# Validade e fidedignidade nos testes coletivos de inteligência*

Murilo Braga

Palavras-Chave: teste de inteligência; validade; fidedignidade.


Ilustração: Fabiano Yoshiyuki Higashiyama

Os testes classificam-se segundo o objeto da prova e a modalidade de aplicação.
Os primeiros resultados de testes coletivos de inteligência foram divulgados em 1913 e desde então têm sido fortemente atacados; todavia, eles desempenham um papel importante na administração e

* Esta monografia foi escrita em 1938 e apresentada ao concurso para a carreira de Técnico de Educação do Ministério da Educação. O autor não fez qualquer alteração, embora o trabalho necessite de uma atualização, em virtude dos progressos nesse campo, especialmente com os resultados que os americanos conseguiram durante a guerra.

N.E.: Publicada originalmente na *RBEP*, v. 12, n. 34, set./dez. 1948. O texto foi atualizado de acordo com as normas bibliográficas da Associação Brasileira de Normas Técnicas (ABNT) e normas de redação atuais; sua estrutura formal foi adaptada ao projeto gráfico da revista, sem comprometimento do conteúdo original.

organização escolares.
A coerência de um instrumento de medida é verificada pelo grau de concordância existente entre os índices internos (fidedignidade) e externos (validade). Emprega-se o coeficiente de correlação para verificar a fidedignidade prática do instrumento de teste, isto é, para verificar a sua coerência em sucessivas aplicações.
Os processos de verificação estatística permitem exprimir por índices numéricos o grau de confiança que é possível atribuir ao teste.

## Introdução

### O teste, sua conceituação

Teste, do inglês *test* (exame, verificação, experiência, ensaio, prova) e palavra hoje de uso universal, na técnica psicológica, significa prova em condições objetivas. No próprio inglês há uma acepção mais genérica, que é a que foi antes apontada; e uma acepção mais restrita de padrão, bitola. Do ponto de vista da lógica, é qualquer critério ou processo empregado para determinar-se a verdade ou a falsidade de uma hipótese, tanto pela evidência empírica como pelo raciocínio. Como a objetividade, que é o caráter essencial do teste, em nosso entender, leva à fixação de normas de comparação, a palavra tem sido empregada por alguns autores como *prova já aferida ou padronizada*. Incluem, assim, uma noção que lhe não é própria ou substancial, mas já conseqüência de aplicação. Convém esta distinção, desde o início, porque o assunto escolhido para esta monografia versa, justamente, sobre as *qualidades essenciais de um teste coletivo de inteligência*, para o efeito de sua padronização. Empregar-se-á bem a palavra para significar prova, experiência, ensaio. Poderá ser ainda empregada no sentido de material com que se faz a prova, no sentido de reativo, pois, de fato, com esse material se procura provocar uma reação, uma modificação de comportamento de que se deseja colher a amostra.[1] Desacompanhada, porém, de qualquer qualificativo, não deverá levar a pensar desde logo em instrumento de prova, graduado e aferido. O teste é simplesmente a prova feita em condições de objetividade, de tal modo que qualquer pessoa habilitada que a empregue, nas condições estabelecidas para seu uso adequado, colha sempre os mesmos resultados ou resultados comparáveis e possa interpretá-los, à vista dos mesmos elementos, também do mesmo modo. Em outras palavras: as provas psicológicas podem sofrer a influência da *equação pessoal* do experimentador, em grau mais ou menos elevado, na sua interpretação; como podem também ser aplicadas de modo a provocar sugestão positiva ou negativa, em relação aos resultados que colher; podem, por outro lado, provocar estados emotivos diversos, nas diversas pessoas sobre que forem aplicadas.[2] A consideração desses elementos perturbadores, na colheita da amostra, levou os pesquisadores a fixarem condições próprias para cada exame, a fim de atenuar a sua influência. E como esse trabalho foi realizado *especialmente por autores norte-americanos e ingleses*, a palavra *teste* se universalizou com o sentido não só de prova, mas de prova em condições objetivas.[3]

### Classificação do teste segundo o objeto da prova

A prova em condições objetivas pode ser aplicada na colheita de qualquer material de estudo. É lícito, pois, o emprego de expressões tais como teste físico, teste químico, teste biológico, teste escolar, teste psicológico. A classificação do teste, por seu objeto, é assim variada. Poderão ser eles distribuídos por gêneros e espécies sem conta. No entanto, a palavra tem sido empregada, especialmente em nosso país, para designar prova escolar, exame de conhecimentos ou exame de capacidades de um indivíduo. Neste último sentido, vemos que o termo pode compreender um grande número de coisas. De fato, tal seja o propósito da prova e o ponto de vista em que o examinador se coloque, ao propô-la, assim serão os resultados ou o material colhido. Haverá

[1] Em espanhol e italiano é comum traduzir-se a palavra *test* por reativo. Cf. LAFORA, *Los niños mentalmente anormales*. 2. ed. Madrid, 1933; GONZALEZ. *Diagnostico de los ninõs anormales*. Madrid : El Magistério Espanhol, [s.d.]; AGUAYO. *Pedagogia científica*. Havana : Cultural S.ª, 1930; SANCTE DE SANCTIS. *Psicologia sperimentale.* Torino : Lates, 1930. Em português, LOURENÇO FILHO. *Testes ABC*. 2. ed. São Paulo : Melhoramentos, 1937, também emprega a palavra "reativo".

[2] O interesse mais acentuado pelos estudos das variações individuais nas observações começou depois que os astrônomos verificaram diferenças em suas observações. De um para outro havia sempre uma diferença de tempo no registro da passagem de um astro pelo *fio de cabelo* posto em uma das lentes do telescópio. Foi na Inglaterra, em 1795, que Maskeline, astrônomo do Observatório de Greenwich, verificou pela primeira vez diferenças entre os seus registros e os de seu assistente Kinnebrook. Julgando-o incapaz de exercer o cargo, despediu-o e fez um relato do incidente. Sabedor do fato na Alemanha, Bessel passou a estudar as causas desses erros. Depois de algum tempo formulou a hipótese de que em toda observação há um erro pessoal, e, em 1822, ao publicar os primeiros resultados, deu o nome de *equação pessoal* a essa diferença individual de observação. Ver PIÉRON. *Psicologia experimental.* Tradução de Lourenço Filho. São Paulo : Melhoramentos, [s.d.]; BORING. *An history of experimental Psychology*. Appleton : Century, 1929; MURPHY. *An historycal introduction to modern Psychology*. 4. ed. rev. New York : Harcourt, Brace, 1938.

[3] Já em 1845, Horace Mann clamava por provas em condições objetivas em substituição aos antigos exames. Em 1864, na Inglaterra, o reverendo Fischer tentava objetivar o julgamento dos trabalhos de seus alunos com o emprego de *Scale-Book*. Cattell, em 1890, emprega pela primeira vez a expressão *mental-test*, e a partir de então formou-se a consciência de que era necessário o emprego de provas em condições objetivas para medir tanto a inteligência e aptidões como o rendimento do trabalho escolar (cf. Ruch; Lincoln e Workman; Monroe).

testes de sensibilidade, testes de inteligência, testes de aptidão, testes de maturidade, testes de fadiga, testes de emoção... Desde que eles compreendam, nos seus resultados, a colheita de material do comportamento, por qualquer que seja o seu aspecto, aí teremos um teste psicológico. O teste psicológico não é, assim, apenas o teste de inteligência, nem só o teste mental, por mais amplas que sejam as acepções dadas a estes adjetivos. Teste psicológico é, assim, um gênero; e teste de inteligência é uma espécie desse gênero.

## Teste de inteligência

A caracterização de teste de inteligência exige uma definição de inteligência. Se fôssemos, porém, discutir, do ponto de vista teórico, o que é inteligência, fugiríamos do nosso objetivo. É certo que não se pode dispensar um ponto de vista teórico. Uma concepção geral, larga e esclarecida, não do que seja inteligência, mas do que é a atividade inteligente, torna-se necessária. Isso não significa o abandono da discussão teórica, que a seu tempo será levantada. Adotaremos para caracterização do teste de inteligência um ponto de vista objetivo e funcional. Com efeito, com o auxílio de certas provas, procuramos verificar não a *inteligência em si*, mas os seus efeitos. E o que interessa é a consideração do *ato inteligente*. Se esse ato pode ser medido ou graduado por provas convenientes, resultará daí que teremos testes de inteligência.[4]

## Classificação dos testes segundo a modalidade de aplicação

Segundo a modalidade de sua aplicação a um indivíduo ou a grupos de indivíduos, simultaneamente, o teste pode ser classificado como individual ou coletivo. O recurso normal para aplicação de um teste coletivo, em geral, é o apelo ao trabalho gráfico, pela simples razão de que este deixa um registro permanente, que pode ser verificado depois, em qualquer tempo, e estudado por qualquer especialista. Tendo diante de si um só examinando, o experimentador pode anotar as reações que esteja observando, com maior ou menor minúcia. O mesmo não seria possível, em face de um grupo de examinandos. E neste caso, o papel e o lápis são os instrumentos necessários. O examinador ou dá as ordens verbalmente, ou as apresenta escritas, em modelos que expõe, à vista de todo o grupo; ou ainda, as entrega já escritas, em cada folha de trabalho, impressas ou mimeografadas, para que cada examinando as leia e as resolva, no mesmo papel. Na disposição do material de exame, na hora da prova, deverá haver o maior cuidado para que todos os examinandos estejam sensivelmente nas mesmas condições de trabalho, para que os resultados não venham a ser influenciados por essa causa de erro, tão comum nos testes. De outra forma, não teríamos um bom teste, por falta de certas condições de objetividade. Os testes de inteligência podem ser apresentados individual ou coletivamente.

