Medidas de precisão e de validade dos testes

OCTAVIO A. L. MARTINS
Técnico de educação

O emprêgo dos testes para avaliar a aprendizagem escolar — para só falar dêsse aspecto das medidas educacionais — é, desde muito, processo corriqueiro na América do Norte. Entre nós, embora seja possível dizer que não há testes padronizados, o uso de testes objetivos em alguns sistemas educacionais e em concursos para admissão de pessoal — a princípio para o Instituto de Industriários e logo depois sistematicamente adotado pelo DASP em quasi todos os concursos para admissão de funcionários públicos — já tornou largamente divulgado êsse processo que hoje não é mais desconhecido, mesmo do público em geral. É pois natural supor que haja especialistas interessados em certos problemas que os testes apresentam. Esta suposição me faz publicar o presente estudo, reprodução desenvolvida e sob forma modificada de trabalho apresentado em abril do corrente ano no *Advanced course of educational statistics*, dirigido pela prof. H. M. Walker no *Teachers College* da Universidade de Columbia. Dêste trabalho, só constituem contribuição original as noções de índice e de êrro de validade e sua interpretação : sob formas às vezes diversas, as demais noções já são encontradas em livros e monograafias sôbre o assunto, merecendo especial menção a tese de Cureton publicada em 1931 (referência 1) e a monografia litografada de Thurstone (ref. 10).

Conquanto distintas, as noções de *precisão*[1] e de *validade* dos testes são intimamente ligadas. No dicionário de termos estatísticos de Kurtz (ref. 5), *precisão* (reliability) é a exatidão (accuracy) com que um teste (ou outro instrumento de medida) mede a função por êle realmente medida, qualquer que seja essa função ; a *validade* pode ser paralelamente definida como a exatidão com que o teste mede a função que desejamos medir, função esta que só aproximadamente se confunde com a função efetivamente medida pelo teste. A distinção entre as duas noções é suficientemente nítida, mas como se trata de um ponto fundamental, não será inútil ilustrá-la com um exemplo. Suponhamos um teste construído com o intúito de medir o aproveitamento em física dos alunos do curso secundário ; na construção dêsse teste houve o propósito de avaliar até que ponto os estudantes atingiram determinados objetivos do ensino : conhecimento de certos fatos, compreensão de determinadas relações, aquisição de certos hábitos de pensamento, etc. Suponhamos ainda que o estudo estatístico dos resultados da aplicação repetida dêsse teste a um grupo de estudantes tenha provado

[1] Emprego *precisão* para traduzir *reliability*, usada em relação a testes e estatísticas. E' realmente esta expressão que traduz a noção de *reliability* tal como é definida e usada em estatística educacional em relação a testes ou outros instrumentos de medida. A palavra *précision* é aliás empregada para traduzir *reliability* por autores de língua francesa, como por exemplo Fessard. Êste autor distingue entre as noções de *precisão* e *coerência* (o que me parece de utilidade duvidosa), ambas correspondendo ao inglês *reliability*. Êle usa a expressão *coefficient de cohérence* para traduzir *coefficient of reliability* (cf. ref. 2, pg. 222), e a denominação genérica de *índice de precisão* para designar várias medidas de precisão dos testes. Prefiro acompanhar mais de perto a terminologia americana, já suficientemente fixada. A falta entre nós de uma terminologia estável referente a estatística educacional me leva, para evitar equívocos, a reproduzir os termos ingleses de onde são derivados a maioira das expressões técnicas usadas neste estudo.

que êsses resultados são precisos e estáveis, isto é, que a aplicação repetida do mesmo teste (ou de formas comparáveis do mesmo teste) conduz a resultados sensivelmente equivalentes. Provada estará assim a *precisão* do teste, mas nada sabemos ainda sôbre sua *validade*, isto é, si uma nota[2] alta indica que o aluno tenha realmente alcançado os objetivos visados pelo ensino : nada impede que essa nota elevada decorra simplesmente do fato de ter o estudante decorado um certo número de fórmulas ou de ter descoberto, pela própria redação das questões do teste, qual a resposta mais conveniente ; no primeiro caso, estaria o teste funcionando como teste de memorização mecânica, e no segundo, como teste de inteligência, funções estas que não são as que, no momento, desejamos medir.

Para se verificar a validade de um teste de aprendizagem, é indispensável a existência de um critério independente para avaliação do aproveitamento do aluno ; a verificação da validade resulta da comparação dos resultados obtidos pela aplicação do teste com os resultados obtidos pela aplicação do critério externo. Daí a dificuldade essencial da medida da validade dos testes: a não existência de um critério absoluto para essa comparação. Entretanto, si não há critério externo que possa satisfazer a todas as correntes de filosofia educacional, há pelo menos critérios mais ou menos aceitáveis. Por outro lado, veremos adiante que não há necessidade especial de que êsse critério comparativo seja de alta precisão. Todos sabemos que notas de julgamento de provas escritas ou de outros trabalhos escolares são sujeitas a grandes variações arbitrárias ; o julgamento subjetivo do professor é portanto um critério muito impreciso, mas si admitirmos que êsse julgamento (ou a média de vários julgamentos) é fundamentalmente válido (isto é, que o julgamento, embora impreciso, incide realmente sôbre os objetivos visados pelo ensino), poderá perfeitamente servir como critério externo para determinação da validade de um teste de aprendizagem.

[2] Designo por *nota* (score) o resultado numérico da aplicação de um teste ou de qualquer outro processo de julgamento ou classificação. A expressão não deve ser confundida com nota ou grau, conferido de acôrdo com preceitos legais, para fins de aprovação ou promoção de alunos.

As considerações acima definem as noções de precisão e de validade dos testes. Para que sejam aplicáveis, deve ficar determinada a maneira de medir êsses elementos, a começar pela medida da precisão, noção mais simples e problema já satisfatoriamente resolvido.

A medida mais usada da precisão de um teste é o *coeficiente de precisão* (coefficient of reliability). Não é mais que o coeficiente pearsoniano de correlação[3] entre os resultados da aplicação de duas formas do mesmo teste a determinado grupo de indivíduos. Sua expressão é

$$r_{xx} = \frac{\Sigma (xx')}{N\, s\, s'}, \qquad (1)$$

em que r_{xx} é o coeficiente de precisão ; x e x', as notas obtidas pela mesma pessoa na primeira e na segunda forma do teste (expressas como desvio das médias do grupo respectivo) ; s e s', os desvios padrão observados na distribuição das notas alcançadas pelo grupo nas duas formas do teste ; e N o número de elementos do grupo. Não cabe aqui discutir os processos de obter o coeficiente de precisão quando não se dispõe de duas formas do mesmo teste, bastando assinalar o processo da correlação interna (split half method) e o da administração sucessiva, com intervalo conveniente, da mesma forma do teste.

A quem conhece teoria estatística, ressalta imediatamente o defeito fundamental dêsse coeficiente como medida da precisão do teste : como se dá com todo coeficiente de correlação, seu valor numérico depende, não somente das qualidades intrínsecas do teste, como também da amplitude de variação da habilidade do grupo em que tiver sido experimentalmente determinado ; em outras palavras, pondo de parte as incertezas devidas à flutuação das amostras, o valor dêsse coeficiente será, por exemplo, muito mais alto quando determinado num grupo de alunos de todas as séries do curso secundário do que quando determinado num grupo de alunos da mesma série, e mais baixo numa classe homogeneizada do que numa classe não homogeneizada.

