após um grande número de lançamentos. Existe 50% de chance de você perder cada vez mais e mais."* (p. 24)

distribuição de probabilidades

Ao jogar 10 moedas para o alto ao mesmo tempo, pode-se contar facilmente o número de caras e coroas obtidas. Mas se o que se deseja saber é a probabilidade de se obter "n" caras e "(10-n)" coroas nessa brincadeira, sendo "n" um número inteiro entre 0 e 10, a resposta corresponde a um conjunto de números para cada valor de "n". Estes

números podem ser pensados como uma função de "n", representada por $f(n)$ e denominada distribuição de probabilidade. No caso das 10 jogadas das moedas, se obtivemos 4 caras e 6 coroas, escrevemos:

f(cara) = 0,4
f(coroa) = 0,6

Em outras palavras, distribuição de probabilidade pode ser definida como uma representação do conjunto de probabilidades de todos os eventos associados a um espaço amostral.

Como a distribuição de probabilidade é definida para um intervalo determinado de valores de "n", a soma de todas as probabilidades de todos os valores deve ser igual a 1, ou seja

$$\sum_{n} f(n) = 1$$

Vamos considerar uma turma de estudantes (nomes fictícios), como exemplo de distribuição de probabilidades discreta. Na Tabela 2, apresentamos a lista dos alunos e suas respectivas notas numa prova cujo grau variava de 0 a 10, com intervalos de 0,5 pontos.

Tabela 2. Notas dos alunos (máximo possível 10 pontos)

Número	Nome do aluno	Grau obtido
1	Agnes Paula	3,0
2	Ana Maria	2,0
3	Ana Paula	3,0
4	Arthur Ananias	1,5
5	Débora Mortícia	8,0
6	Felipe Barbosa	8,0
7	Fernanda Lima	6,0
8	Glenda Rodrigues	5,5
9	Helena Laura	7,0
10	João Pedro	9,5
11	Laura Miller	2,0
12	Luiz Felipe	8,0
13	Luiza Helena	5,5
14	Marina Lima	9,5
15	Mauro Rolando	6,0
16	Nilo Egípcio	4,5
17	Roberta Carla	6,0
18	Tatiana Alegre	1,5
19	Vanessa Guimarães	7,5
20	Vitor Peçanha	4,5

Para determinar a distribuição de probabilidades dos dados apresentados na Tabela 2, colocamos as notas na primeira coluna e na segunda coluna o número total de alunos que tiraram aquela nota. Contamos da Tabela 2 quantos alunos tiraram cada nota para preencher a segunda coluna da Tabela 3.

histogramas

Tabela 3. Número de alunos por nota.

Nota	Número de alunos
1,5	2
2,0	2
3,0	2
4,5	2
5,5	2
6,0	3
7,0	1
7,5	1
8,0	3
9,5	2
Total	20

Na Figura 1, mostra-se de forma gráfica a informação contida na Tabela 3. No eixo horizontal, tem-se o valor da nota. Na vertical, o número de alunos que obtiveram aquela nota. Os dados são apresentados na forma de barras – e esse gráfico é chamado de histograma das notas desta turma. Observe que para o valor 8 o número é 3.

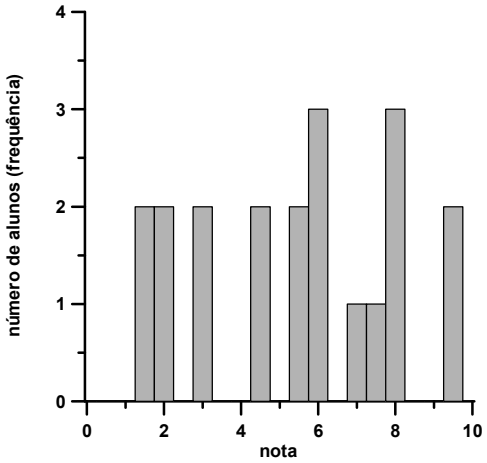


Figura 1. O histograma das notas dos alunos (da Tabela 3).

A distribuição de probabilidade **f(n)** é obtida tomando-se a razão entre o número de alunos com uma determinada nota e o número total de alunos, no caso 20. Então para a nota 1,5, temos que a distribuição de probabilidade será f(1,5) = 2/20 = 1/10, para a nota 6,0 a distribuição de probabilidade será f(6,0) = 3/20. Denominando-se, então, por **n** cada nota obtida no conjunto de provas considerado e por **f(n)** a distribuição de probabilidade para cada valor de **n**, obtém-se a Tabela 4 com a distribuição de probabilidades.

A soma de todos os valores de **f(n)** deve ser 1. Dos dados da Tabela 4:

$$\mathbf{f(0,5) + f(2,0) + f(3,0) + f(4,5) + f(5,5) + f(6,0) + f(7,0) + f(7,5) + f(8,0) + f(9,5) =}$$
$$\mathbf{= 0,1 + 0,1 + 0,1 + 0,1 + 0,1 + 0,15 + 0,05 + 0,05 + 0,15 + 0,10 = 1}$$

Tabela 4. Distribuição de probabilidades das notas.

n (Nota)	f(n)
1,5	0,10
2,0	0,10
3,0	0,10
4,5	0,10
5,5	0,10
6,0	0,15
7,0	0,05
7,5	0,05
8,0	0,15
9,5	0,10
Total	1,00

Pode-se representar a distribuição de probabilidades da Tabela 4 por um histograma de frequências relativas, como na Figura 2.

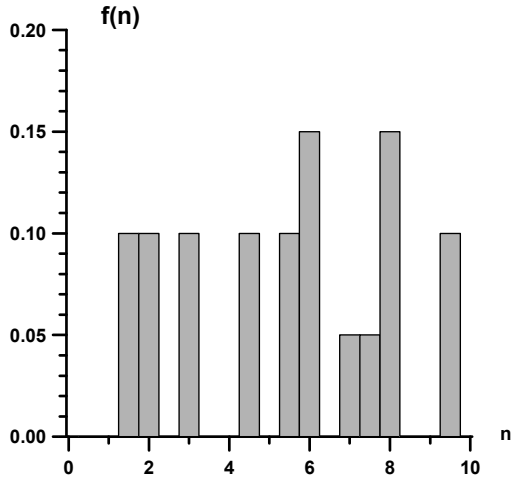


Figura 2. Distribuição de probabilidade de notas dos alunos

Vejamos outro exemplo. Um aluno, no laboratório, faz 10 medidas do período de oscilação de um pêndulo, e os valores por ele obtidos são anotados na Tabela 5.

Tabela 5. Medidas do período de oscilação de um pêndulo.

nº da medida	período (s)
1	3,19
2	3,31
3	3,26
4	3,33
5	3,32
6	3,28
7	3,41
8	3,29
9	3,34
10	3,42

Esses dados podem assumir qualquer valor dentro das limitações do cronômetro (de centésimo de segundo).

Como cada uma das dez medidas obtidas é anotada na tabela uma única vez, a probabilidade de obtenção de cada uma das medidas é igual e corresponde ao valor 1/10, como mostrado na Tabela 6.

Tabela 6. Distribuição de probabilidade dos períodos obtidos.

período (s)	probabilidade
3,19	1/10=0,1
3,26	1/10=0,1
3,28	1/10=0,1
3,29	1/10=0,1
3,31	2/10=0,2
3,32	1/10=0,1
3,33	1/10=0,1
3,34	1/10=0,1
3,41	1/10=0,1
3,42	1/10=0,1

Novamente, a soma de todos os valores dá 1.

Melhor valor: medidas de tendência central

Quando queremos estudar ou apresentar as características de uma medida cujo resultado varia (a nota do aluno na prova, o período do pêndulo em medidas sucessivas), é interessante conseguir apresentar a informação com um pouco mais de simplicidade do que com uma tabela ou um gráfico. Por exemplo, podemos apresentar a idade de um grupo de alunos informando a média dessas idades, como se fosse uma idade representativa do grupo.

Na medida do período do pêndulo, um bom valor para responder à pergunta "qual é o valor do período deste pêndulo?" seria fornecer o valor médio das medidas realizadas. Em outras palavras, resumir um conjunto de dados a um valor representativo, que representa a "tendência central" do grupo de medidas.

medidas de tendência central

Algumas das medidas de tendência central são a *média aritmética* ou abreviadamente *média* (a mais conhecida), a *mediana*, a *moda* e a *média harmônica* ou média ponderada. Cada uma dessas medidas adequa-se a uma situação, com vantagens e desvantagens.

Imaginemos um conjunto de N dados

$$\{x_1, x_2, x_3, x_4, \cdots, x_N\}$$

A média aritmética é obtida somando-se todos os valores e dividindo pelo total de valores; ou seja,

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

média aritmética

Por exemplo, suponha que numa sala de aula de jovens e adultos exista um grupo de 10 alunos com 24, 27, 31, 32, 33, 37, 39, 40, 45 e 47 anos. A média aritmética das idades desse grupo pode ser calculada por:

$$\overline{I} = \frac{24+27+31+32+33+37+39+40+45+47}{10} \Rightarrow \overline{I} = \frac{355}{10} = 35,5$$

Se os valores X_1, X_2, X_3, ..., X_k ocorrem com as frequências f_1, f_2, f_3, ..., f_k, a média aritmética, neste caso, pode ser obtida substituindo os f_1 valores X_1 por $f_1.X_1$ e assim sucessivamente:

$$\overline{X} = \frac{f_1 \cdot X_1 + f_2 \cdot X_2 + f_3 \cdot X_3 + ... + f_k \cdot X_k}{f_1 + f_2 + f_3 + ... + f_k} = \frac{\sum_{i=1}^{k} f_i \cdot X_i}{\sum_{i=1}^{k} f_i}$$

onde n é a soma dos f_i, a freqüência total ou o total de número de eventos.

Como exemplo, tomando-se as notas obtidas pelos alunos e suas respectivas freqüências na Tabela 3, pode-se calcular a média aritmética como se segue:

$$\overline{N} = \frac{2 \cdot 1,5 + 2 \cdot 2,0 + 2 \cdot 3,0 + 2 \cdot 4,5 + 2 \cdot 5,5 + 3 \cdot 6,0 + 1 \cdot 7,0 + 1 \cdot 7,5 + 3 \cdot 8,0 + 2 \cdot 9,5}{2+2+2+2+2+3+1+1+3+2} =$$

$$\overline{N} = \frac{3,0 + 4,0 + 6,0 + 9,0 + 11 + 18 + 7,0 + 7,5 + 24 + 19}{20}$$

$$\overline{N} = \frac{108,5}{20} = 5,3$$

Sendo uma medida de tendência central, o cálculo da média aritmética busca através de um único número apresentar as características de um determinado grupo de números. Porém, em algumas situações, a presença de um valor muito maior ou muito menor que o restante do grupo pode fazer com que ela não consiga traçar o perfil correto desse grupo.