## Resumo histórico dos testes coletivos de inteligência[5]

As dificuldades na aplicação dos testes individuais, o emprego de testes de escolaridade e ainda a necessidade de seleção de grandes grupos, em tempo mínimo, deram como resultado o aparecimento das primeiras tentativas de ensaios para emprego de testes coletivos de inteligência, não sem a oposição dos psicólogos. Dentre os pioneiros podemos assinalar W. Pyle, aplicando vários testes a grupos de crianças, sem todavia combinar os resultados parciais para conseguir um índice global da capacidade. Os resultados desse trabalho foram divulgados em 1913. Por essa época, Thorndike também emprega testes coletivos para examinar os empregados da Metropolitan Life Insurance Co., deixando, porém, de divulgar os resultados conseguidos. Pintner, então professor da Universidade de Ohio, aproveita-se da idéia de Pyle e aplica seis testes a um grupo de crianças, a fim de selecionar os débeis. Tomou como medida da capacidade dos alunos o mediano de seis percentis. Os resultados dessa tentativa foram publicados em 1917. Miller, por outro lado, nos relata que em 1914, quando trabalhava sob a direção de Whipple, preparava uma tese que

[4] Para Piéron, a noção de inteligência é um "conceito de valor". O termo inteligência é empregado, de fato, com acepções muito diversas. Para Claparède, a palavra inteligência tem sido empregada em três sentidos diversos, a saber: a) nome dado à classe de fenômenos psíquicos que têm por objeto o conhecimento. Inteligência, assim, se opõe à afetividade, à reatividade. O adjetivo de inteligência, nesta acepção, é intelectual; b) maneira de ser dos processos psíquicos adaptados com êxito a situações novas. Inteligência será a capacidade de resolver problemas novos pelo pensamento (Stern, Claparède). Nesta acepção, opõe-se ao automatismo, ao instinto, à imbecilidade. O adjetivo será inteligente; c) na linguagem corrente, capacidade superior à média. Ver na bibliografia, Claparède e Pieron.

[5] Para referências históricas sobre os testes coletivos de inteligência, ver na bibliografia Symonds; Yoakum e Yerkes; Levine e Marks; Ballard; E. Smith; Garrett e Schneck; Pintner; Lincoln e Workman; Colvin; e *Memoirs of the National Academy of Sciences*, v. 15.

tratava dos testes coletivos de inteligência, indicando o seu valor e prevendo o próximo emprego dessa modalidade de teste de inteligência como recurso barato, prático e cômodo. Seu teste coletivo foi experimentado na Escola Secundária da Universidade de Minnesota, em 1917. Para esse especialista, até então nenhum instrumento de tal natureza havia sido tentado e nem mesmo normas fidedignas haviam sido estabelecidas para testes simples que pudessem ser usadas em testes coletivos. O teste, porém, de Miller, só foi publicado em 1921. O passo decisivo, no entanto, para o emprego dos testes coletivos de inteligência, foi dado por A. I. Otis, quando assistente de Terman na Leland Standford University, e a ele cabe a primazia de haver organizado o primeiro teste coletivo para medir a capacidade dos alunos. O seu trabalho foi enviado a um comitê da American Psychological Association, que estava tratando, nessa época, do trabalho de seleção dos recrutas para a formação do exército americano que participaria da Grande Guerra. Com auxílio do material de Otis, foi organizado o trabalho preliminar para composição do teste de inteligência destinado a classificar os recrutas. Cerca de quatro testes de escala primitiva organizada pela comissão de que faziam parte Yerkes, Terman, Wells, Whipple, Haines, Goddard e Bingham, eram do trabalho de Otis. Depois das experimentações preliminares, que alcançaram cerca de 80 mil pessoas, apareceu a Army Alpha e logo a seguir a Beta, aquela para alfabetizados e esta para estrangeiros e analfabetos.[6] Em 1918, quando era ainda intenso o trabalho de seleção para formação do exército, Otis traz nova contribuição aos testes coletivos de inteligência, publicando o *Otis Group Intelligence Examination*. Em 1919, aparece o *Intelligence Examination for High School Graduates* de Thorndike e também conhecido por teste CAVD.[7] Ainda no mesmo ano, Thurstone publica a *Psychological Examination for College Freshmen and High School Seniors*. A partir de 1920 começam a aparecer trabalhos de maior vulto e mais precisão científica. Nesse ano, Terman publica o seu famoso e discutido teste coletivo de inteligência – *Group Test of Mental Ability* – cujas normas foram estabelecidas depois de examinados cerca de 40 mil alunos dos graus 7 a 12. Segue-se o trabalho de Haggerty – *Haggerty Intelligence Examination*. Nesse mesmo ano, sob os auspícios do "National Research Council", constitui-se uma comissão para a organização de um teste nacional de inteligência. Dessa comissão fizeram parte Terman, Thorndike, Haggerty, Whipple e Yerkes, que depois de exaustivos trabalhos organizaram o *National Intelligence Test*. Nesse trabalho foram gastos cerca de 25 mil dólares. Em 1920, a Civil Service da Inglaterra, a exemplo de sua similar americana, que então já aplicava testes coletivos de inteligência, cria uma seção denominada "Intelligence Tests" para exame das candidatas aos trabalhos das repartições públicas. Por essa época, eram intensificados os trabalhos de Burt H. Winch, Thomson e Ballard, tendo esses dois últimos organizado alguns testes coletivos de inteligência. Na América, continuava o movimento e cada vez mais intensamente. Vêm os trabalhos de Dearborn (1920), Pintner (1920); Baker (1924); McCall (1925) Goodenough (1925); Bregman (1925); Kuhlmann-Anderson (1927) além de outros. Não podemos deixar de nos referir, ao terminar este resumo, à mais recente tentativa de organização de um teste coletivo de inteligência (1935) baseado na doutrina de Spearman.

## Objeções aos testes coletivos de inteligência

Muitas têm sido as objeções levantadas contra os testes coletivos de inteligência. Desde as primeiras tentativas eles têm sido fortemente atacados, é certo. Não, porém, pelos que os tenham experimentado. Objeções contra os testes – se é que possamos chamar objeções a críticas menos fundadas – ainda são feitas mais pelos leigos que por especialistas. Em nosso meio, por exemplo, é interessante verificar a facilidade com que certas objeções primárias têm curso, e chegam a ponto de afirmar que os testes não são mais empregados nos países de origem. Tal atitude tem apenas duas fontes: a imaginação exaltada do leigo e o primarismo que se forma pela falta de informação. Isso, porém, não deve preocupar a quem se disponha a estudar seriamente o assunto. Sem dúvida, os testes de inteligência e, em especial, os coletivos, são instrumentos que se apresentam ainda com defeitos, somente corrigíveis com o aperfeiçoa-

[6] Foram examinados pela Divisão de Psicologia do Exército Americano 1.726.966 pessoas, tendo a Army Alpha sido aplicada a cerca de 1 milhão e 250 mil. No dia do armistício, os examinadores e ajudantes formavam um verdadeiro regimento: 120 oficiais, 350 soldados e 500 ajudantes.

[7] C significa resolução de problemas de completamento verbal; A, problemas relativos a situações aritméticas; V, problemas referentes ao vocabulário; e D, problemas cujas soluções dependam do cumprimento de ordens verbais. Ver na bibliografia Thorndike e Pintner.

mento progressivo da técnica. Nem por isso devemos abandonar o instrumento. A pouco e pouco, vai ele sendo aperfeiçoado; e dentro de mais alguns anos terá certamente melhorado nas deficiências que ainda apresenta. Seu papel no futuro dependerá tão-somente da compreensão dos que o utilizarem, por emprego adequado, não exigindo dele mais do que realmente possa dar. Muitas vezes, o emprego inadequado do instrumento, ou a utilização por quem desconheça de modo perfeito seu mecanismo, poderá concorrer para que os resultados sejam falhos. É que instrumento dessa natureza só serve para determinados fins e assim mesmo rigorosamente dentro das condições prescritas para sua aplicação.[8]

### Importância dos testes coletivos de inteligência

Os testes coletivos de inteligência desempenham um papel da maior importância na administração e organização escolares, nos departamentos de pesquisas educacionais e na administração em geral. Citaremos como exemplos: Instituto de Educação da Universidade do Distrito Federal, Instituto de Pesquisas Educacionais, Conselho Federal do Serviço Civil.

Dentre outras aplicações citaremos: 1) estudos sobre as diferenças individuais; 2) seleção de alunos, formação de grupos homogêneos e de classes especiais (aceleração, oportunidade); 3) estudos sobre o desenvolvimento mental; 4) orientação profissional; 5) seleção profissional.[9]

## O problema da fidedignidade

### Os testes de inteligência e a técnica geral dos testes[10]

Todas as objeções, fundadas ou não, devem ceder à verificação das qualidades de um teste, como instrumento de medida. Isto é, à verificação técnica de suas próprias qualidades – o *teste* de teste... E como se fará essa verificação? Que requisitos se devem exigir de um instrumento de medida?... A mais simples reflexão nos indica que um instrumento dessa natureza deve ser *sensível e seguro*. Sensível, ou seja, capaz de apreciar as variações do que se quer medir. Seguro, ou seja, capaz de inspirar confiança por sua *coerência*. Coerência interna, isto é, relativa às suas diferentes partes, em que cada uma delas meça *proporcionalmente o que deve medir*. Externa, isto é, que cada uma dessas partes e o seu conjunto apreciem realmente o atributo que o instrumento pretende verificar. De um modo geral, estes problemas envolvem a técnica do que se convencionou chamar de aferição do teste. Nessa aferição, o problema da sensibilidade do instrumento é o mais simples, resolvendo-se pela estatística de distribuição e variação. Não assim, os dois últimos problemas, para cuja solução várias técnicas têm sido apresentadas. Elas envolvem quase sempre verificações de correlação, ou seja a apreciação de fenômenos de observação mais delicada. Numa palavra, envolvem os problemas chamados de *fidedignidade* e de *validade*.

### Coerência: fidedignidade e validade

A coerência de um instrumento de medida é verificada pelo grau de concordância existente entre os índices internos e externos; também pela verificação da concordância nos resultados de sua aplicação repetida. Como se vê, o problema se refere a provar que o instrumento possui certo grau de coerência, tanto pela sua adequabilidade ao atributo que se pretende medir, quanto pela confirmação dos resultados em sucessivas aplicações.

### Coerência interna: fidedignidade

Não resta dúvida que as mensurações feitas com os testes admitem certo número de erros, e os testes coletivos mais que os de aplicação individual. No entanto, quanto mais atenuarmos a interferência das causas de erro, maior será a fidedignidade do instrumento.

A maneira teórica de considerar a fidedignidade será a de admitir a possibilidade de aplicação ao mesmo indivíduo de um número *n* de formas paralelas ou equivalentes do teste. Isso feito, tomar a média

[8] Decroly, em 1923, publicou um interessante trabalho em que condensou os inconvenientes e as vantagens dos testes coletivos e individuais (cf. *L'Année Psychologique*, 1923).