[3] Para a significação dos termos estatísticos não definidos e de algumas fórmulas não demonstradas, consulte-se qualquer tratado elementar de estatística, sendo de recomendar as obras de Yule e de Lindquist (ref. 12, 6 e 7). O tratado de Yule é um dos mais completos de estatística elementar ; as obras de Lindquist têm a vantagem de visar especialmente as aplicações educacionais. Veja-se também Kurtz (ref. 5).

Outra medida — menos frequentemente usada que o coeficiente de precisão, apesar de tersôbre êste certas vantagens teóricas — é o *índice de precisão*[4] (index of reliability), que é o coeficiente de correlação entre as notas obtidas experimentalmente com um teste e os verdadeiros valores (teóricos) das mesmas notas[5]. Embora não possa ser diretamente determinado, o valor numérico do índice de precisão é dado pela fórmula

$$i_{xx} = \sqrt{r_{xx}} \; ; \qquad (2)$$

é simplesmente igual à raiz quadrada do coeficiente de precisão. Da mesma forma que êsse coeficiente, o índice de precisão apresenta o grave inconveniente (para o fim em vista) de variar conforme a amplitude de variação da habilidade do grupo.

Um terceiro elemento característico da precisão dos testes é o *êrro padrão da nota* (standard error of score). Para que se compreenda sua significação, imagine-se que o mesmo indivíduo foi submetido a um número infinitamente grande de formas comparáveis do mesmo teste; a média das notas resultantes será sua nota verdadeira nesse teste, e a diferença entre a nota verdadeira e a nota realmente obtida em uma forma do teste será o êrro desta nota. Como êsses erros estão sujeitos a uma multiplicidade de causas de variação, sua distribuição será aproximadamente normal e o respectivo desvio padrão será o *êrro padrão da nota*. Sua expressão em função do coeficiente ou do índice de precisão (r_{xx} ou i_{xx}) e do desvio padrão (s_x) da distribuição das notas obtidas pelo grupo em que êsse coeficiente foi determinado será

$$e_{xx} = s_x \sqrt{1 - i_{xx}^2} = s_x \sqrt{1 - r_{xx}} . \qquad (3)$$

O êrro padrão da nota tem a vantagem de ser — salvo flutuações de amostra — quasi independente da amplitude de variação da habilidade do grupo em que tenha sido experimentalmente determinado. Essa independência seria absoluta si existisse perfeita normalidade na correlação entre as notas de duas formas do teste. Isto não acontece porque os itens de um teste não formam uma progressão perfeitamente regular na escala da dificuldade nem apresentam todos o mesmo poder discriminante, mas num teste bem construído estas condições são suficientemente atendidas para que o valor do êrro padrão das notas seja aproximadamente uniforme. Quando isto não se dá, é aliás fácil obter o valor do êrro padrão em função do valor da nota.

Estas qualidades fazem com que o êrro padrão da nota caracterize a precisão do teste com muito maiores vantagens que o coeficiente ou o índice de precisão. Tem entretanto ainda um defeito: é expresso em função das notas do teste como unidade, e como as graduações de dois testes diferentes não são em geral comparáveis, o valor numérico do êrro, sendo expresso em unidades arbitrárias, não servirá para comparar as qualidades intrínsecas de precisão de dois testes diferentes. Êsse inconveniente pode ser facilmente remediado pela graduação dos testes em notas comparáveis, como por exemplo as notas padrão de McCall (McCall T scores). Em última análise, isto equivale a exprimir as notas do teste em função do desvio padrão de um grupo perfeitamente determinado e suficientemente estável, como por exemplo o conjunto de todos os escolares de doze anos de idade (ver ref. 5, pg. 497 sqq.)[6]. Quando expresso em unidades padrão, chamarei o êrro padrão das notas de um teste de *êrro de precisão* do teste, que será representado por ε_{xx} ou simplesmente ε. Seu valor numérico será dado pela formula

$$\varepsilon_{xx} = n \frac{e_{xx}}{s_0} = n \frac{s_x}{s_0} \sqrt{1 - r_{xx}} , \qquad (4)$$

na qual s_0 é o desvio padrão das notas obtidas pelo grupo padrão com a aplicação do teste e n é uma constante numérica. Em realidade, salvo o fator constante n, o êrro de precisão do teste nada mais é que o coeficiente de alienação (coefficient of

[4] A expressão índice de precisão tem o defeito de já servir para designar certo parâmetro da curva normal de probabilidades ($h = 1/\sigma\sqrt{2}$). O inconveniente não é grande, pois em estatística educacional êsse parâmetro não é usado (Veja-se também a nota 1 anterior). Em inglês a confusão não se dá, pois h e i_{xx} são respectivamente designados por *index of precision* e *index of reliability*.

[5] Como valores verdadeiros (teóricos) das notas de um teste ou de outro instrumento de medida, entende-se a média das notas que seriam obtidas com a aplicação de um número infinitamente grande de formas comparáveis do mesmo teste.

[6] Não é aliás necessário — embora sempre conveniente — que as notas do teste sejam convertidas em unidades comparáveis. Bastará que essa transformação seja usada para exprimir o êrro padrão da nota. Para uma discussão do problema das unidades de graduação dos testes, consulte-se Smith (ref. 9, cap. vii).

alienation) entre as notas reais do grupo padrão e as notas verdadeiras do mesmo grupo.

Passemos agora à medida da validade dos testes. Já ficou indicado que essa validade só pode ser determinada em relação a um critério externo, como o julgamento do professor, outro teste ou bateria de testes.

A medida de validade mais empregada é o *coeficiente de validade*, que não é mais que o coeficiente de correlação entre os resultados do teste e os do critério independente. Esta medida apresenta todas as desvantagens do coeficiente de precisão e mais ainda : depende também da precisão do critério,fator completamente estranho ao teste, e inconveniente tanto mais sensível quanto são em geral muito imprecisos os critérios independentes de que dispomos como elemento de comparação. Assim, diante apenas de um valor baixo do coeficiente de validade, nada de positivo se poderá concluir sôbre a validade do teste, pois o valor encontrado tanto poderá provir de defeito intrínseco do teste como da baixa precisão do critério externo.

Para suprimir êste inconveniente, seria de grande utilidade introduzir-se no uso corrente, paralelamente à noção de índice de precisão, a de *índice de validade* de um teste, definido como o coeficiente de correlação entre as notas experimentalmente obtidas no teste e os verdadeiros valores (teóricos) das notas do critério de comparação. Seu valor, em função do coeficiente de validade do teste (r_{xy}) e do coeficiente de precisão do critério (r_{yy}), seria

$$ixx = \frac{rxy}{\sqrt{ryy}} \qquad (5)$$

O índice de validade corresponde ao que Fessard e Piéron têm em vista quando aludem à "validade semi-atenuada" (ref. 3, pg. 221) e é estreitamente ligado ao que Cureton define como coeficiente de validade prática (ref. 1, pg. 29). O valor do índice de precisão é igual à raiz quadrada dêste coeficiente[7].

Pode-se notar que, quando o valor do índice de validade de um teste fôr superior ao do índice de precisão do critério, as notas do teste serão mais válidas que as do próprio critério que serviu para validá-lo.