Por exemplo, se num grupo de seis pessoas que fizeram uma prova valendo 100 pontos, as notas obtidas foram respectivamente 2, 3, 1, 1, 2 e 50, a média aritmética desses valores será 9,8:

$$\overline{N} = \frac{2+3+1+1+2+50}{6} = \frac{59}{6} = 9,8$$

Observe que o valor obtido não representa bem as características desse grupo em termos de nota. Então, neste caso, a média não é uma medida de tendência central apropriada. Provavelmente a forma de comunicar o resultado seria a média eliminando o valor espúrio (50) e o valor separadamente: o grupo tem duas partes, um aluno tirou 50 em 100 e o resto teve média 1,8 em 100.

Em alguns casos deseja-se calcular a média de um conjunto de valores em alguns desses valores são mais importantes que outros, e essa importância é representada por um "peso". Para se realizar, então, essa operação levando-se em conta essas "diferentes importâncias" define-se a <u>média aritmética ponderada</u>.

Consideremos os números X_1, X_2, X_3,..., X_N associados respectivamente aos pesos P_1, P_2, P_3,..., P_N. A média aritmética ponderada desses valores é dada por:

$$\overline{X} = \frac{P_1 \cdot X_1 + P_2 \cdot X_2 + P_3 \cdot X_3 + \ldots + P_n \cdot X_n}{P_1 + P_2 + P_3 + \ldots + P_n} = \frac{\sum_{i=1}^{n} P_i \cdot X_i}{\sum_{i=1}^{n} P_i}$$

média ponderada

Como exemplo, pense em um professor que tenha aplicado duas provas e uma atividade de laboratório durante um bimestre. À primeira prova, objetiva, ele atribuiu peso 1. À segunda, peso 2, por ser discursiva; e atribuiu peso 3 ao desempenho na atividade de laboratório que, além de consumir maior tempo, necessitava um maior grau de atenção e criatividade. Gabriela, uma de suas alunas, obteve para essas avaliações, respectivamente, as notas 6,0, 4,5 e 7,0. Calculando-se a média aritmética ponderada das notas desta aluna obtém-se:

$$\overline{N} = \frac{6,0 \cdot 1 + 4,5 \cdot 2 + 7,0 \cdot 3}{1 + 2 + 3} = \frac{6,0 + 9,0 + 21}{6} = 6,0$$

Antes de prosseguir, é útil observar-se algumas das propriedades das médias aritméticas.

a) A soma algébrica dos desvios de um conjunto de números Xi, em relação à média aritmética, é zero:

$$\sum_{i=1}^{n} (X_i - \overline{X}) = 0$$

b) A soma dos quadrados dos desvios de um conjunto de Xi, em relação a um número qualquer "a", é um mínimo se e somente se $a = \overline{X}$:

$\sum_{i=1}^{n} (X_i - a)^2 \geq 0$ é mínimo se e somente se $a = \overline{X}$.

c) Se f_1 números têm média m_1, f_2 números têm média m_2,..., f_k números têm média m_k, a média de todos os números é dada por

$$\overline{X} = \frac{f_1 \cdot m_1 + f_2 \cdot m_2 + f_3 \cdot m_3 + \ldots + f_k \cdot m_k}{f_1 + f_2 + f_3 + \ldots + f_k} = \frac{\sum_{i=1}^{k} f_i \cdot m_i}{\sum_{i=1}^{k} f_i}$$

Outro valor de tendência central é a <u>mediana</u> de um conjunto de números: se esses números são organizados em ordem de grandeza, é o valor que ocupa a posição central dos números. No caso desse conjunto de números possuir um número de elementos ímpar, a mediana ocupa o valor central dessa distribuição ordenada. No entanto, se o número de elementos for par, a mediana é obtida através da média aritmética dos dois valores centrais.

mediana

Por exemplo, um aluno durante um bimestre obteve as seguintes notas nas atividades avaliativas de uma determinada disciplina: 10, 9, 8, 7 e 10. Para obter-se a mediana desse conjunto de números necessita-se, inicialmente, ordenar esses valores, ou seja: 7, 8, 9, 10, 10.

Com a ordenação feita, a mediana será o valor que ocupa a posição central dos números, ou seja, o número 9.

No entanto, ao se procurar a mediana das notas obtidas pelos alunos na Tabela 3.1, não teremos um valor em posição central. Logo, para determinação da mediana deve-se fazer a média aritmética dos dois valores centrais, ou seja:

1,5 1,5 2,0 2,0 3,0 3,0 4,5 4,5 5,5 5,5 6,0 6,0 6,0 7,0 7,5 8,0 8,0 8,0 9,5 9,5

$$\tilde{X} = \frac{5,5+6,0}{2} = 5,8$$

Geometricamente, a mediana é o valor de "X" (abscissa) correspondente à vertical que divide o histograma (dos dados X_i) em duas partes de áreas iguais.

Chama-se moda de um conjunto de dados o valor que ocorre com maior frequência. A moda pode não existir e, mesmo que exista, pode não ser única. E esse é a última medida de tendência central usual.

moda

Quando uma distribuição de probabilidade possui somente uma moda ela é denominada de unimodal.

Medidas de dispersão

O conhecimento do valor médio da grandeza medida não é suficiente para descrever toda a informação obtida com a medida. Com apenas esta informação apenas, não se é capaz de discernir que tipo de medida foi feita, o que acontecerá se ela for repetida, não será possível dizer se a distribuição de probabilidades é simétrica, se é muito "larga" (isto é, se a faixa de valores medidos é grande), etc.

Quando se possui os valores das medidas Xi e sua respectiva média $\overline{X}$, pode-se saber o quão distante cada uma das medidas encontra-se da média, isto é, o quanto os dados numéricos tendem a se dispersar em torno do valor médio. A descrição quantitativa desse distanciamento, denominado de dispersão, fornece uma idéia da precisão dessas medidas.

Por exemplo, suponha que um professor é contratado para uma substituição num curso livre em que as aulas devem conter atividades recreativas. Como ele não conhece a escola, informam-lhe que a turma é constituída de um grupo de 6 pessoas cuja média de idade é de 20 anos. A informação da média de idades, no entanto, não é suficiente, pois se pode ter grupos com essa média de idade com características completamente diferentes:

Grupo 1: 25 anos; 25 anos; 25 anos; 25 anos; 25 anos; 25 anos.

média aritmética das idades: $\overline{I}_1 = \frac{25+25+25+25+25+25}{6} = \frac{150}{6} = 25\ anos$

Grupo 2: 28 anos; 25 anos; 27 anos; 25 anos; 24 anos; 21 anos.

média aritmética das idades: $\overline{I}_2 = \frac{28+25+27+25+24+21}{6} = \frac{150}{6} = 25\ anos$

Grupo 3: 10 anos; 64 anos; 51 anos; 14 anos; 8 anos; 1 ano.

média aritmética das idades: $\overline{I}_3 = \frac{10+64+51+14+8+1}{6} = \frac{150}{6} = 25\ anos$

Pode-se observar que a medida de tendência central (média aritmética) utilizada no exemplo não é suficiente para caracterizar bem os grupos, em particular o terceiro. Neste caso faz-se necessária a utilização de medidas que expressem o grau de dispersão de um conjunto de dados. Dessas medidas de dispersão, as mais utilizadas são a variância e o desvio padrão.

Não adianta pensar em usar a soma dos desvios, pois essa soma dá zero sempre (veja o comentário ao final da discussão sobre médias, e faça o cálculo nos exemplos acima!)

A idéia fundamental da variância $\left(\sigma^2\right)$ é se tomar os desvios dos valores Xi em relação à média aritmética desses valores. A soma dos desvios simples, $d_i = X_i - \overline{X}$, é nula; poder-se-ia pensar em somar os valores absolutos desses desvios, ou então o quadrado desses desvios. Por motivos de facilidade de cálculo, considera-se a variância como sendo a média dos quadrados dos desvios, ou seja:

variância

$$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$$

Então, calculando-se a variância para os grupos 1, 2 e 3 do exemplo anterior, tem-se:

Grupo 1: 25 anos; 25 anos; 25 anos; 25 anos; 25 anos; 25 anos.

$$\overline{I}_1 = 25\ anos$$

$$\sigma^2 = \frac{(25-25)^2 + (25-25)^2 + (25-25)^2 + (25-25)^2 + (25-25)^2 + (25-25)^2}{6} = 0$$

Observe que pelo fato de todos os valores serem iguais, a variância é 0. Este resultado permite afirmar que não há dispersão, que o grupo é completamente homogêneo no que se refere à idade.

Grupo 2: 28 anos; 25 anos; 27 anos; 25 anos; 24 anos; 21 anos.

$$\overline{I}_2 = 25\ anos$$

$$\sigma^2 = \frac{(25-28)^2 + (25-25)^2 + (25-27)^2 + (25-25)^2 + (25-24)^2 + (25-21)^2}{6}$$

$$\sigma^2 = \frac{(-3)^2 + (0)^2 + (-2)^2 + (0)^2 + (1)^2 + (4)^2}{6}$$

$$\sigma^2 = \frac{9+0+4+0+1+16}{6} = \frac{30}{6} = 5$$

Grupo 3: 10 anos; 64 anos; 51 anos; 14 anos; 8 anos; 1 ano.

$$\overline{I}_2 = 25\ anos$$

$$\sigma^2 = \frac{(25-10)^2 + (25-64)^2 + (25-51)^2 + (25-14)^2 + (25-8)^2 + (25-1)^2}{6}$$

$$\sigma^2 = \frac{(15)^2 + (-39)^2 + (-26)^2 + (11)^2 + (17)^2 + (24)^2}{6}$$

$$\sigma^2 = \frac{225+1521+676+121+289+576}{6} = \frac{3408}{6} = 568$$

Com os valores obtidos da variância para cada grupo é possível diferenciar a dispersão de cada um deles. Observa-se que o grupo 1 não possui dispersão, enquanto o grupo 3 apresenta uma dispersão muito maior que o 2. Costuma-se dizer que o grupo 2 é mais homogêneo que o grupo 3 ou que o 3 é mais heterogêneo que o 2.

No entanto, pelo fato da variância ser calculada com os desvios $d_i = X_i - \overline{X}$ elevados ao quadrado, não é possível expressar a variância na mesma unidade dos valores da variável. Então, para que isso seja possível, define-se a medida de dispersão chamada de desvio padrão.

desvio padrão

O desvio padrão é a raiz quadrada da variância. A sua utilização facilita a interpretação dos dados por se expresso na mesma unidade dos valores observados.

Matematicamente tem-se:

$$\sigma = \sqrt{\sigma^2}$$

Então, para os grupos 1, 2 e 3, tem-se:

Grupo 1: $\sigma_1 = \sqrt{0} = 0$ ano

Grupo 2: $\sigma_2 = \sqrt{5} = 2,2$ anos

Grupo 3: $\sigma_1 = \sqrt{568} = 23$ anos

Em relação a essas medidas de dispersão, observa-se que:
a) A variância e o desvio padrão são sempre positivos.
b) Quando todos os valores da variável são iguais, o desvio padrão é 0.
c) Quanto mais tende a zero o desvio padrão, mais homogênea é a distribuição de valores da variável.
d) O desvio padrão é expresso na mesma unidade da variável.