[9] Sobre as aplicações dos testes de inteligência, Hildreth, da Colúmbia, publicou um interessante trabalho, na *Review of Educational Research* (1935). P. Mort, também na mesma revista, em 1932, fez um resumo das aplicações.

[10] Para a discussão desses problemas, ver na bibliografia: Symonds; Ruch e Stoddard; McCall; Pintner; Kelley; Barthelmess; M. Smith; Long e Sandiford; Monroe e Engelhart; Freeman; Thurstone; Garrett; Otis; Skaggs; Kelley e Shen; Willoughby; Piéron e Fessard; Rey; Monroe; Jordan; Colvin; Rugg; Ruch; Fessard.

dos resultados como o resultado verdadeiro em relação ao indivíduo. Ponto de vista apenas teórico.

A maneira prática consiste em calcular o coeficiente de correlação entre os resultados de duas aplicações sucessivas do mesmo instrumento, num mesmo grupo de indivíduos ou em grupos semelhantes; ou, ainda, entre os resultados de duas formas paralelas ou equivalentes, assim aplicadas. Reconheceu-se, todavia, que esses processos não são os menos influenciados por fatores externos. E por essa razão propôs-se como expressão da fidedignidade o coeficiente de correlação entre os resultados das metades do mesmo teste.

## Coerência externa: validade

Ao que denominamos coerência externa, os autores têm chamado de *validade* – os mais modernos autores a têm definido como o grau de coerência entre um índice fixado e um índice externo – este conhecido por meios objetivos ou preliminarmente fixado, por valores estimativos. A esse índice, dá-se o nome de *critério*. O grau de coerência é encontrado pelo coeficiente de correlação entre os dados representativos de cada índice, que recebe a denominação de *coeficiente de validade*. Quanto maior esse grau de coerência, tanto mais válido será o instrumento. No caso particular do teste de inteligência, o que se pretende medir é o *ato inteligente*. O índice fixado deverá estar de tal modo proposto que realmente verifique esse atributo, muito embora se admita sempre possibilidade de erro.

A validade é geralmente definida como a característica do teste que mede realmente o que pretende medir. Isto é, que o teste tenha a qualidade de verificar o atributo visado, pela adequação das questões nele contidas. E tanto é assim que a National Association of Directors of Educacional Research a definiu como o grau de correspondência existente entre a capacidade medida pelo teste e a capacidade delimitada e medida objetivamente.

As definições poderão variar na forma; não, porém, em essência. Em relação à validade, como diz Monroe, o que procuramos é o grau de constância da relação funcional existente entre os resultados do teste e as capacidades consideradas como medidas no exercício de sua função. Barthelmess assegura que a validade nos testes de inteligência é o grau de concordância verificada pela diferenciação que o teste apresente para os indivíduos, e a diferenciação real em inteligência entre esses mesmos indivíduos.

## Fidedignidade prática

Propusemos o problema de um modo ainda teórico. Vejamo-lo agora, na prática. O grau de coerência interna é, em geral, calculado com o emprego do coeficiente de correlação. A esse coeficiente aplicado para esse efeito, os autores têm denominado *coeficiente de fidedignidade* do teste.

A expressão foi primeiramente usada por Spearman, em 1910. Mas já desde 1904, esse mesmo psicólogo e estatístico inglês, ao propor a *teoria dos dois fatores*, empregava em seus trabalhos o processo. Assim, o chamado coeficiente de fidedignidade diz respeito a duas mensurações do mesmo atributo com o emprego do mesmo instrumento ou de instrumentos equivalentes. E o que se pretender determinar é o grau de coerência interna do instrumento empregado, quaisquer que sejam os verificadores, desde que respeitada a técnica de aplicação que estiver fixada para bom uso do instrumento.

Verifica-se, porém, pelo exame das técnicas empregadas que a denominação poderá levar a equívocos. De fato, três técnicas diversas têm sido empregadas e, quase sempre, dando resultados diferentes. No entanto, a expressão está largamente difundida e aceita na extensa bibliografia americana e inglesa. Um ou outro autor tem proposto denominação específica para uma das técnicas, o que não tem logrado aceitação. E tanto é assim que um recente dicionário de psicologia, a que emprestaram colaboração mais de 100 especialistas, a registra como de aceitação mais geral.[11] Nem por isso deixa de ser equívoca, a não ser que, ao usarmos a denominação, estabeleçamos a técnica empregada para seu cálculo.

Procurando unificar a terminologia sobre os índices estatísticos de um teste, a Comissão de Unificação Terminológica, no Congresso de Psicotécnica de 1931, resolveu adotar, em substituição a coefici-

[11] WARREN, W. *Dictionary of Psychology.* New York : Hougton Mifflin, 1934.

ente de fidedignidade, três novas denominações: *coeficiente de homogeneidade, coeficiente de equivalência e coeficiente de constância*. Não foi melhor o resultado: deram para verificação do grau de coerência de um instrumento três denominações. Por essa forma, evidentemente, não se tornou inequívoca a noção.

A expressão pouco importa no caso. Na verdade, o que desejamos verificar é a coerência do instrumento em sucessivas aplicações: se ele é realmente digno de merecer a nossa confiança; se podemos trabalhar com ele sem que, por inconsistência, venha deformar as nossas conclusões. Por essa razão é que, neste trabalho, denominaremos as três técnicas, que passaremos a analisar, do seguinte modo:

a) fidedignidade por constância de aplicação;
b) fidedignidade por equivalência;
c) fidedignidade por homogeneidade.

*Primeira técnica* (a) – Obtém-se o grau de coerência do instrumento pelo cálculo do coeficiente de correlação entre os resultados de duas aplicações sucessivas da mesma forma do teste ao mesmo grupo ou a dois grupos equivalentes de indivíduos. E uma vez que a fidedignidade de um teste é expressa pela sua autocorrelação, a mais simples e a mais direta será essa técnica.[12]

No entanto, a despeito de ser a mais prática e a mais direta, deve ter emprego limitado, principalmente em relação a testes coletivos de inteligência. Se o grupo de indivíduos for submetido às duas aplicações no mesmo dia, ou com intervalo de uma ou duas semanas, muitos indivíduos lembrar-se-ão de algumas questões e de suas respostas, por ocasião da segunda aplicação; em conseqüência, os resultados aparecerão *sensivelmente* melhorados. Também a atitude dos indivíduos poderá variar; a fadiga e o enfado poderão contribuir para diminuição do interesse que ponham na exatidão do trabalho empreendido. Por outro lado, haverá uma possibilidade de transferência.

Procurando atenuar os efeitos da memória e a possibilidade de *transfer*, alguns autores aconselham o aumento de prazo entre as duas aplicações. As objeções, porém, ficariam de pé.

Por tudo isso e, ainda, pela impossibilidade de controlarmos os fatores externos que, como vimos, influem nos resultados da segunda aplicação, é que não convém o emprego irrestrito dessa técnica.

*Segunda técnica* (b) – Obtém-se o grau de coerência pelo cálculo do coeficiente de correlação entre os resultados de aplicações de duas formas paralelas ou equivalentes de um teste ao mesmo grupo de indivíduos ou a dois grupos equivalentes. Deve-se ter o cuidado de deixar um intervalo apreciável entre as duas aplicações.[13] Mesmo evitando-se o fator memória, não se conseguirá evitar o *transfer*.

Alguns autores, procurando diminuir a transferência, apelaram para um ensaio preliminar, por ocasião da primeira aplicação. Outros, com o mesmo objetivo, mandam dar uma bonificação ao número de pontos atribuídos a cada indivíduo, na primeira aplicação.

Além disso tudo, é evidente que esta técnica exige o trabalho de composição de duas formas paralelas do mesmo teste. E serão elas realmente paralelas ou equivalentes? Kelley propõe como critério a similaridade sem identidade de elementos, o que é vago e inexpressivo. Embora certos autores afirmem que um teste não pode ser considerado bom sem que se apresente com várias formas paralelas, julgamos exagerada essa opinião. Não resta dúvida que será realmente útil dispor de formas paralelas. Isso, porém, quando nos mereçam confiança, por coerência interna e externa.

Não devemos esquecer também que o emprego dessa técnica acarretará diferenças nos resultados, para mais ou para menos, em virtude das diferenças de atitude e esforço dos indivíduos submetidos ao exame, e mesmo em virtude de variação das condições ambientes.

Por essas razões é que não julgamos acertado apenas o emprego dessa técnica, embora seja ela superior à primeira.

*Terceira técnica* (c) – Obtém-se o grau de coerência do instrumento pelo cálculo do coeficiente de correlação entre os resultados das questões pares e ímpares do teste. Se bem que pareça a mais adequada, essa técnica tem sido ultimamente objeto das maiores discussões, chegando R. R. Willoughby, da Universidade de Clark, a afirmar que ela não passa de uma versão da técnica anterior. Também a criticam Ruch e Stoddard.

[12] Truman Kelley usa para essa técnica a denominação "coeficiente de reteste". Outros autores usam ainda "coeficiente de consistência". Uma e outra, porém, não tiveram aceitação.

[13] A expressão *forma paralela de um teste*, ou simplesmente, *forma paralela* é de uso comum em medidas educacionais. Freqüentemente também encontramos *forma equivalente, forma comparável, forma duplicata* e, raramente, *forma igual*.

Em abono dessa técnica há, porém, trabalhos do mais alto valor. Remmers, citado por R. C. Jordan, diz:

> É de importância capital notar que esses métodos, em geral, não dão o mesmo coeficiente de fidedignidade. O coeficiente obtido pelo método das questões pares e ímpares é em geral mais alto do que o conseguido pelo método das formas equivalentes.

E ainda:

> Fatores tais como fadiga, monotonia, distração, etc., influirão mais na última técnica (forma equivalente)...

Uma grande experiência de R. C. Jordan também dá margem a que sejamos favorável a essa técnica, porque dá o grau de fidedignidade do instrumento, independentemente do fator individual em sua segunda aplicação. Essa conclusão também encontra apoio em outros autores. Dentre eles, J. C. Dunlap, a cujo trabalho também se refere Jordan.

Deve-se sempre esperar que por essa técnica o resultado seja maior do que o obtido pela segunda. Foi, aliás, o que já verificou também Foran, citado por Monroe e Engelhart.

