

371.26

F581

6

COMISSÃO DE ESTUDOS DE TESTES
E PESQUISAS PSICOLÓGICAS

CADERNO 2

PRONAPA

A MEDIDA EM PSICOLOGIA
E EM EDUCAÇÃO



FUNDAÇÃO GETÚLIO VARGAS

IAO E ORIENTAÇÃO PROFISSIONAL

P/ISOP

CETPP

C

2

AB

18 JUL

FUNDAÇÃO		C.A.S.	
DATA	371.26	MADA	
83.1.69			
N.º DO VOLUME	F981		
55/69			
REGISTRADO POR			
Sania			

AC. 16323

ID 27423

19034-7

N O T A

Com o objetivo de melhor divulgar conhecimentos e informações a respeito da utilização dos testes e medidas no campo da psicologia e da educação, a Comissão de Estudos de Testes e Pesquisas Psicológicas (C.E.T.P.P.), do Instituto de Seleção e Orientação Profissional (I.S.O.P.), programou uma série de publicações para serem distribuídas nos meios educacionais, atendendo à deficiência de material acessível aos professores, diretores, orientadores, pedagogos e psicólogos de modo geral.

Estes cadernos fazem parte de um programa que está sendo realizado pela Fundação Getúlio Vargas em cooperação com a Fundação Ford, com o propósito de promover pesquisas educacionais, criar um Centro de Testes e Pesquisas Psicológicas, aperfeiçoar pessoal especializado e proporcionar estágios de treinamento a psicólogos e orientadores interessados na pesquisa educacional.

Os temas e assuntos foram selecionados atendendo aos interesses dos profissionais que trabalham no campo da psicologia e da educação.

196901 55

371.26 M489 /f



1000027423

A MEDIDA EM PSICOLOGIA E EM EDUCAÇÃO

1. INTRODUÇÃO

Na prática escolar corrente, o aproveitamento dos alunos é expresso por meio de notas, escores ou graus, seja simplesmente para avaliar seu aproveitamento, seja para fins de promoção ou expedição de diplomas ou certificados. Outras informações são também traduzidas por meio de números, entre as quais cabe assinalar os quocientes intelectuais e os escores de nível mental. Todos esses números são habitualmente tratados do mesmo modo que os números resultantes de medidas como as de comprimento ou de tempo: são freqüentes as comparações entre as notas de diversos estudantes na mesma disciplina ou entre as do mesmo estudante em disciplinas diferentes; são calculadas as médias das notas do mesmo aluno para fins de promoção ou as médias das notas dos alunos de uma mesma turma para compará-las às de outra turma; muitas operações análogas são feitas para fins especiais, como no caso da classificação dos candidatos nos concursos.

Serão legítimas as operações matemáticas que efetuamos dessa maneira com as notas ou os escores? A primeira vista esta pergunta soa como um preciosismo de filósofo desocupado: desde tempos imemoriais que isto se faz e nunca foram levantadas objeções contra esse procedimento... Entretanto, um pouco de reflexão indicará que a pergunta tem cabimento, havendo inquestionáveis motivos para se duvidar da legitimidade dessas operações, que só são válidas dentro de certos limites que importa conhecer; do contrário, poderíamos incorrer na prática da "aritmética

frívola" de que se tem falado ultimamente, e correr o risco de aceitar como verdadeiras conclusões enganadoras.

Essas questões assumem maior importância quando se trata de interpretar ou manipular matematicamente os resultados dos testes, instrumentos com os quais nem todos os professores estão familiarizados, não tendo neste caso para se guiar o bom senso inconsciente que lhe serve de apoio na interpretação das notas, que resulta de longa prática em lidar com elas. Por este motivo, julgou-se conveniente dedicar o presente folheto ao estudo desse assunto e de assuntos correlatos.

Questões desta natureza são raramente focalizadas nos meios escolares e provavelmente muitos leitores ainda não estão convencidos da existência de qualquer problema em torno do assunto. Nada melhor, portanto, que a apresentação de alguns exemplos para fazer sentir a necessidade de um estudo da questão.

Nos cursos superiores, antes da reforma Campos do ensino, as notas finais nas diversas cadeiras eram dadas como hoje, numa escala de zero a dez. Entretanto, contrariamente às regras atuais em que a nota mínima de aprovação é 5, naquele tempo só o grau zero reprovava os alunos, não sendo raras as aprovações com grau 1. Isto não significava de modo algum maior benevolência nos exames, pois o esforço que o estudante tinha então que fazer para obter grau 1 era comparável ao que faz atualmente para obter nota 5; de fato, no primeiro caso o esforço era provavelmente maior, pois prevalecia então geralmente maior rigor quanto às exigências para aprovação. Temos portanto aqui duas escalas, ambas graduadas de zero a dez, mas cujos valores numéricos não têm a mesma significação, nem como índice de conhecimentos nem do ponto de vista dos requisitos para aprovação.

Dir-se-á que isto é uma banalidade, e que tais diferenças são comparáveis às apresentadas por duas escalas termométricas, em nada ficando prejudicada a validade das comparações desde que se estabeleça uma fórmula de conversão, análoga à usada para transformar graus centígrados a Fahrenheit. Poder-se-á acrescentar que, de qualquer modo, o exemplo não atinge as comparações feitas dentro de um mesmo sistema.