Função de distribuição, ou densidade de probabilidade

Um conceito importante na apresentação de dados e resultados de medidas é a forma como são apresentados os resultados. Já vimos que é muito mais ilustrativo apresentar um histograma das notas dos alunos de uma turma do que apresentar a tabela dessas notas. Uma extensão da ideia de apresentação de resultados no formato de histogramas é a apresentação de "funções de distribuição", ou densidade de probabilidade, dos valores ou dados.

Imaginemos que estamos tratando de todas as turmas de um grande colégio, com 1000 alunos. As médias finais desses alunos podem ser apresentadas em intervalos discretos, de 5 em 5 pontos (0; 5; 10; 15; ...), ou de forma quase contínua, de 0,1 em 0,1 pontos (0,1; 0,2; 0,3; ...; 99,9; 100) entre os valores de 0 a 100. Ao fazermos o histograma neste segundo caso, é conveniente agrupar os valores em intervalos maiores; nesse caso, há uma acumulação de dados em cada uma das "barras" que compõem o histograma. Por exemplo, com notas entre 50,0 e 50,4 temos as notas da Tabela 7.

Tabela 7. Médias finais de uma faixa numa escola com 1000 alunos

Nota	Número de alunos
⋮ (de 0 a 49,9)	(253)
50,0	10
50,1	15
50,2	8
50,3	7
50,4	20
⋮ (até 100)	(687)

Na "barra"do histograma que contém os dados entre 50,0 e 50,4 devem ser representados 60 alunos (60=10+15+8+7+20). Em outras palavras, o gráfico deve ser lido

multiplicando-se o tamanho do intervalo de agrupamento com o valor da abscissa (para a interpretação do valor total de alunos).

Distribuição de Gauss ou distribuição normal

Os histogramas de resultados de dados experimentais que são quase contínuos, como o caso das médias da grande escola ou o caso de fazermos 400 medidas de um período de um pêndulo com um cronômetro digital, transformam-se em diagramas de barras que em sua extremidade superior assumem o aspecto de uma curva contínua. Essa curva recebe o nome de distribuição, representando uma distribuição de probabilidades de obter o resultado na faixa considerada. Algumas formas que esta curva assume se repetem muito; em particular, em quase todos os casos que em os erros são instrumentais (aleatórios ou ao acaso), esta curva pode ser descrita por uma função particular, a função gaussiana. À distribuição cuja curva é uma função gaussiana é denominada distribuição de Gauss, ou normal.

A importância desse tipo de distribuição deve-se

- ao fato dela descrever a distribuição dos erros aleatórios em muitos tipos de medidas em situações físicas, biológicas e sociais;

- à possibilidade de se demonstrar que, mesmo que os erros individuais não sigam esta distribuição, as médias desses grupos de erros serão distribuídas de uma forma que se aproximem da distribuição de Gauss quando se trata de grupos com um grande número de elementos;

- ao fato de ser fundamental para a inferência estatística.

Numa descrição histórica sobre a construção da curva normal, um professor conhecido que trabalha em avaliações educacionais, o prof. Luiz Pasquali, descreve:

> "A lei dos grandes números de Bernoulli diz o seguinte: numa situação de eventos casualóides, onde as alternativas são independentes, obter coroa em lances de uma moeda de cara e coroa, tem a probabilidade matemática exata de 50% (porque são somente dois eventos possíveis: cara ou coroa), mas na prática esta probabilidade de 50% é apenas aproximada. E essa aproximação é tanto mais exata quanto maior forem as tentativas que você fizer de lançar a moeda, chegando a quase atingir os exatos 50% se você lançar a moeda infinitas vezes. Isto é, quanto mais lances você fizer, menor será o desvio em relação à média de 50% que o resultado irá produzir. Isso quer dizer que os erros (desvios) serão menores e menores na medida em que sobe o número de lances. Desvios grandes são raros e desvios pequenos frequentes; quanto menores os desvios, mais frequentes eles serão, de sorte que, aumentando as tentativas (os lances), aumenta o número de desvios pequenos, sobrepujando cada vez mais os desvios grandes, de tal sorte que, no limite, haverá quase somente desvios pequenos, sendo o desvio 0 o menor deles e, por consequência, o mais frequente.
>
> Moivre assumiu essa idéia de Bernoulli e disse: erros grandes são mais raros que erros pequenos. Assim, quanto menores os erros, mais freqüentes eles serão e quanto maiores, menos freqüentes. Dessa forma, os erros se distribuem eqüitativamente em torno de um ponto modal, a média, formando uma curva simétrica com pico na média e caindo rapidamente para as caudas à esquerda (erros que subestimam a média) e à direita (erros que superestimam a média)." (Luiz Pasquali, Psicometria, 1996, p. 71 e 72).

A distribuição a que se refere a citação é justamente a distribuição gaussiana ou normal. Ela é uma curva simétrica e contínua, na forma de um "sino", que possui a seguinte expressão matemática:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$$

Esta função possui dois parâmetros: $\bar{x}$, que é o valor médio e σ^2 , que é sua variância.

Plotando-se a curva da função P(x), tem-se:

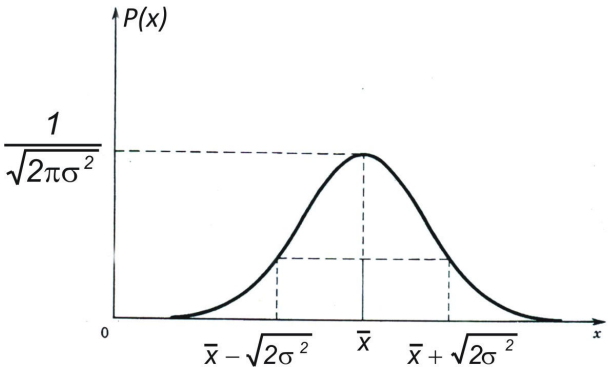


Figura 3. Curva de distribuição de Gauss ou curva normal.

Observa-se que $\frac{1}{\sqrt{2\pi\sigma^2}}$ é o máximo valor da altura (valor mais provável) da função P(x), e que o valor médio $\bar{x}$ representa o valor de "x" para o qual a função P(x) atinge esse valor. Além disso, percebe-se que o termo $\sqrt{2\sigma^2}$ tem a ver com a largura ou estreiteza da curva em forma de boca de sino. Pode-se notar, também, que quanto maior for o expoente da função P(x), mais rapidamente a curva vai caindo para a abscissa, porém, nunca chegando a zero.

Abaixo estão representadas curvas normais com médias M diferentes (a), desvios padrões DP diferentes, porém a mesma média M (b) e com as variações nos dois parâmetros (c).

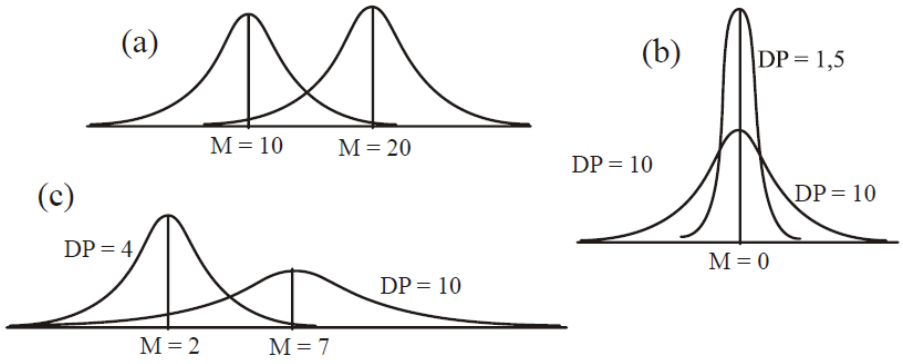


Figura 4. Distribuições normais com diferentes médias e desvios-padrão.

A distribuição de Gauss pode ser usada para se obter a probabilidade de uma medida estar entre quaisquer limites especificados; para isso, basta calcular a área da curva compreendida entre estes limites. Na Figura 5, apresenta-se a representação gráfica desta ideia: o valor da probabilidade de obter para a variável x um valor entre a e b é dada pela área da figura hachurada. Matematicamente, esta área é calculada por meio da integral $P(a,b) = \int_a^b P(x)dx$.

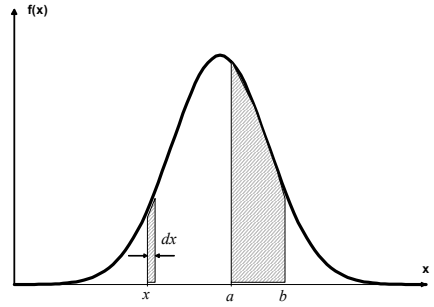


Figura 5. A probabilidade de encontrar o valor da variável entre a e b é dada pela área hachurada na figura do gráfico da distribuição de probabilidade.

Para o caso da distribuição gaussiana, é importante saber o valor da probabilidade de encontrar o resultado entre a média e um ou dois desvio padrão a partir dessa média.

Constata-se, a partir dos cálculos destas áreas, como mostrado na Figura 6, que

- a probabilidade de que a variável assuma um valor entre a média menos um desvio padrão $\left(\overline{X} - \sigma\right)$ e a média mais um desvio padrão $\left(\overline{X} + \sigma\right)$ é de 68 %;

- para região entre $\left(\overline{X} - 2\sigma\right)$ e $\left(\overline{X} + 2\sigma\right)$ é de 95,5%,

- e para a região entre $\left(\overline{X} - 3\sigma\right)$ e $\left(\overline{X} + 3\sigma\right)$ é de 99,7%.

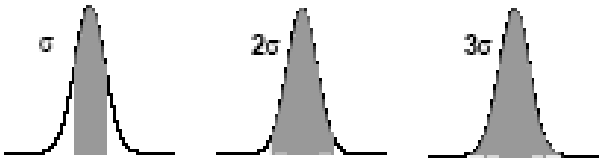


Figura 6. As áreas que representam a probabilidade de encontrar os valores a uma distância de um, dois e três desvios padrão da média.

Em outras palavras, *se uma medida tem como resultado uma distribuição no formato da gaussiana, podemos garantir que 99,7% dos resultados obtidos na faixa que dista 3 desvios padrão da média*.

Ajuste de dados – Regressão Linear

Quando estamos fazendo medidas em um laboratório, muitas vezes queremos descrever os resultados obtidos em termos de modelos e teorias conhecidas. Às vezes, queremos encontrar uma relação entre as variáveis. Uma das formas de fazer isso envolve o que chamos de ajuste de dados por uma curva, e a curva mais fácil de se ajustar é uma reta (ou função afim).

Por exemplo, num laboratório fizemos medidas do movimento de um objeto que achamos estar se movendo em movimento uniforme. Obtivemos medidas de posição como função do tempo que estão representadas na Tabela 8.

Tabela 8. Medidas de posição como função do tempo de um objeto em movimento horizontal sem atrito.

t (s)	x (cm)
0,1	1,0
0,2	1,3
0,3	1,5
0,4	1,9
0,5	2,2
0,6	2,4
0,7	2,8
0,8	3,1
0,9	3,3
1,0	3,5
1,1	3,9

Se colocamos esses pontos sobre um gráfico, obtemos o que está apresentado na Figura 7.