A fórmula usada é a que foi estabelecida simultaneamente, em 1910, por Spearman e Brown, e representa um caso particular da *fórmula* de profecia, desses mesmos autores, como veremos adiante.

$$r_{11} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}}$$

em que $r_{11}$ é fidedignidade por homogeneidade, e $r_{\frac{1}{2}\frac{1}{2}}$ é o coeficiente de correlação entre as metades do teste.

## Fidedignidade virtual

Obtida a fidedignidade prática de um teste coletivo, pelas técnicas indicadas, é possível calcular a fidedignidade virtual do mesmo teste, ou seja, a correlação entre os resultados obtidos e aqueles que, teoricamente, poderiam ser conseguidos. Isto é, aqueles resultados conseguidos com um número *n* de aplicações do teste ou de *n* formas equivalentes, aplicadas ao mesmo grupo de indivíduos ou a dois grupos equivalentes.

O coeficiente obtido tem sido freqüentemente chamado de *índice de fidedignidade*. Na verdade, não se trata de um índice, mas de uma expressão teórica da fidedignidade de que é capaz o instrumento, e por essa razão é que é preferível denominá-lo *coeficiente teórico de fidedignidade* ou, simplesmente, *fidedignidade virtual*, em oposição ao que chamamos de fidedignidade prática.

Passemos, agora, à fórmula que nos dará a fidedignidade virtual.

Sejam a, b, c,... n as formas equivalentes de um teste coletivo de inteligência. Qualquer delas, quando aplicada, está sujeita a erro experimental.

A fidedignidade prática por equivalência entre as formas será:

$$r_{ab}\ r_{ac}\ r_{bc}\ \ldots\ldots\ r_{n-1\ .\ n}$$

O verdadeiro resultado (V) no teste será a média de pontos que o indivíduo conseguir no número *n* de aplicações. É claro que *V* não poderá ser realmente calculado, uma vez que *n* será sempre finito, e por isso haverá a persistência de um erro residual de mensuração, mesmo que desprezemos a influência de fatores sistemáticos: treino, fadiga, *transfer*, etc.

Por definição, *V* não está sujeito a duas espécies de erros; apenas aos sistemáticos. Assim, a fidedignidade virtual será sempre mais alta do que a prática ou real.

Sejam:

$X_a$ os resultados na forma *a*
$X_b$ os resultados na forma *b*
V o resultado verdadeiro

$X_a = V + S$
S e S' os erros
$X_b = V + S'$

A fidedignidade por equivalência será:

$$r_{X_a X_b} = \left( \frac{\Sigma\,(X_a X_b)}{N\ \sigma_{X_a}\ \sigma_{X_b}} \right)$$

mas

$$\Sigma_{X_a X_b} = \Sigma\,(V+S)\,(V+S') = \Sigma\,(V^2 + VS + VS' + SS'') =$$
$$= \Sigma V^2 + \Sigma VS + \Sigma VS' + \Sigma SS'$$



R. bras. Est. pedag., Brasília, v. 79, n. 193, p. 113-134, set./dez. 1998

e

$$\sum VS = \sum VS' = \sum SS' = 0;$$

donde

$$\sum_{Xa\,Xb} = \sum V^2$$

Os desvios-padrão das formas equivalentes serão iguais:

$$\sigma_{Xa} = \sigma_{Xb}$$

$$r_{Xa\,Xb} = \frac{\sum V^2}{N\ \sigma^2_{Xa}}$$

mas,

$$\frac{\sum V^2}{N} = \sigma^2_V$$

sendo $\sigma_V$ o desvio-padrão da distribuição dos resultados verdadeiros:

$$r_{Xa\,Xb} = \frac{\sigma^2_V}{\sigma^2_{Xa}}$$

$$\sqrt{r_{Xa\,Xb}} = \frac{\sigma_V}{\sigma_{Xa}}$$

A fidedignidade virtual será:

$$r_{V\,Xa} = \frac{\sum V_{xa}}{N\ \sigma_V\,\sigma_{Xa}} = \frac{\sum V\,(V+S)}{N\ \sigma_V\ \sigma_{Xa}} =$$

$$= \frac{\sum V^2 + \sum VS}{N\,.\,\sigma_V\,\sigma_{xa}}$$

$$r_{V\,Xa} = \frac{\sum V^2}{N\ \sigma_V\,\sigma_{Xa}} = \frac{\sigma_{V^2}}{\sigma_V\,\sigma_{Xa}} = \frac{\sigma_V}{\sigma_{Xa}}$$

donde

$$r_{V\,Xa} = r_{Xa\,Xb}$$

A fidedignidade virtual será a raiz quadrada da fidedignidade prática, e representará também a correlação máxima de que um teste coletivo de inteligência é capaz aplicado *n* vezes. Quando um teste se apresentar com baixa fidedignidade virtual, deve ser abandonado ou refeito, porque, sendo esses coeficientes menores do que a unidade, a fidedignidade prática será sempre menor do que a virtual. E como instrumento de medida, não merecerá confiança.

## Fidedignidade e extensão do teste

A fidedignidade de um teste aumentará se a esse teste acrescentarmos novas questões, que procurem diagnosticar o mesmo *atributo*? Por outras palavras, se a fidedignidade de um teste não se apresentar como satisfatória, ela melhorará no caso de dobrarmos ou triplicarmos a extensão da prova, desde que as questões acrescidas sejam do mesmo teor que as já existentes? Ainda outra questão: se, ao contrário de dobrarmos ou triplicarmos a extensão do teste, aplicarmos duas ou três formas do teste ao mesmo grupo de indivíduos e tomarmos a média dos resultados das aplicações como o resultado individual, a fidedignidade aumentará?

A essas indagações podemos responder com o emprego da *fórmula de profecia* de Spearman e Brown:

$$r_{nn} = \frac{n r_{ab}}{1 + (n-1) r_{ab}}$$

em que $r_{nn}$ representa a correlação entre *n* formas paralelas do teste; *n*, o número de formas paralelas ou o número de vezes que o teste foi aumentado; e $r_{ab}$ a fidedignidade por constância de aplicação ou por equivalência.

Quando se tratar da duplicação do teste, a *fórmula de profecia* passará a ser esta:

$$r_{nn} = \frac{2\,r_{ab}}{1 + r_{ab}}$$

Essa fórmula é a que se emprega para o cálculo da *fidedignidade por homogeneidade*, em que $r_{ab}$ é o coeficiente de correlação entre os resultados das questões pares e ímpares, como já vimos atrás.

A *fórmula de profecia* pode também ser aplicada para sabermos o número de questões de que deve ser aumentado o teste, a fim de que a sua fidedignidade alcance um valor *x*.

De fato, resolvendo a fórmula para *n* vezes a extensão do teste, teremos:

$$n = \frac{r_{nn}\,(1 - r_{ab})}{r_{ab}\,(1 - r_{nn})}$$

Digamos que um teste de inteligência tenha 30 questões, e sua fidedignidade seja *80*. De quantas questões deverá ser ele aumentado, para que a fidedignidade suba para 90? Calculando *n*, encontraremos 2,5. Donde 2,25 X 30 = 68. Logo, o teste deverá ser aumentado de 38 questões.

É claro que não podemos aumentar indefinidamente a extensão de um teste, com o objetivo de fazer crescer a sua fidedignidade. Se ela for muito baixa, o trabalho não se justificará. Além disso, com o aumento da extensão de um teste, intervirão fatores como a fadiga, o enfado, a diminuição de interesse, etc., que passarão a influir nos resultados. Quando, porém, o material acrescido for bem escolhido, e de tal modo que desperte igual interesse pelo trabalho, poderemos aumentar um teste de três ou quatro vezes, quando ele tiver de 40 a 60 questões; e de cinco a seis vezes, e até mesmo sete, quando tiver de 20 a 30 questões. Alongamentos que ultrapassem os limites referidos tornam a *fórmula de profecia* menos segura. De fato, como notou Garrett, ela dará então resultados acima do valor real da fidedignidade.

## O problema da validade

### Validade e fidedignidade

As relações entre validade e fidedignidade não têm sido suficientemente discutidas, talvez pela crença de que a simples caracterização, com base em cálculos estatísticos, ou definições, por vezes meramente verbais, bastem para que sejam aceitas como conhecidas. Por outro lado, esses dois termos têm sido colocados em planos diferentes, ao estabelecermos as bases para a organização de testes de inteligência. Não nos parece razoável, pelo menos do ponto de vista teórico, a separação linear entre validade e fidedignidade. Qualquer discussão sobre validade, sem a consideração de fidedignidade, será improfícua. Mais ainda: a validade de um teste está condicionada à sua fidedignidade, porque um instrumento só é perfeito quando fidedigno. A fidedignidade será, pois, uma condição necessária; não, porém, suficiente. Sem dúvida que um instrumento fidedigno será sempre válido teoricamente, para certo efeito. Mas poderá não o ser para o fim a que esteja destinado. A validade de tanto quanto a fidedignidade procuram a coerência do instrumento. Se é certo que a fidedignidade não pode ir além dos limites da coerência interna, não é menos certo que a coerência externa dela dependerá sempre.

### As questões do teste de inteligência

As questões de um teste coletivo de inteligência representam a sua pedra de toque. Do cuidado com que as escolhermos e as redigirmos, dependerá, em grande parte, a coerência do instrumento. A aferição do teste não poderá ser feita se, desde os primeiros ensaios de aplicação, não sentirmos que estamos trabalhando com elementos suscetíveis de comporem um instrumento de medida. Por isso, julgamos que as questões devem ficar subordinadas às seguintes condições:

a) cada questão deve incidir sobre matéria que não tenha sido aprendida especificamente na escola (informação sob efeito de treino). Os testes de inteligência não devem verificar conteúdo específico, pois se destinam a hierarquizar indivíduos sob influências educativas diversas;

b) o conteúdo específico, necessário à compreensão e à resolução da questão, deve ser comum à experiência dos indivíduos da idade, ou do grupo de idades, a que o teste de destine;

c) cada questão deve obter um comportamento do indivíduo, de modo que a situação proposta seja nova, muito embora exija o concurso da experiência anterior;

d) as questões, em seu conjunto, devem apresentar variedade de atividades, a fim de que se evite a monotonia do trabalho e a falta de interesse nele;

e) as questões devem variar em dificuldade, a fim de que permitam que os resultados gerais discriminem níveis de desenvolvimento, os quais possam ser atribuídos a idades sucessivas, ou a grupos de idades;

f) cada questão deve ter redação clara e precisa, admitindo uma só resposta.