O argumento parece procedente. Vejamos, porém, se são de fato válidas as comparações dentro do mesmo siste-

ma de notas, pressuposto admitido em geral sem maior exame.

Em certa escola de engenharia, havia um professor de mecânica racional que era de extremo rigor na atribuição de notas. Um aluno brilhante que havia deixado fama na escola tinha, entre suas glórias acadêmicas, a de ter tirado nota 5 em mecânica racional. Por outro lado, o professor de economia política e direito administrativo (matéria que naquela época fazia parte do currículo de engenharia) nunca tinha reprovado aluno algum e, pelo menos na memória dos que tinham cursado a escola durante um longo período, nunca tinha atribuído nota inferior a 8 nos exames finais.

É claro que as mesmas notas em economia política e em mecânica não significavam, nem por sombra, o mesmo grau de conhecimentos nas respectivas disciplinas, e qualquer conclusão baseada na sua equivalência constituiria uma violenta distorção da realidade.

Os casos verídicos desses dois professores foram citados para fins demonstrativos; representam sem dúvida exemplos de extremo rigor e de extrema benevolência, mas não há instituição de ensino onde não se manifestem, em maior ou menor grau, divergências bastante sensíveis entre os critérios de atribuição de notas de um para outro professor, sendo talvez ainda mais acentuadas as diferenças análogas de uma para outra escola.

Note-se que isto não quer necessariamente dizer que as notas nada significam. Quando um professor é competente e consciencioso, suas notas são justas e adequadas *dentro do critério por ele adotado*. O que não se pode admitir *a priori* é que elas sejam comparáveis às atribuídas por outros professores: trata-se de situação análoga à da medida de duas distâncias, uma feita em milhas, outra em quilômetros.

O problema se complica — ou se torna mais técnico — quando se quer interpretar os escores (resultados) de um teste ou compará-los aos de outro teste ou às notas conferidas por um professor. Para abordar convenientemente o problema, deve-se partir da origem, isto é, dos fundamentos lógicos do conceito de medida e de suas aplicações ao caso da psicologia e da educação, assunto do capítulo seguinte.

Certos princípios lógicos do problema da medida estão expostos com certo detalhe no presente capítulo. Para certas pessoas, os problemas fundamentais não despertam grande interesse: bastam-lhes normas práticas de aplicação. Tais pessoas poderão omitir a leitura do capítulo, limitando-se à de seu resumo final. Entretanto, talvez sintam depois a conveniência de voltar para tomar conhecimento das bases em que se apóiam as conclusões apresentadas.

Há diferentes processos de se associar números a determinados fenômenos. Para abreviar a exposição do que se segue, daremos o nome de *sistemas de associações numéricas* a esses processos, empregados às vezes com propósitos bem diferentes.

O mais importante desses sistemas é o constituído pelas medidas, no sentido restrito em que esta palavra é entendida nas ciências físicas. Todos conhecem a definição clássica: medir uma grandeza é compará-la a uma da mesma espécie, tomada como unidade. Nessa definição está implícita uma premissa fundamental que convém enunciar formalmente. Da realização de uma medida, espera-se um resultado indicando *quantas vezes* a grandeza é maior que a unidade. Mas dizer, por exemplo, que uma sala tem 6 metros de comprimento equivale a dizer que seu comprimento é igual a soma de seis comprimentos iguais a um metro, o que exige que se saiba como somar dois comprimentos. Eis aí a premissa fundamental a que nos referimos acima: para que uma grandeza seja mensurável (no sentido restrito dos físicos) é necessário que se saiba somar duas grandezas da espécie considerada. Uma vez sabendo-se como somá-las, não haverá dificuldade lógica em se definir o que seja a diferença entre duas dessas grandezas, nem o que seja a multiplicação ou divisão da grandeza por um número abstrato. *Portanto, é a possibilidade da operação soma, aplicada a determinada espécie de grandeza, que torna legítimas as operações matemáticas habituais efetuadas com os números que resultam da mensuração das grandezas.*

Note-se que isto não está implícito no simples conceito de grandeza. Uma dor de dentes pode ser maior ou menor, incluindo-se assim na idéia empírica que temos de grandeza, mas nenhum tratamento matemático poderá ser aplicado às dores de dentes enquanto não se definir de

modo operacional o que seja a soma de duas dores de dentes, o que não parece tarefa muito fácil.

Convém abrir aqui um parêntesis referente a uma questão de nomenclatura. Para os físicos, as grandezas só são mensuráveis quando se enquadram no tipo indicado acima, e portanto só em relação a elas é que caberia empregar a palavra *medida*. Os psicólogos, bem como aqueles que se dedicam às demais ciências antropológicas, dão em geral um sentido mais amplo a essa palavra. Binet deu a seu famoso teste o nome de *Échelle métrique de l'intelligence* e o título do presente folheto é outro exemplo do emprego da palavra em sentido amplo. Ora, o que determina na realidade o sentido de uma palavra é o uso que dela se faz, e nada obriga os que se dedicam a uma ciência a adotar a mesma nomenclatura usada em outro domínio do conhecimento. Contudo, em parte para evitar disputas estereis sobre palavras, certos autores empregam de preferência a palavra *avaliação*, que inclui, aliás, não só as apreciações cujos resultados são expressos por números, como as classificações em categorias e aquelas que traduzem um julgamento de *valor*, no sentido ético ou social da palavra.

