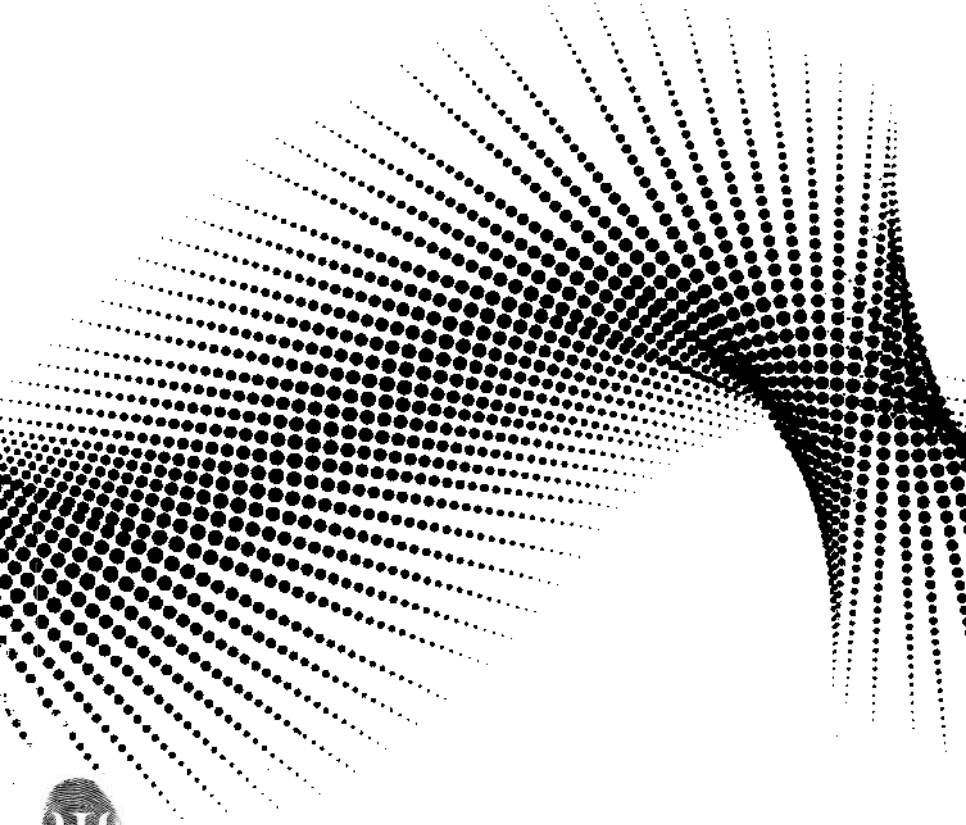
Rodolfo A. M. Ambiel
Ivan Sant'Ana Rabelo
Sílvia Verônica Pacanaro
Gisele Aparecida da Silva Alves
Irene F. Almeida de Sá Leme
(Orgs.)

Avaliação Psicológica

Guia de consulta para estudantes e profissionais de psicologia



Revisado!

Casa do Psicólogo®

Avaliação Psicológica

guia de consulta para estudantes e profissionais de psicologia

Rodolfo A. M. Ambiel
Ivan Sant'Ana Rabelo
Sílvia Verônica Pacanaro
Gisele Aparecida da Silva Alves
Irene F. Almeida de Sá Leme
(Orgs.)

Casa do Psicólogo®
São Paulo
2011

Casa do Psicólogo®

1ª Edição
2011

Editores
Ingo Bernd Güntert e Juliana de Villemor A. Güntert

Revisão Técnica
Rodolfo A. M. Ambiel, Ivan Sant'Ana Rabelo, Sílvia Verônica Pacanaro, Gisele Aparecida da Silva Alves e Irene F. Almeida de Sá Leme

Preparação
Tássia Fernanda Alvarenga de Carvalho

Capa
Marina Takeda

Projeto Gráfico e Editoração Eletrônica
Fabio Alves Melo

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Avaliação psicológica : guia de consulta para estudantes e profissionais de psicologia / Rodolfo A. M. Ambiel...[et al.] . -- São Paulo : Casa do Psicólogo®, 2011.

Outros organizadores: Ivan Sant'Ana Rabelo, Sílvia Verônica Pacanaro, Gisele Aparecida da Silva Alves, Irene F. Almeida de Sá Leme.
Bibliografia.
ISBN 978-85-8040-071-7

1. Avaliação psicológica 2. Psicometria 3. Testes psicológicos
I. Ambiel, Rodolfo A. M. II. Rabelo, Ivan Sant'Ana. III. Pacanaro, Sílvia Verônica. IV. Alves, Gisele Aparecida da Silva. V. Leme, Irene F. Almeida de Sá.

11-04584 CDD-150.287

Índices para catálogo sistemático:
1. Avaliação psicológica 150.287

Impresso no Brasil

Casapsi Livraria e Editora Ltda.
Rua Santo Antônio, 1010
Jardim México • CEP 13253-400
Itatiba/SP - Brasil
Tel. Fax: (11) 4524-6997
www.casadopsicologo.com.br

Sumário

Prefácio PÁG. | 7

Capítulo 1 PÁG. | 11

Da testagem à Avaliação Psicológica: aspectos históricos e perspectivas futuras

Rodolfo A. M. Ambiel

Sílvia Verônica Pacanaro

Capítulo 2 PÁG. | 29

Panorama atual dos testes psicológicos no Brasil de 2003 a 2011

Sílvia Verônica Pacanaro

Gisele Aparecida da Silva Alves

Lívia Sant'Ana Rabelo

Ivonne F. Almeida de Sá Lemme

Rodolfo A. M. Ambiel

Capítulo 3 PÁG. | 49

"E viveram felizes para sempre": a longa (e necessária) relação entre psicologia e estatística

Rodolfo A. M. Ambiel

Jonas Bergmann Moura de Andrade

Lucas de Francisco Carvalho

Clay Vicente Cassepp-Borges

Capítulo 4 PÁG. | 81

Teoria de Resposta ao Item na Avaliação Psicológica

Giulia Valentini

Isabel Ane Faros

Capítulo 5 PÁG. | **109**

Validade e precisão de testes psicológicos
Gisele Aparecida da Silva Alves
Mayra Silva de Souza
Makilim Nunes Baptista

Capítulo 6 PÁG. | **129**

Padronização e normatização de testes psicológicos: simplificando conceitos
Ivan Sant'Ana Rabelo
Leila Brito
Marcia Gabriel da Silva Rego

Capítulo 7 PÁG. | **163**

A ética no uso de testes no processo de Avaliação Psicológica
Maria Cristina Barros Maciel Pellini
Irene F. Almeida de Sá Leme

Sobre os autores PÁG. | **181**

Prefácio

O avanço científico nas diferentes áreas do saber pressupõe avaliação, exigindo esta avaliação instrumentos apropriados para o efeito. A réplica da investigação em diferentes países e em amostras alargadas pressupõe instrumentos ágeis, precisos e válidos. Neste sentido, em qualquer área científica, é fundamental o esforço dos pesquisadores na construção e validação de novos instrumentos de medida. Na Psicologia, enquanto ciência e profissão com responsabilidades sociais, esta avaliação assume papel relevante na tomada de decisões e, como tal, requer instrumentos de avaliação cientificamente aprovados. Em boa medida a qualidade da pesquisa e da prática em psicologia, em ambos os casos sempre dependente da qualidade da informação ou dos resultados que são recolhidos, encontra-se muito associada à própria atualidade, qualidade, confiança e valor dos instrumentos usados. Assim, apesar das críticas relativamente frequentes e universais, algumas vezes justas, e outras vezes injustas, demasiadas vezes incompreensivelmente oriundas do seio da própria psicologia, a consolidação da psicologia em termos de investigação e de exercício profissional acompanha a emergência e o fortalecimento do movimento e história dos testes psicológicos. Não se podendo divinizar os testes, ou lhes dar um estatuto de exclusividade na avaliação psicológica – até porque não precisam de tal estatuto –, os testes psicológicos tiveram e continuam a ter um papel importante no reconhecimento científico e social da psicologia.

Um dos problemas na psicologia, como nas demais áreas, é que ninguém sem sólida formação consegue realizar de forma

apropriada algo de tecnicamente complexo. Por muita psicologia que o psicólogo possa saber, isso não lhe garante por si só a competência no uso dos testes psicológicos. A psicometria como domínio de formação académica e profissional é fundamental ao uso dos testes psicológicos. Algumas das críticas ao método dos testes decorrem da pouca formação por parte dos profissionais e dos próprios críticos. Em qualquer ciência e ramo de atividade, um instrumento é apenas instrumento, e a qualidade de seu uso depende intensamente da sabedoria ou do grau de competência do utilizador. Porque havemos de exigir dos testes psicológicos aquilo que eles não podem dar? É fundamental termos bons testes, contudo é também verdade que um bom teste apenas se rentabiliza nas mãos de um psicólogo competente!

Uma primeira competência por parte do psicólogo é saber quando deve e quando não deve usar um determinado teste psicológico. Esta competência exige conhecer muito bem o teste e conhecer bem o contexto particular em que ele vai ser usado. O uso dos testes psicológicos, por estas razões, exige muito dos psicólogos do ponto de vista técnico e ético, sendo importante que associações científicas e profissionais da psicologia regulem essa utilização. Só com boa formação académica e prática vai o psicólogo reunindo as competências necessárias à realização de boa avaliação, recorrendo, entre outros meios, aos testes psicológicos. Por tudo isto, importa destacar a edição deste livro, entendido como um manual atual na área da psicometria nas mãos de alunos, profissionais e académicos de psicologia. Para além dos académicos que pesquisam em psicometria ou que conduzem estudos de validação de provas psicológicas, este manual serve também aos alunos e aos utilizadores dos testes psicológicos em geral. Também estes precisam conhecer

os procedimentos básicos de estatística utilizados na construção e validação dos instrumentos. Neste livro, tais procedimentos aparecem devidamente enquadrados e justificados pelos conceitos psicométricos de precisão, validade e normas, complementando metodologias clássicas e atuais (teoria da resposta ao item, por exemplo) de sua estimação. Da mesma forma, na organização deste livro atentou-se às considerações éticas na avaliação psicológica. Incluindo-se a avaliação psicológica nos atos psicológicos, seja na investigação, seja no exercício profissional, importa acautelar-se sobre os limites dos instrumentos e da própria avaliação, assim como saber acautelar os direitos dos indivíduos e instituições envolvidos na avaliação.

Pelas ligações pessoais à psicologia e aos psicólogos brasileiros, em particular tendo a avaliação psicológica como um dos motivos dessa aproximação, afirmo o enorme prazer em prefaciar este manual. Felicito os seus autores pela clareza e pelas preocupações pedagógicas colocadas na redação de seus capítulos. Precisamos destes manuais para que alunos, profissionais e acadêmicos ultrapassem as resistências frequentes ao estudo aprofundado da psicometria!

Leandro S. Almeida[1]

[1] Professor Catedrático da Universidade do Minho. Doutorou-se em Psicologia pela Universidade do Porto, em 1987, tendo estagiado na Universidade de Yale, Estados Unidos, e na Universidade Católica de Lovaina, Bélgica, durante a preparação do doutoramento. Leciona e pesquisa sobre inteligência, cognição e aprendizagem, incluindo a construção e validação de provas psicológicas, sendo que algumas dessas provas são estudadas e estão validadas no Brasil. Entre estas destaca-se a Bateria de Provas de Raciocínio (BPR5), em coautoria com Ricardo Primi, editada pela Casa do Psicólogo.

Capítulo 1

Da testagem à Avaliação Psicológica: aspectos históricos e perspectivas futuras

Rodolfo A. M. Ambiel
Sílvia Verônica Pacanaro

A palavra teste, tal como usada em português, originou-se do termo em latim *testis*, que significa testemunha, e, posteriormente, do inglês *test*, com o sentido de prova. Portanto, etimologicamente, realizar um teste é realizar uma prova e dar testemunho de alguma coisa. Nesse sentido, quando se trata de testes psicológicos, seu principal uso é como ferramenta na tomada de decisões que envolvem pessoas, a partir do desempenho ou do autorrelato em provas, questionários ou escalas que avaliem características psicológicas. Embora o surgimento dos testes psicológicos tenha sido registrado no início do século XX, muito antes disso já se fazia verificação e levantamento de características e habilidades

das pessoas, em diversas culturas e em diversos contextos, principalmente em situações relacionadas a selecionar candidatos para algumas funções específicas (Urbina, 2007).

Considerando a longa história dos testes, suas contribuições para o desenvolvimento científico e da prática profissional na psicologia, o presente capítulo pretende fazer uma revisão histórica sobre as origens e as aplicações da testagem, destacando sua importância no processo de reconhecimento da profissão. Em seguida, serão realizadas algumas reflexões sobre a testagem e a avaliação psicológica especificamente no contexto brasileiro.

As origens da testagem psicológica

Há registros de que os procedimentos de avaliação variaram muito ao longo da história, com influências das crenças, das filosofias e das posições políticas próprias de cada época e região, desde o período neolítico, datando de 12.000 a.C., passando pelas culturas egípcias e suméria (10.000 a.C.) até os dias atuais (Barclay, 1991; Van Kolck, 1981; Urbina, 2007). Por exemplo, em 200 a.C., na China, eram realizados concursos públicos, e as provas para seleção envolviam demonstrações de proficiência em música, uso do arco, habilidades de montaria, exames escritos sobre temas relacionados a leis, agricultura e geografia (Urbina, 2007). Especificamente sobre a testagem psicológica, os antecedentes da utilização de procedimentos de avaliação clínica recaem principalmente sobre a psiquiatria, com estudos na Alemanha e na França no início do século XIX, com foco no desenvolvimento de provas para avaliar o nível do funcionamento cognitivo de pessoas com danos cerebrais e com outros transtornos (McReynolds, 1986).

Do ponto de vista do desenvolvimento científico, os testes tiveram grande importância para a psicologia. Em meados do século XIX, os psicofísicos alemães Weber e Fechner deram os primeiros passos em direção ao reconhecimento da psicologia como disciplina científica, cujo grande precursor foi Wilhelm Wundt, com a criação do primeiro laboratório dedicado à pesquisa psicológica, em Leipzig, na Alemanha. Com o crescimento deste e de outros laboratórios no final do século XIX, a psicologia desenvolveu-se cientificamente de forma acelerada, e sua expansão ocorreu com o treinamento de vários pesquisadores de outros países europeus e dos Estados Unidos. Dentre os pesquisadores, estava Francis Galton, que se interessou pela mensuração das funções psicológicas, organizando um laboratório antropométrico em Londres, com o objetivo de coletar dados sobre características físicas e psicológicas das pessoas. A contribuição de Galton para a área da testagem ocorreu de algumas formas, tais como a criação de testes para medida de discriminação sensorial (barras para medir a percepção de comprimento); apito para percepção de altura do tom; criação de escalas de atitudes (escala de pontos, questionários e associação livre) e o desenvolvimento e a simplificação de métodos estatísticos (Anastasi, 1977).

No mesmo sentido, o psicólogo americano James M. Cattell, influenciado por Galton, acreditava que a chave para a compreensão do funcionamento da mente estava nos processos elementares, e, em seus estudos, deu ênfase nas medidas sensoriais. Cattell elaborou uma bateria com testes que investigava áreas relacionadas à acuidade sensorial, ao tempo de reação, à bisseção visual de linha e aos julgamentos sobre a duração de intervalos curtos de tempo, e, também, era aplicada em estudantes universitários com

o objetivo de predizer-lhes o sucesso acadêmico (Logan, 2006). Contudo, foi em 1900 que Binet e Simon, na França, começaram a tecer uma série de críticas aos testes até então utilizados, afirmando que eram medidas exclusivamente sensoriais. O foco da crítica era sobre o fato de que, embora permitissem maior precisão nas medidas, os testes sensoriais não tinham relação importante com as funções intelectuais, fazendo somente referências a habilidades muito específicas, quando deveriam ater-se às funções mais amplas como memória, imaginação, compreensão, entre outras.

Cinco anos depois, os mesmos pesquisadores franceses publicaram o primeiro teste para a mensuração da capacidade cognitiva geral, a Escala Binet-Simon, constituída por trinta itens (dispostos em ordem crescente de dificuldade) com o objetivo de avaliar algumas funções como julgamento, compreensão e raciocínio, detectando, assim, o nível de inteligência em crianças das escolas de Paris. O teste foi desenvolvido a pedido do departamento de educação do governo francês, a fim de identificar as crianças com deficiência intelectual e compor um sistema diferenciado de educação para elas. Essa escala passou por revisão, ampliação e aperfeiçoamento em 1908 e em 1911, um ano antes de Wilhelm Stern propor o Quoeficiente Intelectual (QI), que viria a ser refinado por Lewis Terman, na Universidade de Stanford, nos Estados Unidos, em 1916. Para a obtenção do QI, que exprime numericamente e de forma padronizada a capacidade intelectual dos avaliandos, Stern e Terman sugeriram um cálculo que se baseava na divisão da idade mental (IM) pela idade cronológica (IC) multiplicada por cem. Seus estudos sugeriram que, quando a idade mental ultrapassasse a idade cronológica, a razão resultante levaria a um escore acima de cem. Por outro lado, quando a idade

cronológica ultrapassasse a idade mental, levaria a um escore abaixo de cem (Sternberg, 2000).

Os estudos de Terman, publicados em 1916, além de terem proposto um avanço para os estudos relacionados ao QI, também foram responsáveis pela adaptação da escala francesa para os Estados Unidos, onde passou a se chamar Escala Stanford-Binet. Naquele momento histórico, outros instrumentos inspirados na Stanford-Binet foram construídos nos Estados Unidos, onde a avaliação cognitiva ganhou grande impulso por conta da Primeira Guerra Mundial, ocorrida entre 1914 e 1918 (Dubois, 1970).

Nesse período, a demanda por instrumentos simples, rápidos, de aplicação coletiva e que pudessem captar diferenças de capacidade intelectual de recrutas foi bastante acentuada, o que levou os pesquisadores a desenvolverem o *Army Alpha* e, mais tarde, o *Army Beta*, que se instituiu como um instrumento não verbal de avaliação da inteligência, ou seja, composto por provas que não exigiam leitura ou escrita dos respondentes, podendo ser utilizado com recrutas analfabetos e que não falassem a língua inglesa (Fancher, 1985). Ao final da Primeira Guerra Mundial, os testes *Army Alpha* e *Army Beta*, que eram somente utilizados no exército, foram liberados para uso civil após várias revisões e estudos com pessoas de diferentes faixas etárias e níveis de escolaridade (Anastasi, 1977).

Com o desenvolvimento da testagem psicológica, ocorrido durante o período da guerra, novos testes foram publicados, e houve uma melhora na qualidade dos instrumentos, nos procedimentos de administração e nas pontuações. No bojo desses avanços, a partir da década de 1920, a testagem educacional também ganhou campo e foram desenvolvidas provas para

avaliação de desempenho escolar e de habilidades escolares e acadêmicas específicas, tais como o *School Aptitude Test* (SAT), o *Graduate Record Exam* (GRE), o *Medical College Admission Test* (MCAT) e o *Law School Admission Test* (LSAT).

Como se pode perceber, até por volta de 1930 a testagem estava em plena expansão nos Estados Unidos, em grande parte devido à sua cientificidade e às contribuições para a sociedade em diversos âmbitos. Também se percebe que, até então, o foco estava exclusivamente sobre as capacidades cognitivas e, não por acaso, foram os estudos sobre a inteligência humana que obtiveram os maiores avanços metodológicos, teóricos e científicos nessa época.

O inglês Charles Spearman, que realizou seu doutorado no laboratório de Wundt, em Leipzig, foi um dos principais teóricos da fase inicial da psicometria. Suas contribuições se deram ao aplicar modelos matemáticos ao estudo do funcionamento mental, especialmente com o refinamento do método de correlação, previamente desenvolvido por Karl Pearson, e com o desenvolvimento da técnica de análise fatorial. Lançando mão dessas técnicas estatísticas, Spearman desenvolveu estudos a partir dos quais sugeriu a teoria de que todas as habilidades cognitivas convergiam para uma capacidade geral, o chamado fator g. Por outro lado, Thurstone, em 1938, utilizando-se dos métodos propostos anteriormente, sugeria a existência de habilidades específicas e independentes que não se organizavam em torno de uma habilidade geral. De acordo com Ribeiro (1998), estava implícito nas teorias que o fator geral dependeria de uma energia mental essencialmente biológica e inata, enquanto que os fatores específicos dependeriam da aprendizagem. Essas questões teóricas foram discutidas ao longo de todo o século XX, por Cattell (1940; 1971),

Horn (1991) e, finalmente, por Carroll (1993), que apresentou uma proposta de integração das teorias da inteligência, por meio de um modelo hierárquico das habilidades.

Assim, conforme exposto até aqui, além do desenvolvimento técnico que os testes ajudaram a implementar na psicologia, eles contribuíram para avanços teóricos importantes, uma vez que permitiam que as teorias fossem testadas na realidade. Esse período inicial de desenvolvimento da testagem psicológica ocorreu na Europa e, principalmente, nos Estados Unidos. Na seção a seguir, serão abordados os avanços da área e das práticas no Brasil.

O caminho dos testes no Brasil

A testagem e a avaliação psicológica no Brasil passaram por diversos avanços e dificuldades ao longo de sua história. Pasquali e Alchieri (2001) destacam que a história dos testes psicológicos, no contexto brasileiro, teve um período inicial de grande empolgação e uso indiscriminado, seguido por fases de críticas (e muitas vezes com sentido), para posterior organização e regulamentação do uso, o que ainda está em processo na realidade brasileira.

Pasquali e Alchieri (2001) destacam que o desenvolvimento da testagem e da avaliação psicológica no Brasil passou por cinco grandes fases, tendo início na primeira metade do século XIX. Tais períodos são: produção médico-científica acadêmica (1836-1930); estabelecimento e difusão da psicologia no ensino nas universidades (1930-1962); criação dos cursos de graduação em psicologia (1962-1970); implantação dos cursos de pós-graduação (1970-1987); e emergência dos laboratórios de pesquisa, de 1987

em diante. Nesta última fase consolidaram-se vários eventos científicos em avaliação psicológica em virtude da preocupação de vários grupos de pesquisadores quanto à produção de instrumentos mais sérios e confiáveis.

Embora a psicologia tenha sido "fundada" oficialmente em 1879, por Wundt, na Alemanha, antes disso já havia profissionais interessados na compreensão de processos psicológicos no Brasil, tanto que, ao longo do século XIX, era comum encontrar disciplinas de psicologia em faculdades de medicina. Assim, a partir de um ponto de vista estritamente positivista, observa-se que, nas décadas de 1830 e de 1840, duas teses em faculdades de medicina, versando sobre inteligência, foram defendidas. No início da década de 1900, foram fundados laboratórios de psicologia e adaptação para a realidade brasileira de alguns testes internacionais, tais como o Binet-Simon, realizada pelo médico Isaias Alves, na Bahia. É importante ressaltar que, assim como ocorreu na Europa e nos Estados Unidos, o interesse inicial dos pesquisadores era conhecer os processos psicológicos básicos, relativos principalmente à percepção, voltando, somente nos passos seguintes, o foco para funções cognitivas superiores (Amendola, 2011).

As pesquisas iniciais e as possibilidades práticas do uso dos testes foram entusiasmantes e muito promissoras, tanto que o período compreendido entre 1930 a 1962 foi marcado pelo estabelecimento e pela difusão do ensino da psicologia nas universidades, fazendo parte da grade curricular de diversos cursos, tais como administração, jornalismo, sociologia, medicina, direito, entre outros. Junto a isso, vários laboratórios de pesquisa e institutos de psicologia aplicada, principalmente de seleção e orientação profissional e de condutores, foram fundados no Brasil, ajudando

a desenvolver a testagem, tanto do ponto de vista da pesquisa quanto da prática. Por isso Andriola (1996) afirmou que, até aquele momento, tal período poderia ser considerado como a "fase de ouro" no que se refere à produção científica e à construção de instrumentos de medida, com o desenvolvimento de testes específicos para a população brasileira, embora ainda se recorresse à prática de usar instrumentos importados sem maiores esforços de adaptação para a realidade do país.

O ano de 1962 foi marcado pela oficialização da psicologia como profissão no Brasil, o que se deu pela aprovação da Lei nº 4.119, de 27 de agosto daquele ano. É evidente que essa lei não aconteceu de repente; ao contrário, foi fruto de uma série de avanços ocorridos ao longo de vários anos, os quais fizeram com que a psicologia deixasse de ser uma disciplina aplicada para ganhar um campo próprio (Pereira & Pereira Neto, 2003). A partir dessa Lei, que culminou com a criação do Conselho Federal de Psicologia e de suas sucursais regionais em 1974, os cursos de formação de psicólogos foram oficializados, tendo currículos mínimos estabelecendo os conteúdos básicos a serem ensinados nas graduações. Além disso, organizações como a Fundação Getúlio Vargas e o Instituto de Seleção e Orientação Profissional (ISOP), sob a direção de Emilio Mira y López, fortaleceram-se consideravelmente, junto com outros laboratórios e departamentos que foram fundados nas universidades.