### O atributo: a inteligência[14]

O problema da inteligência que ficou apenas aflorado exige aqui o mais amplo exame. No caso dos testes de inteligên-

[14] Para a discussão deste problema, ver na bibliografia Pintner, Spearman, Stern, Thorndike, Boyton, Claparède, Piéron, Freeman, Peterson, Rey, Melli, Skaggs, *L'Année Psicologique* (1934) e. Fröbes.

cia, *o atributo é o ato inteligente*. Que é, porém, inteligência? A discussão sobre essa pergunta tem merecido a maior atenção dos psicólogos, e originado uma infinidade de definições, baseadas sobre teorias diversas.

As definições de inteligência, segundo Pintner, podem ser distribuídas por quatro grupos, à exceção das de Thurstone, Spearman e Freeman. Estes grupos são os seguintes:

a) *Definições biológicas* – São as que acentuam o caráter de adaptação do organismo a situações novas. Assim, a inteligência dependerá da plasticidade do organismo (Stern, Wells, Woodworth, Peterson, Edwards, Claparède).

b) *Definições educacionais* – São as que acentuam a capacidade de adquirir conhecimentos com rapidez e facilidade. É mais inteligente aquele que aprende mais rapidamente. Infelizmente, inteligência tem sido confundida com capacidade de memorização. Mais inteligente é o que acumula maior número de fatos. Se analisarmos bem, verificaremos que esse grupo é uma subdivisão do primeiro. Quem aprende mais depressa, fá-lo porque tem maior capacidade de adaptação. Aliás, o próprio Pintner chama a atenção para esse ponto. Diz o mestre da Colúmbia: "Toda aprendizagem pode ser encarada como ajustamento ou adaptação a situações novas". Deram definições educacionais, dentre outros, Colvin, Buckingham, Hemmon.

c) *Definições da inteligência como faculdade* – São as que procuram mostrar em que consiste a inteligência e de que processos mentais ela se compõe. Criticando esta concepção de inteligência, Spearman mostrou o número e teor dos processos mentais que compõem a inteligência: variam de autor a autor e não há acordo sobre o seu número. Seguem esta orientação: Terman, Woodrow, Haggerty, dentre outros.

d) *Definições empíricas* – São as que acentuam o aspecto funcional da inteligência. São, via de regra, definições behavioristas, e que salientam o aspecto dinâmico dos atos inteligentes (Ballard, Thorndike, Pintner, Piéron).

Se atentarmos agora para os quatro grupos, veremos que eles poderão reduzir-se a dois únicos. Um, que inclua as definições biológicas e educacionais; outro, as que distinguem a inteligência como faculdade.

Agora, as teorias. Quatro teorias principais procuram explicar a natureza da inteligência. Delas faremos apenas uma ligeira exposição.

a) *Teoria dos dois fatores* – Em 1904, Spearman, discordando da simples descrição da atividade inteligente, e observando que as correlações entre as medidas de diferentes capacidades tendiam para uma disposição peculiar, propôs, em alguns estudos, a teoria dos dois fatores, *g* e *s*. O fator *g* representa a capacidade geral, que é constante no mesmo indivíduo. O fator *s* representa o aspecto específico, variável no mesmo indivíduo. Para o psicólogo e estatístico, em qualquer trabalho há influência desses dois fatores.[15] Essa teoria mereceu a crítica desfavorável de Binet (muito embora a aplaudisse quando proposta), de Thorndike, Thomson, Kelley e outros. A despeito das críticas, essa teoria tem tido larga aceitação.

b) *Teoria da capacidade geral* – Esta teoria foi proposta por Stern, em 1910. A inteligência, nesse caso, é uma capacidade geral que pode ser dirigida em qualquer domínio da atividade. A especialização depende tão-somente do ambiente. Para Stern, não resta dúvida que a capacidade depende da constituição do organismo.

c) *Teoria dos fatores múltiplos* – Agora não há mais nem dois fatores, nem capacidade geral. A inteligência é uma soma de vários fatores específicos (Thorndike). Segundo Pintner, a teoria não exclui a consideração do fator geral de Spearman. A sua existência, porém, não interessa ao mestre da Colúmbia. Do mesmo parecer é, aliás, Claparède, rebatendo as críticas feitas por Spearman à sua concepção de inteligência. Diz o mestre de Genebra que o seu ponto de vista não exclui de modo algum a hipótese de um fator *g*. Pelo contrário, a presença desse fator é até favorável à concepção funcional de inteligência que defende.

d) *Teoria funcional de Thurstone* – Em 1924, este psicólogo americano publicou a sua teoria funcional, segundo a qual a inteligência é a capacidade de apreensão total, com invenção de um processo adaptativo. Parece-nos que essa teoria está realmente muito próxima do ponto de vista de Claparède. Contudo, em seu trabalho – *The nature of intelligence* – não cita uma só vez Claparède.

De tudo isso se verifica que, com os testes de inteligência, procuramos avaliar uma certa capacidade e segundo a

[15] Não cabe aqui uma exposição minuciosa da doutrina de Spearman e de seus colaboradores. Para maiores esclarecimentos, ver na bibliografia Spearman, Melli.

qual conseguimos, para efeitos práticos, hierarquizar os indivíduos. Essa capacidade é muito influenciada pela ação social. Até que ponto irá essa influência? Fugiríamos ao assunto capital desta monografia se tentássemos discutir o assunto.[16] O que o teste aprecia é um comportamento, um nível de desenvolvimento. É, segundo os diferentes níveis obtidos, nos grupos de indivíduos, que os hierarquizamos. O teste de inteligência tem assim um fim prático, não o de resolver uma questão de cunho tanto psicológico como filosófico.

## O projeto do teste e o primeiro ensaio de aplicação

Uma vez organizadas as questões, de acordo com o que foi exposto, devem elas ser distribuídas pela dificuldade relativa que apresentem. E isso em relação a cada grupo de questões, bem como em relação ao conjunto. Essa distribuição pela dificuldade relativa será naturalmente muito precária, para o primeiro ensaio de aplicação.

O número de questões deve ser o dobro ou mais do que deverá conter o teste em sua forma final. O excesso facilitará a organização de formas equivalentes, bem como a eliminação de questões não adequadas ao fim proposto.

O número total das questões organizadas para o primeiro ensaio de aplicação deve ser dividido em três partes, A, B e C. E o grupo de indivíduos, a que formos aplicar a forma provisória, deverá também ser dividido em três subgrupos *a*, *b* e *c*, cada um deles, com um mínimo de 150 a 200 indivíduos, dentro das idades a que se destinar o teste. A fim de permitir que todas as questões sejam examinadas pelos indivíduos do grupo, convém proceder do seguinte modo:

a) ao subgrupo *a* aplicaremos a forma provisória na ordem ABC; ao subgrupo *b*, na ordem BCA; e ao subgrupo *c*, na ordem CAB;

b) dar tempo suficiente para que mais de 84% do grupo tenham possibilidade de tentar resolver todas as questões. Do contrário, seremos levados a conclusões errôneas.

Concluindo esse trabalho preliminar e aplicado o teste, poderemos passar a estudar os problemas fundamentais da validação.

## Validação das questões do teste

Da validade das questões de um teste depende, sem dúvida, a validade do instrumento, no seu conjunto. Um grande número de pesquisas têm sido feitas a esse respeito. E todas demostram que a validade de uma questão resulta de seu poder de discriminar os indivíduos, quanto a determinado atributo. Esse poder de discriminação diz respeito ao grau em que haja possibilidade de êxito ou fracasso numa resposta, e a porcentagem de discriminação dentro de cada idade ou grupo de idade. Regra geral, o melhor meio para obtermos um bom teste será determinar o grau de validade de um grande número de questões e dentre elas escolher aquelas que se apresentarem com maior validade. No entanto, algumas investigações feitas, e dentre elas a de M. Smith, revelam que um teste organizado com a validação de todas as suas questões pode não se apresentar globalmente tão válido como aquelas. Convém não esquecer que, ao planejarmos a organização de um teste, já validamos as questões que o compõem.

Várias técnicas têm sido propostas para a validação das questões. Até 1923, a validação de testes coletivos pela correlação com o critério se limitava ao teste como um todo, ou às partes de que ele se compunha (subteste). Nenhuma atenção às questões dos subtestes; elas apenas deveriam variar em dificuldade. O aparecimento, em 1923, da Otis Self-Administering marcou uma nova fase, pois cada questão foi validada separadamente. Foi também a primeira vez que os elementos de um teste coletivo de inteligência foram validados com um critério diferente da I. C. Em 1924, Leona Vincent propôs nova técnica para validação das questões. Em 1926, Cleeton empregou duas técnicas, simultaneamente, para validação: a que foi empregada por Otis e uma outra, original. Todavia, não discutiu a eficiência do trabalho empreendido; Thorndike, no mesmo ano, retomou a emprego da correlação *bisserial*, anteriormente também usada por Vincent. Ainda em 1926, McCall publicou sua técnica para validar testes de múltipla escolha, e logo a seguir, Long e Bliss propuseram modificações à técnica de McCall. A partir de então, novas técnicas apareceram. Das de

[16] Há a esse respeito dois trabalhos considerados clássicos, dentre outros: o da Califórnia e o de Chicago.

mais conveniente emprego nos testes coletivos de inteligência, daremos pequeno resumo.

## Cristérios de validação

Organizado o projeto do teste, não podemos afirmar ainda se ele *mede realmente o que pretende medir*, "se o seu objetivo prático, para classificação ou ordenação dos indivíduos, é conseguido numa porcentagem que baste para torná-lo instrumento de confiança" (Lourenço Filho). E essa verificação só poderá ser feita se procurarmos, com o auxílio de outros meios objetivos, um critério seguro para validar o instrumento.