De verbis non disputandum. O que importa saber é que empregamos aqui a palavra *medida* em seu sentido amplo e, quando quisermos usá-la em seu sentido habitual em física, indicaremos explicitamente que a estamos empregando em seu sentido restrito.

Fechando agora este parêntesis talvez desnecessariamente longo, passemos ao estudo de um segundo sistema de associação numérica, exemplificado pelas datas ou pelas temperaturas medidas nas escalas centígrada ou Fahrenheit. Para as escalas deste tipo, não tem sentido dizer, por exemplo, que a data 1830 é o dobro da data 915, ou que a temperatura de 10° é a terça parte da temperatura de 30°, mesmo porque, se essas temperaturas fossem convertidas para a escala Fahrenheit, teríamos respectivamente 50°F e 86°F, números que não estão na relação de um para três.

Entretanto, é perfeitamente cabível dizer que a *diferença* entre as datas 1960 e 1950 é a metade da diferença entre 1840 e 1820. O que é que torna esta afirmação legítima enquanto que a anterior não era? Simplesmente o fato de que uma diferença entre datas é um intervalo de tempo ou uma duração e as durações são grandezas para as quais se pode definir a operação soma, enquanto que não

se pode atribuir nenhum sentido útil à soma de duas datas ou de duas temperaturas.

Analisando-se a questão, verifica-se que a distinção essencial entre os dois sistemas de associação numérica já mencionados é que o primeiro é constituído por uma escala que possui um zero absoluto, enquanto que a do segundo possui um zero arbitrário. Realmente, a distância entre dois pontos que coincidem, ou a velocidade de um corpo em repouso, são grandezas nulas pela própria definição, e é isto que queremos exprimir ao dizer que a escala numérica correspondente possui um zero absoluto. Enquanto isto, a origem das datas (o ano zero) é puramente arbitrária, como a data do nascimento de Cristo ou da realização da primeira olimpíada; do mesmo modo, os zeros da escala centígrada ou da Fahrenheit são puramente arbitrários, como a temperatura do gelo em equilíbrio em presença da água (sob pressão normal, se quisermos ser mais precisos), ou de uma mistura de gelo, sal e amônia, que foi usada por Fahrenheit no século 18 para definir o zero de sua escala.

Essas escalas podem ser chamadas *escalas de intervalos uniformes, desprovidas de zero absoluto*. O que importa sobretudo fixar é que, para os valores numéricos dessas escalas, não são válidas as operações soma nem multiplicação ou divisão por um número abstrato; entretanto, é válida a operação subtração (que representa determinado intervalo da escala) e são válidas tôdas as operações citadas em relação a êsses intervalos, isto é, às *diferenças* entre dois valores quaisquer da escala.

Um terceiro sistema de associação numérica é o constituído pelas *ordenações*, de que daremos dois exemplos. É teoricamente possível, na hipótese de juizes competentes e imparciais, classificar as candidatas a um concurso de beleza feminina e dizer que Heloisa é mais bela que Joana e Joana mais bela que Carlota (salvo protestos vehementes das famílias das colocadas em posição inferior). Outro exemplo dêste sistema de associação numérica, menos sujeito a controvérsias, é o representado pela escala de dureza de Mohs, empregado há mais de um século em mineralogia. Diz-se que um mineral é mais duro que outro quando o primeiro pode riscar o segundo sem ser riscado por êste. Para sua aplicação prática, dez minerais foram adotados como pontos de referência, formando assim uma escala graduada de um a dez. O que caracteri-

za uma ordenação é que, entre dois quaisquer de seus elementos, A e B, é sempre possível estabelecer uma das três relações exclusivas: $A > B$, $A = B$ ou $A < B$.* Além disso, se $A > B$ e $B > C$, podemos concluir que $A > C$. Mas não temos nenhum meio válido para afirmar que a diferença entre A e B é superior, igual ou inferior à diferença entre C e D, nem que o valor de X é tantas vezes superior ao valor de Y. Na escala de dureza de Mohs, a fluorita, a apatita, a safira e o diamante são quatro dos pontos de referência, e seus valores são respectivamente 4, 5, 9 e 10. Disso se pode concluir que a apatita é mais dura que a fluorita e menos dura que a safira ou o diamante, mas nada nos autoriza a afirmar que a diferença entre a dureza do diamante e da safira é igual à da diferença entre a da apatita e a da fluorita, ou que a dureza da safira é igual à soma das durezas da apatita e da fluorita, nem que a dureza do diamante é o dobro da dureza da apatita, embora tais relações existam entre os números que traduzem na escala a dureza dessas substâncias. Por outras palavras, para os números indicativos dos elementos de uma ordenação, não são válidas as operações aritméticas de soma ou de subtração, nem a multiplicação ou divisão por um número abstrato; tais operações, se efetuadas, não teriam mais que um rigor precário ou seriam destituídas de sentido concreto; na melhor hipótese, seriam apenas aproximações da realidade, no pressuposto (não comprovado) de que os intervalos da escala sejam uniformes, ou que a escala possua um zero absoluto. (Esta última hipótese, sobretudo, pode levar aos mais violentos absurdos.)

Há ainda um quarto sistema de associação numérica, mencionado aqui apenas *pro memoria*. Nêle os números servem apenas como símbolos de identificação, como é o caso dos números de catálogos ou dos números de telefones. (No Rio, os dois primeiros algarismos servem para identificar a estação, enquanto que os quatro restantes identificam o assinante.) Este sistema de associação é puramente arbitrário: nenhuma ligação lógica existe entre o valor do número do telefone e qualquer elemento característico do assinante.