Apesar de ter sido um período produtivo dos pontos de vista científico e político, instalou-se uma crise de ordem ideológica na área, com críticas relacionadas ao uso indiscriminado de testes estrangeiros sem adaptação. Junto a isso, com o aparecimento e o fortalecimento de abordagens mais sociais e humanistas na

psicologia, uma forte oposição às práticas e às técnicas positivistas se apresentou, fato que desencadeou alguns prejuízos relativos à pesquisa e ao ensino dos testes (Padilha, Noronha & Fagan, 2007). É importante ressaltar que muitas das críticas que emergiram nesse momento tinham um fundamento lógico e prático e, de fato, algumas questões trouxeram consequências importantes não só para a psicologia, mas também para as pessoas que se submetiam às avaliações. Entretanto, como ressaltaram Noronha e cols. (2002), tais críticas careciam de fundamentos científicos e partiam, por vezes, exclusivamente para argumentos políticos e até emocionais, com um discurso que perdurou ao longo de pelo menos duas décadas.

Como se percebe, em grande parte, tais críticas faziam sentido naquele momento, devido à baixa qualidade da formação dos alunos em avaliação psicológica, uma vez que nos cursos recém-criados ainda não havia docentes especializados no assunto em contraposição à grande procura dos alunos pelos cursos de graduação em psicologia. A esse propósito, o texto de Pasquali e Alchieri (2001) relata que, em dez anos de existência da profissão, a quantidade de cursos oferecidos subiu de seis para 21, contando, em 1970, com oito mil alunos cursando.

Dada a situação da formação dos profissionais e dos pesquisadores no Brasil, foram observadas algumas ações no sentido de promover cursos de pós-graduação em psicologia por diversas importantes universidades, tais como as Pontifícias Universidades Católicas do Rio de Janeiro, do Rio Grande do Sul e de São Paulo, a Universidade de Brasília e a Universidade de São Paulo. Entretanto, isso não foi suficiente para que os testes, que surgiram como ferramentas promissoras e rapidamente foram

difundidos em vários âmbitos, se aprimorassem técnica e cientificamente; ao contrário, com práticas de utilização abusivas e sem as devidas reflexões e formação, os testes psicológicos, em meados das décadas de 1980 e de 1990, foram motivo de reportagens e manifestações públicas contrárias ao seu uso, expondo situações vexatórias de atuações de profissionais da psicologia, principalmente no âmbito da seleção de pessoal. Infelizmente, havia muito de verdade no que foi veiculado. Contudo, chegar a tal situação fez com que a área ganhasse um novo impulso, buscando garantir a qualidade da formação dos profissionais e dos docentes, por um lado, e dos instrumentos e dos testes, por outro.

Uma das iniciativas tomadas nesse sentido foi a fundação do Instituto Brasileiro de Avaliação Psicológica (IBAP), por parte de psicólogos pesquisadores que tinham em comum o fato de conduzirem estudos relacionados à construção, à adaptação e à validação de testes psicológicos no Brasil. Desde seu surgimento, o IBAP tem promovido ações em prol de uma melhor qualidade dos testes e da avaliação psicológica no Brasil, por meio de publicação de uma revista científica (*Revista Avaliação Psicológica*), bem como da promoção de congressos e eventos que fomentam a produção científica e a reunião de profissionais em torno do tema (Gomes, 2003; Hutz & Bandeira, 2003).

Outra tendência observada foi o oferecimento de linhas de pesquisa em avaliação psicológica em cursos de pós-graduação *stricto sensu*, ou seja, mestrado e doutorado. Segundo Primi (2010), dos 65 programas existentes em universidades de psicologia no Brasil, há linhas de pesquisa na área na Universidade Federal de Minas Gerais (UFMG), na Universidade Federal de Uberlândia (UFU), na Universidade Federal do Rio Grande do

Sul (UFRGS), na PUC-RS, na Universidade Federal de Santa Catarina (UFSC), na USP (*campi* de Ribeirão Preto e São Paulo), na PUC de Campinas, na UnB e na Universidade São Francisco (USF), sendo que esta última foi a primeira instituição cuja área de concentração é específica em Avaliação psicológica e desenvolvimento de testes.

Na história recente da avaliação psicológica no Brasil, o fato que se destaca é a publicação da resolução 02/2003, que instituiu critérios mínimos de qualidade para se considerar um teste psicológico apto para o uso profissional. Basicamente, esses critérios dizem respeito à fundamentação teórica do teste, evidências empíricas, ou seja, obtidas a partir de pesquisas científicas, da validade e da precisão do teste (esses conceitos serão abordados no capítulo 4 deste livro), dos sistemas de correção e interpretação dos resultados (veja o capítulo 5) e da compilação de todas essas informações em um livro chamado *Manual Técnico*, que compõe o material do teste.

Essa resolução foi importante por alguns motivos. Como já citado, até o início da década de 2000, os instrumentos careciam fortemente de pesquisas que atualizassem seus conteúdos. É possível imaginar que, se o Conselho estabeleceu os critérios mínimos de qualidade citados anteriormente, muitos dos testes disponíveis antes dessa data não apresentavam muitas dessas informações. O leitor, ao estudar este livro na íntegra, vai perceber que é bastante difícil (e arriscado) pensar no uso de testes psicológicos que não apresentem os requisitos mínimos para considerar um teste psicológico aprovado.

Dessa forma, a Resolução 02/2003 disciplinou a construção e a adaptação de testes psicológicos no Brasil, fornecendo diretrizes

claras aos pesquisadores. Isso fez com que muitos testes utilizados até aquele momento e que não se enquadravam no novo padrão de qualidade fossem retirados do mercado, o que provocou uma série de críticas por parte de profissionais que estavam perigosamente acostumados ao uso desses instrumentos. Por outro lado, conforme poderá ser observado no próximo capítulo, o aumento nas pesquisas a partir da resolução foi considerável, e por um motivo muito simples: a partir daquele momento ou se fazia pesquisa com os instrumentos, de acordo com as novas diretrizes, ou não haveria mais instrumental disponível para os psicólogos realizarem avaliação.

Considerações finais

Como se pode observar, a história da testagem e da avaliação psicológica é muito rica e legitima os investimentos atuais e os desenvolvimentos recentes, enquanto área prática de atuação do psicólogo. Porém, o desenvolvimento deve continuar e ainda há muito o que fazer. Nesse sentido, o ano de 2011 promete ser bastante produtivo, uma vez que o Conselho Federal de Psicologia o instituiu como o Ano Temático da Avaliação Psicológica.

Nessa ação, o objetivo é promover debates ao longo de todo o ano em eventos em todo o país, abertos a todos os psicólogos interessados, que poderão propor e discutir melhorias para área. Para tanto, foram definidos três eixos que organizarão os debates, quais sejam:

1) Qualificação, que versa sobre a formação do psicólogo em avaliação psicológica;
2) Relações institucionais, que visa a debater a inserção da avaliação nos diversos âmbitos de atuação da psicologia;
3) E o terceiro eixo, que organizará as reflexões a respeito da relação entre os dois eixos anteriores.

A partir de tais discussões, deverão ser publicados documentos com os resultados e, talvez, até novas resoluções.

Não é possível prever o futuro da avaliação psicológica no Brasil, mas o estudo do passado certamente pode auxiliar a compreender o estado atual e a indicar alguns possíveis passos futuros. Por exemplo, em acréscimo às contribuições de Pasquali e Alchieri (2001), não seria exagero propor mais um período para explicar a história da área, relacionando os fatos ocorridos desde 2003. O fato é que a preocupação crescente com a formação na área parece estar no cerne da continuidade dos desenvolvimentos até então observados, considerando que profissionais bem formados poderão optar por instrumentos e técnicas de forma mais crítica, utilizá-los de forma mais responsável e contribuir para que a psicologia como ciência e profissão continue desenvolvendo-se, tanto do ponto de vista técnico quanto ético.

Questões

1) Por que os testes psicológicos foram importantes para o desenvolvimento científico da psicologia?
2) Explique o conceito de quoeficiente intelectual (QI).
3) Por que houve o interesse inicial dos pesquisadores acerca dos processos psicológicos básicos?
4) Descreva brevemente os critérios mínimos de qualidade de testes psicológicos, segundo a Resolução 02/2003.
5) Por que pode ser perigosa a utilização de testes que não se enquadrem nos critérios mínimos de qualidade?

Referências

Amendola, M. F. (2011). *Panorama da História dos Testes Psicológicos no Brasil.* Recuperado em 23 de Fevereiro de 2011 de www.canalpsi.psc.br/artigos/artigo12.htm.

Anastasi, A. (1977). *Testes psicológicos.* (2a ed.). São Paulo: EPU.

Andriola, N. B. (1996). Avaliação Psicológica no Brasil: considerações a respeito da formação dos psicólogos e dos instrumentos utilizados. *Psique,* 6 (2), 99-108.

Barclay, J. R. (1991). *Psychological Assessment: a theory and systems approach.* Malabar: Krieger.

Carroll, J. B. (1993). *Human cognitive abilities: a survey of factor-analytic studies.* Nova Iorque: Cambridge University.

Cattell, R. B. (1940). A culture-free intelligence test. *Journal of Educational Psychology, 31,* 162-179.

Cattell, R. B. (1971). *Abilities, their structure, growth, and action.* Boston: Houghton Mifflin.

Conselho Federal de Psicologia (2003). *Resolução n.º 002/2003.* Recuperado em 26 de abril de 2011 de http://www.pol.org.br.

Dubois, P. H. (1970). *A history of psychological testing.* Boston: Allyn & Bacon.

Fancher, R. (1985). *The intelligence men: Makers of the IQ controversy.* Nova Iorque: Norton.

Gomes, W. B. (2003). Pesquisa e práticas em Psicologia no Brasil. In O. H. Yamamoto. & V. V. Gouveia (Orgs.). *Construindo a Psicologia Brasileira: desafios da ciência e prática psicológica* (pp. 23-59). São Paulo: Casa do Psicólogo.

Horn, J. L. (1991). Measurement of intellectual capabilities: a review of theory. In K. S. McGrew, J. K. Werder & R. W. Woodcock (Orgs.). *Woodcock-Johnson Technical Manual* (pp. 197-232). Chicago: Riverside.

Hutz, C. S., & Bandeira, D. R. (2003). Avaliação Psicológica no Brasil: situação atual e desafios para o futuro. In O. H. Yamamoto & V. V.

Gouveia (Orgs.). *Construindo a psicologia brasileira: desafios da ciência e prática psicológica* (pp. 261-275). São Paulo: Casa do Psicólogo.

Logan, T. P. (2006). *Introdução à Prática de Testes Psicológicos*. Rio de Janeiro: LTC.

McReynolds, P. (1986). History of assessment in clinical and educational setting. In R. O Nelson & S. C. Hayes (orgs.). *Conceptual foundations of behavioral assessment* (pp. 42-80). Nova Iorque: Guilford.

Noronha, A. P. P., Ziviani, C., Hutz, C. S., Bandeira, D. R., Custódio, E. M., Alves, I. C. B. et al. (2002). Em defesa da avaliação psicológica. *Avaliação Psicológica, 1*, 173-174.

Padilha, S., Noronha, A. P. P., & Fagan, C. Z. (2007). Instrumento de avaliação psicológica: uso e parecer de psicólogos. *Avaliação Psicológica*, 6 (1), 69-76.

Pasquali, L., & Alchieri, J. C. (2001). Os testes psicológicos no Brasil. In L. Pasquali (Org). *Técnicas de Exame Psicológico - TEP: fundamentos de técnicas psicológicas*. (pp. 195-221). São Paulo: Casa do Psicólogo.

Pereira, F. M., & Pereira Neto, A. (2003). O psicólogo no Brasil: notas sobre seu processo de profissionalização. *Psicologia em Estudo*, 8 (2), 19-27.

Primi, R. (2010). Avaliação Psicológica no Brasil: fundamentos, situação atual e direções para o futuro. *Psicologia: Teoria e Pesquisa*, 26 (especial), 25-36.

Ribeiro, I. S. (1998). *Mudanças no desempenho e na estrutura das aptidões: contributos para o estudo da diferenciação cognitiva em jovens*. Braga: Universidade do Minho.

Sternberg, R. J. (2000). *Psicologia Cognitiva*. Porto Alegre: Artmed.

Urbina, A. (2007). *Fundamentos da testagem psicológicos*. Porto Alegre. Artmed.

Van Kolck, O. L. (1981). *Técnicas de exame psicológico e suas aplicações no Brasil*. Petrópolis: Vozes.

Capítulo 2

Panorama atual dos testes psicológicos no Brasil de 2003 a 2011

Sílvia Verônica Pacanaro
Gisele Aparecida da Silva Alves
Ivan Sant'Ana Rabelo
Irene F. Almeida de Sá Leme
Rodolfo A. M. Ambiel

Sabe-se que o uso de testes psicológicos, juntamente com a investigação de outros dados, integra o processo de avaliação psicológica. Pasquali (2001) define os testes como um conjunto de tarefas predeterminadas que o sujeito precisa realizar em uma determinada situação, do qual resultam em alguma forma de medida. Posteriormente, Urbina (2007) descreveu os testes como procedimentos para a obtenção de amostras de comportamentos e respostas de indivíduos com o objetivo de descrever e/ou mensurar características e processos psicológicos, compreendidos tradicionalmente nas áreas emoção/afeto, cognição/inteligência,

motivação, personalidade, psicomotricidade, atenção, memória, percepção, entre outras.

Destaca-se que, nas primeiras cinco décadas do século XX, os testes psicológicos, independentemente do seu tipo, rapidamente atenderam às necessidades da sociedade na época e foram inseridos nos contextos militar, industrial e institucional. Assim, é pertinente lembrar que o progresso da ciência psicológica e o fortalecimento dos pilares básicos para o desenvolvimento dos testes colaboraram com a expansão de seu uso. Nas décadas de 1960 e 1970, houve largo descrédito na área de testagem psicológica, sendo que os instrumentos foram criticados e o seu uso diminuído e menosprezado na atuação do profissional de psicologia. Um dos motivos para esse movimento no Brasil foi a associação dos modelos de avaliação com a cultura técnica norte-americana (Pasquali & Alchieri, 2001). No final dos anos oitenta, surgiram processos judiciais em decorrência de decisões referentes ao psicotécnico na área da seleção, bem como a descrença da prática de alguns psicólogos despreparados para a utilização de testes psicológicos.

Desde então, ocorreram alguns movimentos para que fossem criadas soluções para a melhoria da qualidade dos serviços relativos à área de avaliação psicológica, como a criação da Comissão Nacional sobre Testes, em 1980, bem como sua segunda edição em 1986; o surgimento da Câmara Interinstitucional de Avaliação Psicológica em 1997; e a criação do Manual para Avaliação Psicológica de candidatos à Carteira Nacional de Habilitação e condutores de veículos automotores em 2000; as resoluções que regulamentaram a ação profissional no tocante aos laudos e aos instrumentos de avaliação psicológica em 2001;

e a Resolução CFP 002/2003, que divulgou os requisitos mínimos e obrigatórios que os instrumentos psicológicos precisam ter para o uso profissional adequado (Noronha, Primi & Alchieri, 2004).

Foi a partir da Resolução Nº 002/2003 do Conselho Federal de Psicologia, que foram definidos com um pouco mais de clareza os requisitos mínimos e obrigatórios que os instrumentos psicológicos precisam ter para o uso profissional adequado. Entre os principais requisitos, pode-se mencionar:

- Apresentação da fundamentação teórica do instrumento, com especial ênfase na definição do construto;
- Apresentação de evidências empíricas de validade e precisão das interpretações propostas para os escores do teste, justificando os procedimentos específicos adotados na investigação;
- Apresentação de dados empíricos sobre as propriedades psicométricas dos itens do instrumento;
- Informações sobre os procedimentos de correção e interpretação dos resultados, comunicando detalhadamente o procedimento e o sistema de interpretação no que se refere às normas brasileiras, relatando as características da amostra de padronização de maneira clara e exaustiva, preferencialmente comparando com estimativas nacionais, o que possibilita o julgamento do nível de representatividade do grupo de referência usado para a transformação dos escores;
- Apresentação clara dos procedimentos de aplicação e correção, bem como das condições nas quais o teste deve ser

aplicado, para que haja a garantia da uniformidade dos procedimentos envolvidos na sua aplicação.

Também em 2003, o Conselho Federal de Psicologia (CFP) instituiu a Comissão Nacional de Avaliação Psicológica, chamada de Sistema de Avaliação dos Testes Psicológicos (SATEPSI), que, inicialmente, teve como objetivo central a análise das principais dificuldades que o psicólogo enfrenta diante da avaliação psicológica e da utilização de testes (Pasquali & Alchieri, 2001). Essa Comissão é integrada por psicólogos convidados, de reconhecido saber em testagem psicológica, que analisam e emitem pareceres sobre os testes psicológicos. É importante salientar que todos os instrumentos considerados psicológicos ou não, encaminhados para o Conselho Federal de Psicologia (CFP), passam por uma avaliação do SATEPSI.

Após o recebimento do instrumento por esta comissão, o trâmite envolve alguns procedimentos internos, tais como o recebimento do material a ser analisado, a análise propriamente dita, a avaliação, a comunicação da avaliação aos requerentes, com prazo para recurso, a análise de recurso e a avaliação final. Segundo a Resolução CFP 002/2003, um instrumento psicológico recebe parecer favorável quando, por decisão do Plenário do CFP, o teste é considerado em condições de uso, pois cumpriu os requisitos mínimos para a comercialização de um teste psicológico; o parecer emitido será desfavorável quando a análise indicar que o teste não apresenta as condições mínimas para uso. Nesse caso, o parecer deverá especificar as razões da reprovação, bem como as orientações futuras para mudanças e melhorias nos procedimentos para o uso do instrumento. Após a revisão

e a reformulação, o pesquisador poderá reapresentar o material a qualquer tempo, e todos os procedimentos de análise serão seguidos novamente.

Quando o instrumento é aprovado para publicação, comercialização e utilização profissional, cabe observar que, segundo a Resolução n.º 006/2004 do CFP, os dados empíricos das propriedades de um teste psicológico devem ser revisados periodicamente. Para os dados referentes à padronização, o intervalo entre um estudo e outro não pode ultrapassar quinze anos e, para validade e precisão, o período deve ser de vinte anos. Não sendo apresentadas as revisões nos prazos estabelecidos, o teste psicológico perderá a condição de favorável e será excluído da relação de testes em condições de comercialização e uso.

Quanto à quantidade de testes psicológicos comercializados atualmente, o *Buros Institute of Mental Measurements* fornece aos profissionais interessados, por meio de sua página virtual, informações acerca de testes publicados e comercializados disponíveis em inglês. Essas informações incluem área de aplicação, dados da editora responsável pela publicação e pela comercialização dos testes, ano de publicação, autores, título dos testes, acrônimos e revisões disponíveis. Noronha e Reppold (2010) relataram que, em 2002, foi possível identificar aproximadamente dois mil títulos disponíveis nesta base de dados. Atualmente, em 2011, a página inicial da base de dados relata mais de 3500 testes disponíveis para consulta.

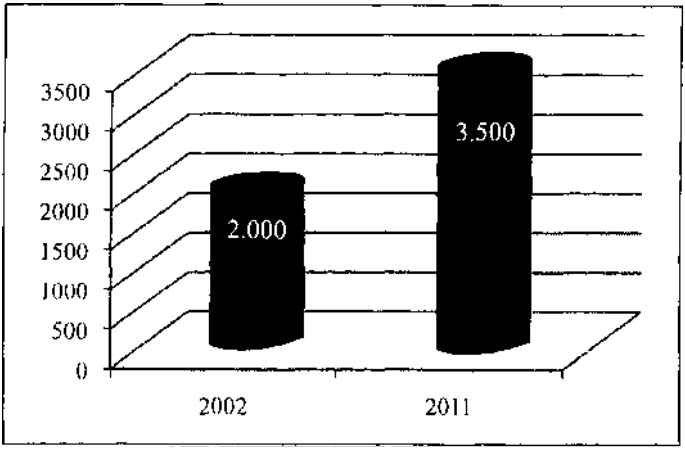


Figura 1. Quantidade de testes psicológicos em língua inglesa comercializados segundo o *Buros Institute of Mental Measurements* (Noronha & Reppold, 2010).

O Sistema de Avaliação dos Testes Psicológicos[1] (SATEPSI) consiste em um sistema brasileiro de certificação de instrumentos de avaliação psicológica para uso profissional, o qual avalia e qualifica os instrumentos psicológicos como aptos (pareceres favoráveis) ou inaptos (pareceres desfavoráveis) para uso. A lista completa dos testes informa título, ano de publicação, requerente (quem submeteu o teste à avaliação do CFP), datas de recepção, análise, avaliação e recurso e avaliação final, que informa o resultado da apreciação. Na época de sua criação, a lista contava com cerca de trinta instrumentos com pareceres favoráveis para uso. Os resultados de um levantamento realizado no mês de abril de 2011 demonstraram que, até essa data, 121 instrumentos

[1] www2.pol.org.br/satepsi

psicológicos aprovados constavam no sistema, sendo que um instrumento ainda mencionado como aprovado passou recentemente para a classificação desfavorável.

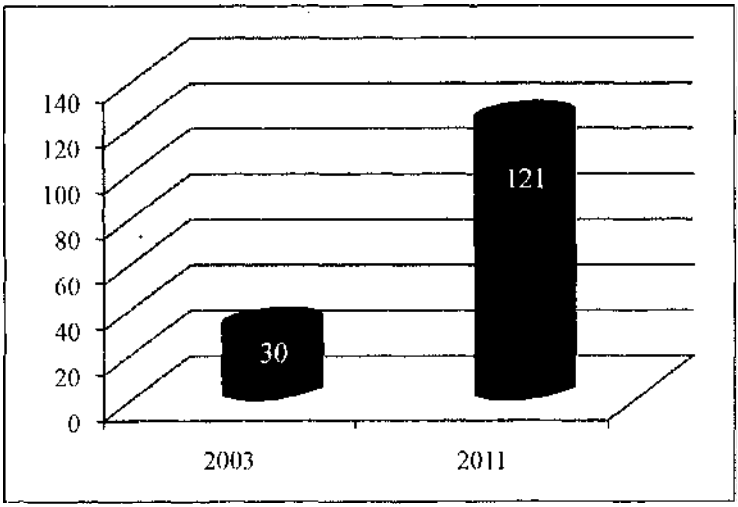


Figura 2. Quantidade de testes psicológicos com parecer favorável no Brasil (consulta ao SATEPSI em abril/2011).

Dos 120 aprovados, 33 são instrumentos para avaliação cognitiva, que contempla inteligência, funções executivas e raciocínio; 27 para avaliação da personalidade; 16 instrumentos para avaliação da atenção; 6 para memória; 6 instrumentos para avaliação de interesses profissionais; 4 para avaliação das habilidades sociais; 4 para avaliação do estresse; 3 para o contexto familiar; 3 para criatividade; 2 para a avaliação da agressividade; 2 para a avaliação da depressão e 1 instrumento para cada tema referenciado a seguir (categoria "Outros"): destreza, autoconceito,

autocontrole, assertividade, avaliação ocupacional, saúde geral, expressão de raiva, lateralidade, avaliação visomotora, expectativa acerca do álcool, TDAH, ansiedade e ideação suicida.

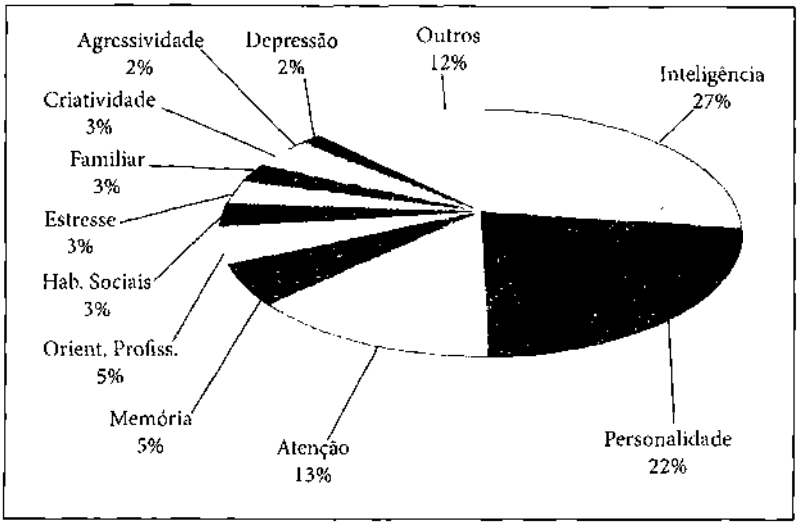


Figura 3. Tipos de testes psicológicos, segundo construto avaliado (consulta ao SATEPSI em abril/2011).