Embora muito superior ao coeficiente de validade, o índice de validade sofre de um de seus defeitos : variar conforme a amplitude de variação da habilidade do grupo no qual tiver sido experimentalmente determinado. Entretanto, assim como do índice de precisão podem-se derivar medidas que não dependem da amplitude da variação do grupo (o êrro padrão da nota e o êrro de precisão do teste), do índice de validade também é possível derivar medidas independentes dessa amplitude. Suponhamos conhecidos os resultados de uma infinidade de aplicações sucessivas do critério y ao mesmo indivíduo ; a média dos resultados será sua nota verdadeira no critério. Consideremos agora o conjunto de estudantes que obtiveram a mesma nota verdadeira no critério e procuremos as notas obtidas por êsses estudantes com a aplicação do teste em estudo. Estas últimas notas terão teoricamente uma distribuição normal ; sua média determinará a correspondência entre as notas do critério e as notas do teste, enquanto que seu desvio padrão será o *êrro padrão de estimativa* (standard error of estimate) *das notas do teste a partir das notas verdadeiras do critério* y. Êste elemento característico da validade do teste tem a vantagem de ser independente da precisão do critério e da amplitude de variação da habilidade do grupo, o que o torna muito preferível, para o fim que se tem aqui em vista, ao coeficiente ou ao índice de validade. Sua expressão em função do desvio padrão das notas que o grupo obteve no teste (s_x), do coeficiente de validade do teste (r_{xy}) e do coeficiente de precisão do critério (r_{yy}) é

$$exy = rx \sqrt{1 - r_{xy}^2 / ryy} \,. \qquad (6)$$

Entretanto, êste elemento tem ainda o defeito de ser expresso em função da nota do teste como unidade, e sendo esta uma unidade arbitrária, não é possível a comparação direta entre os valores numéricos dêsse êrro relativos a testes diferentes, o que poderá ser remediado, como no caso do êrro de precisão, pelo emprêgo das notas padrão de McCall ou de um sistema análogo de notas comparáveis. Quando expresso em tais unidades será chamado *êrro de validade* do teste (em

[7] Thurstone (ref. 10, pág. 48) dá a fórmula (5) como limite para o qual tende o coeficiente de validade de um teste quando o respectivo critério tende para a perfeita precisão. Não dá porém designação especial a êste limite nem realça seu valor como medida da validade dos testes.

relação a ocritério y) e representado por ε_{xy} ou simplesmente ε_y. Sua expressão será

$$\varepsilon_{xx} = n \frac{exy}{so} = n \frac{sx}{so} \sqrt{1 - r^2_{xy} / ryy}, \quad (7)$$

o que corresponde, salvo o fator constante n, ao coeficiente de alienação entre as notas observadas no teste e as notas verdadeiras no critério, relativas ambas ao grupo padrão.

Embora a definição do êrro de validade tenha feito uso de um número infinitamente grande de aplicações do critério externo, seu valor numérico pode ser obtido em função de elementos todos êles sucetíveis de determinação experimental.

Seria de grande vantagem a adoção do êrro de validade como elemento característico da validade (ou melhor, da invalidade) dos testes, pois reúne as seguintes condições desejáveis :

(a) é independente da precisão do critério externo ;

(b) é independente (ou quasi independente) da amplitude de variação da habilidade do grupo no qual tenha sido experimentalmente determinado ;

(c) é independente da escala de graduação das notas do teste.

Poder-se-á pois afirmar que o êrro de validade representa um semi-invariante dos testes, dependendo unicamente de sua validade, isto é, da exatidão com que medem aquilo que desejamos medir. Convém acentuar que, quando o êrro de validade do teste fôr inferior ao êrro de precisão do critério, as notas do teste serão mais válidas que as do próprio critério que serviu para validá-lo.

Até aquí definí medidas de precisão e de validade dos testes e fiz uma interpretação elementar das grandezas definidas. Foram introduzidas duas noções novas : o índice de validade e o êrro de validade, tendo sido sugerido o emprêgo dêste último para caracterizar a validade de um teste em relação a determinado critério externo. Foram dadas fórmulas sem demonstração ; o leitor poderá aceitá-las sem crítica, limitando-se à leitura do que ficou exposto. Entretanto, os que desejarem conhecer os fundamentos dessas fórmulas, terão vantagem em estudar os desenvolvimentos a seguir, onde encontrarão, além da dedução das fórmulas usadas, questões de interesse para a interpretação de medidas escolares e de uso em certos problemas de construção de testes.

Sejam $x_1, x_2, \ldots x_p$, as notas (expressas como desvio da média do grupo) obtidas pela mesma pessoa em p formas comparáveis do teste x. Como formas comparáveis do mesmo teste entende-se aquí testes que visem a medida da mesma função e que tenham os mesmos desvios padrão e mesmas intercorrelações, isto é, para os quais

$$sx_1 = sx_2 = \ldots = sx_p = sx \quad (8)$$

$$r_{x1\,x2} = r_{x1\,x3} = \ldots = r_{xi\,xj} = \ldots = r_{xx} \quad (9)$$

Sejam semelhantemente $y_1, y_2, \ldots y_q$, as notas da mesma pessoa em q formas comparáveis do teste y. Teremos da mesma maneira

$$sy_1 = sx_2 = \ldots = sy_q = sy \quad (10)$$

$$r_{y1\,y2} = r_{y1\,y3} = \ldots = r_{yi\,yj} = \ldots r_{yy} \quad (11)$$

Suponhamos ainda (o que está compreendido na noção de formas *comparáveis*) que a correlação entre qualquer forma do teste x e qualquer forma do teste y seja sempre a mesma, isto é,

$$r_{x1\,y1} = r_{x1\,y2} = \ldots = r_{xi\,yj} = r_{xy} \quad (12)$$

Procuremos o coeficiente de correlação que existiria entre um teste cuja nota fôsse a soma (ou a média) das notas de todas as p formas do teste x, e um outro teste cuja nota fôsse a soma (ou a média) das notas de todas as q formas do teste y. Por definição

$$r(x_1 + x_2 + \ldots + x_p)(y_1 + y_2 + \ldots + y_q) = \frac{\Sigma(x_1 + x_2 + \ldots + x_p)(y_1 + y_2 + \ldots + y_q)}{N \cdot s(x_1 + x_2 + \ldots + x_p) \cdot s(y_1 + y_2 + \ldots + y_q)} \quad (13)$$

Ainda por definição

$$s^2(x_1 + x_2 + \ldots + x_p) = \frac{\Sigma(x_1 + x_2 + \ldots + x_p)^2}{N} = \frac{\Sigma x_1^2 + \Sigma x_2^2 + \ldots + \Sigma x_1 x_2 + \Sigma x_1 x_3 + \ldots}{N},$$

expressão na qual haverá p elementos da forma Σx_i^2 e $p(p-1)$ elementos da forma $\Sigma x_i x_j$. Tendo em vista as relações (8) e (9), tem-se

$$s^2(x_1 + x_2 + \ldots + x_p) = p\, s_x^2 + p(p-1)\, r_{xx}\, s_x^2$$

$$s(x_1 + x_2 + \ldots x_i) = s_x \sqrt{p + p(p-1)\, r_{xx}} \quad (14)$$

Semelhantemente, teremos :

$$s(y_1 + y_2 + \ldots y_p) = s_y \sqrt{q + q(q-1)\, r_{yy}} \quad (15)$$

Partindo da expressão geral do coeficiente de correlação, é facil de vêr, em virtude das relações (12), que

$$= \Sigma\Sigma x_i y_j = \Sigma N r_{xy} s_x q_y = pq N r_{xy} s_x s_y \quad (16)$$

Substituindo na equação (13) os valores dados pelas equações (14), (15) e (16):

$$r(x_1 + x_2 + \ldots + x_p)(y_1 + y_2 + \ldots + y_q) =$$

$$= \frac{pq N r_{xy} s_x s_y}{N s_x \sqrt{p + p(p-1) r_{xx}} \cdot s_y \sqrt{q + q(q-1) r_{yy}}} =$$

$$= \frac{r_{xy}}{\sqrt{\frac{1}{p} + \frac{p-1}{p} r_{xx}} \sqrt{\frac{1}{q} + \frac{q-1}{q} r_{yy}}} \quad (17)$$

Esta equação fundamental se transforma, conforme valores particulares dos elementos que envolve, em fórmulas de grande importância.