Outros sistemas existem, como o que associa um número complexo a um vetor ou a um ponto situado num pla-

* Em certos tipos de ordenação, exclui-se a possibilidade da relação $A = B$.

no, mas o que ficou dito acima já basta para o que temos em vista. Muito provavelmente o leitor achará que já foi mesmo demais.

EM RESUMO:

Entre os diferentes sistemas que podem ser adotados para associar números a coisas ou a observações de fenômenos, quatro foram examinados neste capítulo, a saber:

- O constituído pelas medidas propriamente ditas, como são entendidas nas ciências físicas. *Exemplos de grandezas mensuráveis por esse sistema:* comprimentos, áreas, volumes, tempos, velocidades, massas, densidades, cargas elétricas, etc. *Característica essencial:* possibilidade da definição da operação soma aplicada a essas grandezas. *Propriedades dos números do sistema:* com eles podem ser efetuadas válidamente, além da soma, outras operações aritméticas, a saber: subtração e multiplicação ou divisão por um número abstrato. (Outras operações podem ser efetuadas, desde que sejam aceitas definições complementares.)
- O constituído por escalas desprovidas de zero absoluto, mas com intervalos regulares. *Exemplos:* datas, escalas termométricas, potencial elétrico, altitudes acima de um nível arbitrário. *Característica essencial:* possibilidade de definição da operação subtração aplicada aos elementos da escala. *Propriedades dos números do sistema:* não são válidas a soma nem a divisão ou multiplicação por um número abstrato, mas essas operações podem ser válidamente aplicadas a quaisquer intervalos da escala, isto é, à diferença entre dois de seus valores.
- O constituído pelas ordenações. *Exemplos:* classificação de candidatos num concurso, escala de dureza de Mohs. *Característica essencial:* entre dois quaisquer elementos do sistema pode-se estabelecer uma das seguintes relações: $A > B$, $A = B$ ou $A < B$; além disso, se $A > B$ e $B > C$, pode-se deduzir que $A > C$. *Propriedades dos números do sistema:* Nenhuma operação matemática lhes pode ser aplicada com rigor.
- O constituído pelos números usados como símbolos de identificação. *Exemplos:* números de telefones, números de automóveis. *Característica essencial:* a associação do número ao objeto numerado é

puramente arbitrária, nenhuma ligação lógica existindo entre o número e as características desse objeto (salvo possíveis convenções para facilitar as identificações). *Propriedades dos números do sistema*: qualquer operação aritmética com esses números é inteiramente destituída de sentido.

III. ESCALAS E NORMAS

Na introdução deste folheto, levantamos dúvidas sobre a legitimidade do tratamento matemático dado habitualmente às notas de exames ou aos escores dos testes. No capítulo seguinte, examinamos, de modo geral, vários sistemas de associação de números às observações de um fenómeno; estudamos ainda, para os números que formam cada um desses sistemas, a validade das operações matemáticas a eles aplicadas.

Para dar resposta agora às questões levantadas na introdução, bastará portanto saber em qual desses sistemas se enquadram as notas dadas pelos professores às provas de exame do tipo clássico ou os escores obtidos pela aplicação dos testes. Examinemos primeiro o caso das notas dadas pelos professores.

Admitamos por hipótese que as notas tenham sido dadas por um professor experiente e cuidadoso, e que a prova tenha coberto de modo adequado o conjunto dos conhecimentos que se pretendia avaliar, de modo que, se o exame fosse repetido, os estudantes receberiam notas assaz próximas das anteriores. Podemos mesmo fazer abstração das variações de notas que ocorreriam entre os dois exames, pois esse assunto será tratado com mais detalhe no próximo capítulo.

Creio que o leitor não terá dúvida em enquadrar este caso no sistema constituído pelas ordenações. Se tiver hesitação, leia novamente os exemplos citados na introdução e considere ainda a exposição a seguir.

Todo professor digno deste nome se esforça, consciente ou inconscientemente, para obter que sua escala de notas se aproxime de uma escala de intervalos regulares. Entretanto, como tem a guiá-lo apenas seu julgamento subjetivo, pode-se adiantar com segurança que esse ideal não pode ser atingido. Se fizermos experiências sobre a capacidade de percepção da diferença entre as intensidades de dois

focos de luz, verificaremos que qualquer pessoa nota a diferença entre um foco de 3 e outro de 5 velas, mas que ninguém conseguirá distinguir um de 500 de outro de 502 velas. Se estendermos as experiências a outros sentidos e a outras situações em que se possa ter como contróle uma medida objetiva do fenômeno subjetivamente avaliado, encontraremos sempre discrepâncias análogas, havendo casos em que a discriminação é mais fácil para certa faixa de intensidades do fenômeno que para intensidades maiores ou menores, e são inúmeras as pesquisas no campo da psicologia experimental que, sem exceção, confirmam essas características do julgamento subjetivo. Se isto acontece em todos os casos em que os julgamentos podem ser controlados por medidas objetivas independentes, não há nada que autorize a admitir que o mesmo não se dê no caso das avaliações subjetivas do aproveitamento escolar. Conclui-se portanto que:

- as notas de exames constituem, na melhor das hipóteses, simples ordenações, isto é, escalas desprovidas de zero absoluto e de intervalos uniformes;

- as operações matemáticas efetuadas com notas de exames (somas ou subtrações, multiplicações ou divisões por números abstratos) são de legitimidade duvidosa, e os resultados de tais operações, que representam simples aproximações destituídas de fundamentos sólidos, devem ser interpretadas com cautela e podem levar facilmente a conclusões falaciosas.