Durante algum tempo, no Brasil, foi utilizada uma grande quantidade de instrumentos produzidos em outros países, ocorrendo, em alguns casos, uma tradução sem os cuidados necessários, sendo utilizadas as tabelas de normas estrangeiras que levavam a resultados e conclusões errôneos (Duarte, Miyazaki, Ciconelli & Sesso, 2003). Com a implementação do SATEPSI, buscou-se mudar essa realidade por meio do estabelecimento de padrões para os testes e, indiretamente, para a prática em avaliação, ao impedir que instrumentos sem o devido reconhecimento científico fossem

utilizados profissionalmente. Esse sistema, apesar de algumas controvérsias, estimulou o desenvolvimento de pesquisas, tanto por parte da comunidade de pesquisadores, quanto pelas editoras, que passaram a atender a uma série de exigências técnicas antes de disponibilizarem instrumentos psicológicos para comercialização.

Primi e Nunes (2010) relatam que o SATEPSI, ao longo dos nove anos de existência, foi gradativamente ganhando a aceitação dos profissionais, e os psicólogos foram compreendendo as propostas e os objetivos desse sistema. Entre outros aspectos, o sistema também provocou o aumento na qualidade dos manuais de testes, que atualmente são bem mais completos e detalhados do que há dez anos.

O Departamento de Pesquisa e Produção de Testes da Editora Casa do Psicólogo realizou um levantamento referente a artigos publicados sobre o tema avaliação psicológica, após a Resolução do CFP nº 02/2003. Consultada a base de dados do PEPSIC (Periódicos Eletrônicos em Psicologia), mencionando a palavra-chave "avaliação psicológica", foram registrados todos os artigos que continham essa informação. Do total, 148 registros não possuíam o estabelecimento do período e 139 publicações foram realizadas após o ano de 2003. O ano de 2008 foi o que obteve mais publicações (N=35), seguido de 2007 (N=25) e 2006 (N=24).

Foi observado também que as revistas que mais publicaram sobre o assunto foram a *Revista Avaliação Psicológica*, que obteve maior número de publicações (N=45), seguida da Revista *Psico-USF* (N=28). Diante desse levantamento, pôde constatar-se que, após a resolução do conselho de 2003, ocorreu um aumento nas publicações relacionadas à construção de instrumentos, a estudos de evidências de validade e a outras propriedades psicométricas,

na busca de tornar os instrumentos mais seguros para a população estudada, bem como relacionadas a atualizações de pesquisas que realizem estudos para normas brasileiras.

Com o propósito de analisar o conhecimento dos estudantes do curso de psicologia no que se refere aos conteúdos relacionados à avaliação psicológica, Noronha, Baldo, Barbin e Freitas (2003) realizaram um estudo com 180 alunos do primeiro ao quinto ano do curso de psicologia, de uma instituição particular de ensino do interior do estado de São Paulo. A idade dos participantes variou de dezessete a 51 anos (M=23,8; DP=7,2).

A pesquisa consistiu na aplicação de um instrumento contendo 45 itens, com questões dicotômicas, cujo objetivo era investigar quatro áreas de conhecimento: conceito de avaliação; conceito de instrumento de avaliação; uso de instrumento; e aprendizagem de instrumentos. O instrumento foi aplicado coletivamente, com os seguintes resultados: os itens que faziam parte do instrumento que obtiveram boas porcentagens de acerto foram o item 8 - a aprendizagem de testes deveria acontecer apenas no último ano do curso (97,2%); o item 45 - os testes não servem para nada (96,9%); o item 32 - os testes são realmente aprendidos na prática clínica (96,6%); os itens 2 e 21, respectivamente - a avaliação psicológica pode ser utilizada em muitos contextos de atuação profissional, e os testes psicológicos são instrumentos pouco importantes na prática profissional do psicólogo (94,9%); o item 20 - a entrevista e a observação são técnicas de avaliação (93,9%); o item 41 – consigo aprender um teste pela leitura do manual (93,8%); o item 42 - o psicólogo não deve usar testes, pois eles reduzem o homem a números (93,3%); e o item 12 - a avaliação é um processo de coleta de dados (91,7%).

Os dados revelaram um melhor desempenho dos alunos que estavam no último ano, quando comparados com os alunos do primeiro ano, que nunca tiveram nenhum contato de instrução formal de avaliação, havendo diferença significativa entre os grupos em quase metade dos itens do instrumento. Observa-se ainda que, entre os próprios estudantes de psicologia, há um desconhecimento da prática e da importância em se utilizar a avaliação psicológica corretamente.

Neste sentido, Paula, Pereira e Nascimento (2007) realizaram um levantamento por meio de um questionário sobre a opinião dos alunos de psicologia a respeito da avaliação psicológica, principalmente sobre a utilização dos testes psicológicos. Participaram 358 alunos de psicologia que cursavam o último ano da graduação, pertencentes a quatro instituições, sendo uma pública e três particulares da cidade de Belo Horizonte. Quanto à idade dos participantes, 59% deles tinham até 24 anos; 15%, entre 25 e 26 anos; 25% tinham mais de 26 anos e 1% não relatou a idade.

O questionário era composto por dezenove itens que contemplaram as seguintes informações: formação acadêmica, articulação entre teoria e prática na graduação e na identificação dos problemas mais frequentes no uso dos testes psicológicos. No que se refere aos instrumentos aprendidos durante a graduação, os mais citados foram a Escala de Inteligência Wechsler para Crianças - WISC (N=259), seguida do Desenho da Figura Humana – DFH (N=204) e das Matrizes Progressivas de Raven (N=194), e os menos citados foram o Teste de Atenção Concentrada - D2 (93), as Figuras Complexas de Rey (N=95) e o Teste não verbal de inteligência R1 (N=100). Os instrumentos WISC e as Matrizes Progressivas de Raven também foram os mais citados

como utilizados nos estágios, bem como o HTP (casa, árvore e pessoa).

Dos instrumentos informados como os mais citados, todos estão aprovados pelo CFP, exceto o Teste do Desenho Wartegg e o DFH de *Goodenough-Harris*, o que indica que as universidades também procuram escolher instrumentos que contenham informações relacionadas à normatização, à padronização, à validade e à precisão. Também foram verificadas informações sobre os pontos positivos e negativos dos aspectos da formação acadêmica na área da avaliação psicológica, sendo que os pontos positivos mais citados foram a capacitação profissional dos professores e o conhecimento sobre as técnicas e os testes ensinados.

Sobre os pontos negativos, os mais mencionados foram a insuficiência de informações no conteúdo do ensino e nos números de disciplinas ofertadas. No que se refere ao conhecimento dos alunos sobre a Resolução Nº 002/2003 do CFP, o estudo indicou que 83% dos alunos desconheciam essa resolução, o que é preocupante, pois muitos deles já estão trabalhando em diversas áreas da psicologia. Conclui-se, portanto, uma grande necessidade de aperfeiçoamento da formação acadêmica do psicólogo, visando à aquisição de habilidades suficientes para capacitá-lo ao exercício profissional de melhor qualidade na área em questão.

Joly, Silva, Nunes e Sousa (2007) investigaram a produção científica de assuntos relacionados à avaliação psicológica, pesquisando os resumos de painéis publicados nos Anais dos Congressos Brasileiros de Avaliação Psicológica, nos anos de 2003, 2005 e 2007. Esses resumos estavam disponíveis em CD, e a amostra foi composta por 934 resumos de diversas áreas, sendo 264 do

primeiro congresso (2003), 322 do segundo (2005) e 348 do terceiro (2007).

Os painéis foram analisados com base em alguns critérios, a saber, tipo de estudo (psicométrico, de aplicação e descritivo); tipo de pesquisa (documental, empírica, revisão bibliográfica, estudo de levantamento, estudo de caso); quantidade e sexo dos autores; tipo de instituição às quais os autores estavam vinculados (universidades públicas, particulares, institutos de pesquisa ou empresas); região do país na qual os autores atuam; procedimento de avaliação adotado (teste, entrevista, observação, misto); área de aplicação; nome do construto avaliado; tamanho da amostra, faixa etária e tipo de grupo (estudantes, profissionais ou pessoas institucionalizadas, com e sem distúrbio psicológico diagnosticado); nome; tipo do instrumento utilizado (projetivo, objetivo ou mais de um) e tipo de aplicação (lápis e papel, informatizada, relato verbal); incluem-se os seguintes estudos psicométricos: validade (que tipo), precisão (que tipo), adaptação, normatização e padronização; e procedimento de análise utilizado (qualitativo, quantitativo ou ambos). Destacam-se alguns itens relevantes e observa-se que, no ano de 2007, ocorreu um aumento nas publicações de 37,2%. Sobre os autores dos resumos, houve uma predominância na participação do gênero feminino, havendo um total de 873 mulheres, em comparação com 492 homens.

No que se refere às distribuições de estudos por regiões e ano de congresso, percebeu-se uma maior participação da região Sudeste (64%) em todos os anos avaliados. Observou-se também que os anos de 2005 e 2007 concentraram mais pesquisas envolvendo adultos e crianças, com relevância estatística ($22\ [12] = 142{,}539$; $p < 0{,}001$).

Foi observado que a utilização de testes obteve uma grande prevalência nas técnicas utilizadas, e que, no ano de 2007, se optou por uma maior utilização de instrumentos do Tipo Lápis e Papel, com 71,8%, quando comparado aos demais anos. Além disso, observa-se que, em um número significativo de resumos do congresso de 2003 (34,5%), não constava essa informação, e que isso foi diminuindo ao longo dos outros anos.

Houve diferenças altamente significativas entre os tipos dos instrumentos utilizados em função dos anos dos congressos ($\chi 22$ [10] = 193, 410; $p < 0{,}001$), de modo que ocorreu prevalência pela utilização de instrumentos objetivos em todos os anos do congresso, sendo que, em 2007, essa prevalência foi de 78,4%. No que se refere aos instrumentos mais utilizados nas apresentações dos painéis, pode mencionar-se o Rorschach (N=25), seguido do WISC (n=12) e Bender (N=11), que avaliam respectivamente personalidade, inteligência e psicomotricidade, como os mais citados nos resumos.

Quanto a teses e dissertações disponíveis na Base de Dados da Biblioteca Virtual em Psicologia Brasil (BVS-Psi Brasil), Joly, Berberian, Andrade e Teixeira (2010) realizaram um levantamento utilizando as seguintes palavras-chaves: avaliação psicológica, psicometria, validade, precisão e testes psicológicos, sendo a busca realizada até setembro de 2007. Obteve-se um total de 141 resumos referentes à pesquisa realizada. As análises de frequências dos resumos revelaram que 54,6% eram dissertações de mestrado; 43,3%, teses de doutorado, e 2,1% eram teses de pós-doutorado, sendo que 19,15% dos estudos foram defendidos em universidades da região sul do País; 80,14% no Sudeste, e 0,71%, no Nordeste.

No que se refere ao gênero das autorias desses trabalhos, observou-se que 88,9% eram do sexo feminino. Desses estudos, 66,7% foram realizados em universidades públicas e 33,3%, em universidades privadas.

Além disso, foi verificado também o tipo de construto estudado, o que revelou que os construtos personalidade e inteligência ainda permaneceram como os mais pesquisados, como destacado também na pesquisa anterior, sendo que a análise de frequência da área de aplicação em que os instrumentos desenvolvidos ou utilizados podem ser empregados revelou que a área clínica (26,3%) foi a que mais teve estudos direcionados, seguida pela área de psicologia escolar e educacional (25,6%).

Alves, Alchieri e Marques (2002) relatam uma crítica atual atribuída aos testes psicológicos, a qual se refere à função de rotular o examinando durante o processo de avaliação. Diante disso, vale ressaltar o uso dos testes como ferramentas integrantes do processo de avaliação psicológica e que nunca devem ser utilizados de forma isolada, bem como o fato de que os resultados oriundos da testagem são comparados a normas criadas para aquela população e contexto específicos.

Complementando, há a argumentação de Noronha e cols. (2003) sobre a escolha de um instrumento de avaliação. As autoras insistem na importância de se respeitar a idade, o sexo e a referência a normas peculiares à população avaliada. Ao lado disso, para que os testes sejam úteis e eficientes, devem passar por processos que comprovem suas qualidades psicométricas e também atender a especificações que garantam o reconhecimento e a credibilidade por parte da sociedade e da comunidade científica.

Quanto à expansão de pesquisas relacionadas à testagem psicológica, ela pode ser decorrente também da Resolução nº 002/2003, do Conselho Federal de Psicologia, pois há, a partir dela, a obrigatoriedade de estudos psicométricos para os testes psicológicos utilizados para fins profissionais. É importante salientar que, durante as últimas décadas, são percebidos avanços importantes na área, e, conforme destaca Gouveia (2009), não é possível mais pensar na área de avaliação psicológica como amadora, pois cada vez mais é necessária a preparação de um material de testagem de melhor qualidade, possibilitando estudos cada vez mais representativos dos parâmetros psicométricos.

Em acréscimo, destaca-se a necessidade de mostrar aos alunos ingressos no curso de psicologia das diversas universidades brasileiras a utilidade da área de avaliação psicológica e suas ramificações, pois há um mercado promissor nesses campos. Dessa forma, faz-se necessário um conhecimento técnico e amplo a respeito da testagem e do processo de avaliação psicológica por parte do psicólogo, o que norteará o profissional para a melhor forma de realização de uma intervenção.

Questões

1) Nas décadas de 1960 e 1970, houve amplo descrédito na área de testagem psicológica; os instrumentos foram criticados e o seu uso diminuído e depreciado na atuação do profissional de psicologia. Aponte alguns motivos para esse movimento no Brasil.
2) A partir da Resolução Nº 002/2003 do Conselho Federal de Psicologia, foram definidos com um pouco mais de clareza os requisitos mínimos e obrigatórios que os instrumentos psicológicos precisam ter para o uso profissional adequado. Descreva alguns desses principais requisitos.
3) Também em 2003, o Conselho Federal de Psicologia (CFP) instituiu a Comissão Nacional de Avaliação Psicológica, chamada de Sistema de Avaliação dos Testes Psicológicos (SATEPSI). Qual o objetivo principal dessa comissão?
4) Sabendo-se que os dados empíricos das propriedades de um teste psicológico devem ser revisados periodicamente, indique o período (em anos) de regularidade de um teste quanto à padronização e à validade/precisão.
5) Algumas pesquisas relatam uma crítica atribuída aos testes psicológicos no que se refere à função do rótulo que será plantado no sujeito submetido ao processo de avaliação. Diante disso, explique quais cuidados podem ser tomados para uma avaliação psicológica coesa.

Referências

Alves, I. C. B., Alchieri, J. C., & Marques, K. C. (2001). As técnicas de exame psicológico ensinadas nos cursos de graduação de acordo com os professores. *Psico-USF, 7* (1), p. 77-88.

Conselho Federal de Psicologia. (2003). *Resolução CFP nº 002/2003*. Brasília, DF. Recuperado em 21 de julho de 2010, de http://www.pol.org.br/pol/export/si-tes/default/pol/legislacao/legislacaoDocumentos/resolucao2003_02.pdf

Conselho Federal de Psicologia. (2004). Resolução CFP nº 006/2004. Brasília, DF. Recuperado em 24 de outubro de 2010, de http://www.pol.org.br/pol/export/si-tes/default/pol/legislacao/legislacaoDocumentos/resolucao2003_02.pdf

Duarte, P. S., Miyazaki, M. C. O. S., Ciconelli, R. M. & Sesso, R. (2003). Tradução e adaptação cultural do instrumento de avaliação de qualidade de vida para pacientes renais crônicos (KDQOL-SF). *Revista da Associação Médica Brasileira, 49* (4), 375-81.

Gouveia, V. (2009). A avaliação Psicológica no Brasil: caminhos, desafios e possibilidades. *Psicologia em foco, 2* (1), 110 – 119.

Joly, M. C. R. A., Silva, M. C. R., Nunes, M. F. O. & Sousa, M. S. (2007). Análise da Produção Científica em Painéis dos Congressos Brasileiros de Avaliação Psicológica *Avaliação Psicológica*, 6 (2), 239-252.

Joly, M. C. R. A, Berberian, A. A, Andrade, R. G. & Teixeira, T. C. (2010). Análise de Teses e Dissertações em Avaliação Psicológica Disponíveis na BVS-PSI Brasil. *Psicologia Ciência e Profissão, 30* (1), 174-187.

Noronha, A. P. P., Reppold, C.T. (2010). Considerações Sobre a Avaliação Psicológica no Brasil. *Psicologia: ciência e profissão, 30* (num. esp.), 192-201.

Noronha, A. P., Vendramini, C. M. M., Canguçu, C., Souza, C. V. R., Cobêro, C, Paula, L. M. et al (2003). Propriedades psicométricas apresentadas em manuais de testes de inteligência. *Psicologia em Estudo, 8* (1), 93-99.

Noronha, A. P. P., Primi, R., & Alchieri, J.C. (2004). Parâmetros Psicométricos: uma Análise de Testes Psicológicos Comercializados no Brasil. *Psicologia: Ciência e Profissão, 24* (4), 88-99.

Noronha, A. P. P., Baldo, C. R., Barbin, P. F. & Freitas, J. V. (2003). Conhecimento em avaliação psicológica: um estudo com alunos de Psicologia. *Psicologia: teoria e prática, 5* (2), 37-43.

Oakland, T. (2009). Ethics on assessment: International perspectives. [Conferência]. Em *Congresso Brasileiro de Avaliação Psicológica*. São Paulo: IBAP.

Paula, A. V., Pereira, A. S. & Nascimento, E. (2007). Opinião de alunos de psicologia sobre o ensino em avaliação psicológica. *Psico-USF, 12* (1), 33-43.

Pasquali, L. (Org.) (2001). *Técnicas de exame psicológico - TEP. Manual. Vol. I: Fundamentos das técnicas psicológicas*. São Paulo: Casa do Psicólogo.

Pasquali, L. & Alchieri, J. C. (2001). Os testes psicológicos no Brasil. In L. Pasquali (Orgs). *Técnicas de exame psicológico - TEP* (Manual, Vol. I: Fundamentos das técnicas psicológicas, pp. 195-221). São Paulo: Casa do Psicólogo.

Primi, R., & Nunes, C. H. S. (2010). O Satepsi: desafios e propostas de aprimoramento. In Conselho Federal de Psicologia (Orgs). *Avaliação psicológica: diretrizes na regulamentação da profissão* (pp. 129-148). Brasília: CFP.

Urbina, S. (2007). *Fundamentos da Testagem Psicológica*. Porto Alegre: Artmed Editora.

Werlang, B. S. G., Amaral, A. E. V., & Nascimento, R. S. G. F. (2010). Avaliação Psicológica, testes e possibilidades de uso. In Conselho Federal de Psicologia (Orgs). *Avaliação psicológica: diretrizes na regulamentação da profissão* (pp. 87-100). Brasília: CFP.

Capítulo 3

"E viveram felizes para sempre": a longa (e necessária) relação entre psicologia e estatística

Rodolfo A. M. Ambiel
Josemberg Moura de Andrade
Lucas de Francisco Carvalho
Vicente Cassepp-Borges

Ao escolher um curso de psicologia, inevitavelmente, o estudante carrega consigo uma série de interesses e expectativas, sendo comum encontrar neles gosto por atividades diretamente ligadas ao contato com as pessoas e a ajuda a elas, e, não por acaso, a preferência por atuações psicológicas no contexto clínico é quase unânime entre os estudantes ingressantes (Bueno, Lemos & Lomé, 2004; Meira & Nunes, 2005; Noronha & Ambiel, 2008). A ideia de que a psicologia é somente clínica, muito frequentemente compartilhada por pessoas leigas, tem sido flexibilizada, fazendo

com que outras áreas tenham ganhado destaque, tais como a organizacional, a comunitária e a hospitalar. Considerando a finalidade deste livro, qual seja, de apresentar conceitos básicos sobre avaliação e testes psicológicos, é importante grifar que, no bojo dessa "expansão social" da psicologia, desde o início da década de 2000 a área também tem crescido em credibilidade, eficiência e qualidade (CFP, 2003).

Entretanto, apesar do crescente interesse por outras áreas e por novos métodos psicológicos, há uma ferramenta que parece ainda não ter "caído nas graças" de estudantes e profissionais, embora sua utilização seja tão antiga quanto a própria psicologia enquanto ciência. Sim, aqui se está falando sobre a temida (e mal compreendida) *estatística*!

É bastante comum ouvir nos corredores das faculdades de psicologia murmúrios (e muitas vezes lamentações) a respeito dos conteúdos matemáticos que parecem não fazer sentido em meio a outras disciplinas específicas da psicologia. Por conseguinte, não raro, percebe-se uma certa associação de "aversões" entre estatística, testes psicológicos e pesquisa em psicologia, como se tudo o que fizesse uso da estatística fosse igualmente difícil e chato.

Se você está lendo este texto e identificando-se com tais afirmações, concordando com a maioria delas, saiba que não é o único. Preocupados com as dificuldades dos alunos nesse assunto, vários pesquisadores têm se empenhado em compreendê-las.

Por exemplo, Yunis (2006) pesquisou as principais dificuldades em estatística de estudantes de psicologia egípcios e descobriu cinco principais fontes de dificuldade: (1) o conteúdo do curso, (2) o professor, (3) os exames, (4) o próprio estudante e (5) a distância

do material ensinado da realidade. Esse mesmo autor encontrou dados mostrando que, quanto mais a estatística causava ansiedade nos alunos, mais eles consideravam a matéria difícil.

No Brasil, Silva e Vendramini (2005) pesquisaram o autoconceito estatístico (uma variável afetiva relacionada ao julgamento que a pessoa faz de si mesma em relação à estatística) de estudantes de psicologia e pedagogia. Entre os itens com os quais os estudantes concordaram mais estava este: "Eu me sinto incapaz na aula de estatística". Por outro lado, entre aqueles com mais discordância estavam: "Eu gosto de estudar estatística em casa" e "Eu acredito que eu posso ser um estatístico ou um cientista futuramente". Em outro estudo, que avaliou a atitude de estudantes de psicologia em relação à estatística, Vendramini, Silva e Dias (2009) verificaram que o desempenho na disciplina de estatística estava bastante relacionado com a afirmação: "A estatística me faz sentir como se estivesse perdido em uma selva de números e sem encontrar saída".

Também no Brasil, Noronha, Nunes e Ambiel (2007) observaram que os estudantes de psicologia atribuem pouca importância para o uso da estatística nas práticas de avaliação psicológica. Além disso, os dados sugerem que os estudantes de primeiro ano relataram ter mais domínio em estatística do que os do quinto ano. Ou seja, parece ocorrer algum fenômeno que faz os alunos perceberem que "desaprendem" estatística ao longo do curso.

Com os estudos citados, pode perceber-se que a estatística é mesmo percebida por futuros psicólogos como um "bicho de sete cabeças". Mas talvez você não tenha percebido que, nos últimos dois parágrafos, várias informações e vários conceitos estatísticos foram passados. Possivelmente, você tenha lido os parágrafos e

compreendido as informações, sem que isso lhe causasse nenhuma ansiedade ou lhe fizesse sentir-se em uma selva sem saída. E esse é o objetivo deste capítulo: apresentar conceitos estatísticos básicos e essenciais para uma boa utilização e compreensão de manuais e de testes psicológicos de uma forma simples e clara.

Contudo, antes de prosseguir, é necessário dar-se uma boa e uma má notícia. A má é que, por maior que tenha sido o esforço, não foi possível livrar você, leitor, da apresentação de algumas fórmulas. Mas não se preocupe. A boa notícia é que você não precisa fazer cálculos ou grandes operações matemáticas para utilizar a estatística no seu dia a dia de estudante de psicologia ou de pesquisador. Existem softwares que farão o trabalho por você, tais como o Microsoft Excel ou o *Statical Package for Social Sciences*, o popular SPSS.

A estatística na psicologia

A "parceria" entre psicologia e estatística não é nova. Historiadores apontam que no século XIX o caminho da psicologia rumo ao seu reconhecimento como ciência demandou a adoção de métodos que viabilizassem a quantificação de características psicológicas. Concomitantemente, os primeiros pesquisadores interessados em conhecer os processos psicofísicos das pessoas começaram a fazer uso de procedimentos estatísticos para atribuir validade científica aos seus achados (Sass, 2008).

No momento histórico inicial da psicologia como ciência, um dos pesquisadores que melhor utilizaram a estatística em seus estudos foi Galton. Em seus experimentos, ele, que era biólogo,

estudava as diferenças individuais das pessoas, com a preocupação de compreender como a hereditariedade e o ambiente poderiam influenciar no desenvolvimento e na manifestação de traços característicos de cada um. Nessa empreitada, Galton teve a ajuda de Cattell, que mais tarde viria a se tornar um dos principais cientistas no campo da personalidade (Memória, 2004, Schultz & Schultz, 2007).

No início do século XX, surgiram os primeiros testes psicológicos, tal como são conhecidos atualmente. Urbina (2007) afirma que, nessa época, as sociedades urbanas, industriais e democráticas começavam a se consolidar e, em consequência desses novos conceitos sociais, tornou-se imperativo tomar decisões sobre pessoas de forma justa e considerando suas características pessoais em diversas áreas, tais como nos contextos laboral, educacional e da saúde.