Índice de precisão. Desejamos conhecer o coeficiente de correlação entre as notas de um teste, obtidas experimentalmente, e os verdadeiros valores dessas notas (índice de precisão do teste). Na equação (17) façamos com que x e y sejam formas comparáveis do mesmo teste (e não de testes diferentes, como no caso geral); façamos ainda $p = 1$ e $q = \infty$. Teremos: $r_{yx} = r_{yy} = r_{xy}$; $1/p = 1$; $1/q = 0$; $(q-1)/q = 1$; donde:

$$i_{xx} = r_{x_1 x\infty} = \sqrt{r_{xx}} \quad (18)$$

A equação (18) coincide com a equação (2), dada anteriormente sem demonstração.

Índice de validade. Conhecidos os coeficientes de precisão de um teste e do respectivo critério (r_{xx} e r_{yy}) e o coeficiente de validade do teste em relação ao mesmo critério (r_{xy}), deseja-se conhecer o coeficiente de correlação entre as notas do teste e os verdadeiros valores (teóricos) das notas do critério (índice de validade do teste). Façamos na equação (17) $p = 1$ e $q = \infty$. Teremos:

$$i_{xy} = \frac{r_{xy}}{\sqrt{r_{yy}}} \quad (19)$$

A equação (19) é idêntica à equação (5) que fica assim justificada.

Correção da atenuação (correction for attenuation). Conhecemos a correlação entre as notas, sujeitas a erros de medida, de dois testes que medem funções psicológicas diferentes e desejamos saber a correlação intrínseca entre as funções medidas pelos testes, isto é, a correlação que existiria entre as notas verdadeiras dos dois testes. Fazendo na equação (17) $p = \infty$ e $q = \infty$ teremos:

$$r_{x\infty y\infty} = \frac{r_{xy}}{\sqrt{r_{xx} r_{xy}}} \quad (20)$$

Os erros casuais nas medidas das duas funções fazem com que o valor observado do coeficiente de correlação seja inferior a o que se observaria caso as medidas fôssem isentas de êrro: é o que se chama atenuação dêsse coeficiente. De acôrdo com a fórmula (20), para corrigir essa atenuação, bastará dividir o coeficiente observado pela média geométrica dos coeficientes de precisão das duas medidas (ou, o que dá no mesmo, pelo produto de seus índices de precisão).

Êrro padrão da nota e êrro de precisão do teste. Suponhamos conhecidas as notas realmente obtidas por um grupo em determinado teste, assim como as notas verdadeiras que seriam obtidas pelo mesmo grupo no mesmo teste. Suponhamos traçado o diagrama de correlação entre essas duas variáveis, ficando as notas verdadeiras em abscissas (fig. 1). Si entre as variáveis existir correlação normal[8], para determinado valor da nota verdadeira, terão as notas observadas uma distribuição nor-

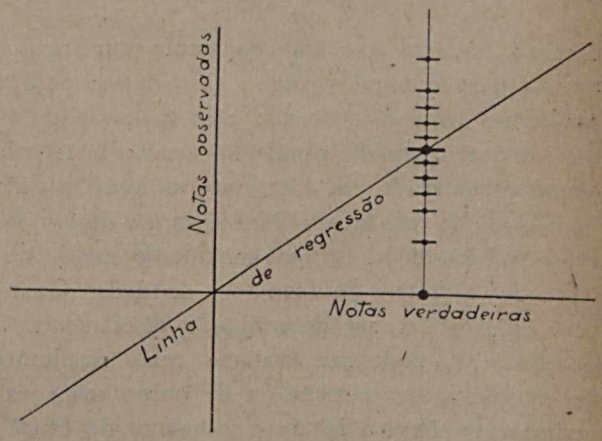

Fig. 1 — Diagrama de correlação entre as notas observadas no teste e as notas verdadeiras no teste (ou no critério), mostrando a dispersão das notas observadas que correspondem a determinado valor da nota verdadeira.

[8] A rigor não será necessário existir correlação normal. Para o que se segue, bastará que haja homocedasticidade e retilineidade da regressão das notas observadas em relação às notas verdadeiras, condições necessárias mas não suficientes para a normalidade da correlação.

mal cuja média estará na linha de regressão das notas observadas em relação às notas verdadeiras; além disso, o desvio padrão dessas notas será o mesmo, qualquer que seja a abscissa considerada. Em outras palavras, o êrro da nota (diferença entre a nota observada e a nota verdadeira) terá uma distribuição normal com desvio padrão constante. Êste desvio padrão (que numa tábua de correlação corresponde ao desvio padrão de uma coluna) será dado, de modo geral, pela fórmula

$$s = s_x \sqrt{1 - r^2}, \qquad (21)$$

na qual s_x é o desvio padrão das notas observadas em todo o grupo e r é o coeficiente de correlação entre as duas variáveis. Ora, no nosso caso, o coeficiente de correlação entre as notas observadas e as notas verdadeiras não é mais que o índice de precisão do teste. Teremos pois :

$$r = i_{xx} = \sqrt{r_{xx}} \qquad \text{e}$$

$$e_{xx} = s_x \sqrt{1 - i_{xx}^2} = s_x \sqrt{1 - r_{xx}} \qquad (22)$$

o que é a reprodução da fórmula (3). Si quizermos exprimir êste êrro, não em função da escala arbitrária das notas do teste, mas como uma função linear do desvio padrão das notas de um grupo padrão, teremos o êrro de precisão do teste ;

$$\varepsilon_{xx} = n \frac{e_{xx}}{s_o} = n \frac{s_x}{s_o} \sqrt{1 - r_{xx}}, \qquad (23)$$

fórmula na qual n é uma constante numérica, a mesma para qualquer grupo ; s_0, o desvio padrão obtido no teste considerado pelo grupo padrão ; e s_x, o desvio padrão obtido no mesmo teste pelo grupo considerado. Si adotarmos as notas padrão de McCall, n será igual a 10 e s_0 será o desvio padrão verificado no grupo constituído pelo conjunto dos escolares de doze anos de idade. Quando s_0 não puder ser determinado diretamente, a relação s_x/s_0 pode ser avaliada como resultante de considerações teóricas ou de observações experimentais sôbre a lei de crescimento da função medida.