Encaremos agora o caso dos escores obtidos pela aplicação de testes objetivos. Quanto a estes, dois casos devem ser distinguidos: o de uma prova objetiva organizada para fins imediatos sem experimentação prévia das questões, e o caso dos testes tecnicamente bem padronizados.*

* A padronização de um teste consiste num conjunto de operações de natureza psicológica e estatística, com a múltipla finalidade de: (a) estabelecer regras fixas para sua aplicação e avaliação de seus resultados, de modo que essas operações se realizem sempre de modo tão idêntico quanto possível; (b) obter que o teste conduza a uma avaliação tão exata quanto possível daquilo que se pretende avaliar; (c) fornecer aos que devem aplicá-lo informações detalhadas sobre o método empregado em sua construção, sobre suas finalidades e condições em que sua aplicação é recomendada, sobre a interpretação de seus resultados, bem como sobre as qualidades essenciais do teste, experimentalmente verificadas.

No primeiro caso, a situação é análoga à das notas conferidas pelos professores nas provas de tipo clássico: seus resultados representam simples ordenações. Entretanto (e no caso de se tratar de teste composto de numerosas questões, que cubram adequadamente os conhecimentos a examinar e que sejam bem distribuídas quanto ao nível de dificuldade) poder-se-á dizer que os intervalos da escala apresentarão provavelmente maior uniformidade que a escala constituída pelas notas do tipo clássico. É claro que o teste poderá ter (ou não) outras vantagens, mas no momento estamos considerando apenas a natureza da escala constituída pelos seus resultados numéricos.

Caso diferente é o dos testes tecnicamente bem padronizados, cujas escalas podem ser consideradas, com suficiente aproximação, como escalas de intervalos iguais. Infelizmente esta afirmação não pode ser aqui devidamente fundamentada, pois teria que lançar mão de conhecimentos de psicomетria descabidos num folheto de iniciação. A alternativa é a de nos basear no argumento não muito convincente do *magister dixit*. De fato, alguns testes, como o Stanford-Binet, foram objetos de numerosos estudos e pesquisas experimentais por parte de especialistas de grande autoridade, com resultados que confirmam plenamente a tese aqui enunciada. Mesmo para outros testes que foram devidamente padronizados mas ainda não foram sujeitos a investigações tão minuciosas, a tese pode ser aceita em face de certos estudos estatísticos mais simples, bem como pelo fato de terem sido construídos e padronizados por processos iguais ou equivalentes aos que foram seguidos por Terman nas suas revisões do teste de Binet.

Pode-se portanto dizer que, para testes bem padronizados, são válidas as operações usuais em relação às diferenças de escores, mas só a operação subtração pode ser diretamente aplicada com legitimidade aos escores desses testes. Note-se que isto não invalida a obtenção de médias aritméticas ou desvios padrão, embora suas fórmulas incluam operações proibidas aplicadas diretamente aos escores, pois tais fórmulas podem ser transformadas em outras onde só aparecem diferenças entre escores. Por outro lado, como a escala não possui zero absoluto, não tem sentido calcular o coeficiente de variação da distribuição dos escores de um teste, mesmo padronizado, pois a respectiva fórmula, $CV = m/s$, pressupõe que se trate de uma escala com zero absoluto.

Estabelecidos êsses pontos fundamentais, vejamos de quais artificios podemos lançar mão para interpretar notas de exames e escores de testes e compará-los entre si.

Se alguém lhe disser que um menino de 9 anos tem altura de 1,48 m e pêso de 26 kg, você terá provavelmente dificuldade na interpretação imediata dêsses dados, a menos que esteja especialmente familiarizado com o assunto. Entretanto, se lhe disserem que as médias de estatura e pêso para meninos dessa idade são respectivamente 1,355 m e 29,94 kg, você concluirá que êle é mais alto e menos pesado que a média dos meninos de 9 anos, e portanto alto e magro para sua idade. Se lhe disserem ainda que, nessa idade, só 2% dos meninos atingem sua altura e que 75% ultrapassam seu pêso, você concluirá imediatamente que se trata de um varapau magricela e que há motivos para submetê-lo a um exame médico para decidir se será indicado algum tratamento ou regime alimentar especial.

Em outro campo, se um engenheiro lhe dissesse que, em seu curso, tinha tirado nota 5 em mecânica racional e nota 8 em economia política, você não teria dúvidas em concluir que êle tinha sido bom aluno de economia e aluno medíocre de mecânica. Esta conclusão seria entretanto radicalmente alterada se você viesse a saber que os professores que lhe tinham conferido essas notas tinham sido os mencionados na introdução dêste folheto, e passaria a considerá-lo como tendo sido excelente aluno de mecânica e péssimo aluno de economia.

Uma das vantagens importantes dos testes padronizados sobre as notas de exame ou sobre os testes não padronizados é a existência de normas que facilitam a interpretação de seus escores. A inexistência de normas no caso das notas de exames ou de escores de testes não padronizados torna impossíveis certas comparações ou exigem cuidados especiais (nem sempre observados) nas comparações e nas conclusões que delas se pode tirar.