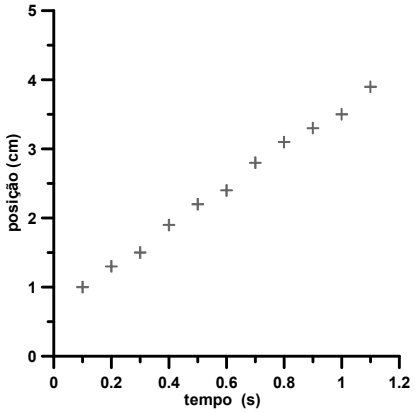


Figura 7. O gráfico da posição como função do tempo para os dados da Tabela 8.

Ao olharmos para este gráfico, observamos nitidamente que os pontos estão alinhados como sobre uma linha reta. Então, é razoavel supor que esses dados podem ser bem representados por uma função linear do tipo f(x)=at+b. O trabalho, agora, é descobrir qual é a melhor reta que descreve esses dados.

O procedimento para determinação dessa melhor reta é denominado regressão linear. De fato, o que falta é definir um critério para utilizar-se a palavra melhor. Costuma-se usar como critério calcular a diferença quadrática entre o ponto experimental (t_i, x_i) e o valor esperado (t_i, at_i+b), somar as diferenças para todos os pontos e calcular o que chamamos de chi quadrado (χ^2). A melhor reta é aquela que têm os valores de a e b tais que o valor de χ^2 é o mínimo possível. É claro que esse procedimento é matematicamente trabalhoso, mas a maior parte dos softwares gráficos já o faz automaticamente.

Testes: a Teoria Clássica de Testes

Nos processos de avaliação de aprendizagem, precisamos atribuir um escore (uma nota), em geral comparativa, aos estudantes. O que usualmente é feito é atribuir uma pontuação a uma série de questões ou problemas, e somar a pontuação obtida pelo aluno para obter uma nota final. Este mecanismo de obtenção do escore a ser atribuído ao aluno é característica da denominada Teoria Clássica dos Testes (TCT).

Alguns dos problemas enfrentados na Teoria Clássica dos Testes (TCT), quando se pensa em avaliações de larga escala, é a viabilidade de comparação entre os grupos respondentes e a dependência existente entre as características do teste e do examinando. É conhecido de todos os professores que existem provas ou testes "fáceis" ou "difíceis", turmas "fracas" e "fortes". Para construir uma forma de elaboração de testes que permita comparações com significados entre grupos e épocas diferentes, utiliza-se a denominada Teoria da Resposta ao Item, que transpõe algumas das limitações observadas na Teoria Clássica de Testes.

Na Teoria Clássica de Testes o foco está na produção de testes de qualidade que no final resultaram em testes válidos. Isso se deve ao fato de que ela se ocupa da explicação do resultado final total do mesmo, ou seja, o escore total que é dado pela soma das respostas corretas de um dado conjunto de itens.

Por exemplo, num questionário de física térmica, tem-se 26 questões a serem respondidas pelos alunos. Um determinado aluno (por exemplo o de número 46), ao respondê-lo, acertou 14 itens. Como foi atribuído grau 0 para cada erro e grau 1 para cada um dos acertos, pode-se afirmar que este aluno, neste teste, apresentou um escore total igual a 14. A TCT, então, se ocupa em dar significado ao que representa esse escore para o referido aluno.

Esta técnica porém apresenta dificuldades bem conhecidas por todos os professores. Para comparar grupos diferentes em momentos diferentes utilizando testes diferentes, a TCT apresenta limitações, das quais algumas são:

1) Os parâmetros clássicos utilizados para analisar um item (uma questão) de um teste (uma prova), a dificuldade do item e a capacidade de discriminação do item, dependem diretamente da amostra de sujeitos utilizada para estabelecê-los. Isto é, é difícil preparar testes ou provas equivalentes para aplicação até mesmo dentro uma escola.

Por exemplo, quando um professor deseja determinar o nível de dificuldade de um teste feito por ele, normalmente ele pede a um colega de sua área para avaliar este teste ou então verifica o resultado do mesmo numa turma que já o tenha feito. No entanto, ao aplicá-lo em outros grupos de alunos, ele percebe que o teste apresenta níveis de dificuldade diferentes, de acordo com as características de cada grupo, ou seja, o resultado da amostra (dos alunos que fizeram o testes) não é válido para análise

das outras turmas que não fizeram o teste. Esta característica da TCT extrapola os muros das instituições de ensino, ocorrendo, por exemplo, nas pesquisas de opinião.

2) O teste que é aplicado interfere diretamente na avaliação das aptidões dos estudantes que realizam este teste. Ao se aplicar testes diferentes para medir a mesma aptidão, obtém-ser escores diferentes da mesma aptidão para sujeitos idênticos. Os escores também serão diferentes quando se aplica testes de dificuldades diferentes. No caso das formas paralelas de testes, é preciso observar que, em primeiro lugar, conseguir formas estritamente paralelas é uma tarefa quase impossível e, em segundo lugar, mesmo conseguindo formas paralelas, é difícil pressupor que elas produzam o mesmo montante de erro, o que vem afetar a estimação do escore verdadeiro dos sujeitos.

Observe, por exemplo, os dois itens apresentados na Figura 8, constantes do vestibular da UFRJ, respectivamente nos anos de 1999 e 2004.

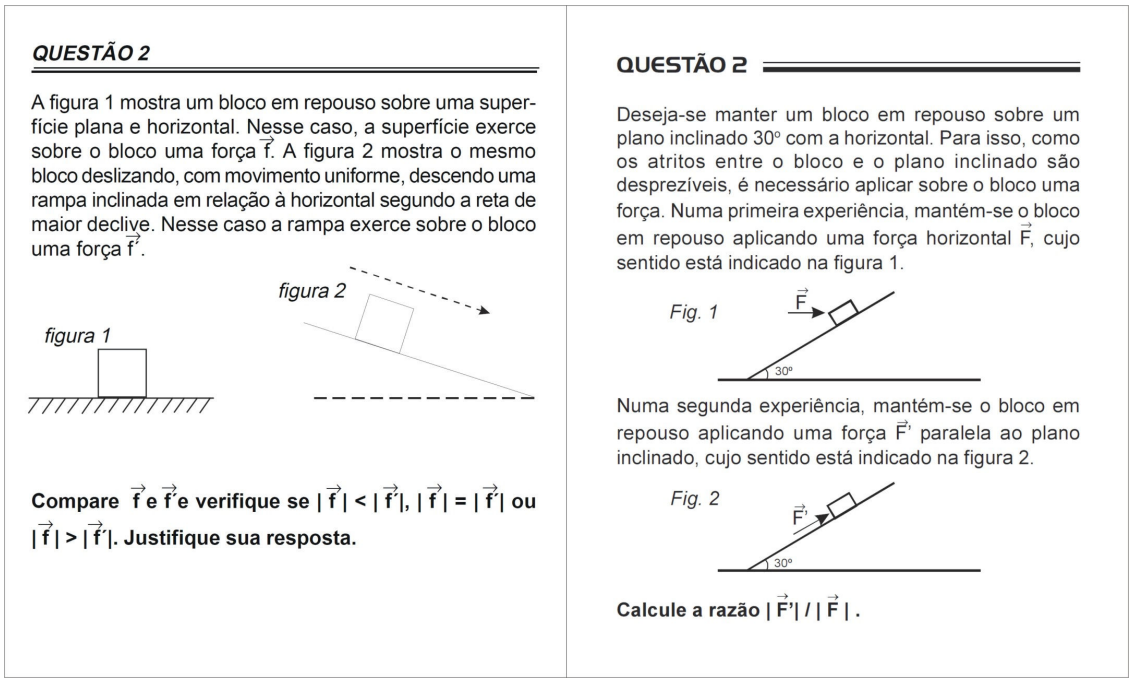


Figura 8. Itens do vestibular da UFRJ 1999 (à esquerda) e 2004 (à direita).

Os dois itens compreendem a aprendizagem de conceitos relacionados à decomposição de grandezas vetoriais e o equilíbrio da partícula. No entanto, no item de 2004, a colocação da força $\vec{F}$ numa direção não comumente utilizada nas resoluções de planos inclinados torna o item mais difícil que o de 1999. Supondo, então, que essas duas questões fossem aplicadas para determinar a aptidão dos alunos em relação a estes conceitos, o fato de a segunda questão ser mais difícil faria com que um mesmo aluno tivesse valores diferentes para sua aptidão dependendo de qual dos dois itens lhe fosse proposto.

Em outras palavras, construir dois testes que apresentem exatamente as mesmas características, avaliem as mesmas aptidões e ainda mais, apresentem a mesma dificuldade para todos os estudantes, afim de que sejam comparáveis, não é uma tarefa possível dentro desta teoria.

Dentro dessas limitações, ao se aplicar um teste, a característica do examinando na qual se está interessado é a "aptidão" medida pelo instrumento de avaliação. A TCT expressa essa "aptidão" pelo escore verdadeiro que é definido como o valor esperado de desempenho observado no teste de interesse. Porém, na determinação desta aptidão na TCT enfrenta-se um grande problema que é o fato de que as características do examinando e do teste não podem ser separadas, ou seja, a aptidão do examinando é definida em termos de um teste particular. Isto é evidenciado quando se aplica um teste considerado difícil ao examinando, que neste caso pode obter um resultado que indique que o ele apresenta uma aptidão baixa. No entanto, se for aplicado outro teste que exija a mesma aptidão, mas este teste possuir um nível de dificuldade menor, pode-se obter, para o mesmo estudante, uma aptidão alta.

Tem-se, então, que na TCT as características do teste e do item mudam quando o contexto do examinando muda, e as características do examinando mudam quando o contexto do teste muda. Portanto, é muito difícil comparar examinandos que fazem testes diferentes, é muito difícil comparar itens cujas características são obtidas usando grupos diferentes de examinandos. Esse problema é constante na atuação dos professores que precisam fazer avaliações diferentes para turmas que estão cursando a mesma série, avaliações substitutas ou de segunda chamada. Por exemplo, um aluno que faz uma prova na data prevista e outro que por algum motivo necessita fazer a segunda chamada da mesma, mesmo tirando a mesma nota, não terão a mesma aptidão.

Testes: a Teoria da Resposta ao Item

Buscando-se superar as limitações existentes na TCT, outras metodologias de avaliação foram propostas e, dentre elas, a Teoria de Resposta ao Item (TRI). Como características inerentes a esta teoria tem-se que:

1) os itens (questões) não são dependentes do grupo que faz o teste.

2) os escores descrevem a proficiência que independe do teste aplicado.

3) esse modelo de avaliação é expresso a partir do item e não do teste como um todo, permitindo que se analise cada item individualmente, independentemente dos demais itens do teste.

4) não há exigência de testes estritamente paralelos para a verificação de sua confiabilidade.