Vários critérios existem para validação. Cada um deles, porém, não é satisfatório por si só. O emprego isolado de um poderá concorrer para deformação dos resultados. Por outro lado, da coerência, interna e externa, do critério de validade dependerá, em grande parte, a validade do instrumento em organização, o que tanto basta para demonstrar o cuidado que devemos ter presente na escolha do critério ou de um grupo destes critérios.

a) *Idade cronológica* – É o mais antigo critério para validação do teste de inteligência. Foi empregado por Binet na organização de seus testes e, bem assim, por dois outros experimentadores de renome: Terman e Kuhlmann. Este critério se baseia na hipótese de que a inteligência cresce no mesmo indivíduo à medida que ele fica mais velho; e ainda na hipótese de trabalho que a distribuição da inteligência em um grupo numeroso homogêneo e não selecionado seja igual a de um outro grupo, nas mesmas condições. E é por essa razão que, nos testes de inteligência, o valor da norma cresce em valor absoluto de idade a idade.

Esse critério, porém, não deve ser o utilizado como exclusivo. Apresenta falhas e dificuldades de execução. Não é também o mais empregado hoje.

b) *Grupos conhecidos* – Este critério também foi empregado por Binet. Por grupos conhecidos, entendemos aqueles que foram classificados por meio de outras provas de inteligência ou pelo consenso geral. Aplica-se o teste sucessivamente a três grupos conhecidos: inframédio, médio e supramédio. O teste deverá discriminar, e as diferenças de resultados entre os grupos deverão ser significativas. Acontece que uma objeção poderá ser prontamente levantada: quem garantirá a validade do consenso geral? Servirá ele de critério? Ninguém pode afirmar com segurança.

Quando nos utilizamos de outras provas de inteligência, e fundamentamos o julgamento nos seus resultados, então, o critério terá valor menos discutível. No entanto, o simples fato da discriminação de três grupos não será o bastante para garantia do critério. Estamos, pois, em face de um critério que não pode ser empregado sem restrições.

c) *Julgamento de especialistas* – Este é um critério muito em uso. Dentre um grande número de questões, alguns especialistas escolhem as questões que devem medir a inteligência. Compõe-se, em seguida, o teste. Segundo Ruch e Stoddard, este método é muito usado e aconselham mesmo o seu emprego. Segundo eles, já verificou por experiência que, no julgamento do verdadeiro valor e dificuldade das questões, a média entre os julgamentos de um grupo de três a dez juízes cuidadosos é superior ao de um único. Devem os especialistas distribuir as questões em três categorias: satisfatória, regularmente satisfatória, e não aproveitável. Em seguida, distribuir as primeiras e as segundas, respectivamente, pela ordem de dificuldade.

Ora, tal critério supõe os julgamentos dos especialistas como uma espécie de elementos iguais e adicionáveis. Ainda mais: quando as questões forem em grande número, darão uma amostra que poderá ser mais representativa do comportamento inteligente. No entanto, o julgamento dos especialistas já representava uma tentativa, pelo menos, de validação. Seria comparar a coisa a ser julgada com a própria coisa.

Por outro lado, sabemos que esse julgamento não é de valor notável, nem pela sua constância nem pela correlação com os resultados do teste. Isso vem justamente demonstrar que esse critério pode ser usado mais como ponto de referência do que como denominador comum.

d) *Julgamento dos professores* – O julgamento dos professores sobre a inteligência dos seus alunos tem também sido usado como critério para validação de testes, na suposição de que esses julgamentos mereçam confiança. Tal não é o nosso parecer. Em geral, os professores conhecem a inteligência de *alguns* alunos. Na maio-

ria dos casos, porém, a sobreestimam. E, na mesma série, os julgamentos entre os diversos professores divergem muito. Para o de Geografia, mais inteligentes poderão ser aqueles que melhor souberem desenhar mapas; para o de Português, poderão ser os mais imaginosos... Há assim uma infinidade de classificações, cada uma dependendo do critério subjetivo do professor. O exame dos resultados das experiências levadas a efeito não nos autoriza a levar em consideração este critério.

e) *Rendimento escolar* – Como critério para validação dos testes de inteligência, o rendimento escolar tem sido muitas vezes empregado. Este critério está baseado na suposição de que os mais inteligentes são aqueles que obtêm os melhores resultados, e os menos inteligentes são aqueles que apresentam baixos resultados no aproveitamento escolar. Várias objeções poderão desde logo ser levantadas. Merecerão fé as notas atribuídas pelos professores? Não. As notas atribuídas pelos professores, desde que para tal se utilizem de provas clássicas, não são dignas de confiança. São numerosas e altamente probantes as verificações a esse respeito.[17]

Se examinarmos as pesquisas de Symonds, Jordan e Wilson em relação ao emprego desse critério, citadas por Pintner, verificaremos que ele não merece confiança.

Quando, em vez de provas clássicas, os professores empregam provas objetivas, o critério passa a ser mais digno de merecer confiança. De fato, o coeficiente de correlação entre os resultados de provas objetivas e testes de inteligência é alto. Pintner nos dá os resultados de 14 coeficientes de correlação calculados entre aproveitamento escolar, aferido pelo julgamento do professor e provas clássicas, e testes de inteligência. Deles, apenas um é superior a 50. O mesmo especialista nos dá os resultados de 15 coeficientes de correlação calculados com os resultados de provas objetivas e de testes de inteligência. Deles, apenas três estão abaixo de 50. Mas, mesmo que se fundamente em provas objetivas, este critério deve ser de uso limitado. Não se pode usar uma prova objetiva como critério único para validação de testes de inteligência. Ademais, esses dois tipos de provas verificam coisas diversas, não se podendo concluir de uma pelos resultados da outra.

f) *Provas já validadas* – Este critério é de grande emprego na validação de testes de inteligência. Para efeitos práticos, dividiremos em duas partes: teste Binet-Simon (BS) e qualquer outro teste de inteligência já validado.

I) Binet-Simon – O teste BS é muito empregado como critério, e principalmente se estivermos convencidos de que a BS é a melhor medida da inteligência. Nesse caso, o coeficiente de validade deve ser superior a 70.

II) Outro teste – Outro teste de inteligência poderá ser empregado como critério, desde que mereça confiança.

Julgamos também de boa técnica o emprego simultâneo de vários critérios, porque isso nos habilitará a chegar a uma determinação mais segura da validade do teste.

## Técnicas de validação[18]

a) *Correlação bisserial* – É um método que se aplica a dados em que uma variável é quantitativa e contínua e a outra é apresentada em classificação dicotômica. Assim, aplicamos essa técnica para calcular a correlação entre os resultados do critério e o acerto ou erro nas questões de um teste. Os resultados do critério dão uma variável contínua; as respostas a uma questão constituem a segunda variável: *certo ou errado*.

A fórmula é a seguinte:

$$r_b = \frac{(M_c - M_e)_{Pq}}{D.\,P. \times Z}$$

$M_c$ = média do resultado do critério do grupo que acertou a resposta.

$M_e$ = média do resultado do critério do grupo que errou a resposta.

D.P. = desvio-padrão de todos os resultados do critério.

p = porcentagem dos que acertaram a resposta.

q = 1 – p

Z = ordenada da curva normal, sem a consideração dos q.

Esta técnica tem a desvantagem de ser muito demorada quando tivermos um número de observações muito elevado.

b) *Técnica de W. McCall* – Esta técnica foi a empregada por McCall para validação das questões de sua *"Multi-mental Scale"*. Como se verifica, é para validação dos testes de múltipla escolha, pois em

[17] As pesquisas de Starch, Elliot e outros, na América, demonstram, de maneira impressionante, a inexatidão e a variabilidade das notas atribuídas pelos professores. Essas pesquisas cobriram principalmente as matérias do curso secundário, como Matemática, Inglês, História (cf. Starch. *Educational Measurements*. New York : Macmillan, 1918). A esse respeito foram também levados a efeito trabalhos dessa natureza por E. Siqueira, em São Paulo, e Lourenço Filho, no Rio.

[18] Para estudo das técnicas de validação, ver na bibliografia Symonds; Long e Sandiford; Barthelmess; Ruch; e Smith.

sua escala todas as questões são de múltipla escolha. Essa técnica está baseada na hipótese de que a questão mais válida é aquela que divide os candidatos de acordo com os resultados do critério, em dois grupos nitidamente homogêneos. Deve-se notar que essa técnica é defeituosa, pela razão de não aceitar a predeterminação da resposta, entre as três, quatro, ou cinco possíveis.

A fórmula é a seguinte, de acordo com H. M. Barthelmess:

$$V = \frac{\Sigma\ [\ Fx\ (Y'x - Y')\ ]}{N}$$

Y' é a média dos resultados do critério do grupo total;

Y'x é o desvio-padrão em relação à média do resultado do critério com referência à resposta própria do grupo;

Fx é a frequência dessa resposta particular;

N é o número total de alunos.

Segundo Long e Sandiford, a fórmula de McCall é absurda, pois leva a valores negativos, em virtude de não ser possível a operação indicada pelo fator Y'x – Y'. Considerando tal defeito, os autores acima sugeriram o seguinte: modificar aquele fator apenas por Y'x, que será o desvio, sem atenção ao sinal, do resultado do critério de um grupo de resposta própria ou característica, em relação à média do resultado do critério de todo o grupo.

Conforme se disse antes, a técnica de McCall é para testes de múltipla escolha. Long e Sandiford adaptaram-na para outros tipos de testes:

$$V = \frac{f_2\ (M_2 - M) + f_1\ (M - M_1)}{N \times D.P.}$$

$M_2$ = média do resultado critério do grupo que acertou

$M_1$ = idem, idem, que errou

M = média do resultado critério de todo o grupo

$f_2$ = frequência dos que responderam acertadamente

$f_1$ = idem, idem, erradamente

D.P. = desvio-padrão dos resultados do critério

c) *Técnica de Vincent* – Em 1924, Leona Vincent estabeleceu uma técnica para validação das questões de um teste, consistindo na comparação de dois grupos, por meio da medida de superposição dos resultados respectivos. O valor de validade de uma questão é a porcentagem daqueles que erram a questão e que tenham mais altos resultados critérios do que o resultado critério mediano daqueles que acertam a mesma questão. Quanto menor a superposição, maior será o grau de validade da questão. Long e Sandiford demonstraram que essa técnica leva muitas vezes a resultados absolutamente inexatos, embora tenha ela a vantagem da facilidade de cálculo.

d) *Técnica corrigida de Vincent* – Não se trata, na verdade, de uma técnica original, mas de uma adaptação da técnica anterior. No presente caso, o valor de validade é a porcentagem daqueles que acertam a resposta e que tenham resultados mais baixos do que o resultado critério mediano daqueles que erram a mesma resposta. A mesma crítica feita por Long e Sandiford à técnica anterior aplica-se igualmente neste caso.