A partir desse avanço inicial, ao longo do século XX as testagens psicológica e educacional se desenvolveram sobremaneira, com a contribuição e o refinamento das análises estatísticas (Urbina, 2007). Além disso, a necessidade de selecionar soldados para as grandes guerras mundiais e o surgimento de *softwares* e pacotes estatísticos possibilitaram tornar os instrumentos de avaliação cada vez mais válidos e precisos.

Dessa forma, após uma breve contextualização histórica sobre o uso da estatística pela psicologia, alguns conceitos básicos serão expostos. É importante notar que a intenção deste capítulo, bem como da disciplina de estatística nas faculdades de psicologia, não é formar estatísticos e, sim, instrumentalizar os estudantes e os profissionais da psicologia para uma boa utilização de manuais de testes psicológicos e aplicação em pesquisa.

População e amostra

A estatística é um ramo de conhecimento formado por um conjunto de métodos matemáticos que ajudam as pessoas a tomar decisões. O termo é derivado de *status* (estado) e pode ter duas interpretações: estado, enquanto condição atual de determinada situação (por exemplo, "meu estado financeiro atual está péssimo"); ou Estado, enquanto administração pública, ou seja, métodos adotados pelo Estado para monitorar o desenvolvimento de alguma característica da população (por exemplo, "a renda per capita do brasileiro subiu 5% nos últimos anos") (Memória, 2004).

Ao trazer essa ideia para a psicologia, é necessário lembrar que geralmente se falará sobre pessoas. A esse respeito, é importante entender que a estatística, sendo um conjunto de métodos, vai informar sobre os dados que estiverem disponíveis, seja qual for a fonte da coleta. Por exemplo, em exames médicos clínicos, a estatística ajuda os médicos a entender a condição de saúde do examinado a partir de uma amostra de alguma substância biológica, como o sangue. Na psicologia, os dados dizem respeito a comportamentos coletados de uma parte da população (Pasquali, 2010).

Essa noção é primordial para a compreensão dos próximos passos: dificilmente será possível para um psicólogo fazer uma pesquisa com toda uma população e, por isso, seleciona-se uma amostra para a realização da pesquisa. Portanto, uma amostra é uma parte de uma população, selecionada com base em algum critério. População, por sua vez, é o conjunto de todos os indivíduos de uma determinada classe. Por exemplo, um pesquisador quer verificar os níveis do traço de personalidade *Extroversão* em estudantes

de psicologia brasileiros; a população em questão seria composta por todos os estudantes de psicologia do Brasil, de todas as universidades, de todas as cidades, de todos os estados brasileiros no momento da pesquisa. Como isso seria muito complicado e caro, o pesquisador seleciona uma amostra de estudantes de psicologia para realizar seu estudo, a qual pode ser uma turma, algumas turmas de uma universidade, algumas universidades de um estado ou, aleatoriamente, uma parte dos estudantes de psicologia do Brasil.

É claro que, quanto menos aleatória for a seleção da amostra, maior a possibilidade de tendenciosidade dos dados, ou seja, as informações são relativas apenas àquela pequena amostra, e o pesquisador não pode generalizar os dados. Por exemplo, seria errado que um estudo cujos dados foram coletados em apenas uma turma de psicologia da Universidade Estadual de Londrina (UEL), no Paraná, concluísse que os resultados encontrados, naquela amostra, refletem as características dos estudantes de psicologia do Brasil. Aliás, não é possível nem generalizar para os futuros psicólogos do estado do Paraná, quiçá nem mesmo da própria UEL, mas apenas para as pessoas em questão (podendo ou não ser verdade para o restante da população estudada).

Afinal, quando se fala em seleção da amostra, deve ser considerada a ideia de representatividade. Ou seja, já que em geral não se tem dinheiro, tampouco tempo, para avaliar todos os representantes de uma certa população, é necessário que a amostra selecionada (a) tenha uma quantidade de pessoas suficiente, (b) não tenha características próprias que a diferenciem do "normal" da população e (c) seja escolhida de maneira aleatória; desse modo, o pesquisador não deve escolher os sujeitos que interessam para chegar ao resultado que deseja.

Medidas de tendência central e variabilidade

Basicamente, o uso da estatística em psicologia tem a finalidade de descrever e resumir dados provindos de observações de comportamento, que podem ser feitas de diferentes formas, como testes, questionários e entrevistas. Tais descrições e conjuntos de dados são realizados, especificamente, por meio de números, que expressam e ajudam a entender o significado dos resultados. A função de descrever e resumir resultados é do domínio da estatística descritiva. Já a função de interpretar resultados, especificamente quando se deseja generalizar os resultados de uma amostra de respondentes para a população alvo, é do domínio da estatística inferencial (Glassman & Hadad, 2008; Urbina, 2007).

Especificamente nesta seção, vamos concentrar-nos na estatística descritiva. Podemos observar, nos vários livros de estatística disponíveis, que a forma mais comum de estatística descritiva são as medidas de tendência central. Como assinalam Dancey e Reidy (2006), uma medida de tendência central de um conjunto de dados fornece uma indicação do escore típico (mais comum) deste conjunto de dados. Em outras palavras, uma maneira vantajosa de caracterizar um grupo de sujeitos como um todo é achar um número único que represente o que é médio, ou típico daquele conjunto de dados. Assim, podemos dizer, por exemplo, que o tempo médio de realização da prova de psicologia social da turma A do curso de psicologia de uma universidade qualquer foi de uma hora e dez minutos. Por outro lado, podemos dizer que a média de idade desses alunos é de 23,5 anos. Embora as idades

dos alunos do curso de psicologia variem, a média de idade de 23,5 anos oferece uma noção geral das idades deles. As medidas de tendência central comumente utilizadas para descrever dados são a *moda*, a *mediana* e a *média*, que serão discutidas a seguir.

A medida de tendência central mais simples de se obter é a moda (Mo), que é simplesmente o valor mais frequente, mais típico ou mais comum em uma distribuição de dados. Suponha que dez candidatos fizeram uma prova de conhecimentos gerais para provimento de uma vaga de analista de sistemas em uma determinada empresa. Dois candidatos obtiveram a nota 6,0; três candidatos, a nota 7,5; quatro candidatos, a nota 8,0 e um candidato, a nota 10. Qual seria a moda? O valor modal é 8,0, já que quatro candidatos obtiveram essa nota. Em outras palavras, 8,0 é a moda porque é a nota que ocorre com maior frequência. Algumas distribuições de frequência podem conter duas ou mais modas; por exemplo, quando se têm duas modas, falamos que a distribuição é bimodal.

A moda é a única medida de tendência central que podemos utilizar para representar variáveis do tipo nominal. Nesse nível de medida, os números são utilizados de forma arbitrária, simplesmente como símbolos de identificação de grupos a que os elementos pertencem (Bunchaft & Cavas, 2002). Exemplos de variáveis no nível nominal são sexo (masculino e feminino), religião (católica, evangélica, espírita, etc.) e curso universitário (psicologia, medicina, direito, cinema, etc.). Vale ressaltar que a moda pode, entretanto, ser utilizada para descrever o escore mais

comum em qualquer distribuição, independentemente do nível de mensuração[1] (Levin & Fox, 2004).

Uma segunda medida de tendência central é a mediana (Mdn), a qual diz respeito à pontuação que está no meio da distribuição da frequência. Quando uma distribuição de frequências é disposta em ordem de tamanho, torna-se possível localizar a mediana, o ponto do meio de uma distribuição. A mediana é encarada como uma medida de tendência central, pois separa a distribuição de frequências em duas partes iguais (Levin & Fox, 2004). Considere a seguinte distribuição de frequência: 12, 11, 18, 16, 15, 13 e 17. Quando se vai determinar a mediana, o primeiro passo é ordenar os dados do menor para o maior, como segue: 11, 12, 13, 15, 16, 17 e 18. Cada valor tem um posto; por exemplo, o valor 11 assume o 1º posto, o valor 12 assume o 2º posto e assim por diante. O posto do valor da mediana pode ser determinado por inspeção (valor do meio em uma distribuição de frequência ímpar) ou pela fórmula: $\frac{N+1}{2}$. Nesse caso, temos $\frac{7+1}{2}$, que resulta em 4. Assim, no posto 4, temos o valor da mediana igual a 15. No caso de uma distribuição de frequências par, a mediana será a média entre os dois valores centrais.

Compreendidas a moda e a mediana, passaremos a explicar a média. A medida de tendência central mais comum, intensa e extensivamente utilizada, é a média aritmética, geralmente denominada de média (M) (Ferreira, 2005). Para calcular-se a média, deve-se somar o escore de cada sujeito e dividir o resultado pelo número de sujeitos. A título de informação, a fórmula da média é: $\bar{X} = \frac{\Sigma X}{N}$, na qual:

[1] Para maiores informações sobre níveis de mensuração, consultar Bunchaft e Cavas (2002), Hogan (2006) ou Pasquali (2003).

- $\overline{X}$ = média (lê-se 'X barra');
- Σ = soma (letra grega maiúscula *sigma*);
- X = escore bruto em um conjunto de escores;
- N = número total de escores no conjunto (Levin & Fox, 2004).

Segundo Glassman e Hadad (2008), a média é comumente utilizada devido a duas características ou vantagens. Primeiro, não é necessário dispor as pontuações em uma ordem sequencial para calcular a média; segundo, ao contrário da mediana ou da moda, a média reflete todas as pontuações, ou seja, se você mudar uma pontuação, a média também vai mudar. Um dos problemas da média é que ela é sensível aos casos extremos, os famosos *outliers* no jargão da estatística (Dancey & Reidy, 2007). Suponha que você queira saber a média da renda mensal dos sujeitos da sua pesquisa e, por coincidência, Bill Gates esteja na sua amostra. Nesse caso, a média da renda mensal será maximizada. Em outro caso aberrante, você quer saber a média de idade dos sujeitos da sua pesquisa, e o homem mais velho do mundo, certificado pelo livro dos recordes, está na sua amostra. Mais uma vez, o valor da média de idade será maximizado. É claro que esses exemplos são exagerados, mas servem para ilustrar que a média sofre influências dos *outliers*. Em casos similares, devemos utilizar a mediana em vez da média.

Também não devemos utilizar a média quando a distribuição de frequências não é normal. A distribuição normal, ilustrada na Figura 1, tem a forma de uma curva simétrica que parece um sino

de perfil. Essa distribuição indica que os sujeitos da sua amostra se distribuem normalmente em torno de um valor modal.

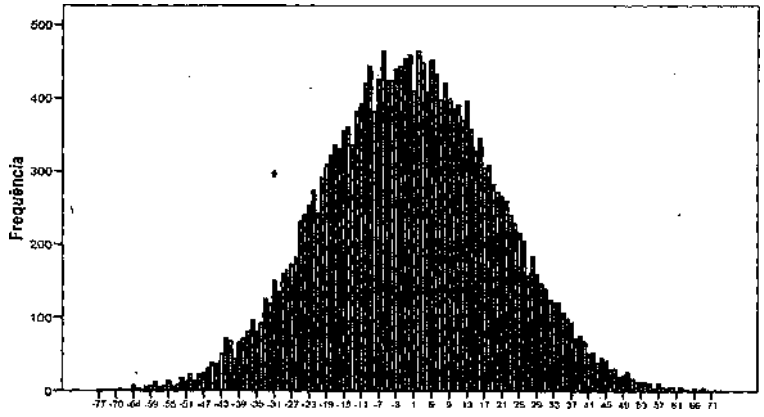


Figura 1. Curva aproximadamente normal.

Em uma distribuição perfeitamente normal, o ponto mais elevado ocorre no meio da distribuição, ou seja, a moda, a mediana e a média são iguais. Assim, em uma distribuição normal, não importa qual medida de tendência central você usará, porque todas produzem o mesmo resultado. Por outro lado, quando se tem uma distribuição assimétrica, é improvável que a média seja representativa da maioria das pontuações. Nesses casos, a maioria dos pesquisadores prefere utilizar a mediana como uma maneira de descrever o resultado típico (Glassman & Hadad, 2008). Alguns exemplos de variáveis com distribuições assimétricas são renda mensal e nível de escolaridade da população.

As medidas de tendência central, todavia, nunca devem ser utilizadas sozinhas, pois, quando o são pelo pesquisador inexperiente, elas representam os dados de maneira incompleta. As medidas de variabilidade, por outro lado, nos informam o quanto as pontuações estão distribuídas em torno do centro. Nesse sentido, uma medida de tendência central deve vir sempre acompanhada de alguma medida de variabilidade. As medidas de variabilidade mais utilizadas são a amplitude, a variância e o desvio-padrão, as quais serão discutidas a seguir.

Um indicador simples da variabilidade é a amplitude, que nada mais é do que a diferença entre a pontuação mais alta e a pontuação mais baixa. Por exemplo, para calcular a amplitude da seguinte distribuição de dados 0, 12, 13, 14, 25, 30, 35, basta calcular 35 - 0 = 35, ou seja, o valor mais alto menos o valor mais baixo é igual a 35. Como a amplitude só reflete as duas pontuações mais extremas, é apenas uma medida bruta da variabilidade (Glassman & Hadad, 2008). Embora a amplitude forneça uma ideia da variação total dos valores, ela, de fato, não fornece uma ideia global da distribuição dos valores de uma amostra (Dancey & Reidy, 2004). Por esse motivo, devemos recorrer a outras medidas de variabilidade.

Uma medida mais informativa da variabilidade dos dados é o desvio padrão, caracterizado pela medida de quanto os valores da nossa amostra variam em torno da média. Para calculá-lo, é necessário, antes, calcular a variância da distribuição de dados. Procuramos explicar aqui esses conceitos da forma mais intuitiva possível.

De forma simplificada, se subtrairmos a média de cada valor observado, obteremos os desvios, que são uma indicação de quão

longe cada um desses valores está da média. Cada um dos desvios deve ser elevado ao quadrado para evitar valores negativos. Feito isso, pode calcular-se a média dos desvios ao quadrado para obter uma indicação da variabilidade do conjunto como um todo. Esse resultado é conhecido como variância e, embora seja usada de várias maneiras pelos estatísticos, ela nos dá um número inflado, pois é baseada nos quadrados dos desvios, e não nos próprios desvios em si. Para obtermos uma medida compatível com os valores originais das nossas variáveis, utilizamos a raiz quadrada da variância, que é denominada desvio padrão (Dancey & Reidy, 2004). Considerados juntos, a média e o desvio padrão nos dizem muito sobre o nosso conjunto de dados, pois, em geral, quanto maior o desvio padrão, maior a variabilidade das respostas (Glassman & Hadad, 2008).

Diferenças entre grupos

Uma vez compreendido que as medidas de tendência central servem para resumir e organizar um conjunto de dados, é necessário entender como isso funciona na vida real. Não é raro encontrar, nos diferentes estudos realizados em psicologia, perguntas tratando de diferenças entre grupos. Por exemplo, as mulheres são mais ciumentas do que os homens? Essa pergunta poderia ser apresentada, em termos de diferenças entre grupos, da seguinte maneira: Existem diferenças entre mulheres e homens em relação ao ciúme? Muitos são os delineamentos possíveis para responder a essa pergunta, isto é, podem-se utilizar distintas técnicas e métodos para responder a ela.

Uma das possibilidades é aplicar um instrumento (teste) que avalie o ciúme, em homens e mulheres, e, então, comparar a pontuação que cada um dos grupos obteve no instrumento em questão. Contudo, para tornar possível essa comparação, deve-se calcular a média de cada um desses grupos. Só para relembrar, como já foi explicado, a média nada mais é do que a soma dos valores dividida pelo número de valores. Por exemplo, se o grupo de mulheres era composto por 3 mulheres e cada uma obteve uma pontuação igual a 2 no teste de ciúme, então teremos (2 + 2 + 2)/3, isto é, a soma dos valores que cada mulher teve dividida pelo número (quantidade) de valores. O resultado, no caso, seria 2.

Outro exemplo, em um estudo realizado por Lopes e cols. (2006), os pesquisadores buscaram verificar possíveis diferenças entre homens e mulheres em relação à neofobia alimentar, que se refere a uma relutância em ingerir alimentos novos. Portanto, o objetivo do estudo foi verificar se homens apresentam mais relutância do que mulheres, ou vice-versa. No estudo, participaram 266 pessoas, sendo 109 homens e 157 mulheres. Eles utilizaram um teste que avaliava a neofobia, e, a partir de um procedimento estatístico amplamente conhecido na área de diferenças entre grupos, o teste *t* de Student, verificaram uma tendência de as mulheres serem mais neofóbicas que os homens. Ou seja, nessa amostra estudada (já que não participaram todos os homens e todas as mulheres do mundo), as mulheres tiveram médias maiores do que os homens no teste aplicado.

O teste *t* pode ser utilizado em diferentes condições; para o caso do estudo de Lopes e cols. (2006), procedeu-se ao teste *t* para amostras independentes (já que as pessoas de um grupo não são as mesmas pessoas do outro grupo). Basicamente, para o cálculo

desse procedimento estatístico, consideram-se diferenças intragrupo, ou seja, quanto as respostas das pessoas de um dos grupos varia, diferenças entre grupos, isto é, se há uma tendência (média) para um grupo obter pontuações (escores) inferiores ou superiores em relação ao outro grupo.

Geralmente, dois grupos distintos não apresentarão médias e variações idênticas, por isso, na maior parte dos casos, existirão diferenças entre dois grupos. Contudo, será que as diferenças entre os grupos são suficientes ou significativamente grandes do ponto de vista estatístico? E, se forem, qual a chance de serem reais (e não terem ocorrido ao acaso)? Vamos por partes. Primeiro, para verificar se as diferenças são suficientes, calcula-se o *t*. De modo superficial, ele é a medida da variância entre os grupos dividido pela variância dentro dos grupos. Assim, quanto maior a variância entre grupos em relação à variância intragrupos, maior é o valor de *t* (Dancey & Reidy, 2006). Isso significa que, para que se tenha um maior valor de *t*, e, consequentemente, maior probabilidade de o teste ser significativo, a diferença entre as médias dos grupos deve ser grande, mas os desvios padrão dentro dos grupos devem ser baixos.

Após o cálculo do *t*, isto é, depois de verificar a diferença entre os grupos, calcula-se a probabilidade de a diferença encontrada ser devido ao acaso ou ser real. A comunidade científica adotou o seguinte critério: para os resultados encontrados nas pesquisas científicas serem considerados significativos do ponto de vista estatístico, deve garantir-se que aquele dado será encontrado novamente (em condições similares à pesquisa realizada) em pelo menos 95% dos casos. Assim, a dúvida se o dado será ou não encontrado novamente não deve passar de 5% (ou 0,05).

Chama-se esse cálculo de índice de significância (indicado pela letra *p*).

Outra possibilidade para o uso do teste *t*, ou seja, de procedimentos estatísticos para comparação de dois grupos (importante: o teste *t* possibilita a comparação de dois grupos e não mais que isso), é quando se quer comparar um grupo de pessoas com ele mesmo. A fórmula subjacente ao teste *t* pareado (comparação de um grupo com ele mesmo) é similar à do teste *t* para amostras independentes, contudo, considera o fato de que as mesmas pessoas estão sendo comparadas (e, por isso, tende a ser um procedimento mais sensível).

Esse tipo de procedimento estatístico pode ser utilizado para casos, por exemplo, em que se quer comparar as pessoas antes de um determinado fato e depois desse fato. Outra possibilidade de aplicação desse procedimento é para verificar o quanto um determinado grupo varia em um determinado construto. Por exemplo, Evangelista, Saldanha, Balbinotti, e Barbosa (2010) verificaram o quanto o grupo de homens atletas (e, posteriormente, o grupo de mulheres) se diferenciava (intragrupo) em relação a construtos relacionados às atitudes morais. Os dados apontaram para diferenças significativas do ponto de vista estatístico (ou seja, com pelo menos 95% de segurança de não serem ao acaso) em quatro das seis dimensões avaliadas no estudo.

Em ambos os casos, amostras independentes e pareado, o teste t funciona com base em algumas suposições, que são válidas para o grupo de procedimentos estatísticos categorizados como testes paramétricos. Basicamente, a suposição dos testes paramétricos refere-se à distribuição da amostra. Em outros termos, no caso de procedimentos estatísticos que utilizam a média (como o

teste *t*), supõe-se que os grupos de pessoas apresentam uma distribuição normal ou próxima a isso (já discutido anteriormente). Essa exigência decorre exatamente do uso da média, isto é, se as variâncias entre os grupos forem muito desiguais, então o resultado verificado não representará nenhum dos grupos (Dancey & Reidy, 2006).

Quando essa suposição não é atendida, sugere-se o uso de testes não paramétricos. Usualmente, o teste Mann-Whitney está para o teste *t* para amostras independentes, assim como o teste Wilcoxon está para o teste *t* pareado. Ambos tendem a ser bem mais simples que o teste *t*, já que não utilizam médias, desvios padrões e erros padrões. Subjacente a esses testes, o raciocínio imbuído é a comparação entre as distribuições dos grupos.

Ao lado disso, ainda em relação à comparação entre grupos, em algumas situações, torna-se necessária a comparação entre mais de dois grupos. Para esses casos, nenhum dos testes mencionados anteriormente é válido. Por isso, utiliza-se a análise de variância (ANOVA), que analisa duas ou mais médias (ou grupos), buscando verificar se há ou não alguma diferença significativa do ponto de vista estatístico (Tabachnick & Fidell, 2007). De modo similar ao teste *t*, o conjunto de procedimentos estatísticos que representam a ANOVA considera tanto a variância de respostas dentro do grupo quanto a diferença entre grupos (médias). Contudo, para o caso desses procedimentos, não se calcula o *t*, mas sim o *F*.

Do procedimento ANOVA, derivam-se resultados de diferenças estatísticas (que podem ser ou não significativas) entre grupos, nos quais as médias dos mais de dois grupos são comparadas, e intragrupos, nas quais as variâncias entre participantes dentro dos grupos são verificadas. Por exemplo, no estudo realizado por

Diniz e Zanini (2010), foram verificadas diferenças entre médias de três grupos diferenciados pela idade nos construtos personalidade e *coping*. Os resultados apresentaram diferença significativa (portanto, $p < 0,05$, indicando uma chance menor de 5% dos dados serem atribuídos ao acaso) para três dimensões relacionadas à personalidade e ao *coping* das treze possíveis.

Cabe ressaltar que a gama de procedimentos que representam as análises de variância é bastante ampla e abarca diversas possibilidades de delineamentos. Um critério inicial que justifica o uso dessas análises é a existência de mais de dois grupos (de outro modo, pode utilizar-se o teste *t*, por exemplo). A partir disso, dependendo do tipo de estudo sendo realizado, é possível considerar um número amplo de grupos e de variáveis independentes (isto é, construtos que estão sendo avaliados), e também diferentes variáveis dependentes (no caso, aquelas que dividem as pessoas em grupos). Ainda é possível verificar efeitos entre variáveis independentes e dependentes.

Verificam-se, a partir da breve explanação acerca das diferenças entre grupos, as diversas possibilidades de procedimentos estatísticos que podem ser utilizados para observação de diferenças entre grupos distintos. Essas análises se diferenciam em complexidade, isto é, o número de grupos de pessoas que são capazes de comparar e o número de variáveis independentes e dependentes.

Correlações

Voltando à introdução, o estudo de Vendramini, Silva e Dias (2009) afirma que aqueles estudantes que encaram a estatística de

maneira mais positiva tendem a ter melhores notas na disciplina. Assim, ler este capítulo com uma atitude positiva deve melhorar a sua aprendizagem (assim esperam os autores!). Pode ter certeza de que não vai doer. De qualquer forma, como as autoras do estudo citado chegaram a essa conclusão? Como você chegaria a essa conclusão? O teste estatístico que responde a essa pergunta chama-se correlação. Uma primeira maneira de verificar a relação entre duas coisas (na linguagem estatística, duas variáveis), mais racional, seria analisar cada caso individualmente, para depois ver o todo. Imaginemos que Rafael é alguém que adora estatística e tirou uma nota muito boa. Leonardo, por outro lado, odeia estatística, e foi muito mal na disciplina. Se houvesse somente alunos como Rafael e Leonardo, teríamos evidências de que alunos que gostam de estatística tendem a se sair melhor na prova.

Todavia, quando se lida com muita gente, existem pessoas que fogem à regra. Podemos ter um aluno como Michelangelo, por um lado, que ama estatística, mas, coitado, ficou doente um dia antes da prova e não teve um bom desempenho. Por outro lado, Donatello, que odeia estatística, com medo de ter de fazer duas vezes sua tão temida matéria, estudou demais e tirou uma excelente nota. Casos como esses estariam mostrando que não há relação entre gostar de estatística e a nota na disciplina.

Mas, como são muitos alunos, possivelmente existirão vários Leonardos, Rafaéis, Donatellos e Michelangelos. Devemos descobrir se temos mais Rafaéis e Leonardos do que Donatellos e Michelangelos? Existe, entretanto, uma pergunta anterior a essa: Como saber se um estudante é Leonardo, Rafael, Donatello ou Michelangelo? Ocorre que não há um critério claro para rotular se a pessoa gosta ou não de estatística. Fazer isso seria perder

a riqueza do dado, pois o que fazer com os que gostam mais ou menos da matéria? E os que gostam mais ou menos para mais? Enfim, é mais interessante pegar a quantificação deste gostar, em vez de criar rótulos, e o mesmo vale para a nota na matéria. Para não se perder, é mais fácil colocar dois eixos em um gráfico, no qual um representa a nota de cada sujeito, e o outro, o quanto tal sujeito gosta de estatística (Figura 2). Cada ponto é uma pessoa, e as letras representam as iniciais de onde estariam os alunos Rafael (R), Leonardo (L), Donatello (D) e Michelangelo (M).