5) esse modelo fornece uma medida de precisão para cada escore de habilidade.

Nesse modelo de análise, busca-se medir o que se denomina por traço-latente ou variável não-observável. Para uma melhor compreensão do que isso significa, suponha que um professor de física, do ensino médio ou superior, ao avaliar o quanto seus alunos aprenderam (e isso é um conceito abstrato), utiliza como instrumento para obter essa informação uma prova que contém alguns problemas; neste caso, a variável observada por ele é a capacidade de cada estudante resolver os problemas apresentados ou determinado tipo de problema e não há dúvidas que aprender algum tópico de física não é equivalente a saber resolver problemas daquele tópico.

Outro exemplo de medida de traço latente ou variável não observável pode ser encontrado na Matriz de Referência do ENEM[1]. Suponha que, ao se construir a prova de Ciências da Natureza, deseja-se medir a competência de se apropriar de conhecimentos da física para, em situações problema, interpretar, avaliar ou planejar intervenções científico-tecnológicas (competência 6), que nesse documento representa uma variável não-observável ou traço latente. Para isso, constróem-se itens que contemplem as habilidades (variável observável) listadas abaixo que, em conjunto, possibilitam determinar essa competência (traço latente).

Tabela 9. Habilidades (variáveis observáveis) pertencentes à competência 6.

Variáveis observáveis ou habilidades referentes à competência 6
H20 – Caracterizar causas ou efeitos dos movimentos de partículas, substâncias, objetos ou corpos celestes.
H21 – Utilizar leis físicas e (ou) químicas para interpretar processos naturais ou tecnológicos inseridos no contexto da termodinâmica e (ou) do eletromagnetismo.
H22 – Compreender fenômenos decorrentes da interação entre a radiação e a matéria em suas manifestações em processos naturais ou tecnológicos, ou em suas implicações biológicas, sociais, econômicas ou ambientais.
H23 – Avaliar possibilidades de geração, uso ou transformação de energia em ambientes específicos, considerando implicações éticas, ambientais, sociais e/ou econômicas.

A Teoria da Resposta ao Item corresponde então a uma proposta psicométrica para determinação de aptidões (traços latentes) que possui dois postulados básicos:

Primeiro Postulado

O desempenho de um examinando num item do teste pode ser previsto (ou explicado) por um conjunto de fatores denominados de traços, traços latentes, ou aptidão.

Segundo Postulado

A relação entre o desempenho em um item e o conjunto de traços que definem este desempenho no item pode ser descrita por uma função monótona-mente crescente da aptidão chamada função característica do item, ou curva característica do item (ICC=item characteristic curve).

Observe que o primeiro postulado traz consigo a idéia básica de que, através da realização de testes, é possível se medir a aptidão do examinando.

O segundo postulado, por sua vez, apresenta uma idéia bastante razoável que consiste no fato de que ao se construir uma curva da probabilidade de acerto de um item para um aluno com determinada aptidão, essa curva deve assumir a forma de um S deitado, isto é, se este aluno possuir uma nota alta, a probabilidade de ele acertar qualquer item é também alta, se este aluno possuir nota média, a probabilidade de acertar qualquer item é média e finalmente, se o aluno tem nota baixa, a probabilidade também será baixa.

[1] Neste caso se utiliza o termo competência como sinônimo de traço latente (aptidão) e o termo habilidade como sinônimo de comportamento ou variável observável.

Considere, por exemplo, que um mesmo teste seja aplicado a duas turmas distintas que possuam níveis de aptidão diferentes, isto é, uma turma com menor nível de aptidão (turma 1) e a outra com maior nível de aptidão (turma2). Podemos construir uma curva que represente a distribuição dos escores do teste para as duas turmas. Certamente a turma 2 apresentará um valor médio mais alto do que a turma 1. Na Figura 8, apresenta-se um gráfico onde a abscissa corresponde ao escore do aluno, ou melhor, à habilidade ou aptidão do aluno. Na parte inferior do gráfico, estão representadas (de forma invertida) as duas distribuições de notas para cada uma das duas turmas. Na parte superior da curva, representa-se a ICC dos dois grupos, isto é, a curva que representa, para um item do teste, a probabilidade de acerto no item para o candidato cujo escore é dado.

Observe que a ICC para as duas turmas é a mesma, sendo que se pode observar que os respondentes com maior aptidão (localizados mais à esquerda no eixo das abscissas) possuem também maior probabilidade de responder o item corretamente.

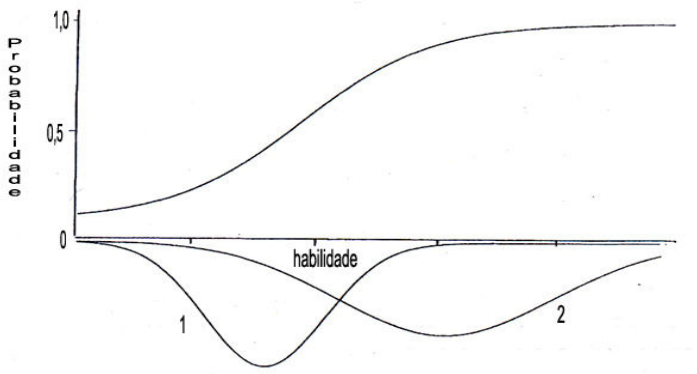


Figura 9. ICC e curvas de distribuição de aptidão para as turmas 1 e 2 de examinandos.

Como na utilização da metodologia da TRI os itens não são dependentes do grupo que é submetido ao teste, os escores que descrevem a aptidão de cada respondente independem do teste aplicado, e os parâmetros estimados em diferentes grupos também devem ser os mesmos, tanto o item quanto os parâmetros de aptidão são ditos invariantes. Isto pode ser observado na figura 3.10, onde dois examinandos que possuem a mesma aptidão, seja na turma 1 ou na turma 2, possuirão a mesma probabilidade de acertar um item.

Os modelos matemáticos que fundamentam a TRI especificam que a probabilidade de um examinando marcar a alternativa correta depende de sua aptidão ou do conjunto formado por esta aptidão e as características do item. A aplicação destes modelos estão alicerçados em três hipóteses:

1. A hipótese de unidimensionalidade do item: uma única aptidão é medida por cada item que compõe o teste. Esta característica é muito difícil de ser verificada, sendo que normalmente basta que exista uma componente ou fator dominante da referida aptidão que influencie no processo.

2. A independência local que ocorre quando as aptidões que influenciam no desempenho no teste são mantidas constantes e as respostas a qualquer par de itens são estatisticamente independentes (ou seja, os únicos fatores que influenciam as respostas aos itens do teste são as aptidões especificadas no modelo)

3. E finalmente, o fato que a função característica do item reflete a relação verdadeira entre as variáveis não observáveis (aptidões) e as variáveis observáveis (respostas aos itens).

Construir uma ICC (curva característica do item) e o modelo de Rasch[2]

Quando se deseja estimar a aptidão de um atleta de salto em altura, esta pode ser obtida através:

de um recorde individual;

de um recorde individual durante um evento oficial ou internacional;

da média do desempenho do atleta durante um determinado período de tempo; ou

do desempenho mais freqüente num determinado período de tempo.

Suponha, então, que para realização dessa tarefa, tomou-se os resultados obtidos por dois atletas de salto em altura durante um ano e construiu-se um diagrama da probabilidade de sucesso desses atletas em função da altura saltada.

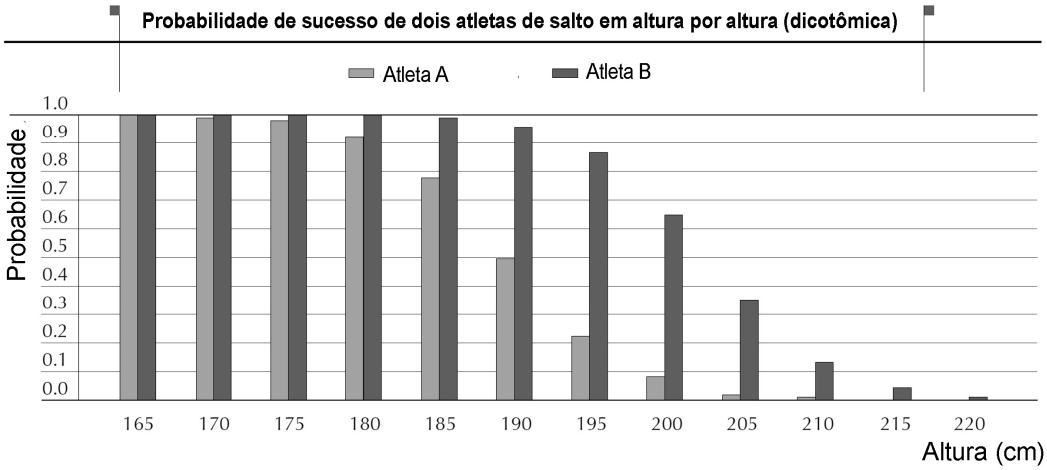


Figura 10. Probabilidade de sucesso de dois atletas de salto em altura pela altura saltada no período de um ano.

Observe que os dois atletas sempre têm sucesso em 165 cm. A partir daí, a probabilidade de sucesso cai até alcançar 0 (zero) para ambos quando a altura a ser saltada é de 225 cm. Para o primeiro atleta, essa diminuição da probabilidade se inicia a altura de 170 cm, enquanto para o segundo, o mesmo fato só ocorre a partir de 185 cm.

Estas informações podem ser tratadas com o modelo de regressão logística. Esta análise estatística consiste em transformar a variável dicotômica (sim/não, sucesso/fracasso, 1/0) em uma variável contínua. Neste exemplo, a variável contínua indicará o sucesso ou fracasso de um atleta em particular em função da altura do salto. Os resultados desta análise permitirão estimar a probabilidade de sucesso em qualquer

[2] Exemplo retirado de PISA 2009 Data Analysis Manual, OECD.
http://www.oecd.org/document/19/0,3746,en_2649_35845621_48577747_1_1_1_1,00.html

altura. O diagrama da Figura 11 apresenta a possibilidade de sucesso para os dois atletas.