As normas dos testes se apresentam sob várias formas e permitem interpretações análogas às indicadas acima sobre o pêso e altura do menino de 9 anos. Sua apresentação e a maneira de usá-las serão assunto de futura publicação desta série.

IV. PRECISÃO DAS MEDIDAS PSICOLÓGICAS

Talvez o caminho mais conveniente para dar uma idéia exata do que seja a precisão (ou fidedignidade) dos

instrumentos de medida psicológica seja sua comparação com o conceito análogo da precisão das medidas físicas, assunto com o qual o leitor já deverá estar de certo modo familiarizado. Assim, abrir-se-á aqui um parêntesis para tratar d'êste último problema, pois tudo que se disser a seu respeito tem paralelo no campo das medidas psicológicas.

Ao medirmos qualquer grandeza física, por mais precisos que sejam os instrumentos empregados, existe sempre um *erro experimental* ou erro de medida. Em consequência, o número que exprime o resultado dessa operação será sempre um número aproximado: só a contagem de unidades descontínuas poderá fornecer números exatos.

Para se tirar conclusões válidas com base nos resultados das medidas, importa conhecer a margem de erro a que elas podem estar sujeitas. Com uma régua graduada em milímetros e com o auxílio de uma lente poder-se-á tentar avaliar comprimentos até décimos de milímetro, porém, por mais cuidadoso e experiente que seja o observador, dificilmente poderá garantir a exatidão do resultado até êsse limite, havendo uma margem de erro de cerca de um ou dois décimos de milímetro. Com instrumentos de alta precisão, poder-se-á reduzir consideravelmente êsse erro, mas êle sempre existirá. É óbvio que de muito pouco nos poderá servir os resultados da medida se não tivermos pelo menos uma indicação aproximada do limite superior do erro de que os resultados poderão estar afetados.

Na teoria dos erros — capítulo importante da ciência experimental — chama-se *erro absoluto* a diferença entre a medida de uma grandeza e seu verdadeiro valor. Entretanto, o que importa em geral para caracterizar a precisão de um instrumento ou de um processo de medida é o *erro relativo*, isto é, a relação entre o erro absoluto e o valor da grandeza medida (o que pode ser expresso de várias maneiras, como por exemplo em termos de porcentagem). Se eu dou a distância entre duas cidades com um erro da ordem de cem metros, estou fornecendo êsse dado com precisão mais que suficiente para a maioria dos casos em que a informação é utilizada. Mas o mesmo erro de cem metros seria muito grosseiro se se tratasse do comprimento de uma rua, e não teria nenhum sentido dar o comprimento de uma sala com um erro de cem metros. Se, no exemplo acima, a distância entre as duas cidades fôsse 20 quilômetros e o comprimento da rua 500 metros, ter-se-ia no primeiro caso um erro relativo de meio por cento e no se-

gundo um de vinte por cento, o que traduz melhor a idéia da precisão das medidas respectivas, precisão essa razoavelmente satisfatória quanto à distância entre as cidades, mas bem grosseira quanto ao comprimento da rua.

O modo pelo qual se determina experimentalmente a margem de erro de um processo de medida consiste em repetir muitas vezes a determinação da mesma grandeza, procurando exprimir o resultado com a maior exatidão possível. Verifica-se então que os resultados obtidos não coincidem exatamente. A partir das discrepâncias observadas, a teoria dos erros nos ensina a deduzir certos valores numéricos (erro padrão, erro provável, etc.) que servem para caracterizar a ordem de grandeza do erro absoluto do processo de medida empregado, podendo-se daí deduzir, se necessário, a ordem de grandeza do erro relativo.

Mutatis mutandi, tudo isto se aplica às medidas psicológicas. Quando uma pessoa é submetida a um teste de inteligência ou de escolaridade, o escore resultante está sujeito a um erro experimental, como se dá com qualquer outro instrumento de medida. A maneira de determinar a margem de erro em psicologia é ligeiramente diferente da empregada nas ciências físicas, pois não teria sentido aplicar grande número de vezes o mesmo teste ao mesmo indivíduo: não somente sua paciência se esgotaria rapidamente, como ainda, depois de certo número de repetições, o sujeito teria decorado as respostas e passaria a reproduzi-las de memória, de modo que o resultado da primeira aplicação teria significação psicológica diferente da do resultado da décima.

Um dos meios de contornar a dificuldade consiste em aplicar o teste a um grupo numeroso e, depois de um intervalo conveniente (que pode variar conforme a natureza do teste), aplicá-lo novamente ao mesmo grupo. Haverá, de uma para outra aplicação, discrepâncias entre os escores obtidos pela mesma pessoa, e é fácil de ver que, a partir dessas discrepâncias, é possível tirar conclusões sobre a margem de erro a que estão sujeitos os escores por um sistema análogo ao empregado nas ciências físicas.*

* Não cabe neste folheto de iniciação entrar em detalhes técnicos sobre este assunto. Bastará indicar aqui que essa margem de erro é traduzida pelo valor numérico do *erro padrão dos escores*, conceito definido em estatística educacional. Além do processo mencionado para sua determinação, há ainda outros que podem ser aplicados, inclusive, para o caso de uma única aplicação do teste a um grupo.