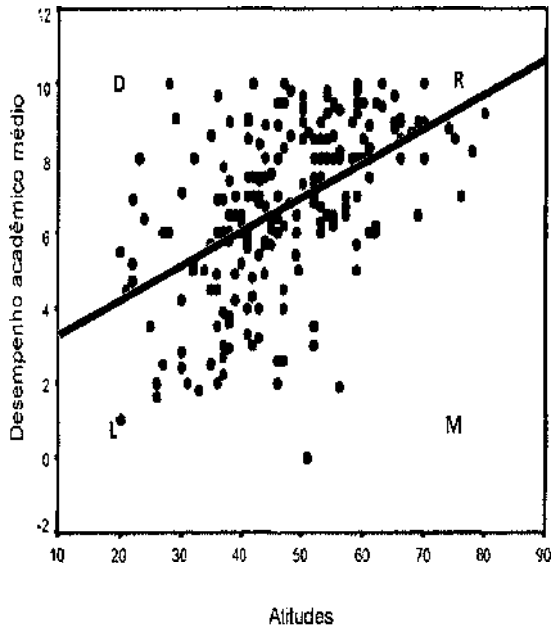


Figura 2. Diagrama de dispersão entre o desempenho em estatística e as atitudes com relação à matéria (adaptado de Vendramini, Silva & Dias, 2009).

Conforme foi observado na amostra, o Michelangelo, ou alguém que gosta de estatística e se sai mal na disciplina, é um aluno hipotético, que não existe. Ou você está observando algum ponto próximo ao "M" no gráfico? Os Donatellos também são raros. O que se observa no gráfico é que, geralmente, quem tem atitudes negativas com relação à estatística não se sai bem na prova, e quem tem atitudes positivas obtém sucesso. Olhando os pontos espalhados no gráfico, conseguimos chegar a essa conclusão, ainda que se trate de uma conclusão de olho. Desse modo, precisamos de algo mais simples e mais preciso para indicar o quanto a relação entre duas coisas é forte.

Para isso serve aquela linha que está no gráfico, chamada de *best fit line*, ou linha de melhor ajuste. A ideia é que essa seja uma linha reta que represente todos os pontos do gráfico da melhor maneira possível. Os pontos devem estar próximos a essa linha para indicar que há correlação positiva, ou seja, que a tendência é encontrarmos alunos como Rafael e Leonardo na turma de estatística. Ambas as variáveis crescem juntas, ou seja, quanto mais positiva a atitude com relação à estatística, melhor o desempenho na disciplina. Se essa linha, que melhor representa os pontos do gráfico, estivesse no sentido Donatello-Michelangelo, essa correlação seria negativa, pois o aumento em uma variável se relaciona com a diminuição na outra. Isso significaria que, quanto mais positiva a atitude com relação à estatística, pior a nota na matéria. Tal conclusão, entretanto, não faz sentido, talvez por isso não tenha sido encontrada nos dados empíricos de Vendramini, Silva e Dias (2009), porém esperamos que tenha servido para você compreender o conceito de correlação negativa.

Mas ainda não está tão simples. É muito ruim trabalhar com gráficos de dispersão, como o da Figura 2, pois eles ocupam muito espaço, e temos de nos concentrar muito para entendê-lo. Além disso, apenas observando a linha e os pontos, não é possível ter uma noção muito clara do quão relacionadas são as duas variáveis. Assim, resume-se toda essa informação em um número, que é alcançado por meio de um cálculo que considera o quanto cada ponto do gráfico está distante da linha de melhor ajuste. Esse número varia de -1 até +1, passando pelo zero. Uma correlação igual a zero, conforme pode ser deduzido na própria linguagem, significa que não existe relação entre duas variáveis. Assim, se quando a atitude com relação à estatística se tornasse mais positiva, a nota aumentasse ou diminuísse aleatoriamente, ou se houvesse Rafaéis, Leonardos, Donatellos e Michelangelos em igual proporção, teríamos uma correlação próxima de zero. Quando os pontos estão espalhados aleatoriamente, qualquer que seja a linha de melhor ajuste, haveria dificuldade de representar bem todos os pontos. Por outro lado, se todos os pontos do gráfico estivessem em cima da linha, ou seja, se não houvesse distância alguma entre a linha e os pontos, a correlação seria perfeita, indicando que o aumento das notas na estatística estaria ligado a gostar da matéria, sem espaço para exceções. O valor de uma correlação perfeita é +1, ou -1 se essa correlação perfeita ocorrer em sentido oposto (o aumento em uma variável se relaciona perfeitamente com a diminuição da outra). Para interpretar valores intermediários a 0 e a ±1, torna-se importante a experiência do pesquisador. Diferentes livros trazem diferentes critérios; Duffy, Mclean e Monshipouri (2011) interpretam as correlações da maneira descrita na Tabela 1.

Tabela 1. Interpretação do coeficiente de correlação de acordo com Duffy, Mclean e Monshipouri (2011).

Valor	Interpretação
0,00 a 0,19	Sem relação ou relação desprezível
0,20 a 0,29	Relação fraca
0,30 a 0,39	Relação moderada
0,40 a 0,69	Relação forte
0,70 a 1,00	Relação muito forte

Contudo, pode relevar-se esse critério, considerando que áreas como a psicometria, por medir o psíquico, trabalham com maiores margens de erro do que as engenharias por exemplo. Uma correlação não tão forte nas ciências exatas pode ter alguma relevância para as humanas. Embora a palavra "moderada" pareça indicar algo de pouco valor, correlações moderadas são valiosas para a psicologia. Caso você tenha curiosidade, a correlação encontrada por Vendramini, Silva e Dias (2009) foi de 0,618. Assim, se você entendeu o que é uma correlação e os demais conceitos deste capítulo, existem dados empíricos sugerindo que você pode estar começando a se interessar por estatística e, desse modo, está entrando no grupo do Rafael.

Análise Fatorial

Correlações são conceitos importantes na estatística aplicada à psicologia, mas constituem apenas a base de uma análise fatorial. Tudo aquilo que você viu foi apenas a relação entre duas coisas (variáveis). Agora, imaginemos a correlação entre diversas

coisas ao mesmo tempo. Por exemplo, a Tabela 2 mostra diversas frases para as quais os sujeitos hipotéticos que responderam à pesquisa hipotética devem dizer como determinada frase se aplica a eles, em um contínuo que vai de "não se aplica de jeito nenhum" a "se aplica totalmente".

Tabela 2. Matriz de correlações hipotéticas da Escala de gosto por eventos culturais e esportivos.

Frase	1	2	3	4	5	6	7	8
1. Eu gosto de ir a exposições em museus.	1							
2. Eu frequentemente vou ao teatro.	+	1						
3. Apresentações circenses me deixam alegre.	+	+	1					
4. Adoro filmes com enredos esportivos.	+	+	+	1				
5. Acompanho jogos de basquete no ginásio.	0	0	0	+	1			
6. Torço para meu time de futebol no estádio.	0	0	0	+	+	1		
7. Assisti pelo menos a uma partida de vôlei neste ano.	0	0	0	+	+	+	1	
8. A corrupção na política me revolta.	0	0	0	0	0	0	0	1

Caso as afirmações acima fossem reunidas em uma escala, seria esperado que sujeitos que gostam de ir ao cinema também gostam de teatro e de circo, ocasionando correlações positivas entre as

frases desse grupo (por isso o símbolo "+"). As pessoas que gostam de basquete também possuem maior tendência a gostar de futebol e vôlei, criando outro grupo de frases. A frase 4 (Adoro filmes com enredos esportivos) pode relacionar-se com afirmativas dos dois grupos. A frase 8 (A corrupção na política me revolta), no entanto, parece não ter relação com nenhum dos dois grupos. Por outro lado, não se espera que frases de um grupo não se correlacionem com as de outro, pois um sujeito que vai ao estádio de futebol pode ir ou não ao teatro. Note que a correlação entre uma afirmativa com ela mesma está representada pelo número 1, pois se trata de uma correlação perfeita. Um sujeito que gosta de ir a exposições em museus sempre vai gostar de ir a exposições em museus. Note também que metade da tabela é vazia, pois preenchê-la seria repetir informação.

Uma vez que percebemos que existem dois grupos distintos, o próximo passo, sempre seguindo o espírito de simplificar, é verificar o que cada grupo de frases tem em comum e refazer a Tabela 2. Em vez de dizer como cada item se relaciona com cada item, podemos dizer como cada item se relaciona com o grupo de itens. A Tabela 3 batiza os grupos e mostra quais frases tem a ver (+) com quais grupos.

Tabela 3. Matriz Fatorial da Escala de gosto por eventos culturais e esportivos.

Frase	Fator 1. Eventos culturais	Fator 2. Eventos esportivos
1. Gosto de ir a exposições em museus.	+	0
2. Frequentemente vou ao teatro.	+	0
3. Apresentações circenses me deixam alegre.	+	0
4. Adoro filmes com enredos esportivos.	+	+
5. Acompanho jogos de basquete no ginásio.	0	+
6. Torço por meu time de futebol no estádio.	0	+
7. Assisti pelo menos a uma partida de vôlei neste ano.	0	+
8. A corrupção na política me revolta.	0	0

Análise fatorial tem um nome estranho, mas no fundo é algo simples. Cada grupo de frases que se correlacionam pode se chamar fator. O objetivo da análise é determinar quantos fatores existem e quais frases (geralmente itens de um teste psicológico) se relacionam com quais fatores. Cabe salientar que a relação item-fator (chamada de carga fatorial), a exemplo da correlação, também é quantificada com um número que varia de - 1 a + 1, obviamente que passando pelo zero. Uma carga fatorial de ±0,32 (Pasquali,

2005), geralmente arredondada para ±0,30, indica que existe alguma relação entre o item e o fator.

Outro indicador importante é a precisão do grupo de frases (fatores). Quando falamos de escalas, estamos falando de instrumentos de medida, e sempre é importante verificar a precisão de uma medida. A precisão de um instrumento é indicada pela relação entre seus itens; se todos os itens estiverem medindo a mesma coisa de maneira semelhante, então esse instrumento é preciso. Você teria alguma hipótese de como verificar essa relação entre os itens? A resposta está debaixo do seu nariz: é por correlações. Nesse caso específico, o Alfa de Cronbach (α) é o teste mais famoso, pois indica em um número as relações entre todos os itens. Esse indicador vai de zero a um e, embora haja controvérsias sobre qual seria um valor aceitável para o α, convencionou-se que um teste preciso possui α superior a 0,80, sendo que alfas maiores que 0,70 podem ser considerados aceitáveis.

Considerações finais

Com este texto, visou-se a apresentar informações e conceitos técnicos, próprios da estatística, de forma leve e acessível, para estudantes de psicologia e para profissionais que, por algum motivo, tenham desenvolvido atitudes não muito positivas em relação à estatística, mas que necessitem agora de informações a respeito. Deve deixar-se claro que este capítulo não esgota as possibilidades de uso dessas técnicas estatísticas, tampouco explora de forma aprofundada qualquer conceito matemático que esteja por trás dessas análises. O objetivo foi unicamente apresentar os

procedimentos e dar alguns exemplos de utilização prática, questionando alguns mitos em relação à estatística na psicologia. O leitor interessado pode recorrer a uma vasta literatura na área. Inclusive, é possível encontrar livros de estatística aplicados especificamente à psicologia.

Dessa forma, espera-se ter contribuído com a formação de futuros usuários de testes ou pesquisadores, ao demonstrar que a estatística pode ser uma grande aliada dos psicólogos para tomar decisões sobre o futuro das pessoas, seja por meio de pesquisas, seja por meio da correta interpretação e compreensão das informações provindas de manuais de testes. As pesquisas mais atuais usam números para demonstrar seus achados, e espera-se que um psicólogo esteja por dentro do conhecimento atual na sua área. Se você não gosta de números porque prefere ajudar as pessoas, deve compreender que o conhecimento das técnicas estatísticas fará com que ajude as pessoas de uma maneira mais qualificada e eficaz.

Questões

1) Diferencie amostra de população.
2) Liste e discuta as medidas de tendência central e de variabilidade.
3) Como identificar se os resultados encontrados não ocorreram ao acaso?
4) Em que casos devem-se utilizar procedimentos estatísticos não paramétricos?
5) Explique a lógica por trás do conceito de correlação positiva.

Referências

Bueno, J. M. H., Lemos, C. G., Tomé, F. A. M. F. (2004). Interesses profissionais de um grupo de estudantes de psicologia e suas relações com inteligência e personalidade. *Psicologia em estudo*, 9 (2), 271-278.

Bunchaft, G & Cavas, C. S. T. (2002). *Sob medida: um guia sobre a elaboração de medidas do comportamento e suas aplicações*. São Paulo: Vetor Editora.

Conselho Federal de Psicologia - CFP (2003). *Resolução nº 02/2003*. Recuperado em 28 de fevereiro de 2011, de http://www.pol.org.br

Dancey, C. P. & Reidy, J. (2007). *Estatística sem matemática para psicologia usando SPSS para Windows*. 3a ed. Porto Alegre: Artmed.

Diniz, S. S., & Zanini, D. S. (2010). Relação entre fatores de personalidade e estratégias de coping em adolescentes. *Psico-USF*, *15* (1), 71-80.

Duffy, S. P., McLean, S. L. & Monshipouri, M. (2011). *Pearson's r correlation*. Recuperado em 20 de fevereiro de 2011, de http://faculty.quinnipiac.edu/libarts/polsci/Statistics.html

Evangelista, P. H. M., Saldanha, R. P., Balbinotti, C. A. A., Balbinotti, M. A. B., & Barbosa, M. L. L. (2010). Atitudes morais de jovens atletas praticantes de modalidades esportivas coletivas: um estudo comparativo segundo a variável "sexo". *Motriz*, *16* (2), 379-86.

Ferreira, D. F. (2005). *Estatística básica*. Lavras: Editora UFLA.

Glassman, W. E. & Hadad, M. (2008). *Psicologia: abordagens atuais*. 4a ed. Porto Alegre: Artmed.

Hogan, T. P. (2006). *Introdução à pratica de testes psicológicos*. Rio de Janeiro: LTC Editora.

Levin, J & Fox, J. A. (2004). *Estatística para Ciências Humanas*. 9a ed. São Paulo: Prentice Hall.

Lopes, F. A., Cabral, J. S. P., Spinelli, L. H. P., Cervenka, L., Yamamoto, M. A., Branco, R. C. et al (2006). Comer ou não comer, eis a questão: diferenças de gêneros na neofobia alimentar. *Psico-USF*, *11* (1), 123-5.

Meira, C. H. M. G. & Nunes, M. L. T. (2005). Psicologia clínica, psicoterapia e o estudante de psicologia. *Paidéia*, *15* (32), 339-343.

Memória, J. M. P. (2004). *Breve história da estatística*. Brasília: Embrapa Informação Tecnológica.

Noronha, A. P. P. & Ambiel, R. A. M. (2008a). Estudo correlacional entre Escala de Aconselhamento Profissional (EAP) e Self Directed Search (SDS). *Interação em Psicologia, 12* (1), p. 21-33.

Noronha, A. P. P., Nunes, M. F. O. & Ambiel, R. A. M. (2007). Importância e domínios de avaliação psicológica: um estudo com alunos de Psicologia. *Paidéia, 17* (37), 231-244.

Pasquali, L. (2003). *Psicometria: Teoria dos testes na Psicologia e na Educação*. Petrópolis: Vozes.

Pasquali, L. (2005). *Análise Fatorial para Pesquisadores*. Brasília: LabPAM.

Pasquali, L. (2010). Teoria da medida. In L. Pasquali (org.), *Instrumentação psicológica: fundamentos e práticas*. Porto Alegre: Artmed.

Sass, O. (2008) Controle social na sociedade industrial: aproximações entre psicologia e estatística. *InterMeio, 14* (28), 41-56.

Schultz, D. P., & Schultz, S. E. (2007). *História da Psicologia Moderna*. 8a ed. São Paulo: Thomson Learning.

Silva, M. C. R. & Vendramini, C. M. M. (2005). Autoconceito e desempenho de universitários na disciplina Estatística. *Psicologia Escolar e Educacional*, 9 (2), 261-268.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics*. 5a ed. Boston: Allyn & Bacon.

Urbina, S. (2007). *Fundamentos da Testagem Psicológica*. Porto Alegre: Artmed Editora.

Vendramini, C. M. M., Silva, M. C. R. & Dias, A. S. (2009). Avaliação de atitudes de estudantes de psicologia via modelo de crédito parcial da TRI. *Psico-USF, 14 (3)*, 287-298.

Yunis, F. A. (2006). Factors influencing the psychology student in dealing with statistics courses. *ICOTS, 7*, 1-5.

Capítulo 4

Teoria de Resposta ao Item na Avaliação Psicológica

Felipe Valentini
Jacob Arie Laros

Maria, psicóloga recém-formada, na sua primeira experiência profissional é requisitada para realizar uma avaliação das habilidades cognitivas de um cliente. Após selecionar alguns testes com pareceres favoráveis pelo Sistema Satepsi (disponível em www2.pol.org.br/satepsi) e estudar os respectivos manuais, Maria continua com algumas dúvidas: Um dos manuais indica que os itens possuem dificuldade média de 0,50 e boa capacidade discriminativa (correlações bisseriais entre 0,40 e 0,80). O que isso indica? Outro manual sugere, por meio da curva de informação, que o teste é mais eficiente ao avaliar pessoas com escores baixos no construto. O que isso significa? O último teste é aplicado via computador, com base na TAC (Testagem Adaptiva

Computadorizada), por meio da qual as pessoas podem ser comparadas entre si, mesmo respondendo a questões diferentes. Como isso é possível?

Este capítulo foi construído visando a apresentar os principais conceitos e discussões acerca da Teoria Clássica dos Testes (TCT), bem como na Teoria de Resposta ao Item (TRI). Espera-se, desse modo, auxiliar os psicólogos e estudantes de graduação a "resolver" as dúvidas de Maria, assim como outros problemas que surgem na prática da Avaliação Psicológica e que envolvem conceitos básicos de psicometria. Primeiramente, serão exploradas as definições básicas da TCT. Posteriormente, será apresentada a TRI e suas principais aplicações.

Teoria Clássica dos Testes (TCT)

Em qualquer área do conhecimento (até mesmo para as ciências exatas), o problema da medida é inerente à ciência. Como medir a distância entre duas estrelas, por exemplo, localizadas a anos-luz da Terra? Nesse contexto, é fácil perceber que as medidas produzidas (mesmo para o tamanho do parafuso utilizado em naves espaciais) não estão isentas de erros. Na psicologia, a ciência que se ocupa da medida e do seu erro é a psicometria.

A medida, de acordo com a psicologia, pode ser definida como a utilização de números e de categorias para representar um comportamento (Andrade, Laros & Gouveia, 2010; Nunnally & Bernstein, 1994). Além dos comportamentos, a psicometria moderna interessa-se pela medida do traço latente ou *theta* (θ). O traço latente pode ser definido como a habilidade, aptidão ou

fator hipotético que organizam e agem sobre os comportamentos (Pasquali, 2003). Neste sentido, a psicometria moderna ocupa-se tanto da medida dos comportamentos quanto dos traços latentes.

Um dos aspectos importantes da medida e do erro na TCT, que se baseia na Teoria do Escore Verdadeiro (TEV), diz respeito à fidedignidade. A teoria parte do pressuposto de que, a despeito do erro de medida, uma parte dos escores dos examinandos é genuinamente verdadeira. Sendo assim, os escores totais observados (e a variância) de um grupo de pessoas são equacionados pela soma dos escores verdadeiros e do erro. Ou seja, escore observado = escore verdadeiro + erro (Crocker & Algina, 1986; Hogan, 2006; Nunnally & Bernstein, 1994).

A fidedignidade refere-se à estabilidade dos escores dos sujeitos em administrações repetidas do mesmo teste ou formas paralelas. Em outras palavras, a fidedignidade (ou precisão) é o grau em que os escores z de um sujeito permanecem consistentes em administrações repetidas de um mesmo teste. Em termos práticos, a fidedignidade, na TCT, indica a porcentagem que representa o escore verdadeiro sobre o escore total. Suponha-se, por exemplo, que o manual de um teste relate um coeficiente de fidedignidade de 0,85. Por meio da equação da TEV, é correto afirmar que 85% da variância dos escores observados são atribuíveis à variância verdadeira. Além disso, 15% devem-se à variância de erro. Ressalta-se, no entanto, que o coeficiente de fidedignidade para um conjunto de escores é um conceito puramente teórico (Crocker & Algina, 1986).

No que se refere ao item, as principais qualidades da medida avaliadas são a dificuldade, a capacidade de discriminação e a possibilidade de resposta ao chute (Nunnally & Bernstein, 1994).

O parâmetro da dificuldade do item na TCT é operacionalizado por meio do cálculo da proporção de acerto (Nunnally & Bernstein, 1994). Um item, por exemplo, submetido a cem participantes e respondido corretamente por oitenta deles, teria o seu parâmetro de dificuldade igual a 0,80 (ou seja, 80/100=0,80), indicando que, em média, 80% das pessoas acertam o item. Nota-se que esse indicador de dificuldade é apresentado em uma escala invertida: quanto maior o parâmetro de dificuldade, mais fácil é o item, pois uma proporção maior de pessoas consegue responder-lhe corretamente. Por esse motivo, tem-se dito que o parâmetro de dificuldade na TCT, na realidade, é um indicador de facilidade do item.

O parâmetro discriminação diz respeito à qualidade do item em separar os examinandos em grupos conforme suas capacidades (ou escores). Ou seja, o poder do item está em distinguir sujeitos com escores relativamente parecidos. Um item pouco discriminativo tem sua utilidade diminuída, uma vez que pouco auxilia o teste a separar as pessoas mais habilidosas das menos habilidosas. Na TCT, a discriminação é avaliada, principalmente, por meio da correlação bisserial ou pela correlação ponto-bisserial entre o item e o escore total. Neste aspecto, é esperada uma correlação item-total positiva, refletindo o fato de que as respostas corretas ao item são mais frequentes nos examinandos com escores totais altos. Quanto maior a correlação, maior a discriminação (Hambleton, Swaminathan & Rogers, 1991; Nunnally & Bernstein, 1994; Pasquali, 2003). Correlações negativas indicam que as respostas corretas ao item são mais frequentes nos examinandos com os menores escores totais. Neste caso, há indícios de que o

item apresenta algum problema, normalmente relacionado à troca de gabarito (Andrade, Laros & Gouveia, 2010).

Finalmente, na TCT, a avaliação da possibilidade de acerto ao acaso (ou "chute") é dada em função do número de alternativas de resposta. Por exemplo, um item de múltipla escolha com quatro alternativas (A, B, C e D) possui 25% (1 item / 4 alternativas = 0,25) de chance de acerto devido ao acaso (Nunnally & Bernstein, 1994). O pressuposto é que o examinando que não tem a habilidade suficiente para dar a resposta chutará cegamente, sem avaliar qual das respostas é a mais provável. Destaca-se que este pressuposto da TCT foi amplamente criticado, uma vez que, em geral, as alternativas incorretas não têm a mesma atratividade (Embretson & Reise, 2000; Crocker & Algina, 1986; Hambleton, Swaminathan & Rogers, 1991).

Uma das principais limitações das medidas obtidas por meio da TCT é que as estatísticas de pessoas (os escores) são dependentes das características psicométricas dos itens, e as estatísticas psicométricas de itens são dependentes das características do grupo de examinandos. Suponha-se que um mesmo teste seja aplicado a dois grupos de pessoas com características distintas, e um dos grupos é bem mais habilidoso do que o outro. Neste caso, o parâmetro da dificuldade do teste receberia diferentes indicadores para o primeiro e para o segundo grupo. No entanto, o parâmetro da dificuldade deveria ser uma característica do teste e não dos examinandos (Andrade, Laros & Gouveia, 2010; Hambleton, Swaminathan & Rogers, 1991; Nunes & Primi, 2009).

A Teoria de Resposta ao Item (TRI) surgiu como uma proposta para lidar com estes e outros problemas da TCT. A seguir, serão discutidos os principais aspectos e as aplicações da TRI.

Teoria de Resposta ao Item (TRI)

Conceitos Básicos

A TRI é conjunto de modelos que procuram representar a probabilidade de uma pessoa apresentar uma determinada resposta a um item, considerando os parâmetros do item e o nível de habilidade da pessoa avaliada (Andrade, Tavares & Valle, 2000). O surgimento da TRI representou avanços em alguns aspectos da psicometria, pois, por intermédio dela, é possível estimar os parâmetros dos itens de maneira independente do grupo avaliado, assim como estimar as habilidades dos participantes de maneira independente das características psicométricas dos itens. Além disso, outro avanço é que o item pode ser considerado individualmente em vez do teste como um todo. Finalmente, a fidedignidade do teste pode ser avaliada para diferentes níveis de habilidade, gerando indicadores mais precisos (Andrade, Laros & Gouveia, 2010; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991; Pasquali & Primi, 2003).