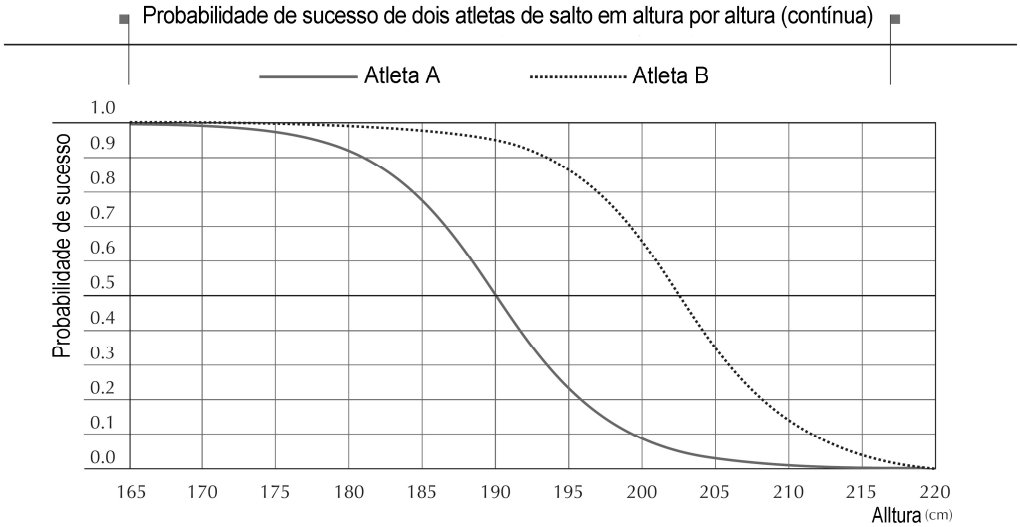


Figura 11. Probabilidade de sucesso de dois atletas de salto em altura pela altura saltada no período de um ano.

Estas duas curvas representam a probabilidade de sucesso para os dois atletas. A curva sólida representa a probabilidade de sucesso do atleta A e pontilhada do atleta B. Por definição, o nível de desempenho pode ser definido como a altura em que a probabilidade de sucesso é 0,5. Isso faz sentido uma vez que abaixo deste nível a probabilidade de sucesso passa a ser menor do que a de fracasso e acima dele, o inverso.

Neste exemplo em particular, o nível de desempenho dos dois atletas é respectivamente 190 cm e 202,5 cm. Observe na Figura 11 que o nível de desempenho do atleta A é visto diretamente no gráfico, enquanto o do atleta B é estimado do modelo. Uma propriedade fundamental deste tipo de abordagem é que o nível (ou seja, a altura) do sarrafo a ser saltado e o desempenho dos atletas são expressas na mesma medida ou escala.

A mesma idéia até aqui explorada está por trás do modelo de Rasch (um tipo de modelo logístico de um parâmetro) para a TRI. A dificuldade dos itens em um teste é análoga à dificuldade do salto com base na altura da barra. Além disso, assim como um salto em particular possui dois resultados possíveis (sucesso ou fracasso), a resposta de um aluno a um determinado item também possui duas possibilidades (acerto ou erro). Finalmente, assim como o desempenho do atleta foi definida no ponto onde a probabilidade de sucesso é de 0,5, o desempenho ou aptidão do aluno é medida no ponto onde sua probabilidade de sucesso no item é de 0,5.

Uma das características importantes do modelo de Rasch é que ele cria um contínuo no qual tanto o desempenho do estudante quanto a dificuldade do item estarão localizados em uma função probabilística que relaciona essas duas componentes. Alunos de baixo desempenho ou aptidão e itens fáceis estão localizados na parte inferior do contínuo ou escala enquanto alunos de alto desempenho ou aptidão e itens difíceis estarão localizados na parte superior do continuo ou escala. Na Figura 12, está representada a probabilidade de sucesso e de fracasso num item de dificuldade zero.

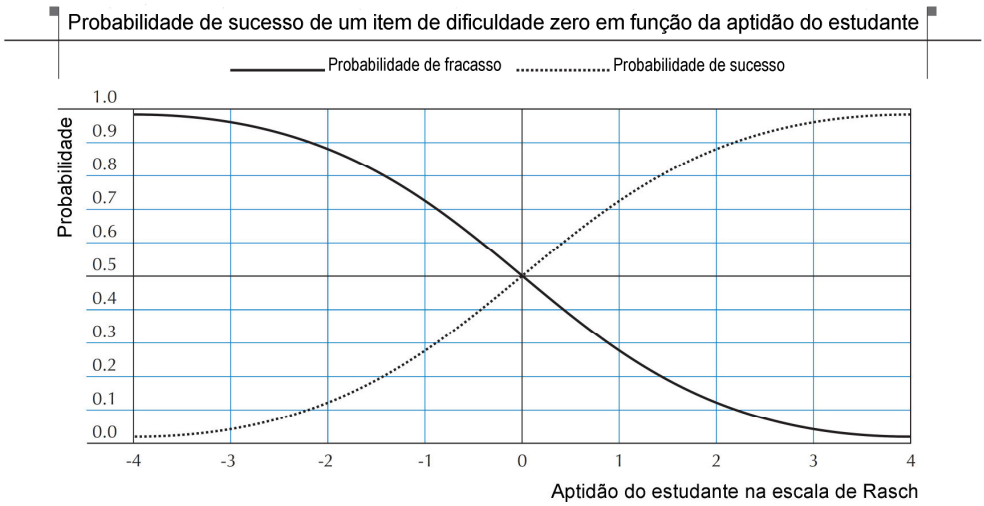


Figura 12. Probabilidade de sucesso de dois atletas de salto em altura pela altura saltada no período de um ano.

Como se pode observar acima, um estudante com uma aptidão zero tem a probabilidade de 0,5 de sucesso ou fracasso num item de dificuldade zero. Um estudante com aptidão – 2 tem a probabilidade de um pouco mais de 0,10 de sucesso e um pouco menos de 0.90 de fracasso no mesmo item de dificuldade zero. Mas, este último, teria probabilidade 0,5 de sucesso num item de dificuldade – 2.

Do ponto de vista matemático, a probabilidade de um estudante com uma aptidão "θ" responder corretamente um item de dificuldade "b" é:

$$"odds\ ratio" = \frac{probabilidade\ (acerto)}{probabilidade\ (acerto)} = \frac{P}{1-P}$$

$$\frac{P}{1-P} = e^{(\theta-b)} \quad \rightarrow \quad P(\theta) = \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}} = \frac{1}{1+e^{-(\theta-b)}}$$

De forma similar, a probabilidade de fracasso P(θ) =0 é dada por:

$$P(\theta_{fracasso}) = 1 - \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}} \quad \rightarrow \quad P(\theta_{fracasso}) = \frac{1+e^{(\theta-b)} - e^{(\theta-b)}}{1+e^{(\theta-b)}} = \frac{1}{1+e^{(\theta-b)}}$$

Observe que P(θ) + P(θfracasso) = 1, ou seja:

$$P(\theta) + P(\theta_{fracasso}) = \frac{e^{(\theta-b)}}{1+e^{(\theta-b)}} + \frac{1}{1+e^{(\theta-b)}} = \frac{1+e^{(\theta-b)}}{1+e^{(\theta-b)}} = 1$$

Em outras palavras, considera-se que a curva característica do item pode ser ajustada (no sentido discutido antes, de regressão) por uma curva do tipo mostrado para P(θ).

A seguir, apresentam-se alguns exemplos do cálculo da probabilidade de sucesso para aptidões e níveis de dificuldades estabelecidos.

1º exemplo) Cálculo da probabilidade de sucesso quando a habilidade do estudante é igual a dificuldade do item.

Tabela 10. Relação de aptidão do estudante, dificuldade do item e probabilidade de sucesso quando a aptidão do estudante é igual a dificuldade do item.

Aptidão do estudante (θ)	Dificuldade do item (b)	Probabilidade de sucesso
-2	-2	0,5
-1	-1	0,5
0	0	0,5
1	1	0,5
2	2	0,5

$\theta - b = 0$ ➔ $P(\theta) = \frac{1}{1+e^{-(\theta-b)}} = \frac{1}{1+e^{-(0)}} = \frac{1}{1+1} = 0,5$

Logo, a probabilidade de fracasso será: P(θfracasso) = 1- 0,5 = 0,5

2º exemplo) Cálculo da probabilidade de sucesso quando a aptidão do estudante é uma unidade menor do que a dificuldade do item.

Tabela 11. Relação de aptidão do estudante, dificuldade do item e probabilidade de sucesso quando a aptidão do estudante é uma unidade menor do que a dificuldade do item.

Aptidão do estudante (θ)	Dificuldade do item (b)	Probabilidade de sucesso
-2	-1	0,27
-1	0	0,27
0	1	0,27
1	2	0,27
2	3	0,27

$\theta - b = -1$ ➔ $P(\theta) = \frac{1}{1+e^{-(\theta-b)}} = \frac{1}{1+e^{-(-1)}} = \frac{1}{1+(2,718)^1} = 0,27$

Logo, a probabilidade de fracasso será: P(θfracasso) = 1- 0,27 = 0,73

3º exemplo) Cálculo da probabilidade de sucesso quando a aptidão do estudante é uma unidade maior do que a dificuldade do item.

Tabela 12. Relação de aptidão do estudante, dificuldade do item e probabilidade de sucesso quando a aptidão do estudante é uma unidade maior do que a dificuldade do item.

Aptidão do estudante (θ)	Dificuldade do item (b)	Probabilidade de sucesso
-2	-3	0,73
-1	-2	0,73
0	-1	0,73
1	0	0,73
2	1	0,73

$\theta - b = 1$ ➔ $P(\theta) = \frac{1}{1+e^{-(\theta-b)}} = \frac{1}{1+e^{-(1)}} = \frac{1}{1+(2,718)^{-1}} = 0,73$

Logo, a probabilidade de fracasso será: P(θfracasso) = 1- 0,73 = 0,27

4º exemplo) Cálculo da probabilidade de sucesso quando a aptidão do estudante é duas unidades menor do que a dificuldade do item.

Tabela 13. Relação de aptidão do estudante, dificuldade do item e probabilidade de sucesso quando a aptidão do estudante é duas unidades menores do que a dificuldade do item.

Aptidão do estudante (θ)	Dificuldade do item (b)	Probabilidade de sucesso
-2	0	0,12
-1	1	0,12
0	2	0,12
1	3	0,12
2	4	0,12

$\theta - b = -2$ ➔ $P(\theta) = \frac{1}{1+e^{-(\theta-b)}} = \frac{1}{1+e^{-(-2)}} = \frac{1}{1+(2,718)^{2}} = 0,12$

Logo, a probabilidade de fracasso será: P(θfracasso) = 1- 0,12 = 0,88

5º exemplo) Cálculo da probabilidade de sucesso quando a aptidão do estudante é duas unidades maior do que a dificuldade do item.

Tabela 14 – Relação de aptidão do estudante, dificuldade do item e probabilidade de sucesso quando a aptidão do estudante é duas unidades maior do que a dificuldade do item.

Aptidão do estudante (θ)	Dificuldade do item (b)	Probabilidade de sucesso
-2	-4	0,12
-1	-3	0,12
0	-2	0,12
1	-1	0,12
2	0	0,12

$\theta - b = 2$ ➔ $P(\theta) = \frac{1}{1+e^{-(\theta-b)}} = \frac{1}{1+e^{-(2)}} = \frac{1}{1+(2,718)^{-2}} = 0,88$

Logo, a probabilidade de fracasso será: P(θfracasso) = 1- 0,88 = 0,12

Observe que:

* se a aptidão do estudante é igual a o nível de dificuldade do item, a probabilidade de sucesso será sempre igual a 0,5, independentemente da posição da aptidão e da dificuldade do item no contínuo;

* se a dificuldade do item excede aptidão do estudante em uma unidade, a probabilidade de sucesso será sempre igual a 0,27, independentemente da posição da aptidão e da dificuldade do item no contínuo;

* se aptidão do estudante excede a dificuldade do item em uma unidade, a probabilidade de sucesso será sempre igual a 0,73, independentemente da posição da aptidão e da dificuldade do item no contínuo;

* se duas unidades separam a aptidão do estudante e o nível de dificuldade do item, a probabilidade de sucesso será igual a 0,12 e 0,88, respectivamente, independentemente da posição da aptidão e da dificuldade do item no contínuo.