Barthelmess, já referida, propôs o emprego simultâneo dessas duas técnicas para validação das questões, calculando-se a média dos dois valores de validade encontrados.

e) *Técnica de Long-Bliss* – Essa técnica foi idealizada por dois discípulos de McCall, Long e Bliss, num esforço para eliminar os defeitos da técnica de seu mestre.

A fórmula é a seguinte:

$$C.\ L.\ B. = \frac{(m_1 - m_2)\ f_1\ f_2 + (m_1 - m_3) f_1\ f_3 + \ldots}{D.P. \times N^2} + \frac{(m_{n-1} - m_a)\ f_{n-1}\ f_n}{D.P. \times N^2}$$

Sendo $m_1$, $m_2$, $m_3$... m as médias dos resultados critérios, em ordem de grandeza do mais alto para o mais baixo, das várias respostas da questão; $f_1\ f_2\ f_3$ as freqüências das respectivas respostas; DP o desvio-padrão de todos resultados do critério em questão e N, o número de resultados do critério.

Essa fórmula se aplica apenas aos testes de múltipla escolha. Quando houver erro ou acerto, a fórmula passará a ser

$$V = \frac{(M_2 - M_1)\ f_2\ f_1}{D.P. \times N^2}$$

f) *Técnica de Clark* – A técnica de Clark foi proposta para validar testes de conhecimentos em psicologia. No entanto, ela pode ser empregada para validar questões de testes de inteligência:

$$V = \frac{P - D}{1 - D}$$

D = proporção dos que erram a resposta

P = proporção dos indivíduos que erram, no grupo critério

g) *Técnica de Long* – Esta técnica foi publicada em 1934, e é muito semelhante à de Vincent em seus fundamentos, eliminando, no entanto, defeitos desta.

$$V = 1 - \frac{2 \Sigma \text{ acertos sob erros}}{N_a \times N_e}$$

Esta técnica, como se vê, é de fácil aplicação. E seus resultados satisfazem perfeitamente o objetivo.

Muitas outras técnicas existem para validação das questões de um teste de inteligência, tais como a de Henry, Cook, Symonds e Kelley. Contam-se por 22 técnicas. Julgamos, porém, que as apresentadas são as que se aplicam mais adequadamente aos testes coletivos de inteligência.

## Formas equivalentes

Depois do estudo preliminar da validação das questões do teste, podemos verificar a possibilidade de organização de *formas equivalentes* para o instrumento em questão. Muito se tem discutido sobre a verdadeira significação de forma equivalente. Por vezes, apela-se exageradamente para sua organização. Autores há que julgam que um teste deve ter, pelo menos, três formas equivalentes para que possa ser considerado bom. Julgamos que bastem duas, nos testes coletivos de inteligência.

Organizamos as formas equivalentes do seguinte modo: 1) depois de terminado o trabalho inicial de validação das questões, serão retiradas ou substituídas as que apresentarem baixa validade; 2) as questões restantes deverão ser distribuídas em ordem de dificuldade crescente; 3) constituir-se-ão, então, duas formas-teste do seguinte modo (Ruch e Stoddard):

| Forma A | Forma B |
|---|---|
| 1 | 2 |
| 4 | 3 |
| 5 | 6 |
| 8 | 7 |
| 9 | 10 |
| 12 | 11 |
| 13 | 14 |
| | etc. |

Depois desse trabalho, reaplicaremos o teste, nas duas formas. Verificamos se as diferenças entre as duas distribuições não são significativas. Neste caso, as duas formas podem ser chamadas de equivalentes. Devemos ter cuidado neste passo da organização porque, em caso contrário, o treino na resolução da *Forma A* poderá influir nos resultados da *Forma B.* Para evitar possibilidade dessa causa de erro, devemos dividir o grupo em dois subgrupos semelhantes. Ao subgrupo A daremos primeiro a *Forma A*, e a seguir a B. Ao subgrupo B, *Forma B* e, depois, a *Forma A*. Atenuaremos, desse modo, a possível influência do treino. E os resultados de uma forma poderão ser comparados aos da outra.

## Tempo, sua fixação

O problema da fixação do *tempo-limite* nos testes de inteligência é ainda um problema em aberto. E, sem dúvida, muito trabalharão os especialistas antes do acordo geral. Para uns, a fixação do tempo-limite é fonte de injustiças com relação aos indivíduos vagarosos, não obstante capazes. Não há o que negar a esse res-

peito. Contornaremos essa dificuldade se levarmos a fixação do tempo-limite, quando cerca de 90% ou 95% tiverem *tentado* todas as questões (Ruch e Stoddard). W. Lippmann, citado por esses especialistas, julga que mesmo os menos capazes obterão resultados superiores desde que tenham tempo suficiente. A experiência tem demonstrado que mesmo com tempo suficiente os menos capazes não apresentam resultados muito superiores aos que dariam sem o mesmo limite de tempo. Sugerimos uma solução para fixação do tempo-limite. Trata-se de uma adaptação de uma proposta de Ruch e Stoddard:

a) separam-se os indivíduos em dois grupos, A e B.

b) o grupo A começará o trabalho pelo início do teste; e o grupo B pelas questões do fim;

c) a cada grupo distribuem-se lápis de diversas cores: preta, azul, vermelha, etc.

d) dado o sinal de início do trabalho, os grupos começarão a trabalhar com um ou dois lápis; dez minutos depois, todos mudarão o lápis; decorridos mais cinco minutos, novo lápis, e assim sucessivamente. Desse modo poderemos ter o resultado de cada indivíduo na base de 10, 15, 20, 25 minutos, e bem assim o resultado de cada grupo. E isso nos permitirá o estudo-velocidade dentro de cada grupo, o que facilitará a fixação do tempo-limite para o teste. Por tempo-limite entendemos o tempo máximo para resolução de um teste. E esse tempo deve ser o necessário para que uma porcentagem entre 70% e 80% tentem todas as questões do teste.

## Correção das questões

O problema da correção das questões tem grande importância para a coerência do instrumento. A questão se apresenta da seguinte forma – como devemos penalizar? A correção, em alguns testes, deve ser feita do seguinte modo:

Sejam:

N – Número de questões do teste
C – Resultado do critério
$R_c$ – Respostas certas
$R_e$ – Respostas erradas
R – Resultado final

$$R = R_c + KR_e \qquad (a)$$

em que K é o peso aos erros e tem sinal negativo. Assim,

$$N = R_c + R_e$$
$$R_e = N - R_e$$

substituindo $R_e$ em (a) temos:

$$R = R_c + K(N - R_c)$$
$$R = R_c(1 - K) + KN$$

Mas KN é uma constante e, desse modo, o coeficiente de correlação não será alterado se adicionarmos uma constante a uma das variáveis, e assim

$$r_{CR} = r_C [R_c (1 - K) + KN] = r_C [R_c (1 - K)]$$

Mas 1–K é também uma constante e uma vez que a correlação não é alterada multiplicando-se uma das variáveis por uma constante, teremos

$$r_{CR} = r_C [R_c (1 - K)] = r_{CR_c}$$

Isso significa que a correlação entre o resultado critério e o resultado *R* é o mesmo que entre C e $R_c$; portanto, *R* deve ser $R_c$.

Quando se tratar de testes de múltipla escolha, a correção deverá mudar.

Seja N o número de pontos em um teste de múltipla escolha, e t, o número total de questões tentadas. Representará (t – N) o número de questões respondidas ao acaso; *n*, o número de alternativas em cada questão; $\frac{(t - N)}{n}$ será a média de questões respondidas corretamente ao acaso; C, as respostas certas; e E, as erradas.

Então,

$$C = N + \frac{t - N}{n}$$

$$E = t - \left[ N + \frac{t - N}{n} \right]$$

mas, t = C + E,
donde

$$E = E + C - \left[ N + \frac{C + E - N}{n} \right]$$

$$nC - nN - C - E + N = 0$$

$$C(n - 1) - E = N(n - 1)$$

$$N = C - \frac{E}{n - 1}$$

## Dificuldades das questões

A verificação da dificuldade das questões não é indiferente ao trabalho de validação do teste e da procura de sua fidedignidade. Pelo contrário, são simultâneos. É erro freqüente dos organizadores classificar as questões em médias, fáceis e difíceis. Essa classificação só poderá *decorrer de verificação experimental*. Segundo Monroe e Engelhart, Thurstone julga que uma questão tem valor discriminativo quando for respondida por uma porcentagem compreendida entre 30% e 70% dos indivíduos. Symonds julga que a melhor questão é aquela que apresentar 50% de acertos.

## Validade e extensão do teste

A validade de um teste aumentará, se a esse teste acrescentarmos novas questões, que procurem diagnosticar o mesmo atributo? Por outras palavras, se a validade de um teste não se apresentar como satisfatória, ela melhorará no caso de alongarmos a extensão do teste, desde que as questões acrescidas sejam do mesmo teor que as já existentes?

Vimos que podemos elevar a fidedignidade de um teste aumentando a sua extensão. Do mesmo modo aumentará a validade do teste. O efeito sobre a validade acrescentando mais questões ao teste pode ser calculado pela seguinte fórmula:

$$r_{cnx} = \frac{r_{cx}}{\sqrt{\frac{1 - r_{xx}}{n} + r_{xx}}}$$

em que $r_{cx}$ é o coeficiente de validade; $r_{xx}$, é a fidedignidade do mesmo teste; e *n*, o número de vezes de que ele foi aumentado.

Muitas vezes, desejamos saber da *validade virtual* de um teste. Isto é, o limite para que tenderá a validade, no caso de aumentarmos o teste indefinidamente, ou aplicarmos um número infinito de formas equivalentes. Bastará que, na fórmula acima, façamos *n* tender para o infinito, e, então, teremos:

$$r_{(xx)c} = \frac{r_{cx}}{\sqrt{r_{xx}}}$$

# Conclusões

Com o presente trabalho não podemos ter a pretensão de haver esgotado os problemas levantados, na teoria e na prática, com relação à *fidedignidade* e à *validade* dos testes coletivos de inteligência.