Êrro padrão de estimativa das notas do teste a partir das notas verdadeiras do critério e êrro de validade do teste. Por meio de considerações análogas às que fizemos em relação ao êrro padrão das notas do teste, e imaginando-se, no diagrama da figura 1, que as abscissas são as notas verdadeiras do critério externo e as ordenadas são as notas observadas no teste, teremos a mesma fórmula geral (12), mas agora r será o coeficiente de correlação entre as notas do teste e as notas verdadeiras do critério, isto é, será o índice de validade do teste. O valor dêsse índice é dado pela fórmula (19) e êsse valor, introduzido na fórmula (21), dará, para o êrro de estimativa das notas do teste a partir das notas verdadeiras do critério y :

$$e_{xy} = s_x \sqrt{1 - i_{xy}^2} = s_x \sqrt{1 - r_{xy}^2 / r_{yy}} \qquad (24)$$

Exprimindo êste êrro em unidades padrão, teremos o êrro de validade do teste :

$$\varepsilon_{xy} = n \frac{e_{xy}}{s_o} = n \frac{s_x}{s_o} \sqrt{1 - i_{yx}^2} = n \frac{s_x}{s_o} \sqrt{1 - r_{xy}^2 / r_{yy}} \qquad (25)$$

Procuremos agora verificar como variam r_{xx}, r_{xy}, i_{xx}, i_{xy}, e_{xx}, e_{xy}, ε_{xx} e ε_{xy} quando varia o comprimento do teste a que se referem. Quando imaginamos que o comprimento do teste aumenta ou diminue, devemos supor que o faz pelo acréscimo ou supressão de itens equivalentes aos primitivos (como conteúdo, poder discriminante e validade). As conclusões a que chegarmos só terão valor quando preenchida, pelo menos aproximadamente, esta condição.

Variação do coeficiente de precisão. Seja um teste de coeficiente de precisão r_{xx} ; para obtermos o coeficiente de precisão R_{xx} que terá êsse teste si seu comprimento fôr aumentado, bastará que consideremos, na fórmula (17), x e y formas comparáveis do mesmo teste e que façamos $p = q = k$ igual ao fator pelo qual o comprimento do teste primitivo deve ser multiplicado para atingir o comprimento do novo teste. O coeficiente procurado será :

$$R_{xx} = \frac{r_{xx}}{\frac{1}{k} + \frac{k-1}{k} r_{xx}} = \frac{k\, r_{xx}}{1 + (k-1)\, r_{xx}}. \qquad (26)$$

A equação (26) é conhecida sob o nome de fórmula de previsão de Spearman-Brown (Spearman-Brown prophecy formula) e pode ser resolvida para k, obtendo-se assim

$$k = \frac{R_{xx}\,(1 - r_{xx})}{r_{xx}\,(1 - R_{xx})}. \qquad (27)$$

Para o valor $k = 2$, a fórmula (26) dá

$$R_{xx} = \frac{2\, r_{xx}}{1 + r_{xx}} \qquad (28)$$

e é empregada, quando não se dispõe de duas formas do mesmo teste, para obter-se o coeficiente de precisão pelo processo da correlação interna[9].

Variação do índice de precisão. As variações do índice de precisão em função do comprimento do teste podem ser obtidas diretamente da fórmula (17); a fórmula (2) combinada com a fórmula (26) dá porém imediatamente :

$$I_{xx} = \sqrt{R_{xx}} = \sqrt{\frac{k\, r_{xx}}{1 + (k-1)\, r_{xx}}} \qquad (29)$$

Variação do coeficiente de validade. Sejam, na fórmula (17), x o teste e y o critério externo. Para conhecermos o coeficiente de correlação que existiria entre as notas do teste, com o comprimento multiplicado pelo fator k, e as notas do critério, bastará fazermos $p = k$ e $q = 1$. Teremos :

$$R_{xy} = \frac{r_{xy}}{\sqrt{\frac{1}{k} + \frac{k-1}{k}\, r_{xx}}} = \frac{r_{xy}\sqrt{k}}{\sqrt{1 + (k-1)\, r_{xx}}} =$$

$$= \frac{r_{xy}}{\sqrt{r_{xx}}}\, I_{xx} \qquad (30)$$

donde se deduz que as variações do coeficiente de validade são proporcionais às do índice de precisão. A equação (30) pode ser resolvida para k, obtendo-se :

$$k = \frac{1 - r_{xx}}{\frac{r^2_{xy}}{R^2_{xy}} - r_{xx}} \qquad (31)$$

Na equação (30), si fizermos k tender para ∞ (isto é, si aumentarmos indefinidamente o comprimento do teste), R_{xy} tenderá para um limite, sempre inferior à unidade ;

$$\lim R_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}}} \qquad (32)$$

Variação do índice de validade. A expressão do índice de validade em função do comprimento aumentado do teste pode ser obtida fazendo-se na equação (17) $p = k$ e $q = \infty$ ou, o que dá no mesmo, fazendo-se na equação (19) $r_{xy} = R_{xy}$ e introduzindo-se o valor de R_{xy} dado pela equação (30). Em qualquer caso, tem-se :

$$I_{xy} = \frac{R_{xy}}{\sqrt{r_{xy}}} = \frac{r_{xy}\,.\,\sqrt{k}}{\sqrt{1 + (k-1)\, r_{xx}}\,.\,\sqrt{r_{yy}}} =$$

$$= \frac{r_{xy}}{\sqrt{r_{xx}\, r_{yy}}}\, I_{xx} \qquad (33)$$

donde se conclue que as variações do índice de validade são proporcionais às variações do coeficiente de validade ou às do índice de precisão. Quando k tende para ∞, I_{xy} tende para o valor

$$\lim I_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}\, r_{xy}}} \qquad (34)$$

que, como era de prever, não é mais que o coeficiente de validade corrigido da atenuação. Êste limite é idêntico ao coeficiente de validade fundamental definido por Cureton (ref. 1, pg. 28).

Resolvendo-se a equação (33) em relação a k, obtem-se :

$$k = \frac{I^2_{xy}\, r_{yy}\, (1 - r_{xx})}{r^2_{xy} - I^2_{xy}\, r_{xx}\, r_{yy}} \,. \qquad (35)$$

As variações dessas quatro estatísticas estão representadas em função de k no gráfico da figura 2, para cujo traçado foram tomados os seguintes valores numéricos para o comprimento unitário do teste ($k = 1$): $r_{xx} = 0{,}50$; $r_{xy} = 0.32$; $r_{yy} = 0{,}30$, condições que não se afastam muito das que seriam encontradas em um teste escolar não padronizado de pequena extensão.

---

[9] Para isso, dividem-se os itens do teste de modo a formar dois sub-testes comparáveis ; procura-se o coeficiente de precisão de meio teste (correlação entre as duas metades) e pela fórmula (28) chega-se ao coeficiente de precisão do teste inteiro. O valor assim obtido é em geral mais elevado do que o obtido pela correlação entre duas formas do mesmo teste ou pelo processo da administração sucessiva, com intervalo, do mesmo teste. Compreende-se aliás o motivo disto, pois a imperfeição do coeficiente de correlação depende dos erros de medida inerentes ao instrumento e da variação que sofre o objeto medido (neste caso o estudante) entre as duas medições : o processo da correlação interna elimina sensivelmente a segunda causa de variação. Êste processo evidencia pois, melhor que os demais, as qualidades intrínsecas do teste. Dá porém resultados exagerados quando o que se tem em vista é prever, a partir dos resultados obtidos com uma forma do teste, os que seriam obtidos com a aplicação posterior de outra ou da mesma forma do teste, pois neste caso entram em operação as variações individuais não levadas em conta no coeficiente obtido pelo processo da correlação interna.