Desses exemplos, fica evidente que somente a distância entre a aptidão do estudante e o nível de dificuldade do item, no continuo de Rasch, que será fator determinante da probabilidade de sucesso deste aluno na execução de um determinado item.

Os exemplos também ilustram a simetria das escalas. Se um estudante possui aptidão uma unidade abaixo do nível de dificuldade do item, a probabilidade de sucesso deste será de 0,27, ou seja, 0,23 abaixo da probabilidade de sucesso quando essa aptidão e dificuldade do item são iguais. Agora, se ao contrário, o estudante possuir aptidão uma unidade acima do nível de dificuldade do item, sua probabilidade de acerto será de 0,73, ou seja, 0,23 acima da probabilidade de sucesso quando essa aptidão é igual ao nível de dificuldade do item. Da mesma forma, a diferença é de duas unidades gerará um fator de simetria igual a 0,38.

Modelos para a Teoria da Resposta ao Item

Quando falamos de modelo, estamos pensando em ajustar os dados obtidos por uma curva. No caso da Teoria da Resposta ao Item, esse ajuste deve ser feito de forma a fornecer uma função que descreva adequadamente o comportamento da curva característica do item, na forma de um S deitado de lado.

Os modelos existentes para a TRI diferenciam-se basicamente pelo número de parâmetros utilizados para descrever os itens. Os modelos mais populares são os modelos logísticos de um, dois e três parâmetros, sendo que estes são apropriados para dados de respostas de itens dicotômicos.

Modelo Logístico de um Parâmetro

Este modelo para a TRI é um dos mais utilizados. Sua curva característica é dada por

$P_i(\theta) = \frac{1}{1+e^{-y(\theta)}} = \frac{e^{y(\theta)}}{1+e^{y(\theta)}}$, sendo $y(\theta) = \theta - b_i$ com $i = 1, 2, ..., n$,

Nesta expressão,

i refere-se a um item de um total de n itens que constituem o teste; esses itens são dicotômicos, isto é, só admitem respostas do tipo certo (1) ou errado (0); $P_i(\theta)$ é a probabilidade que um examinando, escolhido aleatoriamente, com aptidão θ (parâmetro do modelo), responda ao item i corretamente; a probabilidade tem sempre um valor entre 0 e 1;

b_i é um parâmetro do modelo, que indica o nível de dificuldade do item.

Observa-se que $y(\theta) = \theta - b_i$ corresponde a uma função linear, com coeficiente angular igual a 1. Pode-se verificar que, quando $y(\theta = b_i) = 0$, a probabilidade de acerto $P_i(\theta = b_i) = 0,5$, ou seja, b_i define a aptidão para a qual o candidato tem probabilidade 50% de acertar o item.

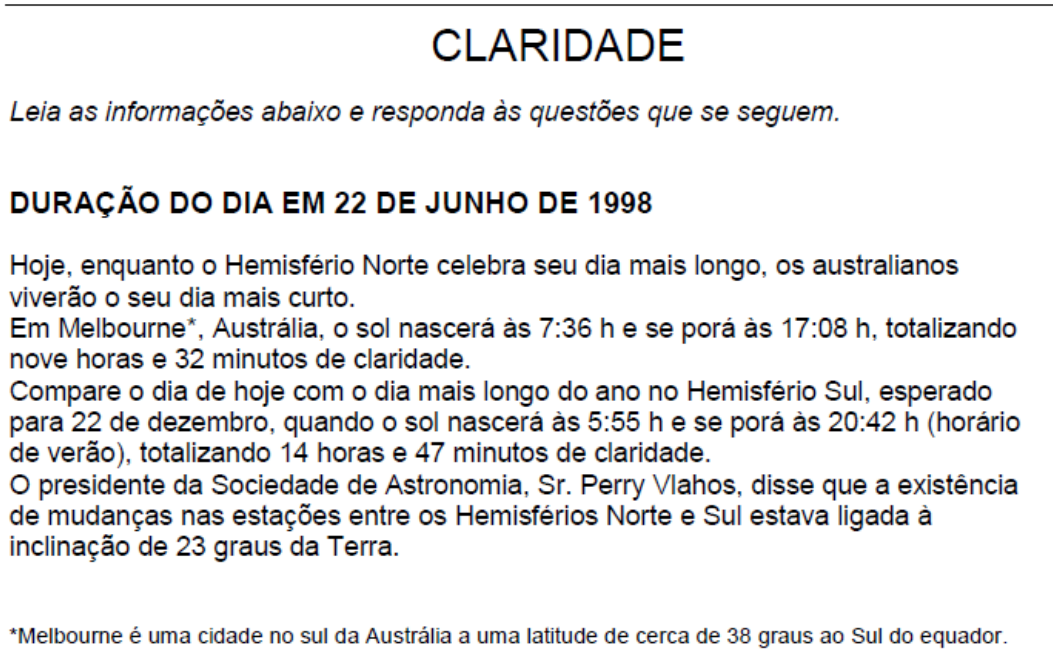
Essa parametrização corresponde ao modelo de Rasch, sendo que esse modelo logístico para a TRI é o modelo utilizado pelo PISA.

Veja agora, um exemplo da curva característica do item, da questão intitulada "Claridade" (que se constitui de dois itens), integrante da prova de Ciências, na edição de 2000.

A questão "Claridade" foi tornada pública pelo PISA3 e seu texto é apresentado a seguir. Como em todas as unidades do PISA, uma "questão" é constituída de um texto

[3] As questões públicas do PISA podem ser consultadas na página do INEP: http://portal.inep.gov.br/internacional-novo-pisa-itens. consultado em

introdutório seguido de um ou mais itens. A unidade Claridade tem dois itens associados. Na figura 3.14 está apresentado o texto introdutório.



CLARIDADE

Leia as informações abaixo e responda às questões que se seguem.

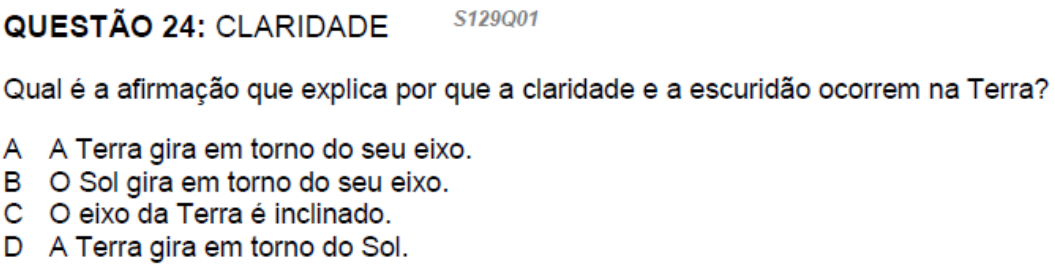
DURAÇÃO DO DIA EM 22 DE JUNHO DE 1998

Hoje, enquanto o Hemisfério Norte celebra seu dia mais longo, os australianos viverão o seu dia mais curto.
Em Melbourne*, Austrália, o sol nascerá às 7:36 h e se porá às 17:08 h, totalizando nove horas e 32 minutos de claridade.
Compare o dia de hoje com o dia mais longo do ano no Hemisfério Sul, esperado para 22 de dezembro, quando o sol nascerá às 5:55 h e se porá às 20:42 h (horário de verão), totalizando 14 horas e 47 minutos de claridade.
O presidente da Sociedade de Astronomia, Sr. Perry Vlahos, disse que a existência de mudanças nas estações entre os Hemisférios Norte e Sul estava ligada à inclinação de 23 graus da Terra.

*Melbourne é uma cidade no sul da Austrália a uma latitude de cerca de 38 graus ao Sul do equador.

Figura 12. Texto introdutório da questão "Claridade" do PISA 2000 [PISA 2000].

O primeiro item apresentado é uma questão de múltipla escolha referente ao conhecimento de fenômenos astronômicos básicos, no caso, especificamente a explicação para a existência de dia e noite (claridade e escuridão). Seu texto está apresentado na Figura 13.



QUESTÃO 24: CLARIDADE *S129Q01*

Qual é a afirmação que explica por que a claridade e a escuridão ocorrem na Terra?

A A Terra gira em torno do seu eixo.
B O Sol gira em torno do seu eixo.
C O eixo da Terra é inclinado.
D A Terra gira em torno do Sol.

Figura 13. Texto do item 1 da questão "Claridade" do PISA 2000 [PISA 2000].

A construção das ICC´s deste item pode ser feita através da utilização dos dados brutos do PISA (disponíveis em www.pisa.oecd.org). Considerando-se o escore bruto dos participantes do exame como sendo a aptidão, pode-se reconstruir as ICC´s para diversas sub-amostras dos participantes. Na Figura 14 apresentamos as ICC´s para a item "Claridade – Q1" da figura 3.15 para todos os participantes do PISA, do Brasil, do Japão e do Reino Unido.

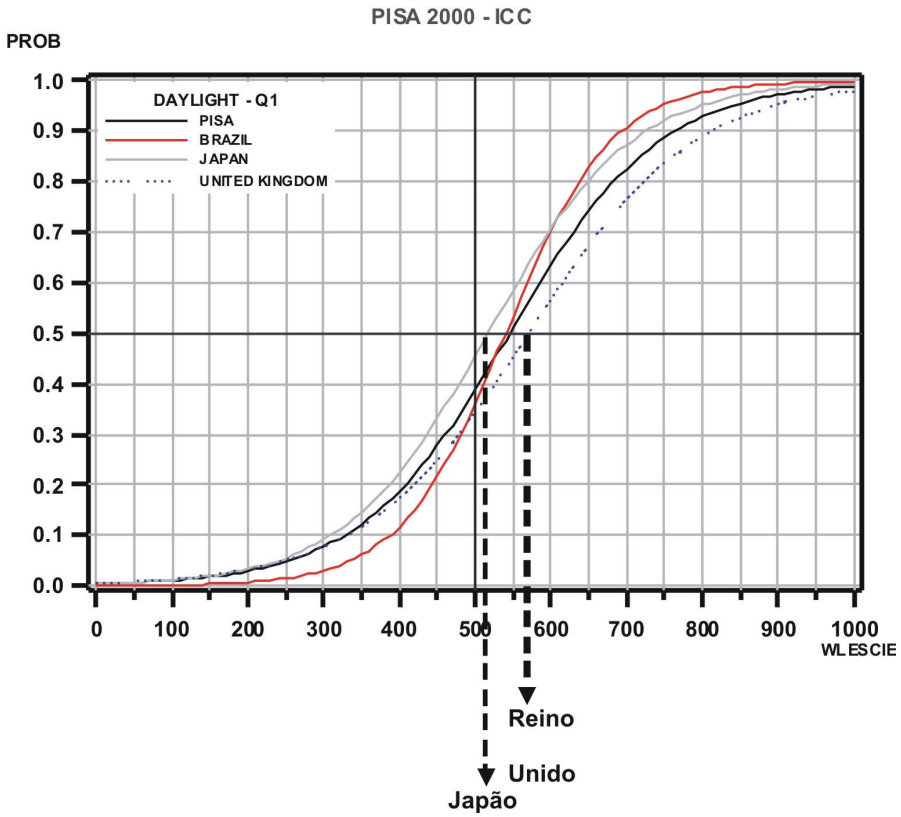


Figura 14. ICC de Todos participantes do PISA 2000, do Brasil, do Japão e do Reino Unido.