O que se obtém desta maneira é uma avaliação da ordem de grandeza do erro absoluto. No campo das ciências físicas, obtém-se o erro relativo dividindo o erro absoluto pelo valor da grandeza medida. Se o leitor se recorda do que foi dito nos dois capítulos anteriores, compreenderá que não será legítimo fazer tal operação em relação aos escores dos testes, que são traduzidos por escalas desprovidas de um zero absoluto. Há várias maneiras de contornar a dificuldade, porém a mais usual consiste no emprego do *coeficiente de precisão* (ou coeficiente de fidelidade) do teste. Para os fins aqui visados, bastará indicar que esse coeficiente teria valor numérico igual à unidade na hipótese (impossível de ser realizada na prática) de serem nulos os erros experimentais, e que seu valor seria igual a zero se nenhuma relação existisse entre o escore de uma pessoa na primeira aplicação do teste e seu escore na segunda aplicação (hipótese que também nunca se verifica, por mais ordinário que seja o teste). Para fixar melhor as idéias, diremos que um coeficiente igual a 0,90 indica um teste de boa precisão e um coeficiente superior a 0,95 um teste com excelentes qualidades sob esse ponto de vista. O leitor não deverá porém considerar isto uma afirmação categórica, pois o valor numérico do coeficiente de precisão é um pouco influenciado pela heterogeneidade do grupo ao qual o teste foi aplicado, bem como por certas circunstâncias outras dessa aplicação.

Estes conceitos também se aplicam aos testes organizados sem experimentação prévia, bem como (com certas dificuldades) às notas de exame conferidas pelos professores. Na divulgação dos testes padronizados, como já foi notado, devem ser fornecidas informações exatas sobre sua precisão, por meio do erro padrão dos escores, do coeficiente de precisão ou de ambos. Uma das vantagens importantes dos testes padronizados é que são conhecidos de antemão os erros a que estão sujeitos os seus resultados, enquanto que no caso das provas clássicas ou dos testes não padronizados essa informação é de difícil obtenção ou só pode ser obtida *a posteriori*.

V. VALIDADE DOS TESTES

A validade de um teste é em geral definida como a exatidão com que êle mede aquilo que desejamos medir. Isto, à primeira vista, parece ser uma coisa muito fácil de

determinar; as dificuldades começam quando temos de aplicá-la a um caso concreto.

A primeira dessas dificuldades está na falta de unanimidade entre os psicólogos quanto à definição exata daquilo que queremos medir, como acontece mesmo com um conceito importante e fundamental como o da inteligência, ou, dentro do âmbito da rotina escolar, com a proficiência de um aluno em matemática. Esta dificuldade não existe em geral nas ciências físicas: a resistência de um condutor à passagem da corrente elétrica, por exemplo, tem uma definição precisa e aceita sem discussão por todos os cientistas, ao passo que é raro o psicólogo que aceita sem restrições qualquer das numerosas definições de inteligência que têm sido adiantadas por diversas autoridades na matéria.

Além disso, não podemos medir diretamente a inteligência de um sujeito ou sua competência em matemática: o que se pode é tentar avaliar a maneira mais ou menos inteligente com que ele se comporta em determinadas situações, ou a maior ou menor facilidade com que resolve certas questões de matemática.

Note-se que isto não é uma situação peculiar à psicologia, mas antes um caso normalmente encontrado nos diversos domínios científicos: não é diretamente que se mede a distância da terra à lua ou a pressão atmosférica. Mas há uma diferença muito importante entre a medida indireta em física e a medida indireta em psicologia, pois no primeiro caso são em geral perfeitamente conhecidas as relações existentes entre as grandezas que se quer medir e as diretamente observadas, o que quase nunca acontece em psicologia. Vale a pena ilustrar essa diferença pela comparação entre duas medidas anteriormente mencionadas: a da pressão atmosférica e a da inteligência, o que nos servirá, aliás, para esclarecer melhor o conceito de validade das medidas psicológicas.

Todos sabem que, no barômetro comum, a pressão atmosférica é avaliada pela altura da coluna de mercúrio no tubo barométrico. Se a pressão que a coluna exerce na sua base é equilibrada pela pressão atmosférica, e uma vez que a pressão da coluna de mercúrio depende de sua altura sobre a superfície livre, bastará medir essa altura para se conhecer a pressão procurada.

Se quisermos ser mais exatos, teremos que considerar que a pressão exercida pela coluna de mercúrio depende

de fato de sua altura, mas é também ligeiramente influenciada pela temperatura. Entretanto, nada mais fácil que levar em conta essa influência (que constitui o que se chama uma *perturbação*), pois há uma fórmula que permite eliminar facilmente a influência perturbadora da temperatura.

Comparemos esse caso com o da medida da inteligência por meio da aplicação do Stanford-Binet, um dos mais justamente famosos testes de inteligência. O escore que uma pessoa obtém nesse teste depende fundamentalmente de seu nível mental, mas é também influenciado pelos antecedentes culturais da pessoa examinada, que constituem uma perturbação análoga à da temperatura no caso da pressão atmosférica. Mas a dificuldade no caso da medida da inteligência é que não temos maneira fácil de medir os antecedentes culturais e, mesmo que pudéssemos fazê-lo, não dispomos de uma fórmula simples para eliminar a perturbação causada por este fator.