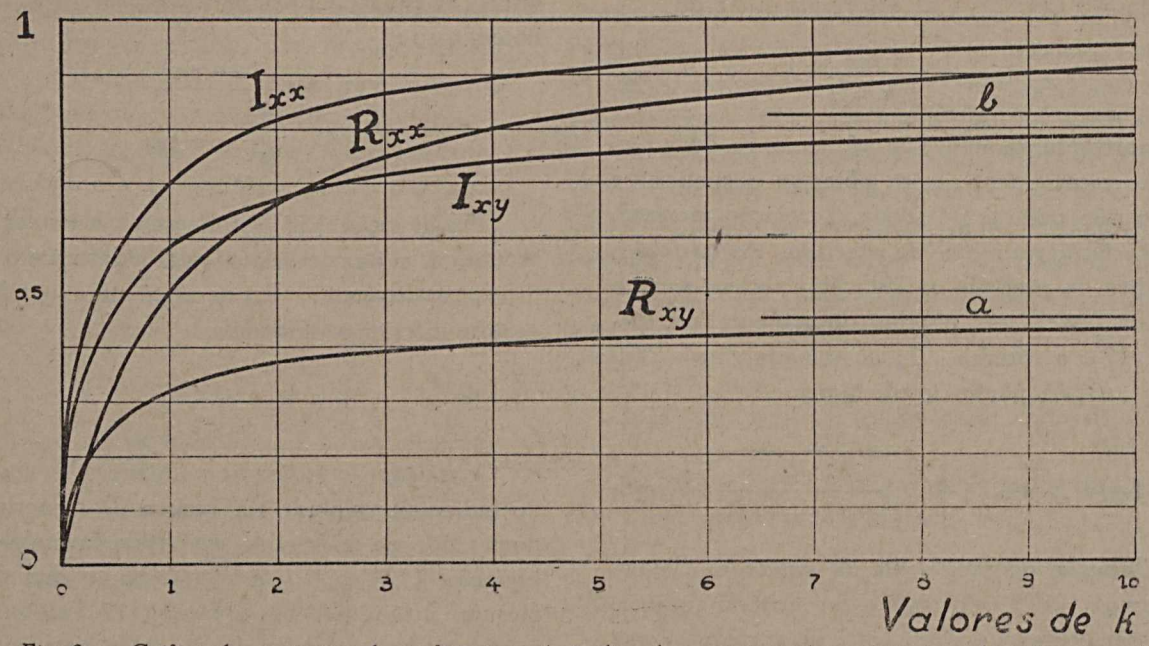

Fig. 2 — Gráfico das variações do coeficiente e do índice de precisão e do coeficiente e do índice de validade, em função do com primento do teste.

As curvas representativas de R_{xx} e de I_{xx} têm para assíntota a reta $y = 1$. R_{xy} e I_{xy} tendem respectivamente para os limites 0,452 e 0,825, representados pelas retas *a* e *b*, assíntotas das curvas respectivas. As posições dessas retas podem variar, mas quaisquer que sejam os valores numéricos (possíveis) dos dados iniciais, observam-se sempre as seguintes propriedades das curvas representadas : (a) as ordenadas de I_{xx} são sempre superiores às ordenadas correspondentes de R_{xx}, pois o índice de precisão é a raiz quadrada do coeficiente de precisão e êste, sendo um coeficiente de correlação, é sempre inferior à unidade ; (b) as ordenadas de R_{xy} são proporcionais às ordenadas correspondentes de I_{xx} e sempre menores que estas; (c) as ordenadas de I_{xy} são também proporcionais às de I_{xx} e seus valores são intermediários entre os das ordenadas de I_{xx} e R_{xy} ; (d) a partir de um valor suficientemente grande de *k*, as ordenadas de R_{xx} se conservam superiores às de R_{xy} e de I_{xy} ; (e) para valores suficientemente pequenos de *k*, R_{xx} será sempre menor que R_{xy} ou I_{xy}.

Esta última propriedade conduz ao seguinte paradoxo : Por definição, a precisão de um teste é a exatidão com que êle mede *aquilo que realmente mede*, enquanto que sua validade é a precisão com que mede *aquilo que deveria medir* ; a partir dessas definições, é óbvio que a *validade* de um teste não pode ser superior a sua *precisão* : donde a afirmação que se encontra, mesmo em autores de grande reputação, de que o *coeficiente de validade* de um teste não pode ser superior ao seu *coeficiente de precisão* (cf. Thurstone, ref. 10, pg. 109), *o que entretanto não é verdade*. Na realidade, nos testes padronizados, o coeficiente de precisão é sempre superior ao coeficiente de validade, isto porque, na construção de um teste, é relativamente facil aumentar seu coeficiente de precisão, o mesmo não se dando com seu coeficiente de validade ; em outras palavras, os testes em uso correspondem, no gráfico dado, à região em que a curva R_{xx} é superior à curva R_{xy} ; mas, si houvesse êste propósito, seria facílimo obter-se um teste suficientemente impreciso para que seu coeficiente de precisão fôsse inferior ao seu coeficiente de validade. Êste paradoxo resulta apenas do fato de não serem os coeficientes de precisão e de validade medidas convenientes da precisão e da validade dos testes ; considerando os índices respectivos, vemos que o de precisão é sempre superior ao de validade e inversamente que o êrro de precisão é sempre inferior ao êrro de validade.

Variação do êrro padrão da nota. A fórmula (14) nos dá diretamente a variação do desvio padrão da distribuição das notas do teste em função de seu alongamento. Usando, como temos feito, maiúsculas para caracterizar os valores das esta-

tísticas quando o comprimento do teste se torna k vezes maior, teremos :

$$S_x = s_x \sqrt{k + k\,(k-1)\,r_{xx}} \qquad (36)$$

Combinando as equações (3), (26) e (36), obtemos as variações do êrro padrão da nota :

$$E_{xx} = S_x \sqrt{1 - R_{xx}} =$$

$$= s_x \sqrt{k + k(k-1) r_{xx}} \sqrt{1 - \frac{1 + (k-1)\, r_{xx}}{k\, r_{xx}}} =$$

$$= s_x \sqrt{1 - r_{xx}} \cdot \sqrt{k} = e_{xx} \sqrt{k} \qquad (37)$$

Variação do êrro padrão de estimativa da nota do teste a partir da nota verdadeira do critério.

As equações (6), (30) e (36) conduzem imediatamente a :

$$E_{xy} = S_x \sqrt{1 - \frac{R_{xy}^2}{r_{yy}}} =$$

$$= s_x \sqrt{k + k\,(k-1)\, r_{xx} - \frac{k^2\, r_{xy}^2}{r_{yy}}} \qquad (38)$$

As variações de S_x, E_{xx} e E_{xy} em função de k estão representadas no gráfico da fig. 3, no qual fiz arbitrariamente $s_x = 1$ e empreguei para r_{xx}, r_{xy} e r_{yy} os mesmos valores do gráfico anterior.