No geral, essas importantes questões têm sido discutidas para as provas objetivas ou testes. Mas é evidente que problemas particulares existem para modalidades especiais de certas provas, com objetivos também específicos. E, dentre elas, não há dúvida que a de maior importância prática, pelo vulto de suas aplicações, é a dos testes coletivos de inteligência, o que justifica a escolha do assunto desta monografia.

Da bibliografia variada e abundante, sobre a matéria e de que damos aqui apenas um resumo, bem como dos ensaios e experiências do autor, obtivemos as seguintes conclusões:

a) o problema geral da aferição dos testes não é independente da questão de sua validade e fidedignidade;

b) já do ponto de vista teórico, já do ponto de vista da prática, as questões de validade e de fidedignidade também não se separam de modo completo;

c) a fidedignidade, ou coerência interna, consiste na qualidade que um teste pode ter de medir em cada uma de suas partes o que outra parte equivalente também mede;

d) essa equivalência pode não depender da forma de apresentação das questões, de sua posição relativa e da extensão de cada uma das partes do próprio teste, sendo certo, porém, que esses fatores devem ser apreciados na composição dos ensaios preliminares;

e) a validade, ou coerência externa, consiste na eficiência prática com que um teste realmente meça o atributo para cuja apreciação esteja preparado;

f) a avaliação do grau dessa eficiência dependerá, antes de tudo, da autenticidade de um critério; e a perfeição dessa avaliação, do emprego hábil de uma fórmula de correlação;

g) os processos de verificação estatística permitem, desde que convenientemente aplicados, exprimir por índices numéricos o grau de confiança que podemos atribuir a um teste coletivo de inteligência, verificando-se, porém, e de modo especial, quanto aos problemas de validação, que a interpretação desses índices não deve resultar simplesmente da aplicação automática de fórmulas;



R. bras. Est. pedag., Brasília, v. 79, n. 193, p. 113-134, set./dez. 1998

h) a aferição geral de um teste e, em particular, de um teste coletivo de inteligência dependerá, portanto, não só de um tratamento quantitativo de amostras representativas de um universo, mas também da acuidade com que o especialista atenda à variedade e à complexidade dos problemas que a questão apresenta.

## Referências bibliográficas


L'ANNÉE psycologique, v. 35, 1934.

BALLARD, P. B. *Mental tests.* Londres : Hodder and Stoughton, 1920.

_____. G*roup tests of intelligence.* Londres : Hodder and Stoughton, 1922.

BARTHELMESS, H. M. *The validity of intelligence test elements.* New York : Columbia University, 1931.

BINGHAM, W. V. D. *Aptitudes and aptitude testing.* New York : Harper and Brothers, 1937.

BOYTON, P. L. *Intelligence* : its manifestations and measurement. New York : Appleton, 1933.

BROWN, W., THOMSON, G. H. *The essentials of mental measurements.* 3. ed. Cambridge University, 1925.

CHAPMAN, J. C., DALE, A. B. Further criterion of the selection of mental test elements. *The Journal of Educational Psychology*, n. 5, p. 267-276, 1922.

CLAPARÈDE, Edouard. *La genèse de l'hypothèse.* Genève : Kundig, 1934.

_____. *La educación funcional.* Tradução de M. Rodrigo. Bilbáo : Espasa-Calpe, 1932.

COLVIN, S. Principles underlying the construction and use of intelligence tests. In: NATIONAL SOCIETY FOR THE STUDY OF EDUCATION. *21th yearbook*. Bloomington, 1923.

EDGERTON, H. A., TOOPS, H. A. A table for predicting the validity and reliability coefficients of test when lengthened. *Journal of Educational Research*, n. 3, p. 225-234, 1928.

FESSARD, A. Precision et coherénce dans les éxamens par tests. *L'Année Psychologique,* n. 28, 1927.

FREEMAN, F. N. *Mental Tests* : their history, principles and applications. New York : Houghton Mifflin, 1926.

_____. The individual in school : special abilities and their measurements. In: MURCHISON, Carl A. (Ed.). *Foundations of Experimental Psychology.* Worcester : Clark University, 1929.

FRÖBES, Joseph. *Tratado de Psicologia Experimental.* Tradução de José A. Menchaca. 2. ed. Madrid : [Huelves y Compañia], 1933.

FRYER, D., HENRY, E. *An Outline of General Psychology.* New York : Barnes and Nobles, 1936.

GARRETT, H. E. *Statistics in Psychology and Education.* 2. ed. New York : Longmans, Green, 1937.

GARRETT, H. E, SCHNECK, M. R. *Psychological tests, methods and results.* New York : Harper and Brothers, 1933.

HULL, C. L. *Aptitude testing.* New York : World Book, 1928.

JORDAN, R. C. An empirical study of the reliability coefficient. *The Journal of Educational Psychology*., n. 4, p. 307-311, 1935.

KELLEY, T. L. *Interpretation of educational measurements.* New York : World Book, 1927.

KELLEY, T. L, SHEN, E. The statistical treatment of certain typical problems. In: MURCHISON, Carl A. (Ed.). *Foundations of Experimental Psychology.* Worcester : Clark University, 1929.

LEVINE, A. J., MARKS, L. *Testing intelligence and achievement.* New York : Macmillan, 1928.

LINCOLN, Edward, WORKMAN, L. L. *Testing of test results.* New York : Macmillan, 1935.

LONG, J. A., SANDIFORD, P. *The validation of test items.* Toronto : University of Toronto Press, 1935.

LOURENÇO FILHO. *Testes ABC*. 2. ed. São Paulo : Melhoramentos, 1937.

McCALL, W. A. *How to measure in education.* New York : Macmillan, 1922.

_____. *How to experiment in education.* New York : Macmillan, 1923.

MELLI, R. *Recherches sur les formes d'intelligence.* Genève : Kundig, 1930.

MONROE, W. S. *An introduction to the theory of educational measurements.* New York : Houchton Mifflin, 1923.

MONROE, W. S, ENGELHART, M. D. *The scientific study of educational problems.* New York : Macmillan, 1936.

ODOROFF, M. E. A correlational method applicable to the study of the time factor in intelligence tests. *The Journal of Educational Psycology*, n. 4, p. 307-311, 1935.

OTIS, A. I. *Statistical method in educational measurement.* New York : World Book, 1925.

PENNA, J. B. D. Iniciação ao estudo da medida da inteligência. *Revista de Educação*, São Paulo, n. 5, p. 7-85, 1934.

PETERSON, J. *Early conceptions and tests of intelligence.* New York : World Book, 1925.

PIÉRON, H. Le problème de l'intelligence. *Scientia*, v. 12, n. 1, 1927.

_____. *Le dévelopment mental et l'intelligence.* Paris : Alcan, 1929.

PIÉRON, H., FESSARD, A. La notion de valité. *L'Année Psychologique,* n. 31, 1930.

PINTNER, R. The individual in school: general ability. In: MURCHISON, Carl A. (Ed.). *Foundations of Experimental Psychology.* Worcester : Clark University, 1929.

_____. *Intelligence testing* : methods and results. New York : Henry Holt, 1932.

REY, A. *Réflexions sur le problème du diagnostic mental*. Genève: Université de Genève, 1935.

RUCH, G. M. *Objective or new type examination*. New York : Scott, Foresman, 1929.

RUCH, G. M., STODDARD, G. D. *Tests and measurements in high-school instruction.* New York : World Book, 1927.

RUGG, H. Statistical methods applied to educational testing. In: NATIONAL SOCIETY FOR THE STUDY OF EDUCATION. *21th yearbook*. Bloomington, 1923.

SKAGGS, E. R. *An elementary textbook of mental measurements.* Michigan : G. Wahr, 1923.

_____. Some critical comments on certain prevailing concepts and methods used in mental testing. *The Journal of Applied Psychology*, n. 6, p. 503-508, 1927.

SMITH, C. E. *The construction and validation of a group test of intelligence using the Spearman technique.* Toronto University, 1935.

SMITH, M. *The relationship between item validity and test validity.* New York : Columbia University, 1934.

SPEARMAN, C. *The nature of "intelligence" and the principles of cognition.* 2. ed. New York : Macmillan, 1927.

_____. *The abilities of man* : their nature and measurements. New York : Macmillan, 1927.

STERN, W. *The psychological methods of testing intelligence.* Tradução de G. M. Whipple. New York : Warwick and York, 1914.

SWINEFORD, F. Biserial r versus Pearson as measures of test : item validity. *The Journal of Educational Psychology*, n. 6, p. 471-472, 1936.

SYMONDS, P. M. *Measurements in secondary education.* New York : Macmillan, 1934.

_____. Choice of items for a test on the basis of difficulty. *The Journal of Educational Psychology*, n. 7, p. 481-493, 1929.

THORNDIKE, Edward L. *An introduction to the theory of mental and social measurements.* 2. ed. New York : Columbia University, 1922.

_____. *The measurement of intelligence.* New York : Columbia University, 1926.

THURSTONE, L. L. *The reliability and validity of tests.* Ann Arbor : Ed. Bros, 1937.

WEST, P. V. The significance of weighted scores. *The Journal of Educational Psychology*, n. 5, p. 302-308, 1924.

WILLOUGHBY, R. R. The concept of reliability. *Psychological Review*, n. 2, p. 153-165, 1935.

YERKES. R. M. (Ed.). *Memoirs of the National Academy of Science*s. Washington : Government Printing Office, 1921. v. 15.

YOAKUM, C., YERKES, R. M. *Army Mental Tests.* New York : Henry Holt, 1923.

Murilo Braga (1912-1952). Sucedeu a Lourenço Filho na direção do Inep, à frente do qual atuou de 1946 a 1952. Sua gestão concentrou-se em dois setores: desenvolvimento de um plano destinado a expandir a rede primária e normal e cursos de aperfeiçoamento para professores do magistério primário.

## Abstract

*The tests are classified by its objectives and application modalities. The first results on collective intelligence tests were published in 1913 and since then they have been strongly attacked. However, the tests play na important role in school administration and organization. The coherency in a measurement instrument is verified by the level of congruity shown between internal reports. Coefficient correlation is applied to verify the practical reliability and authenticity as well as to verify its coherency in consecutive applications. The statatistical verification procedures allow the numerical reports to expose the level of asssurance attributed to the tests.*

*Key-Words: intelligence test; validity; reliability.*