A TRI também é conhecida como a Teoria do Traço Latente, pois considera que as respostas observadas de um teste estão embasadas em traços latentes ou habilidades não observáveis (Hambleton & Swaminathan, 1985; Pasquali, 2007), sendo as habilidades representadas pelo θ (*theta*), o maior foco da teoria. Neste modelo teórico, adota-se uma escala padronizada para mensuração deste valor. Essa escala possui distribuição semelhante ao escore z, podendo assumir valores entre $-\infty$ (infinito) a $+\infty$ (Hambleton & Swaminathan, 1985; Pasquali, 2007). Todavia, na maior parte das vezes, o *theta* varia entre -3 e +3.

A TRI possui dois postulados gerais. O primeiro indica que o desempenho (ou escore) de uma pessoa num determinado item pode ser explicado unicamente pelo traço latente da pessoa e pelas características do item. Ou seja, a partir do θ do participante e das características do item, é possível estimar a probabilidade de ele acertar o item (ou endossá-lo, em escalas de preferência). O segundo postulado indica que é possível expressar a probabilidade do desempenho em função do θ, por meio de uma curva ascendente, denominada Curva Característica do Item - CCI (Andrade, Laros & Gouveia, 2010; Baker, 2001; Hambleton & Swaminathan, 1985; Nunes & Primi, 2009). A CCI especifica que a probalidade de resposta correta é maior conforme o aumento da habilidade. Entretanto, digno de nota é que esta relação não é linear, conforme a Figura 1 (Andrade, Tavares & Valle, 2000; Baker, 2001; Hambleton & Swaminathan, 1985).

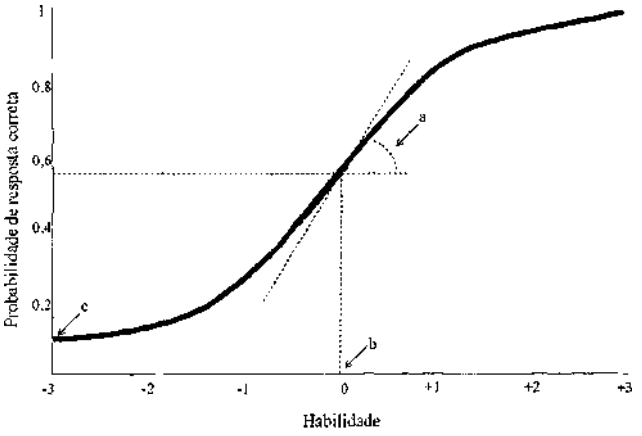


Figura 1. Curva Característica do Item (CCI).

Por meio da curva de CCI, também é possível identificar os parâmetros de dificuldade (*b*), discriminação (*a*) e acerto ao acaso (*c*) dos itens. O valor do parâmetro *b* representa o tamanho de θ necessário para que o item seja respondido corretamente ou endossado. Mais especificamente, o *b* representa a habilidade θ necessária para uma probabilidade de acerto igual a $(1 + c)/2$. No exemplo da Figura 1, $c = 1$, portanto $(1+0,1)/2 = 0,55$. Sendo assim, um θ de aproximadamente 0 é suficiente para que a probabilidade de acerto seja de 0,55 (o que indica que se trata de um item de dificuldade média). Assim, quando maior o valor de *b*, mais difícil é o item (Andrade, Tavares & Valle, 2000; Hambleton & Swaminathan,1985).

Para a discriminação do item, considera-se a inclinação da curva. Por esse motivo, o parâmetro *a* também é conhecido como *slope* (inclinação). Mais especificamente, o parâmetro *a* representa a inclinação da derivada da tangente (linha pontilhada e inclinada na Figura 1) da CCI no momento em que ela incide sobre o parâmetro *b*, isto é, o ponto de inflexão. Quanto maior a inclinação, maior a discriminação do item. Baixos valores de *a* indicam que o item tem pouco poder para diferenciar examinandos com habilidades θ semelhantes (Andrade, Tavares & Valle, 2000; Hambleton & Swaminathan, 1985).

Nos testes de múltipla escolha, é esperado que algumas pessoas, mesmo com habilidades muito baixas, lhes respondam corretamente devido ao acaso (ou "chute"). Para indicar o tamanho deste efeito, utiliza-se o parâmetro *c*, que se refere simplesmente à probabilidade de acerto ao acaso. Em outras palavras, o parâmetro *c* indica a probabilidade de um aluno com baixa habilidade responder corretamente ao item. Na CCI, este parâmetro

corresponde ao valor no qual a curva intercepta o eixo das ordenadas "Y" (eixo da probabilidade de acerto, neste caso). Na Figura 1, o item apresenta o parâmetro *c* estimado em 0,10. Ou seja, uma pessoa, mesmo com habilidade muito baixa, possui cerca de 10% de chances de responder corretamente ao item devido ao acaso (Andrade, Tavares & Valle, 2000; Hambleton & Swaminathan, 1985).

Uma das vantagens da TRI, em comparação com a TCT, refere-se à estimação dos parâmetros de itens e da habilidade do examinando. Na TRI, pelo menos teoricamente, a estimativa de habilidade do examinando não depende dos parâmetros dos itens, assim como os parâmetros dos itens independem das habilidades dos participantes. Ressalta-se que esta independência somente ocorre quando os pressupostos da TRI são satisfeitos e quando os dados se adequam ao modelo da TRI.

Na parte inferior da Figura 2, são indicadas as distribuições de habilidades dos grupos A e B. Nota-se que o grupo B é, em média, um pouco mais habilidoso do que o grupo A. Caso este item fosse analisado pela TCT, a proporção de questões respondidas corretamente seria maior para o grupo B. Consequentemente, o parâmetro dificuldade acompanharia esta tendência, e o item seria estimado como mais fácil para o grupo B do que para o grupo A. Entretanto, na TRI, embora as curvas de distribuição das habilidades do grupo A e B sejam diferentes, conforme Figura 2, as estimativas vão resultar na mesma CCI (se os pressupostos da TRI forem satisfeitos). Sendo assim, o item e a habilidade são compreendidos como invariantes (Hambleton & Swaminathan, 1985).

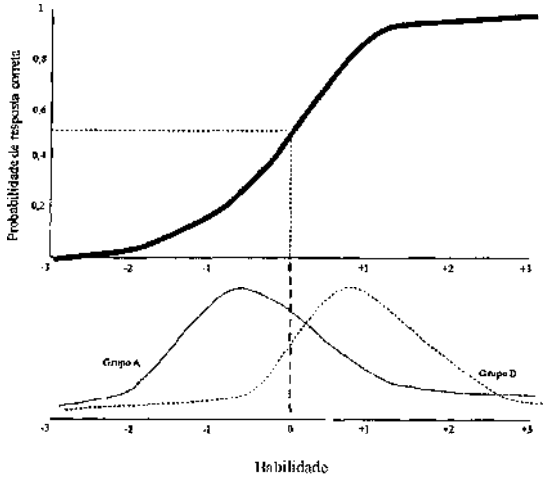


Figura 2. Curva Característica do Item estimada para dois grupos com distribuições de habilidades θ diferentes.

Outro conceito básico importante para a TRI é o da Curva de Informação (CI). Existem dois tipos de CI: Uma para o item (CII - Curva de Informação do Item) e uma para o teste (CIT - Curva de Informação do Teste). Ambas são definidas como a quantidade de informação fornecida pelo item (ou teste) para avaliação de uma habilidade θ (Andrade, Laros & Gouveia, 2010; Pasquali, 2007). Por meio dela, é possível avaliar para quais intervalos de habilidade o item é mais útil, considerando a maior quantidade de informação, bem como para qual faixa eles agregam mais erro do que informação. Ressalta-se que cada item de um teste possui uma CII, contribuindo para a CIT geral do teste. Alguns aspectos podem impactar na informação de um teste: (a) quanto maior a discriminiação dos itens (parâmetro a), maior será a informação; (b) a CI é maior quando o parâmetro b for igual ao θ médio do grupo de examinandos; (c) a CI diminui em função do acerto ao

acaso (parâmetro c); (d) a CI aumenta conforme o acréscimo de itens (Baker, 2001).

A Figura 3 reproduz uma CIT ilustrativa de um teste. A curva de linha contínua indica a informação do teste para os diferentes níveis de θ, enquanto a curva de linha pontilhada indica o erro padrão de medida. Neste exemplo, a maior quantidade de informação é relativa às habilidades entre -1 e 0. Para os níveis de θ inferiores a -2 e superiores a +1, o teste produz mais erro do que informação legítima. Em suma, é possível avaliar que o teste, neste exemplo, seria mais indicado para pessoas com habilidades médias.

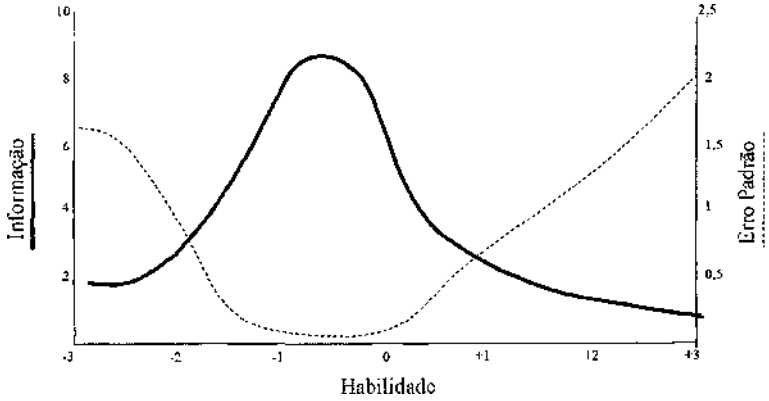


Figura 3. Curva de informação do teste (CIT).

Modelos da TRI

Hambleton, Swaminathan e Rogers (1991) salientaram que, embora seja possível conceber um número bastante grande de modelos

de TRI, poucos são utilizados na prática. O número de parâmetros estimados e o tipo de resposta ao item (dicotômico e escala Likert, por exemplo) são as principais características que diferenciam os modelos. De maneira geral, os mais utilizados são os modelos logísticos de um, dois e três parâmetros para itens dicotômicos, bem como os modelos para itens politômicos (Andrade, Laros & Gouveia, 2010).

O modelo de um parâmetro foi criado por Rasch (1960) e, posteriormente, adaptado para um modelo logístico por Wright e Stone (1979). Nesta perspectiva, é avaliada somente a dificuldade dos itens, ou seja, o parâmetro *b*. Sendo assim, assume-se que a dificuldade do item é a única característica deste que influencia o desempenho do examinando. Além disso, pressupõe-se que todos os itens possuem níveis iguais de discriminação. Na escala θ, itens com *b* inferiores a -2 podem ser considerados fáceis e itens com *b* superiores a +2, difíceis (Hambleton & Swaminathan, 1985). A equação do modelo logístico de um parâmetro é definida como segue:

$$P(U_{ij} = 1 \mid \theta_j) = \frac{1}{1 + e^{-D(\theta_j - b_i)}}$$

Onde:

- $P(U_{ij} = 1 \mid \theta_j)$ = é a probabilidade de um indivíduo com habilidade θj responder corretamente ao item *i*;

- D = é uma constante e igual a 1 (quando se deseja comparar o modelo logístico com os resultados obtidos pelo modelo da função ogiva normal, utiliza-se D = 1,7);
- E = é uma constante igual a 2,718;
- b_i = é o parâmetro de dificuldade do item i.

Espera-se que essa equação não cause pesadelos à Maria, que está buscando auxílio a uma demanda de avaliação psicológica prática. Entretanto, esta equação diz respeito simplesmente à curva que representa o modelo de um parâmetro. Retirando as constantes E e D, restará apenas o parâmetro b e o θ, que determinam a probabilidade de acertar o item.

O modelo de dois parâmetros foi desenvolvido por Lord (1980), sendo adaptado posteriormente por outros autores. Neste modelo, além do parâmetro de dificuldade b, também se avalia a capacidade discriminativa dos itens (parâmetro a). Segue a equação do modelo de dois parâmetros:

$$P(U_{ij} = 1 \mid \theta_j) = \frac{1}{1 + e^{-D\, a_i(\theta_j - b_i)}}$$

Maria deve ter notado que a única alteração da equação do modelo de um parâmetro para o de dois parâmetros é a inserção da letra $a_j$que corresponde, exatamente, à discriminação do item j. Na CCI, este parâmetro se refere à inclinação da curva, sendo que as curvas mais inclinadas indicam itens mais úteis para diferenciar as habilidades dos avaliandos (Andrade, Laros & Gouveia, 2010).

Teoricamente, o parâmetro *a* pode assumir valores entre $-\infty$ e $+\infty$. Entretanto, itens com discriminação menor do que 0 devem ser excluídos, pois algo deve estar errado com eles. Isso ocorre, geralmente, em razão de um gabarito errado ou de itens confusos, nos quais há mais de uma alternativa correta. Portanto, na prática, a discriminação varia de 0 a 2. Baker (2001) afirma que os valores de *b* entre 0,65 e 1,34 indicam um poder discriminativo moderado do item; entre 1,35 e 1,69, alto; acima de 1,70, muito alto. Embora sejam desejáveis itens com discriminação minimamente moderada, itens extremamente discriminativos não são úteis para algumas situações de avaliação psicológica, pois separam os indivíduos basicamente em dois grupos (com e sem habilidade θ).

O modelo logístico de três parâmetros, também desenvolvido por Lord (1974, 1980), acrescentou a probabilidade de acerto ao acaso, ou pseudoprobabilidade de acerto (representado pela letra *c*). Para a equação deste modelo, Maria já espera que seja inserido apenas o valor do parâmetro *c* (e alguns ajustes, obviamente). Sendo assim, a equação do modelo de três parâmetros pode ser escrita da seguinte forma:

$$P(U_{ij} = 1 \mid \theta_j) = c_i + (1 - c_i) \frac{1}{1 + e^{-D\,a_i(\theta_j - b_i)}}$$

Este modelo é bastante útil para os testes nos quais são oferecidas as alternativas de respostas (ou opções de marcação). O parâmetro *c* pode variar de zero a um. Espera-se que, para itens bons com cinco alternativas, o parâmetro *c* não ultrapasse 0,20 (1 / 5 alternativas = 0,2); para quatro alternativas, o valor máximo seria de 0,25 (Andrade, Laros & Gouveia, 2010). Para

itens problemáticos, o parâmetro c aumenta com a presença de alternativas não atraentes ou que estejam obviamente incorretas. Ou seja, para os itens nos quais é bastante fácil detectar uma ou duas alternativas incorretas, restarão apenas outras duas ou três alternativas para o "chute", o que aumenta a probabilidade de acerto ao acaso.

Os modelos anteriormente apresentados pressupõem a utilização de itens dicotômicos. Na prática, estes são os mais utilizados pelos profissionais da área. Todavia, é necessário destacar que uma grande parte dos testes psicológicos utilizam itens politômicos (escala Likert de quatro pontos, por exemplo). Para que a TRI seja utilizada na construção destes instrumentos, alguns modelos para escalas politômicas foram propostos (veja, por exemplo, Andrich, 1978; Samejima, 1974).

Nos modelos de TRI para itens politômicos, são estimadas as probalidades de um participante dar a resposta da categoria x ao item *i*. Sendo assim, passa-se a avaliar a probabilidade de endosso da categoria x, em vez da resposta certa. O tipo de escala utilizada pelo item é a diferença básica entre os dois principais modelos. O modelo de resposta gradual de Samejima (1974) assume que as categorias do item podem ser ordenadas entre si, enquanto o modelo de escala gradual proposto por Andrich (1978) assume que, além da possibilidade de ordenar as categorias, os escores das categorias são igualmente espaçados.

Nota-se que os modelos politômicos são representados, graficamente, por meio das CCI's, em relação às categorias, além da habilidade e da probabilidade de endosso (Andrade, Tavares & Valle, 2000). Na Figura 4, é apresentado um exemplo de CCI para itens politômicos (modelo de escala gradual).

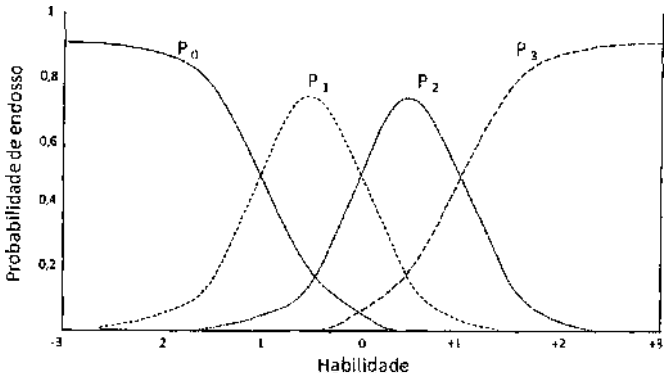


Figura 4. CCI's para um item politômico.

A Figura 4 refere-se a um item politômico com quatro possibilidades de resposta. Portanto, são apresentadas quatro curvas, cuja ordem indica que os participantes com maiores níveis de habilidade (ou traço latente, neste caso) tendem a endossar as categorias que representam os valores mais altos (Andrade, Tavares & Valle, 2000). Suponha-se, por exemplo, que o item corresponda à avaliação do sentimento de felicidade na última semana. A categoria 0 indicaria a resposta de pouco feliz e a categoria 3, muito feliz. Na Figura 3, pessoas com θ até -1 tenderiam a endossar a categoria 0. Já as pessoas com θ entre -1 e 0 tenderiam a alcançar a categoria 1. A maior probabilidade de resposta para as pessoas com com θ entre 0 e +1 é a categoria 2. Finalmente, as pessoas com θ maior do que +1 tenderiam a endossar a categoria 3.

Pressupostos

Da mesma forma que outros modelos de estimação, a TRI possui alguns pressupostos. Um deles diz respeito à unidimensionalidade,

ou seja, para estimação dos parâmetros, é condição necessária que o conjunto de itens avalie apenas um único traço latente. As respostas a um teste de matemática, por exemplo, não podem sofrer influência da compreensão de língua portuguesa dos participantes. O problema deste pressuposto, entretanto, é que ele dificilmente pode ser plenamente satisfeito. As respostas a um teste de inteligência, por exemplo, com frequência não estão isentas de uma série de outros aspectos, tais como motivação, experiência prévia em outras testagens, calma para responder, etc. Para resolver este paradoxo, alguns autores têm ressaltado que a presença de um traço latente dominante é o suficiente para satisfazer este pressuposto (Hambleton, Swaminathan & Rogers, 1991; Condé & Laros, 2007; Laros, Pasquali & Rodrigues, 2000; Vitória, Almeida & Primi, 2006).

Destaca-se que, a despeito das dificuldades, é importante avaliar a unidimensionalidade. Condé e Laros (2007) concluíram, por meio de um estudo empírico com aproximadamente 19 mil alunos, que a estimação da habilidade apresenta menor relação com a dificuldade do conjunto de itens conforme aumenta a unidimensionalidade. Em outra palavras, a invariância das estimativas da habilidade (uma das vantagens da TRI sobre a TCT) aumenta em função da unidimensionalidade. No intuito de avaliar este pressuposto, a Análise Fatorial de Informação Plena (*Full Information Factor Analysis – FIFA*) é frequentemente utilizada para itens dicotômicos. Para os testes que avaliam mais de uma dimensão (por exemplo, testes de personalidade), uma solução adotada é estimar os parâmetros dos itens para cada fator. Neste caso, deve garantir-se a unidimensionalidade dos fatores por meio de uma análise fatorial prévia.

Outro pressuposto importante é o da independência local. Ele visa a garantir que os itens sejam respondidos exclusivamente em função da habilidade θ dominante. Em outras palavras, a resposta de um examinando a um item x não afeta sua resposta aos demais itens (Andrade, Laros & Gouveia, 2010; Hambleton & Swaminathan, 1985; Lord, 1980). Em caso de violação do pressuposto, as respostas dos participantes passam a sofrer influências de aspectos não controlados na análise, por exemplo, se o próprio item oferecer sugestões da resposta correta, algumas pessoas as perceberão e outras não. Neste caso, o fator "perceber as dicas" estaria influenciando as respostas dos examinandos, sem o controle do pesquisador. A melhor maneira de lidar com este problema é prevenir-lo durante a construção dos itens (Embreston & Reise, 2000). Em geral, o pressuposto da independência local pode ser assumido uma vez que o pressuposto da unidimensionalidade é satisfeito (Condé & Laros, 2007; Pasquali & Primi, 2003).

O tamanho da amostra, embora não seja exatamente um pressuposto, também é um aspecto importante da TRI. A amostra tem sido um "calcanhar de Aquiles" para a psicologia e para as ciências humanas em geral. Na TRI, não há concenso sobre o seu tamanho ideal, ainda que alguns autores sugiram um número mínimo entre cem e trezentos participantes (Comrey & Lee, 1992; Pasquali, 2007). A despeito de tal divergência, um ponto de corte testado empiricamente foi apresentado por Nunes e Primi (2005). Eles avaliaram diversas subamostras de um banco com 44 mil participantes, e as conclusões indicaram que as estimativas dos parâmetros tendem a ser mais estáveis quando geradas com amostras maiores de duzentos sujeitos.

Principais aplicações

Tendo discutido os conceitos básicos, modelos e pressupostos da TRI, serão apresentadas, nesta seção, algumas das principais aplicações dela. Destaca-se a construção de instrumentos, a equalização e a testagem computadorizada.

Inegavelmente, a principal utilização da TRI é na construção de instrumentos e nas análise das suas qualidades psicométricas. Nunes e Primi (2009) descatam o frequente uso da TRI na seleção dos itens para as versões finais dos instrumentos. Normalmente, a partir dos parâmetros estimados, selecionam-se os itens mais discriminativos, bem como os que apresentam os níveis de dificuldade mais adequados para a população alvo do teste. Sendo assim, buscam-se eliminar os itens que pouco contribuem para a curva de informação do teste. Por meio desta seleção, também é possível diminuir o número de itens do teste, sem, no entanto, causar impacto negativo na qualidade do instrumento. Algumas pesquisas brasileiras fizeram uso da TRI na construção de instrumentos, como os testes de inteligência WISC-III (Nascimento & Figueiredo, 2002), BPR-5 (Primi & Almeida, 1998) e SON-R 2½-7 (Laros, Tellegen, Jesus & Karino, 2011) e a Bateria Fatorial de Personalidade (Nunes, Hutz & Nunes, 2008).

Outra importante aplicação da TRI diz respeito ao processo de equalização. Este objetiva tornar comparáveis os escores de dois grupos de pessoas (ou mais) submetidos a duas versões distintas de um teste, desde que ambos avaliem o mesmo construto (Kolen & Brennan, 2010). Por exemplo, o escore de Maria no teste de inteligência x torna-se comparável com o escore de João no teste de inteligência y, desde que mantidos alguns pressupostos. Este

procedimento é bastane útil quando não é possível aplicar um mesmo instrumento em dois momentos e/ou grupos. Por exemplo, um pesquisador está interessado em avaliar o desenvolvimento cognitivo de bebês, crianças, adolescente e adultos. Considerando a dificuldade de encontrar um único instrumento que avalie todas estas faixas etárias, o pesquisador constrói quatro testes e os aplica. Obviamente, ele é bastante inteligente e pensou nos pressupostos, o que lhe permitiu comparar o desenvolvimento cognitivo dos quatro grupos submetidos a quatro testes diferentes.

A equalização na TRI é possível ao considerar que a estimação da habilidade θ de um examinando é invariante entre os itens do teste. Ou seja, conhecidos os parâmetros dos itens, dois grupos de participantes que respondem a dois testes distintos terão suas habilidades estimadas na mesma escala, tornando-os comparáveis. A psicóloga recém-formada Maria, agora, deve entender como tal comparação é possível. Entretanto, ainda é necessário informá-la de que isso pode ser realizado somente se alguns pressupostos forem observados. Além dos pressupostos usuais da TRI discutidos anteriormente, é importante que ambos os testes (ou subtestes) avaliem exatamente o mesmo construto psicológico. Caso um teste avalie ansiedade e outro depressão, por exemplo, o processo de equalização torna-se inviável.

Outro pressuposto importante é a presença de um delineamento que permita lincar os diferentes testes (*linking designs*). Embora existam quatro ou cinco delineamentos conhecidos, os mais utilizados são o de itens âncoras e o de grupos em comum. No primeiro (delineamento de itens âncoras), são construídos dois ou mais testes distintos, mas que mantêm alguns itens (âncoras) em comum, por meio dos quais os testes são lincados. Outro

lhamento possível diz respeito à construção de testes distintos para grupos distintos, mas também aplicados, concomitantemente, a um grupo comum. Neste caso, o grupo comum responderia a ambos os testes, permitindo o linque (Hambleton, Swaminathan & Rogers, 1991).