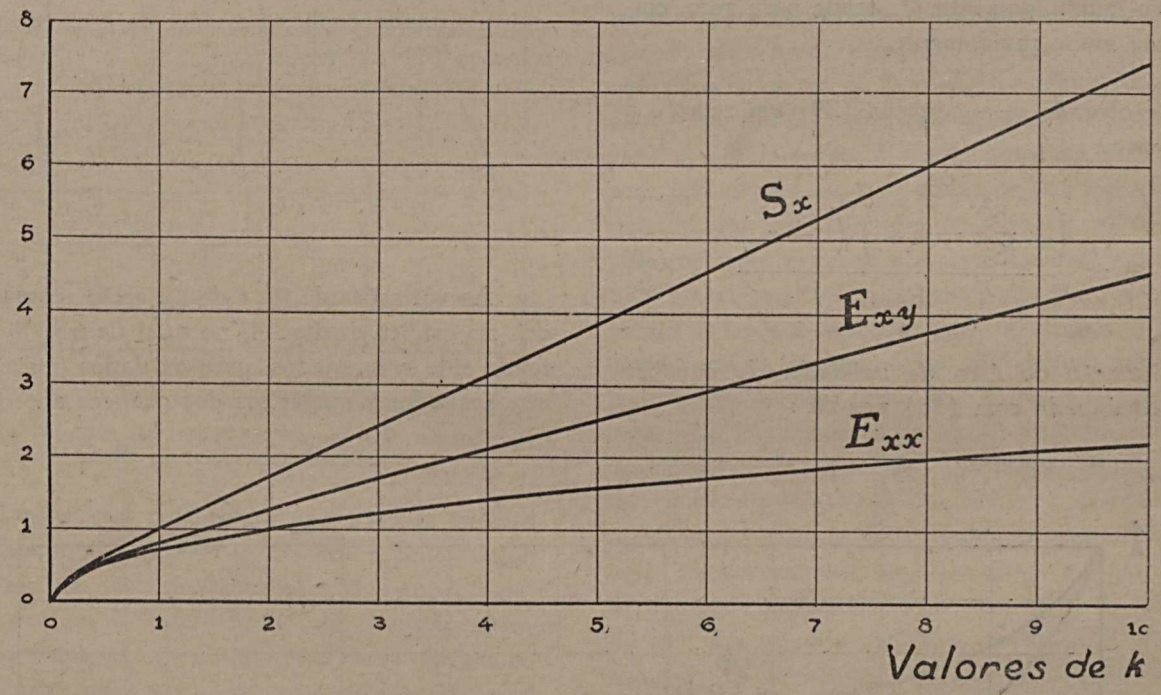

Fig. 3 — Gráfico das variações do desvio padrão do grupo, do êrro padrão da nota e do êrro de estimativa da nota, em função do comprimento do teste.

A curva E_{xx} é uma parábola e as curvas S_x e E_{xy} são ramos de hipérboles que, no caso representado, podem ser, sem êrro sensível, substituídos por segmentos retilíneos a partir de $k = 2$. As ordenadas de S_x e E_{xy} não são porém proporcionais, porque os prolongamentos dos respectivos trechos retilíneos não cortam o eixo horizontal no mesmo ponto.

À primeira vista poderá parecer que, tendo o êrro padrão da nota e o êrro padrão de estimativa aumentado com o alongamento do teste, houve desvantagem nesse alongamento. Isto mostra simplesmente que o êrro padrão da nota e o êrro de estimativa não são elementos convenientes para caracterizar a precisão e a validade dos testes. Realmente, o que importa não é o êrro absoluto e sim o êrro relativo, isto é, a relação entre o êrro da nota e o desvio padrão do grupo, o que equivale a dizer que, para comparar a precisão e a validade de testes diferentes deve-se lançar mão do êrro de precisão e do êrro de validade anteriormente definidos.

Variação do êrro de precisão. Pode ser facilmente obtida a partir da equação (23):

$$\varepsilon_{xx} = n \frac{S_x}{S_o} \sqrt{1 - R_{xx}} =$$

$$= n \frac{S_x}{S_o} \sqrt{\frac{1 - r_{xx}}{1 + (k-1)\, r_{xx}}} =$$

$$= \frac{\varepsilon_{xx}}{\sqrt{1 + (k-1)\, r_{xx}}} \qquad (39)$$

Deve-se notar que, para um grupo determinado, $n\, S_x/S_0 = n\, s_x/s_0$ é uma constante. Quando k cresce indefinidamente, ε_{xx} tende para zero, embora não muito rapidamente.

Resolvendo-se a equação (39) em relação a k, obtem-se:

$$k = \frac{\left(n^2 \frac{S_x^2}{S_o^2} - \varepsilon_{xx}^2\right)(1 - r_{xx})}{\varepsilon_{xx}^2\, r_{xx}} \qquad (40)$$

Variação do êrro de validade. A equação (25) combinada com (30) nos dá:

$$\varepsilon_{xy} = n \frac{S_x}{S_o} \sqrt{1 - \frac{R_{xy}^2}{r_{yy}}} =$$

$$= n \frac{S_x}{S_o} \sqrt{1 - \frac{k\, r_{xy}^2}{r_{yy}[1 + (k-1)\, r_{xx}]}} =$$

$$= n \frac{S_x}{S_o} \sqrt{\frac{k\,(r_{xx}\, r_{yy} - r_{xy}^2) + r_{yy}\,(1 - r_{xx})}{k\, r_{xx}\, r_{yy} + r_{yy}\,(1 - r_{xx})}} \qquad (41)$$

Quando k cresce indefinidamente, ε_{xy} tende para o limite mínimo:

$$\lim \varepsilon_{xy} = n \frac{S_x}{S_o} \sqrt{1 - \frac{r_{xy}^2}{r_{xx}\, r_{yy}}} \qquad (42)$$

Tornando explícito o valor de k em uma das relações (41), obtem-se:

$$k = \frac{r_{yy}\,(1 - r_{xx})\left(1 - \frac{S_o^2}{n^2 S_x^2}\, \varepsilon_{xy}^2\right)}{r_{xy}^2 - r_{xx}\, r_{yy}\left(1 - \frac{S_o^2}{n^2 S_x^2}\, \varepsilon_{xy}^2\right)} \qquad (43)$$

As variações de ε_{xx} e de ε_{xy} estão representadas no gráfico da fig. 4, no qual fiz $nS_x/S_0 = 1$, tendo sido empregados, para os dados iniciais, os mesmos valores numéricos dos gráficos anteriores.

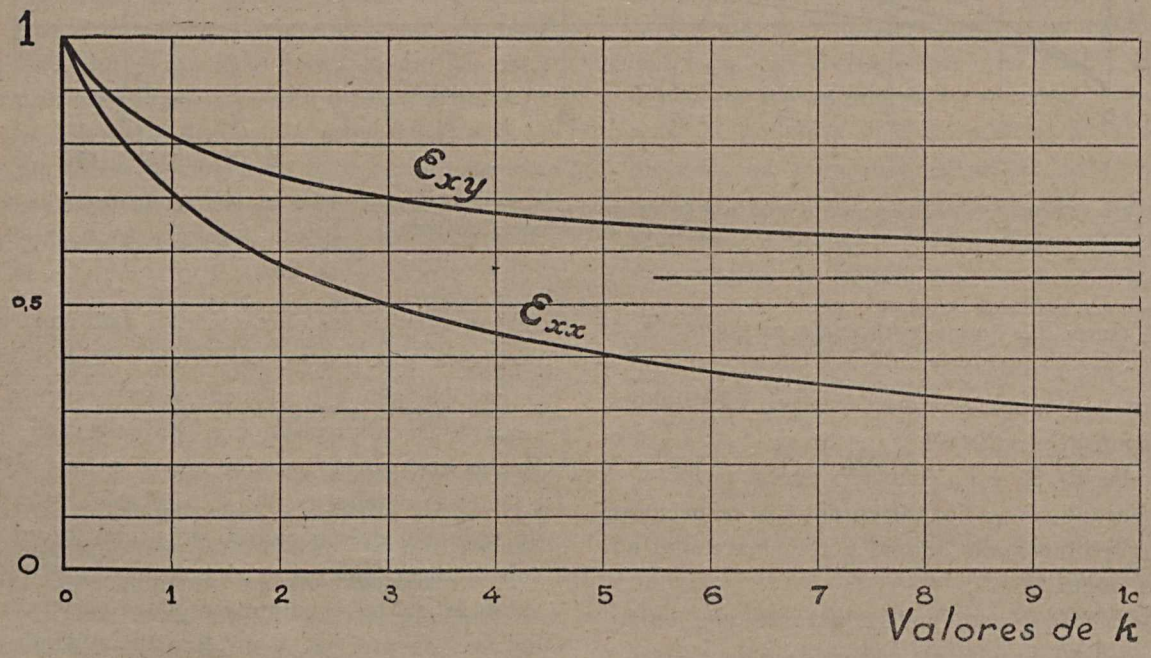

Fig. 4 — Gráfico das variações do êrro de precisão e do êrro de validade, em função do comprimento do teste.