Observa-se, pela posição em relação ao eixo de aptidão (WLESCIE no gráfico) que esse item foi um pouco mais difícil para os estudantes do Reino Unido (linha tracejada azul) do que para os do Japão (linha contínua cinza), já que a aptidão que corresponde a uma probabilidade de acerto do item igual a 50% foi maior para o Reino Unido do que para o Japão. O índice de dificuldade para os participantes brasileiros é intermediário entre esses dois países e praticamente igual ao da amostra global.

Esta constatação revela a idéia básica da construção das ICC´s. O desempenho dos participantes brasileiros (no PISA 2000) revela-se bastante pior do que do restante do mundo quando observamos os histogramas dos escores no exame. O índice de acerto dos estudantes brasileiros neste item é muito mais baixo do que do total de participantes e dos participantes do Japão e Reino Unido, por exemplo[4]. Mesmo assim, as ICC´s são praticamente idênticas.

Na Figura 15 mostramos o item 2 da unidade "Claridade". Este item exige o desenho das posições do eixo da Terra e do Equador, sendo considerada bastante difícil.

[4] Esses gráficos e resultados estão apresentados em artigos de M.F. Barroso.

QUESTÃO 25: CLARIDADE S129Q02- 0 1 2 8 9

A Figura 1 demonstra os raios de luz do sol se refletindo sobre a Terra.

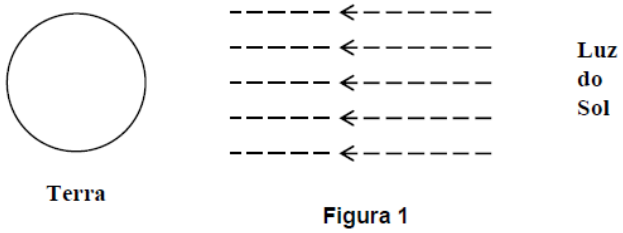


Figura 1

Suponha que seja o dia mais curto em Melbourne. Mostre o eixo da Terra, o Hemisfério Norte e o Hemisfério Sul na Figura 1.

Figura 15. Texto do item 1 da questão "Claridade" do PISA 2000.

A construção das ICC´s deste item para todos os participantes do PISA e para os participantes do Brasil, Japão, Portugal e Reino Unido revela que este item apresenta um grande nível de dificuldade. Como se pode observar na Figura 16, o índice de dificuldade desse item é de 650, ou seja, apenas os participantes cujo escore foi acima da média em 1,5 desvio padrão (menos de ≅10% dos participantes) tem probabilidade de acertar este item superior a 50%.

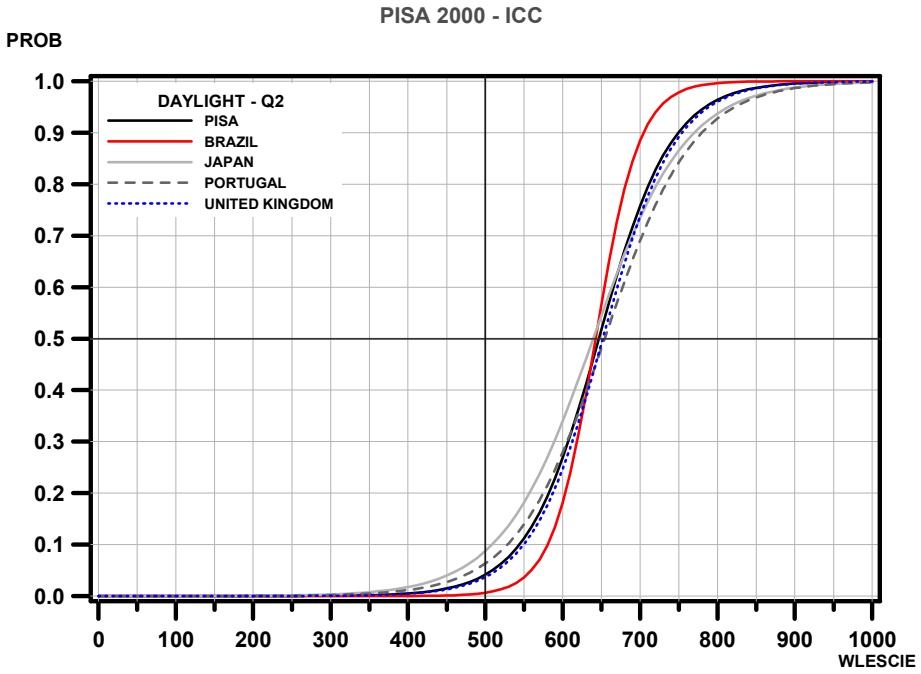


Figura 16. ICC de Todos participantes do PISA 2000, do Brasil, do Japão, de Portugal e do Reino Unido.

Ainda se pode observar na figura o índice de dificuldade deste item 2 é praticamente o mesmo para todos os países representados no diagrama.

Modelo logístico de dois parâmetros

O modelo logístico de dois parâmetros permite avaliar dois parâmetros do item: a dificuldade e a discriminação. A curva característica deste modelo para TRI é dada pela equação

$$P_i(\theta) = \frac{1}{1+e^{-y(\theta)}} = \frac{e^{y(\theta)}}{1+e^{y(\theta)}}$$, sendo $y(\theta) = Da_i(\theta - b_i)$ com $i = 1, 2, \ldots, n$,

onde

D é um fator de escala introduzido em virtude de considerações estatísticas com valor 1,75;

a_i representa o parâmetro de discriminação do item, proporcional à inclinação da ICC no ponto b_i na escala de aptidão e é determinado por $\left.\frac{dP_i}{d\theta}\right|_{\theta=b_i} = \frac{D}{4}a_i$

Esse valor pode variar de 0 a ∞, mas, normalmente, varia entre 0 e 2. Valores negativos não são aceitos pois representariam que quanto menor a aptidão do estudante, maior a possibilidade dele acertar o item.

A Figura 17 mostra os parâmetros de dificuldade (b_i) e de discriminação (a_i) de dois itens.

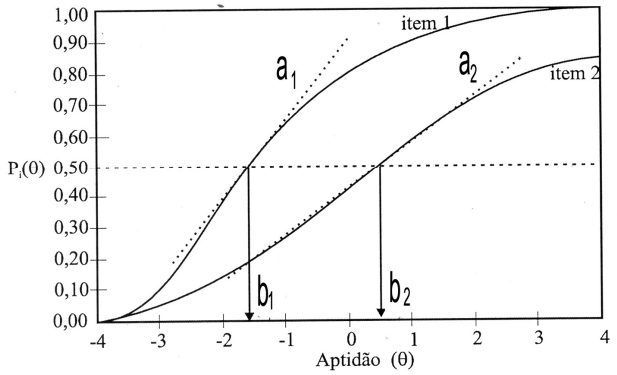


Figura 17. Parâmetros de dificuldade (b) e discriminação (a) de dois itens.

Observe que o item 2 apresenta uma maior nível de dificuldade que o item 1 (valor de b maior). No entanto, o item 1 é mais discriminatório, pois possui maior inclinação no ponto correspondente a probabilidade de acerto igual a 0,5.

Modelo logístico de três parâmetros

O modelo logístico de três parâmetros permite avaliar dois parâmetros do item: a dificuldade, a discriminação e a resposta dada ao acaso (o chute). A curva característica deste modelo para TRI é dada pela equação

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{y(\theta)}}{1+e^{y(\theta)}}$$, sendo $y(\theta) = Da_i(\theta - b_i)$ com $i = 1, 2, \ldots, n$,

onde "c_i" é o parâmetro do item que permite avaliar a resposta correta dada ao item por acaso e é expresso pela assíntota inferior da curva, no nível mais baixo do contínuo de aptidões.

Na Figura 18, apresentam-se os parâmetros de dificuldade (b_i), de discriminação (a_i) e de "chute" (c_i) de três itens.

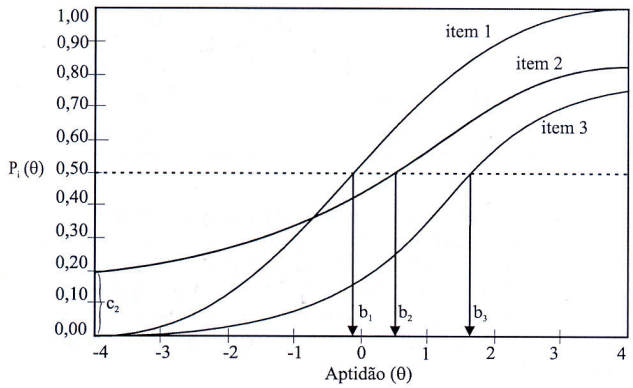


Figura 18. Parâmetros de dificuldade (b) e discriminação (a) de dois itens.

Observe que existe um nível de dificuldade crescente, do item 1 para o 3, que pode ser observado pela posição das ICC (quando Pi(θ) = 0,5) em relação ao eixo das aptidões (b1<b2<b3). É também possível determinar o índice de discriminação de cada uma das curvas, pela inclinação da mesma nesse mesmo ponto (quando Pi(θ) = 0,5).

O que difere o modelo de três parâmetros é que, no item 2 por exemplo, o ponto da ordenada que é cortado pela ICC indica a existência de chute, cuja probabilidade nesse item é de 20%. Observe que nos itens 1 e 2 essa probabilidade é zero. Segundo Pasquali,

"A lógica que fundamenta essa interpretação da assíntota é a seguinte: supostamente o sujeito não tem habilidade praticamente nenhuma, pois ele tem um θ menor que –3, e apesar disso acerta o item; consequentemente, ele só pode ter chutado e teve sorte, porque acertou."

O modelo de três parâmetros é o adotado para análise das provas do Exame Nacional do Ensino Médio

Após essa discussão, pode-se observar que a Teoria de Resposta ao Item e a ideia de construção de curvas características do item constituem-se em poderosas ferramentas para a análise de avaliações de aprendizagem. A TRI, devido à independência do teste em relação ao examinando, permite a avaliação e a comparação de diferentes grupos com diferentes aptidões. Isso é de fundamental importância para que se possa obter resultados confiáveis que permitam a tomada de ações sejam pelos professores ou elaboradores de políticas públicas.