Finalmente, a TRI é utilizada nas principais testagens adaptativas computadorizadas (TAC - *Computerized Adaptative Testing*, CAT). Baseando-se nos parâmetros da TRI, um sistema informatizado seleciona e aplica os itens mais adaptados ao perfil do participante. Maria deve estar lembrada que, por meio da curva de informação, é possível saber para quais níveis de habilidade um determinado item é mais indicado. Por um lado, a aplicação de itens de inteligência muito difíceis, por exemplo, pode facilmente desestimular um examinando com baixas habilidades cognitivas. Por outro lado, itens muito fáceis podem ser ridicularizados por pessoas com altas habilidades cognitivas. Sendo assim, a aplicação de itens de dificuldade condizentes com a habilidade θ do participante pode aumentar a confiabilidade dos escores e diminuir o tempo de aplicação (Hambleton, Swaminathan & Rogers, 1991; Nunes & Primi, 2009).

Embora os algoritmos envolvidos sejam complexos, a lógica da TAC é bastante simples. O computador, primeiramente, seleciona alguns itens de boa discriminação e dificuldade média. Então, é realizada uma estimativa preliminar da habilidade θ do examinando, com base nas respostas aos itens iniciais. Conforme o participante erra ou acerta, o computador seleciona e apresenta, respectivamente, itens mais fáceis ou difíceis, processo durante o qual a estimativa de habilidade é ajustada. O computador encerra o processo ao atingir um número predeterminado de itens e/ou

um valor de erro mínimo, também predeterminado (Hambleton, Swaminathan & Rogers, 1991).

Além das vantagens de maior precissão e menor tempo de testagem oferecidas pela TAC, ela reduz consideravelmente o grave problema relacionado à divulgação dos gabaritos e das respostas dos testes (Nunes & Primi, 2009). Ora, os itens respondidos por um participante são distintos daqueles respondidos por outro participante, e assim por diante. Alguns itens são obviamente repetidos, mas apresentados, normalmente, em ordem diferente. Tudo isso dificulta o processo de memorização, cópia e divulgação dos itens, evitando fraudes nos processos de avaliação psicológica. Obviamente, quanto mais amplo e melhor for o banco de itens disponíveis ao sistema informatizado, mais precisos serão os resultados e menores serão os problemas com a divulgação do teste. Reside aqui a principal dificuldade da TAC: custo! Para montar um banco suficientemente grande e bom, o custo financeiro é razoavelmente elevado.

Não seria justo encerrar este subcapítulo sem mencionar as aplicações nas avaliações educacionais. Sem dúvida, a TRI tem sido bastante útil em provas nacionais, tais como ENEM, SAEB, na avaliação dos estudantes de São Paulo, do Rio de Janeiro e da Bahia, dentre outros, bem como em avaliações internacionais, tais como o SAT (EUA) e o exame TOEFL. Todavia esta discussão extrapola o escopo deste capítulo. Para outras informações, sugerem-se as seguintes referências: Andrade, Laros e Gouveia (2010); Laros, Pasquali e Rorigues (2000); Nunes e Primi (2009), dentre outros.

Considerações finais

Este capítulo teve como objetivo apresentar uma introdução sobre a Teoria Clássica dos Testes e a Teoria de Resposta ao Item. Embora se tenham explicado brevemente os principais aspectos da TCT, nosso foco maior era discutir a TRI, considerando sua crescente utilização, o que, obviamente, não é reflexo da importância das teorias. Embora a TRI represente alguns avanços na ciência psicométrica, a TCT, de forma alguma, deve perder o seu espaço, na medida em que ainda é bastante útil em diversos contextos. A TCT, aparentemente, apresenta a vantagem de maior robustez à violação dos pressupostos da TRI. Além disso, existe um número relativamente restrito de pesquisas sobre o efeito de tais violações nas estimativas dos parâmetros e habilidades geradas pela TRI. Ademais, a TCT e a TRI apresentam correlações altas no que diz respeito às estimativas de dificuldade e de habilidade, indicando que elas são parecidas nesses aspectos (Fan, 1998; MacDonald & Paunonen, 2002).

Espera-se que este capítulo tenha auxiliado a psicóloga Maria a resolver as suas dúvidas. Maria precisa escolher, aplicar e analisar alguns testes psicológicos. Esta tarefa deve estar embasada na adequação dos instrumentos à demanda de avaliação, bem como na qualidade psicométrica dos instrumentos. Portanto, acredita-se que o conhecimento básico de psicometria é fundamental para a utilização adequada dos testes psicológicos.

Questões

1) Quais são as vantagens e desvantagens da TRI? Tais vantagens e desvantagens indicam que se deva abandonar a TCT?
2) Elabore um quadro resumo contendo os principais conceitos e as principais características da estimação dos parâmetros de dificuldade, discriminação e "chute" na TCT e na TRI. Destaque as principais semelhanças e diferenças.
3) Defina a Curva de Informação do Teste e a Curva de Informação do Item. Indique suas principais utilidades.
4) Quais são os pressupostos dos principais modelos da TRI? Por que é importante avaliá-los?
5) Quais são as principais aplicações da TRI? Qual delas você julga ser a mais relevante para a avaliação psicológica (justifique a sua escolha)?

Referências

Andrade, D. F. de, Tavares, H. R., & Valle, R. da C. (2000). *Teoria de resposta ao item: conceitos e aplicações*. São Paulo: Associação Brasileira de Estatística.

Andrade, J. M., Laros, J. A., & Gouveia, V. V. (2010). O uso da Teoria de Resposta ao Item em avaliações educacionais: Diretrizes para pesquisadores. *Avaliação Psicológica, 9*, 421-435.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). New York: Eric Clearinghouse on Assessment and Evaluation.

Comrey, A. L., & Lee, H. B. (1992). *A first course for identifying biased test items*. Hillsdale: Erlbaum.

Condé, F. N., & Laros, J. A. (2007). Unidimensionalidade e propriedade de invariância das estimativas da habilidade pela TRI. *Avaliação Psicológica, 6*, 205-215.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Publishers.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-381.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. London: Sage.

Hogan, T. P. (2006). Introdução à prática de testes psicológicos. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora.

Kolen, M. J., & Brennan, R. L. (2010). *Test Equating, Scaling, and Linking: Methods and Practices* (2nd ed.). New York: Springer.

Laros, J. A., Tellegen, P. J., Jesus, G. R., de, & Karino, C. A. (2011). *SON-R 2½-7[a]: Teste não-verbal de inteligência. Manual com validação e normatização brasileira.* São Paulo: Casa do Psicólogo.

Laros, J. A., Pasquali, L., & Rodrigues, M. M. (2000). *Análise da unidimensionalidade das provas do SAEB. Relatório Técnico.* Brasília: LabPAM.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-267.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Lawrence Eribaum.

MacDonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement, 62*, 921-943.

Nascimento, E., & Figueiredo, V. L. (2002). WISC-III e WAIS-III: Alterações nas versões originais americanas decorrentes das adaptações para uso no Brasil. *Psicologia: Reflexão e Crítica* (15), 603-612.

Nunes, C. H., & Primi, R. (2005). Impacto do tamanho da amostra na calibração de itens e estimativa de escores por Teoria de Resposta ao Item. *Avaliação Psicológica, 4*, 141-153.

Nunes, C. H., & Primi, R. (2009). Teoria de Resposta ao Item: Conceitos e aplicações na psicologia e na educação. In C. S. Hutz (Org.), *Avanços e polêmicas em avaliação psicológica* (pp. 25 -70). São Paulo: Casa do Psicólogo.

Nunes, C. H., Hutz, C. S., & Nunes, M. F. (2008). *Bateria Fatorial de Personalidade (BFP): Manual técnico.* Itatiba: Casa do Psicólogo.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

Pasquali, L. (2003). *Psicometria: Teoria dos testes na Psicologia e na Educação.* Petrópolis: Vozes.

Pasquali, L. (2007). *TRI - Teoria de Resposta ao Item: Teoria, Procedimentos e Aplicações.* Brasília: LabPAM.

Pasquali, L., & Primi, R. (2003). Fundamentos da Teoria da Resposta ao Item - TRI. *Avaliação Psicológica, 2*, 99-110.

Primi, R., & Almeida, L. S. (1998). *BPR-5: Manual Técnico*. São Paulo: Casa do Psicólogo.

...ch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: Mesa Press.

...mejima, F. (1974). Normal ogive model on the continuous response level in the multi-dimensional latent space. *Psychometrika, 39*, 111-121.

...nória, F., Almeida, L. S., & Primi, R. (2006). Unidimensionalidade em testes psicológicos: conceito, estratégias e dificuldades na avaliação. *Revista de Psicologia da Vetor, 7*, 1-7.

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

Capítulo 5

Validade e precisão de testes psicológicos

Gisele Aparecida da Silva Alves
Mayra Silva de Souza
Makilim Nunes Baptista

Introdução

Todos os psicólogos já ouviram falar dos conceitos validade e precisão, mesmo que nunca tenham sentido a necessidade de se aprofundar neles. Um estudo realizado por Vendramini e Lopes (2008), com trinta psicólogos de diversas áreas de atuação e trinta estudantes concluintes do curso de Psicologia, revelou que mais de 50% dos participantes não conseguem ler as informações sobre evidências de validade e precisão, apesar de quase 30% dos profissionais considerarem a leitura desses aspectos importante para se manterem atualizados e aproximadamente 40% dos estudantes avaliarem que esse dado é útil para a utilização segura dos testes.

Ocorre que esses dois conceitos em questão são imprescindíveis na prática do psicólogo que faz uso de testes psicológicos; muitas vezes, entretanto, não são abordados adequadamente no curso de graduação em Psicologia.

Vários autores questionam a formação do psicólogo no Brasil e apontam falhas nesse percurso (Bettoi & Simão, 2000; Dias, 2001; Francisco & Bastos, 1992; Pereira & Carellos, 1995). Dentre as críticas mais citadas, ressalta-se a distância entre teoria e prática, com falta de articulação entre elas, e a limitação de uma formação voltada para a prática em clínica particular, o que dificulta o profissional de psicologia a lidar com demandas diversificadas (Moura, 1999).

Sobre a formação específica em avaliação psicológica, outros pesquisadores expõem dificuldades como as abordagens geral e reduzida do conteúdo, incompatíveis com a demanda prática da área, teoria e técnica precárias no ensinamento de testes e na confecção de relatórios, assim como a não reciclagem de professores e a precariedade de material que estes utilizam (Hutz & Bandeira, 2003; Jacquemin, 1995; Noronha & Alchieri, 2004; Pereira & Carellos, 1995; Sbardelini, 1991; Simões, 1999; Wechsler & Guzzo, 1999).

Os conceitos de validade e precisão popularizaram-se no meio profissional com o movimento do Conselho Federal de Psicologia (CFP), em 2001, que se preocupou com a melhora da qualidade dos testes, que até então estavam sendo utilizados na prática profissional do psicólogo, em avaliação psicológica. Essa ação foi reflexo da crise do uso de testes, na década de 1960, quando estes deixaram de ser utilizados porque se acreditava que não cumpriam os objetivos a que se propunham (Hutz & Bandeira, 2003; Urbina,

2007). Dentre outras exigências do CFP, como a apresentação de fundamentação teórica do instrumento, um teste precisa possuir evidências empíricas de validade e precisão relatadas em seu manual para que seja aprovado para o uso profissional e para que seja também devidamente comercializado.

Um exemplo de teste muito utilizado na prática do psicólogo, e que gerou muita polêmica pelo questionamento da sua validade, é o teste de Wartegg. Trata-se de uma técnica gráfica para avaliação da personalidade segundo a teoria dos arquétipos de Jung. Assim, são apresentados estímulos em oito campos, nos quais, basicamente, o respondente deve continuar o desenho.

Estudos realizados com o teste em questão (como o de Salazar, Tróccoli & Vasconcellos, 2001, e o de Souza, Primi & Miguel, 2007) não conseguiram encontrar evidências de validade suficientes para sustentar as interpretações sugeridas no seu manual, e, dessa forma, não é possível assegurar o que o teste realmente avalia, impossibilitando, desse modo, seu uso na prática profissional. No entanto, isso não impede que novos estudos sejam realizados com este instrumento, com o objetivo de buscar evidências de validade, bem como com outros objetivos relacionados às qualidades psicométricas do teste. O teste Wartegg poderá ser utilizado novamente, após ser considerado aprovado quando da apreciação feita pelo Conselho Federal de Psicologia, caso sejam encontradas evidências de validade, bem como índices de precisão adequados (dos quais se tratará nos próximos tópicos), e outros requisitos sejam atendidos.

A resolução CFP 006/2004 altera o artigo 14 da Resolução CFP n.º 002/2003, estipulando que estudos referentes à validade e à precisão dos instrumentos devem ser realizados de modo a

não ultrapassar um período de vinte anos. Assim, nenhum teste é aprovado ou reprovado para sempre, sendo que apenas pode não haver naquele momento estudos que evidenciem e justifiquem seu uso de forma segura.

Validade

A validade de um teste, basicamente, diz respeito ao cumprimento da tarefa de medir o que este se destinou a medir, ou seja, é a comprovação que o teste mede aquilo a que ele se propõe. É comum também encontrar a definição de validade como o nível em que o teste mede a característica que quer medir. Em muitos testes, encontram-se evidências de validade, porém a pergunta que se faz é: Será que essas evidências são suficientes para essa avaliação? Por isso validade se refere aos questionamentos: "O teste avalia o que ele anteriormente se propôs a avaliar?", e "Quão bem ele faz isso?" (Anastasi & Urbina, 2000; Cronbach, 1996; Hogan, 2006).

O termo "validade" remete a um conceito unitário, como da possibilidade de um teste ser válido ou não, porém as coisas não funcionam dessa forma, num ponto de vista do tudo ou nada (Hogan, 2006; Urbina, 2007). Dessa maneira, passou-se a considerar que um teste pode possuir "evidências de validade", pois se buscam saber as suas qualidades diante de um propósito ou uma utilização particular. Portanto, um mesmo teste pode servir a um objetivo de avaliação e não servir a outro diferente. Esses objetivos diferentes podem ser, por exemplo, populações diferentes (estudantes, pacientes psiquiátricos, população geral, etc.) ou

contextos diversos (clínica, hospital, trânsito etc.). Imagine-se um teste de personalidade usado para propósitos diferentes: a) numa clínica, onde o psicólogo vai explorar as características de personalidade do paciente, com a finalidade de trabalhar esses aspectos em psicoterapia, e b) no trânsito, no qual serão avaliados aspectos da personalidade do candidato à carteira de habilitação, sendo aprovados aqueles considerados aptos, com vistas a diminuir os acidentes no trânsito.

Na realidade, não é o teste que possui essas evidências de validade; são as interpretações feitas a partir dos resultados encontrados numa pesquisa com o teste em questão. Isso se deve ao fato de que as características psicológicas avaliadas não são diretamente observadas; exemplificando, tem-se a possibilidade de avaliar a altura de uma pessoa com uma fita métrica, mas, nos estudos de validade de testes psicológicos, faz-se o uso de números, de análises estatísticas, porém sempre se tem de atribuir um significado para os números encontrados, uma interpretação (AERA, APA & NCME, 1999).

Foi pensando assim que a definição de validade passou a ser concebida como "o grau no qual as interpretações obtidas dos dados empíricos do teste encontram sustentação em base científica sólida" Urbina (2007). Tal autor propôs ainda que sejam consideradas as evidências encontradas de forma acumuladas, de maneira que o grau dessas evidências concorde com os resultados do teste para os objetivos propostos.

Uma vez que a conceituação de validade foi concluída, deve-se pensar em como se buscam evidências de validade. Numa abordagem clássica, a validade foi dividida em três tipos (Anastasi & Urbina, 2000), a saber:

a) ***Validade de conteúdo***: Responde à pergunta "os itens do teste representam adequadamente a característica que se quer avaliar?";

b) ***Validade de critério***: Responde à pergunta "os itens do teste conseguem fazer uma previsão de uma variável externa ao teste no futuro ou no presente?". Por exemplo, um teste vocacional é utilizado com o propósito de avaliar se o indivíduo tem aptidões necessárias para exercer determinada profissão, e, se evidências são encontradas nesse sentido, diz-se de uma evidência de validade de critério;

c) ***Validade de construto***: Responde à pergunta "quanto os itens do teste realmente medem uma determinada característica?". Pensando na evolução do conceito, esse tipo engloba o conceito atual de validade, pois todos os estudos nesse sentido buscam responder a essa questão.

Esta definição de validade de Anastasi e Urbina (2000), chamada de definição tripartite, foi questionada e aprimorada posteriormente. Achou-se importante, mesmo assim, apresentá-la neste capítulo para dar ao leitor uma visão histórica das definições de validade e também porque esta definição ainda é utilizada nos manuais anteriores às novas nomenclaturas e em outras discussões sobre validade feitas pela comunidade científica. Um dos autores que contribuiu significativamente para a reformulação da definição tripartite foi Messick (1989), quem, em um dos questionamentos feitos a esta definição, argumentou que tanto a validade de conteúdo quanto a de critério também apresentam informações sobre o construto, de modo que quase toda a informação sobre o teste contribuirá para sua validade de construto,

de formas diversas. Dessa maneira, validade de construto passou a ser entendida como um conceito abrangente em que se incluem outras formas de validade (Primi, Muniz & Nunes, 2009).

Assim, contemporaneamente são utilizadas outras nomenclaturas, que foram reformuladas pela AERA, APA & NCME (1999), e distinguem-se, não em tipos, mas em fontes pelas quais se é possível encontrar evidências de validade, a saber:

a) ***Evidências de validade baseadas no conteúdo***: Nessa fonte, busca-se uma relação entre o conteúdo do teste (o que seus itens abordam) e o domínio que se quer avaliar. Para se ter uma evidência de validade baseada no conteúdo, é necessário que os itens do teste estejam representando de forma adequada a característica psicológica que se quer avaliar. Por exemplo, um teste para avaliação de depressão precisa conter itens que descrevam a depressão. Pensando no conceito de depressão pelo DSM-IV-TR (APA, 2002), o transtorno depressivo maior envolve sintomas centrais que são o humor deprimido e a anedonia, e outros sintomas somáticos, como alterações de apetite, de sono, dificuldade de concentração, pessimismo, sentimentos de culpa, ideias de morte, dentre outros. Dessa forma, um teste que avalia depressão precisa conter tanto os sintomas centrais, como os outros sintomas, pois é a junção deles (e mais outros critérios, no caso) que vai poder configurar ou não um transtorno depressivo. Para avaliar se o conteúdo dos itens do teste é adequado ou não, geralmente são chamados especialistas na área (chamados juízes), que vão avaliar se a descrição do conteúdo foi feita de maneira cuidadosa, desmembrando seus componentes principais, e julgar a relação entre o que o teste traz em seu conteúdo e o que deveria trazer, de acordo com a literatura. Essa fonte equivale ao tipo de validade de conteúdo.

b) ***Evidências de validade baseadas nas relações com outras variáveis***: Nessa fonte, são buscadas relações entre os escores do teste e outras variáveis medindo a mesma característica, características relacionadas ou características diferentes. As outras variáveis podem ser sexo, idade, desempenho acadêmico, critério diagnóstico e também outros testes. Por exemplo, um teste que visa a avaliar a inteligência de alunos do terceiro ano do ensino fundamental pode ser comparado ao desempenho acadêmico desses alunos nas disciplinas da escola (notas ao final do ano), e, se for encontrada uma boa relação entre os dois (pontuações que indiquem inteligência alta e sucesso escolar, por exemplo), pode-se interpretar que foi encontrada uma evidência de validade baseada na relação com outras variáveis, critérios externos. A relação entre essas duas características, que são as mesmas, está ilustrada na Figura 1, a seguir. As Figuras 2 e 3 ilustram as relações medindo características relacionadas e características diferentes.

A partir dessas relações, portanto, é possível inferir evidências de validade que convergem (mesma característica ou características relacionadas) ou divergem (características diferentes). Quanto às evidências de validade que divergem, ilustradas pela Figura 3, pode-se citar um estudo que relaciona dois instrumentos: um avaliando inteligência e outro, personalidade. Assim, esperam-se relações muito baixas entre esses testes, já que avaliam características (construtos) diferentes. Se as relações encontradas forem de fato baixas, pode-se interpretar que foi constatada evidência de validade divergente.

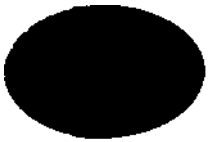


Figura 1

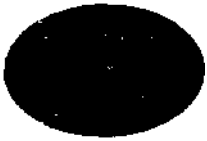


Figura 2

Figura 3

Com essa fonte de validade, também é possível obter dados sobre a capacidade do teste de predição. Um exemplo é um teste utilizado na seleção de pessoal para uma vaga de emprego numa empresa, capaz de predizer o sucesso do indivíduo avaliado no cargo pretendido. Em uma avaliação para cargos hierárquicos mais altos, por exemplo, uma das habilidades a serem avaliadas é a da liderança, que pode predizer o desempenho desses candidatos quando ocuparem o cargo que requer esse tipo de habilidade. Essa fonte equivale à validade de critério.

c) ***Evidências de validade baseadas na estrutura interna:*** Como o próprio nome já sugere, essa fonte de evidência de validade busca relação entre o teste e seus itens. Com o uso de análises estatísticas, é possível saber a contribuição de cada item no resultado total do teste (correlaciona-se um item ao resultado total do teste, e, se essa relação for significativa, supõe-se que o item contribui para o teste no geral, na representação da característica que se pretende medir); assim os itens podem ou

não ser considerados adequados para avaliação do domínio que se quer medir. Outra forma de se avaliar esse tipo de evidência é verificando o agrupamento de itens em fatores já previstos teoricamente. A Bateria Fatorial de Personalidade (BFP), por exemplo, é um instrumento psicológico construído para a avaliação da personalidade a partir do modelo dos Cinco Grandes Fatores (CGF), que foram confirmados pelo procedimento estatístico de análise fatorial, apresentando os fatores: Extroversão, Socialização, Realização, Neuroticismo e Abertura para novas experiências (Nunes, Hutz & Nunes, 2010).

d) ***Evidências de validade baseadas no processo de resposta***: Fornecem dados sobre processos mentais presentes na execução das tarefas propostas pelo teste, atribuindo-se significado psicológico para a realização correta do item a partir das relações entre seus componentes cognitivos. Baseando-se sempre na teoria referência da característica avaliada pelo teste, são criados modelos explicativos do processamento mental que ocorre durante a execução das tarefas propostas nos itens do teste e previsões dos comportamentos de acerto, tempo de reação, etc. As observações dos padrões de resposta são comparadas ao modelo teórico, e, quanto mais próximas, maior a confiança no modelo teórico de base para a interpretação do que o teste avalia. Outra maneira de estudar essa fonte de validade é analisar as respostas do indivíduo, quando questionado sobre suas estratégias para responder aos itens do teste (Primi, Muniz & Nunes, 2009). Cunha e Santos (2009) realizaram um estudo que objetivava a busca por essa evidência de validade, por meio da análise das respostas de crianças ao teste Cloze, em que seriam exploradas diferenças qualitativas nos erros apresentados. A partir dessa análise das respostas, foi possível verificar

que as crianças com médias mais altas cometeram mais erros lexicais, e as com médias mais baixas, erros semânticos. Os resultados encontrados demonstraram o que era conceitualmente esperado, portanto, foram encontradas evidências de validade baseadas no processo de resposta para o Cloze.

e) ***Evidências de validade baseadas nas consequências da testagem***: Essa fonte avalia as consequências sociais do uso do teste para verificar se as implicações de sua utilização coincidem com os resultados desejados de acordo com a finalidade para a qual foi criado. A expectativa é a de que o teste contribua de maneira benéfica em contextos clínicos e escolares, em seleção de pessoal, etc. Porém, para que se obtenha o resultado desejado do uso do teste, não basta apenas que este seja validado, pois existem outras variáveis que podem interferir, de maneira a comprometer as interpretações resultantes da sua utilização. Buscar por essa fonte de evidência de validade implica ter uma visão ampliada da situação, e não somente do teste, na medida em que estão envolvidos, além do psicólogo responsável pela avaliação, os outros agentes (profissionais de outras áreas, governos, dentre outros) que fazem uso desses dados finais para tomada de decisões, e podem utilizar essa informação de maneira equivocada, enviesada, de modo a prejudicar indivíduos e sociedade, de maneira geral.

Uma crítica a essa fonte de validade é que ela se esquiva das propriedades de controle na pesquisa e na construção de testes; por outro lado, ampliar a situação da testagem e avaliar suas consequências pode ser visto como uma atitude ética (Primi, Muniz & Nunes, 2009). Como exemplo, um instrumento diagnóstico é aplicado num indivíduo e, se conseguir detectar precocemente uma doença e indicar uma intervenção adequada

ao caso, é sinal de que o efeito produzido foi benéfico, como desejado, e este resultado agrega evidência de validade consequencial ao teste utilizado; em contraposição, se essa avaliação provocar um diagnóstico equivocado ou indicações desfavoráveis de intervenção, inicia-se um questionamento sobre a validade daquele teste para avaliação naquele contexto.