Enquanto que a curva ε_{xx} tende assintoticamente para o eixo dos k, a curva ε_{xy} tem para assíntota uma reta horizontal cuja ordenada é dada pela equação (42) e cujo valor numérico, no gráfico traçado, é 0,563. Com os valores numéricos do exemplo, tornando-se o comprimento do teste inicial dez vezes maior, seu êrro de precisão fica reduzido a 43 % do valor inicial (para $k = 1$) e seu êrro de validade a 76 % do valor inicial.

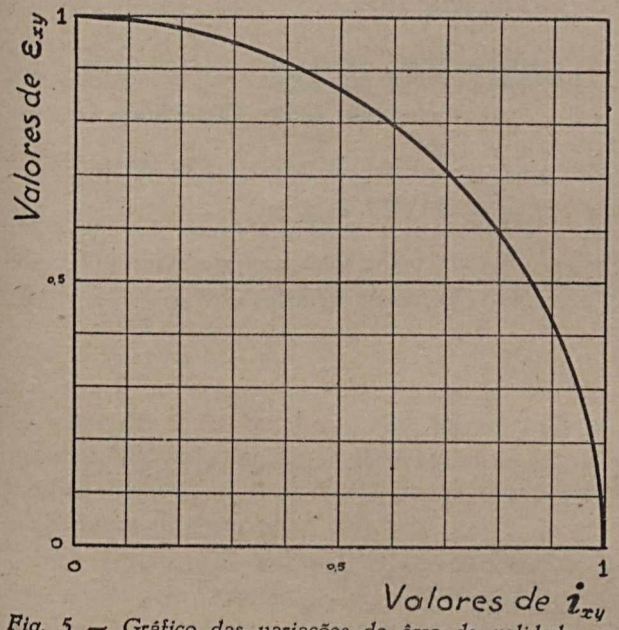

Fig. 5 — Gráfico das variações do êrro de validade em função do índice de validade.

Si quizermos representar as variações do êrro de validade, ε_{xy}, em função do índice de validade, i_{xy}, tendo em vista as relações (25), encontraremos o gráfico da figura 5, em que a curva representada é um arco de círculo. (O valor de $r\, s_x/s_0$ foi suposto igual a 1). Si convertermos em z os valores numéricos do índice de validade por meio da transformação de Fisher (cf. ref. 4, pg. 200), isto é, si fizermos

$$z_{ry} = \tfrac{1}{2}[\log_e (1 + i_{xy}) - \log_e (1 - i_{xy})]$$

e exprimirmos ε_{xy} em função de z_{xy}, teremos o gráfico da figura 6, no qual a curva representativa de ε_{xy} tem o aspecto geral da curva normal de Gauss, embora se aproxime muito mais lentamente de sua assíntota que é o eixo horizontal. No gráfico da fig. 6 está tambem representada a variação correspondente de i_{xy}.

O gráfico da figura 4 deve ser considerado de importância fundamental para o construtor de

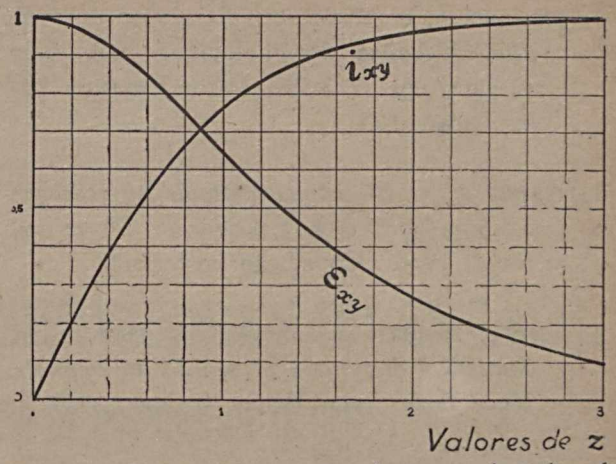

Fig. 6 — Gráfico das variações do êrro e do índice de validade em função de z_{xy}

testes. *Uma vez obtido um critério externo satisfatório, toda a sua atenção deve ser voltada para reduzir o valor de ε_{xy}, pois disso depende a qualidade essencial do teste: sua validade.* Depois dos ensaios preliminares, a forma geral da curva lhe permitirá prever até que ponto poderá esperar um determinado valor de ε_{xy} pelo alongamento de um pequeno teste experimental, e si êsse alongamento não fôr compatível com o tempo que razoavelmente se poderá conceder para aplicação do teste, êle saberá de antemão que não poderá por êste meio obter o fim desejado. Terá portanto que recorrer a modificações no gênero das questões ou no processo de marcação das notas, sendo para isto indicado fazer uso das técnicas empregadas por Tyler (ref. 9). Êste assunto escapa à finalidade do presente estudo, mas como a construção de um teste padronizado envolve dispêndio considerável de tempo e dinheiro, nunca será demais salientar os elementos com que o técnico deverá jogar para obter o mais economicamente possível um resultado de ante-mão visado.

REFERÊNCIAS :

1. CURETON, EDWARD E., *Errors of measurement and correlation.* Archives of Psychology No. 125, New York; Columbia University, 1931.

2. FESSARD, A., *La précision et la cohérence des résultats dans les examens par tests,* L' Année Psychologique, vol xxviii (1927), pgs. 205-235, Paris: Librairie Félix Alcan, 1928.

3. FESSARD, A. e PIÉRON, H., *La notion de validité*, L'Année Psychologique, vol. xxxi (1930), pgs. 217-228, Paris: Librairie Félix Alcan, 1931.

4. FISHER, R. A., *Statistical methods for research workers* (6.ª edição), Londres: Oliver and Boyd, 1936 (Há edição posterior).

5. KURTZ, ALBERT K. e EDGERTON, HAROLD A., *Statistical dictionary of terms and symbols*, New York: John Wiley & Sons, 1939.

6. LINDQUIST, E. F., *A First course in statistics*, Boston: Houghton Mifflin Co., 1938.

7. LINDQUIST, E. F., *Statistical analysis in educational research*, Boston: Houghton Mifflin Co., 1940.

8. MCCALL, WILLIAM A., *Measurement*, New York: The Macmillan Co., 1939.

9. SMITH, B. OTHANEL, *Logical aspects of educational measurement*, New York: Columbia Uniersity Press, 1938.

10. THURSTONE, L. L., *The reliability and validity of tests*, Ann Arbor, Mich.: Edwards Brothers, 1939.

11. TYLER, RALPH W., *Constructing achievement tests*, Columbus, Ohio: Ohio State University, 1934.

12. YULE, G. UNDY e KENDALL, M. G., *An introduction to the theory of statistics* (11.ª edição), Londres: Charles Griffin & Co., 1937.