Não procuremos nos iludir sobre a importância dessas e de outras dificuldades não mencionadas, inerentes aliás ao estágio em que se encontram as ciências psicológicas. Mas não devemos exagerá-las. Desde que, de acordo com o consenso das pessoas competentes no assunto, seja possível afirmar que Aristóteles era mais inteligente que Marco Antônio ou, no plano mais terra à terra de uma situação escolar, que Carlos é mais inteligente que Artur, temos em mãos um *critério* em relação ao qual poderemos aferir a validade de nosso teste de inteligência: ele será tanto mais válido quanto suas classificações mais se aproximarem do julgamento de um grupo representativo de pessoas competentes no assunto, e isto independentemente do fato de não haver perfeito acordo entre os especialistas quanto à definição conveniente do que seja inteligência. Aliás, não precisamos ter uma definição satisfatória do que seja, em sua essência, a eletricidade, para medirmos uma carga elétrica ou a intensidade de uma corrente.

Assim sendo, as dificuldades sobre a validade dos testes de inteligência (ou de qualquer outro aspecto psicológico) serão superadas desde que se consiga um *critério* satisfatório referente ao que se quer medir, em relação ao qual possamos comparar experimentalmente os resultados da aplicação do teste.

Em alguns casos, esse critério é fácil de obter. Suponhamos que se tenha aplicado uma bateria (conjunto) de testes para selecionar operários para uma fábrica. É claro

que a validade da bateria será tanto maior quanto seus resultados mais se aproximarem da eficiência posteriormente demonstrada pelos operários no trabalho que vierem a exercer. Se essa eficiência puder ser avaliada de modo satisfatório, como é o caso de certos trabalhos elementares que podem ser avaliados pelo número de peças aceitáveis produzidas por hora, ter-se-á com facilidade um critério quase perfeito para avaliar a validade da bateria de testes empregada na seleção.

Em outras situações não será tão fácil a obtenção de um bom critério de avaliação, podendo-se lançar mão de vários critérios diferentes ou de outros meios de apreciação da validade. Mas antes de assinalar um desses expedientes, voltemos ao caso da medida da pressão barométrica para elucidar mais nitidamente a diferença que existe entre precisão e validade.

Consideremos a pressão atmosférica avaliada pela altura da coluna barométrica, sem corrigi-la da influência perturbadora da temperatura. Essa coluna pode ter sua altura medida com determinado grau de exatidão, o que dependerá do processo usado para essa operação. No que diz respeito à avaliação da pressão atmosférica (que é o que nos interessa diretamente), aos erros experimentais da medida acrescentam-se os erros devidos à influência perturbadora da temperatura, que não foram eliminados. Temos assim certo grau de exatidão na medida da altura e um menor grau de exatidão na avaliação da pressão. Usando a linguagem habitual em psicometria, poderemos dizer que a exatidão com que medimos a altura da coluna (observação direta) corresponde à precisão do teste, enquanto que a exatidão com que medimos a altura da coluna (observamos pretendemos avaliar) representa a validade do teste.

Este exemplo nos permite encarar de modo um pouco diferente o problema da validade, que apresenta interesse especial para os testes de conhecimentos escolares. Da própria definição do que seja validade, pode-se decompô-la em dois aspectos, traduzidos pelas seguintes perguntas: O que o teste mede é realmente aquilo que pretendemos medir? Com que exatidão ele realiza essa medida? A resposta à primeira pergunta diz respeito ao problema da *relevância*, que procura investigar se os elementos diretamente avaliados pelo teste são fatores importantes e suficientes para a caracterização do traço psicológico que se pretende medir. A segunda pergunta se reduz ao problema da *precisão*, cuja solução não apresenta dificuldades especiais.

A investigação da relevância constitui assim, de certo modo, a chave para a solução do problema da validade. Vários métodos podem ser empregados para isso, inclusive métodos elaborados de análise fatorial. No caso das provas de aproveitamento escolar, uma análise subjetiva do conteúdo do teste permite verificar se os objetivos visados pelo ensino da disciplina estão traduzidos pelas questões formuladas no teste de modo adequado e na justa proporção da importância relativa de cada um deles. Obter-se-á assim o que se costuma chamar a validade aparente (*face validity*) do teste que, embora sujeita às variações de pontos de vista individuais inerentes a todas as apreciações subjetivas, não deixa de ser uma contribuição importante ao estudo da validade do instrumento.

Ao terminar este capítulo, cabe assinalar que a validade não é uma característica inerente ao teste, pois depende evidentemente do uso que dele se pretende fazer. Um teste de inteligência pode ter validade elevada para prognosticar o sucesso em seus estudos universitários dos candidatos ao curso de engenharia, mas uma validade bem mais baixa para prognosticar o sucesso de um candidato ao curso de violino ou para selecionar datilógrafas. Assim, enquanto se pode falar da precisão de um teste sem referência à sua aplicação, não terá sentido falar da validade do mesmo teste sem mencionar o propósito com que ele vai ser utilizado.

N.Cham. P/ISOP CETPP C 2

Título: A Medida em psicologia e em educação.



00027423
16323

FGV - BMHS AB

Nº Pat.:55/69

11 MAI 70
18 JUN 75

PRONAPA

Editado pela Fundação Getúlio Vargas
Praça de Botafogo, 186 — ZC-06 — RIO — GB



GELSA — Composto e impresso na Gráfica Editora Livro S. A.
Rua Prefeito Olímpio de Melo, 1460 - tel. 48-5057 - Rio - GB