Precisão

A precisão (também conhecida como confiabilidade ou, ainda, fidedignidade) refere-se à estabilidade do teste, de maneira que, quanto mais próximas forem as pontuações obtidas por métodos ou em situações diferentes, maior será a consistência do teste (Anastasi & Urbina 2000; Cronbach, 1996). Imagine-se uma só pessoa sendo submetida a um teste que avalia traços de personalidade, realizado em dois lugares diferentes, por pessoas diferentes. Como se trata de uma só pessoa que está sendo avaliada, e como está sendo avaliada nos dois lugares pelo mesmo teste (que avalia a mesma coisa), espera-se que os resultados sejam muito próximos. Se assim ocorrer, como esperado, pode-se conferir precisão aos resultados do teste, e, no caso de dois resultados diferentes, desconfia-se de um erro de medida.

O conceito de precisão opõe-se ao de erro de medida, de maneira que, quanto mais preciso for considerado um teste, significa que mais livre de erros ele se encontra. Dessa forma, a precisão de um teste é determinada pelo nível com que suas pontuações são livres de erros. É necessário considerar qual fenômeno está sendo estudado, avaliar suas particularidades, pois fenômenos

psicológicos diferentes possuem características distintas e sofrem influência de diferentes fatores (Anastasi & Urbina 2000; Cronbach, 1996)

Nenhuma medida está livre de erro, e os erros que interferem no resultado de um teste podem vir de várias fontes, dentre as quais se destacam as relacionadas ao contexto da testagem (incluindo o aplicador, o avaliador, o ambiente de testagem e os motivos da aplicação do teste), ao testando e ao teste em si. Essas fontes de erro incluem condições emocionais, como disposição, ansiedade, fadiga, ou acertos ao acaso em determinadas situações, familiaridade com o conteúdo, subjetividade, ambiente barulhento, dentre outras. Se os devidos cuidados forem tomados no desenvolvimento, na seleção, na aplicação e na correção dos testes, parte dos erros provindos dessas três fontes pode ser anulada ou minimizada. Em contraposição, em situações nas quais o testando não responde às questões ou tenta falsear respostas que pensa ser desejáveis, não é possível manipular o erro, porém pode ser possível detectá-lo. Por isso é importante saber das práticas adequadas e dos procedimentos padronizados no uso dos testes, porque são formas de reduzir os erros na testagem (Urbina, 2007).

Existem diferentes métodos utilizados para se estimarem os coeficientes de precisão, e cada um deles tem suas fontes de erro principais (Anastasi & Urbina 2000, Urbina, 2007). Não se trata de eliminar os erros, mas, sim, de identificar suas fontes e estimar a extensão da sua influência, de modo que, se o erro for muito grande, o teste perde sua utilidade. A seguir são apresentados os métodos e suas fontes de erro centrais:

a) *Método das formas alternadas*: O mesmo indivíduo pode ser avaliado com duas ou mais formas do teste (formas paralelas), no mesmo dia ou em dias diferentes. No caso da aplicação das formas alternadas no mesmo dia (imediato), a principal fonte de erro está ligada ao conteúdo. Cabe pensar em duas formas de um teste matemático. Em uma forma do teste, o item 1 é a resolução da operação 2 + 2, e, em outra forma (paralela) do teste, o item 1 é a resolução da operação 3 + 3. Sabe-se de um dito popular que diz: "Mais certo do que dois mais dois são quatro". Pessoas que conhecem esse ditado podem responder certamente ao item 1 da primeira forma do teste pela familiaridade do resultado dessa operação, e não pela resolução matemática da soma dos dois números, e são, portanto, favorecidas nesse item. Um indivíduo pode, por exemplo, errar o item 1 da segunda forma porque tem dificuldade em operação matemática de soma. Concluindo, se uma das formas do teste está mais suscetível à familiaridade dos respondentes do que a outra, as pontuações nas duas formas podem ser diferentes, gerando erro e diminuindo o coeficiente de precisão. Quando se trata de formas alternadas aplicadas em dias diferentes, além do erro ligado ao conteúdo, essa avaliação pode sofrer influências do tempo. Ao se avaliar personalidade, por exemplo, distinguem-se os traços, que são características relativamente duradouras, e os estados, que são características temporárias. Nesse caso, a aplicação das formas do teste em dias diferentes no mesmo indivíduo pode provocar respostas diferentes, pois seu estado pode estar diferente de tempos em tempos.

b) *Método de teste-reteste*: Consiste na aplicação e reaplicação do mesmo teste ao respondente, mas em ocasiões diferentes. Aqui, a principal fonte de erro é relacionada ao tempo, já que são feitas duas aplicações do teste em momentos distintos. A precisão é obtida por meio de um coeficiente resultante de uma análise de correlação entre pontuações do teste obtidas em duas estimações diferentes. Quanto mais correlacionadas essas pontuações estiverem (coeficiente mais próximo de 1), maior a precisão encontrada, e quanto menos correlacionadas (coeficiente mais próximo de 0), menor precisão da medida (Anastasi & Urbina 2000; Cronbach, 1996; Urbina, 2007).

c) *Modelo das metades (Split-half)*: Consiste na divisão do teste em duas partes homogêneas ou equivalentes, porém ele é aplicado uma única vez ao testando. Essa divisão pode ocorrer de modo que sejam separados com relação aos itens pares e ímpares, ou, ainda, a primeira e segunda metade do teste, etc. Fatores relacionados ao conteúdo dos itens constituem as principais fontes de erro.

d) *Método de coeficientes de Kuder-Richardson e alfa de Cronbach*: Consiste na aplicação do teste uma única vez e no estabelecimento de uma relação entre respostas individuais nos itens com o escore total do teste. A fonte de erro principal está relacionada ao conteúdo dos itens. Pode haver variação na homogeneidade do teste, e, quanto mais homogêneo ele for, com itens homogêneos, maior será o coeficiente de precisão calculado por esse método.

e) *Método de precisão entre avaliadores*: Consiste em solicitar a avaliação de dois ou mais avaliadores diferentes ao mesmo

método respondido pelo mesmo indivíduo, estabelecendo-se uma correlação entre os resultados dos avaliadores. Esse método está sujeito à fonte de erro relacionada à subjetividade do avaliador. Quando os testes dependem em grande parte do julgamento do avaliador, como é o caso das técnicas projetivas, os resultados para o mesmo teste aplicado na mesma pessoa podem ser diferentes, pois nessa avaliação está embutida a subjetividade e a interpretação pessoal do avaliador.

Considerações finais

Pode-se pensar numa relação entre os dois conceitos abordados nesse capítulo. De modo geral, uma boa precisão é condição imprescindível para que um teste seja válido, porém somente esta característica não é suficiente. Mesmo que se obtenha um teste consistente, estável, este pode estar medindo algo que não aquilo para o qual foi destinado a medir, ou seja, válido. Dessa forma, um coeficiente baixo de precisão revela seguramente uma perda na qualidade psicométrica, validade de um teste, mas o contrário não necessariamente acontece, ou seja, um teste sem evidências de validade pode ter bom desempenho na precisão, apesar de ser pouco provável (Hogan, 2006; Urbina, 2007).

Como apontado nos tópicos anteriores sobre validade e precisão, considera-se que esses conceitos não são pura e simplesmente restritos ao processo de construção e estudos, validação e precisão dos instrumentos. A operacionalização dos dois conceitos

abordados depende também do uso adequado do teste, desde sua aplicação até sua correção, e dos cuidados na interpretação dos seus resultados. É necessário o conhecimento da característica que se está avaliando, além dos alcances e limites do teste utilizado, para que possa agregar um valor significativo no processo de avaliação, além do que o processo de avaliação é considerado como muito maior e amplo do que somente a testagem psicológica.

Questões

1) Um teste pode ser considerado aprovado ou reprovado para sempre pelo Conselho Federal de Psicologia? Explique.
2) Qual o período máximo de tempo estipulado pelo Conselho Federal de Psicologia que deve existir entre os estudos de validade e precisão de um instrumento?
3) Qual a definição contemporânea de validade?
4) Cite e explique a definição tripartite (clássica) de validade.
5) Cite e explique as fontes de evidências de validade, segundo AERA, APA e NCME (1999).
6) Qual a definição de precisão?
7) Cite e explique os métodos utilizados para se estimarem os coeficientes de precisão.
8) Qual a relação existente entre validade e precisão?

Referências

American Educational Research Association, American Psychological Association, & Nacional Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.

American Psychiatric Association. (2002). *Diagnostic and Statistical Manual of Mental Disorders- IV- TR*. Washington, DC: American Psychological Association.

Anastasi, A., & Urbina, S. (2000). Validade: Conceitos Básicos. In A. Anastasi & S. Urbina, *Testagem Psicológica* (pp. 107-127). Porto Alegre: Artmed.

Bettoi, W., & Simão, L. M. (2000). Profissionais para Si ou Para Outros? *Psicologia Ciência e Profissão, 20* (2), 20-31.

Conselho Federal de Psicologia (2003). *Resolução nº 002/2003*. Recuperado em 14 de março de 2011, de http://www.pol.org.br

Conselho Federal de Psicologia (2004). *Resolução nº 006/2004*. Recuperado em 14 de março de 2011, de http://www.pol.org.br

Cronbach, L. J. (1996). *Fundamentos da Testagem Psicológica*. Porto Alegre: Artes Médicas.

Cunha, N. B., & Santos, A. A. A. (2009). Validade por processo de resposta no teste Cloze. *Fractal: Revista de Psicologia, 21* (3), 549-562.

Dias, C. A. (2001). Considerações sobre elaboração de currículos para formação de psicólogos: a partir de uma perspectiva didática. *Psicologia Ciência e Profissão, 21* (3), 36-49.

Francisco, A. L., & Bastos, A. V. B. (1992). Conhecimento, formação e prática - o necessário caminho da integração. In A. L. Francisco, C. R. Klomfahs, & N. M. D. Rocha (Orgs.), *Psicólogo brasileiro: construção de novos espaços* (pp. 211-227). Campinas: Átomo.

Hogan, T. P. (2006). *Introdução à prática de testes psicológicos*. Rio de Janeiro: LTC.

Hutz, C. S., & Bandeira, D. R. (2003). Avaliação psicológica no Brasil: situação e desafios para o futuro. In O. H. Yamamoto & V. V. Gouveia

(Orgs.), *Construindo a psicologia brasileira: desafios da ciência e prática psicológica*. São Paulo: Casa do Psicólogo.

Jacquemin, A. (1995). Ensino e Pesquisa sobre testes psicológicos. *Boletim de Psicologia*, XLV(102), 19-21.

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher, 18* (2), 5-11

Moura, E. P. G. (1999). A psicologia (e os psicólogos) que temos e psicologia que queremos: reflexos a partir das propostas de Diretrizes Curriculares (MC/SESU) para os cursos de graduação em psicologia. *Psicologia Ciência e Profissão, 19* (2), 10-19.

Noronha, A. P. P., & Alchieri, J. C. (2004). Conhecimento em avaliação psicológica. *Estudos de Psicologia (Campinas), 21* (1), 43-53.

Nunes, C. H. S. S., Hutz, C. S., & Nunes, M. F. O. (2010). *Bateria Fatorial de Personalidade (BFP). Manual Técnico*. São Paulo: Casa do Psicólogo.

Pereira, A. P. C., & Carellos, S. D. M. S. (1995). Examinando o Ensino das Técnicas de Exame Psicológico. *Cadernos de Psicologia, Belo Horizonte, 3* (4), 33-36.

Primi, R., Muniz, M., & Nunes, C. H. S. S. (2009). Definições Contemporâneas de Validade dos Testes Psicológicos. In C. S. Hutz (Org.), *Avanços e Polêmicas em Avaliação Psicológica* (pp. 243-265). São Paulo: Casa do Psicólogo.

Salazar, A., Troccóli, B., & Vasconcelos, T. (2001). *Investigação das Correlações entre Medidas Projetivas e Objetivas de Personalidade*. In XXXI Reunião Anual de Psicologia (SBP). Rio de Janeiro, RJ.

Sbardelini, E. T. B. (1991). Os mitos que envolvem os testes psicológicos. *Documento CRP-08, 1* (1), 53-57.

Simões, M. R. (1999). O ensino e a aprendizagem da avaliação psicológica: o caso da avaliação da personalidade. *Psychologica, 22*, 135-172.

Souza, C. V. R., Primi, R., & Miguel, F. K. (2007). Validade do Teste Wartegg: correlação com 16PF, BPR-5 e desempenho profissional. *Avaliação Psicológica, 6* (1), 39-49.

Urbina, S. (2007). *Fundamentos da Testagem Psicológica*. Porto Alegre: Artmed.

Vendramini, C. M. M., & Lopes, F. L. (2008). Leitura de manuais de testes psicológicos por estudantes e profissionais de psicologia. *Avaliação Psicológica, 7* (1), 93-105.

Wechsler, S. M., & Guzzo, R. S. L. (1999). Apresentação. In S. M. Weschler e R. S. L. Guzzo (Orgs.), *Avaliação psicológica: Perspectiva internacional* (pp. 9-11). São Paulo: Casa do Psicólogo.

Capítulo 6

Padronização e normatização de testes psicológicos: simplificando conceitos

Ivan Sant'Ana Rabelo
Leila Brito
Marcia Gabriel da Silva Rego

Os testes psicológicos diferenciam-se de outras técnicas de avaliação, por se tratar de procedimentos referenciados a normas e a diretrizes interpretativas padronizadas, com base em categorias preestabelecidas. Outros procedimentos também são utilizados em contextos de avaliação psicológica, como meios de acesso ao universo psicológico do indivíduo, visando à maior compreensão da sua singularidade para melhor adequação das formas de intervenção quando necessárias. Alguns tipos de entrevistas, técnicas de observação, aplicação de atividades lúdicas, entre outros, constituem exemplos de estratégias de avaliação psicológica que não pertencem à categoria de testagem.

A avaliação psicológica pode ser representada como resultante de três critérios, a saber: a medida, o instrumento e o processo de avaliação. Alchieri e Cruz (2003) mencionam que cada um destes critérios possui uma representação teórica e uma metodológica própria e que concebe assim, de forma constitutiva, uma via própria de compreensão do seu objeto de investigação, denominado de fenômenos ou processos psicológicos.

Métodos que envolvem o ato de desenhar, narrar histórias, realizar encenações ou brincar com bonecos, em geral, não se propõem a apresentar estudos normativos ou indicadores metódicos de interpretação, ficando assim também caracterizados fora do que é considerado testagem psicológica. Apesar de não pertencerem à categoria de testes, os resultados de tais instrumentos podem ter credibilidade, desde que as conclusões apresentadas pelo psicólogo estejam condicionadas a um referencial teórico válido, que sustente as interpretações segundo o pressuposto do determinismo psíquico[1].

Seja na testagem psicológica ou em outras formas de avaliação, a experiência do profissional, o conhecimento do constructo que está sendo investigado e o embasamento teórico consistente, acompanhados de outros métodos de observação e de análise, são condições imprescindíveis para garantir a confiabilidade dos

[1] Determinismo Psíquico - Eventos mentais são precedidos de eventos anteriores que os determinam; é uma crença de que a mente funciona como uma máquina em que cada fase se encadeia à anterior e à posterior (Freud, 1901/1960). O termo é mais comumente encontrado na Psicanálise, contudo outras teorias também o aplicam. No Behaviorismo, pode-se compreender que o acontecimento de um comportamento se dá a partir da influência de vários fatores ambientais que o precedem, e este influencia nos comportamentos decorrentes deste. "A força de uma única resposta pode ser, e usualmente é, função de mais do que uma variável e uma única variável usualmente afeta mais do que uma resposta" (Skinner, 1957 apud Chiesa, 1994, p. 113).

resultados que se integrarão de modo a compor uma avaliação coerente (Urbina, 2007). Neste capítulo, abordar-se-ão as questões referentes à padronização e à normatização dos testes psicológicos, mesmo sabendo que diferenciar o que se enquadra como teste psicológico de outros tipos de procedimentos de avaliação psicológica não é algo simples. Focar-se-á, portanto, no teste psicológico considerando-o como um instrumento de mensuração padronizado que avalia características ou processos psicológicos fundamentados em uma teoria e que atende aos requisitos de validade e precisão.

Em resumo, segundo Werlang, Villemor-Amaral e Nascimento (2010):

> Para os testes psicológicos serem confiáveis, devem ser padronizados e atender a requisitos de fidedignidade e validade. A padronização refere-se à necessária existência de uniformidade tanto para a aplicação do instrumento, como nos critérios para interpretação dos resultados obtidos. A fidedignidade diz respeito à coerência sistemática, precisão e estabilidade do teste, e a validade reflete se o teste mede realmente o que pretende medir (p. 92).

Assim, uma avaliação psicológica realizada com qualidade está relacionada principalmente à utilização de técnicas de avaliação reconhecidas pela Psicologia. A Comissão de Avaliação Psicológica do Conselho Federal de Psicologia, em parceria com as instituições de ensino e pesquisa, definiu critérios de adaptação de instrumentos de avaliação para a realidade brasileira, considerando

que a fundamentação teórica e as propriedades psicométricas dos testes disponíveis estejam de acordo com parâmetros internacionais de qualidade, baseados em estudos de precisão, validade e normatização (Werlang e cols., 2010).

Padronização e normatização: conceitos iguais ou distintos?

Alguns autores, tais como Cronbach (1996), Alchieri & Cruz (2003), Pasquali (2003, 2010), entre outros, procuram fazer uma distinção clara entre padronização e normatização, sendo:

- *Padronização*: A uniformidade na aplicação dos testes (material, ambiente, aplicador, instruções de aplicação e correção, etc.);
- *Normatização*: A uniformidade na interpretação dos escores dos testes (tabelas, percentis, escore z etc.).

Já Urbina (2007) considera que um teste psicológico pode ser descrito como padronizado, desde que contemple duas facetas que possibilitam objetividade no processo de testagem. A primeira relaciona-se à uniformidade dos procedimentos, desde a aplicação até a correção e interpretação dos resultados, englobando, inclusive, o local em que o teste é administrado, as circunstâncias de sua administração, o examinador, tudo que pode afetar os resultados, objetivando tornar tão uniformes quanto possível todas as variáveis que estão sob controle do examinador, para que os indivíduos que se submetam ao teste o façam da mesma forma.

A segunda faceta refere-se ao uso de padrões para a avaliação dos resultados. Estes padrões costumam ser normas derivadas de um grupo de indivíduos, denominados amostra normativa ou amostra de padronização, sendo que o desempenho coletivo do grupo, tanto em termos de média quanto de variabilidade, passa a ser um padrão pelo qual o desempenho dos outros indivíduos que se submetem ao teste sejam comparados. Desta forma, apesar de observar-se a diferença entre padronização e normatização existente, não há uma diferenciação dos termos.

A nosso ver, tal distinção se faz relevante, porque trata de duas questões muito importantes, principalmente do ponto de vista didático, para a compreensão destes conceitos na psicometria. Haja vista que a literatura neste sentido não é insistente sobre a nomenclatura, pelo contrário, as duas expressões são utilizadas indistintamente. Contudo, como se trata de questões específicas para a aprendizagem na área, tratar-se-á o tema em duas seções separadas neste capítulo.

Padronização

A padronização, em seu sentido mais geral, refere-se à uniformidade dos procedimentos no uso de um teste válido e preciso, desde os cuidados a serem tomados na aplicação do teste (uniformidade das condições de testagem) até os parâmetros ou critérios para a interpretação dos resultados dos sujeitos submetidos à testagem (Anastasi, 1977).

Precauções a serem tomadas na aplicação dos testes: padronizando as condições de administração dos testes psicológicos

É importante padronizar as condições de aplicação dos testes psicológicos com o intuito de garantir que a coleta de dados sobre o sujeito seja de boa qualidade. Uma má aplicação pode comprometer o resultado dos testes, tornando-os inválidos, mesmo quando da utilização de uma boa ferramenta. Vale ressaltar que uma má aplicação não invalida a qualidade psicométrica do teste, mas, sim, invalida o protocolo do sujeito, ou seja, os dados obtidos na avaliação não serão confiáveis.

O mau uso que se faz de um teste psicológico compromete sua utilidade. Por isso a padronização é tão importante na área de avaliação psicológica, pois pretende garantir o uso adequado e autêntico dos testes psicológicos.

Segundo Pasquali (2001), para se garantir uma boa administração dos testes psicológicos, é preciso acatar alguns requisitos no que se refere ao material da testagem, ao ambiente da testagem, às condições de aplicação e ao aplicador.

O material de testagem

A qualidade e a pertinência do teste são duas condições que devem ser atendidas:

- *Qualidade do teste:* O teste tem de ser válido e preciso. Caso o uso de testes seja feito sem cumprimento destes

parâmetros, corre-se o risco de processos jurídicos e condenação ética, tornando-os inúteis.

- *Pertinência do teste:* Além de ser válido e preciso, o teste tem de ser relevante ao problema apresentado pelo sujeito. O aplicador deve conhecer a utilidade de um dado teste, se este será para uma avaliação de personalidade, cognitiva, para área organizacional, clínica, etc., e, assim, escolher aquele que melhor se aplica à necessidade do sujeito.

Além da pertinência, o teste deve adaptar-se ao nível (intelectual, profissional, etc.) do candidato. É importante seguir à risca as instruções e recomendações que explicitam seus manuais, sem, entretanto, assumir uma postura estereotipada e rígida (Alchieri & Cruz, 2003).

O ambiente de testagem

Com relação ao ambiente de aplicação dos testes psicológicos, algumas condições devem ser atendidas, a saber: o ambiente físico, as condições psicológicas, o momento (tempo de aplicação) e o estabelecimento de *rapport*.

Ambiente físico: Deve proporcionar ao candidato a sensação de estar em suas melhores condições para responder ao teste; para tanto, é importante diminuir ou, se possível, eliminar a presença de distratores ambientais, pois o ambiente não pode tornar-se um fator desfavorável, desmotivador e incômodo para o testando. Entre os fatores a serem considerados, ressalta-se a utilização de uma sala adequada, com móveis igualmente adequados e

confortáveis, iluminação apropriada, ventilação, higiene, ausência de barulho. O ambiente deve reunir condições adequadas tanto em aplicações individuais quanto coletivas.

Condições psicológicas: Devem atender a condições do sujeito e do profissional que aplica o teste. Entre os aspectos relevantes, deve observar-se se o sujeito apresenta-se em condições normais de saúde física e psicológica; é preciso certificar-se de que o sujeito compreendeu exatamente a tarefa a realizar, sempre tomando o cuidado para não mudar as instruções do manual. O nível de ansiedade do sujeito que é submetido ao teste pode ser reduzido com o estabelecimento do *rapport*. Estabelecer *rapport* significa assumir uma atitude que faça o sujeito sentir-se à vontade ao fazer o teste, implicando o examinador mostrar-se motivador, encorajador, não irritadiço, sem gritar ou demonstrar expressões faciais e corporais desagradáveis, durante o contexto da testagem.

Para Silva (2008), o *rapport* permite ao profissional oferecer respaldo informativo fundamental para uma melhor compreensão da dinâmica desse processo, no sentido de colocar o(s) examinando(s) ou o(s) paciente(s) mais próximo(s) desse momento avaliativo, e o próprio psicólogo se faz agente de motivação e solicitude.

O *momento*: De acordo com os ambientes físico e psicológico, será avaliado quanto tempo deverá durar a aplicação dos testes psicológicos, considerando os dois fatores mencionados acima como determinantes para uma boa aplicação. Quando a bateria de testes for muito extensa, é recomendado dividi-la em mais de uma sessão, com o objetivo de evitar fadiga, aborrecimento e outros dissabores que impedem um uso adequado dos testes.

Ainda é preciso considerar, no âmbito do ambiente de testagem, o que denominamos de *situações adversas*, tais como a aplicação de testes para fins periciais e a testagem para seleção. Em alguns casos, o sujeito pode encontrar-se em condições psicológicas e, às vezes, até físicas não satisfatórias, principalmente em situações de grande competição, como em concursos públicos.

Condições de aplicação

Condições desfavoráveis para administração de testes psicológicos podem causar efeitos no desempenho deles. Alguns estudos mostram que os resultados nos testes são afetados pelos procedimentos utilizados durante a administração dos instrumentos.

Treffinger (1987) fez uma revisão de vários estudos sobre o efeito de condições de aplicação e clima psicológico no desempenho em testes de criatividade. Tais estudos indicam que os resultados são afetados pelos procedimentos utilizados durante a administração dos instrumentos. Para os responsáveis pela aplicação de testes, é indispensável orientação ou treinamento, no sentido de afiançar condições adequadas e comparáveis em todas as aplicações.

É de suma importância controlar alguns fatores, a saber:

1) O tempo suficiente para o examinando responder e o examinador fazer as observações necessárias para emitir seu julgamento;
2) O nível de dificuldade das palavras e a maneira de apresentar as instruções;

3) O controle de fatores que podem distrair a atenção do examinando.

Há de se considerar que são muitas as variáveis que afetam uma avaliação, até mesmo o estado físico do examinando ou sua motivação.

O aplicador

O aplicador do teste é um elemento importante da situação, principalmente na testagem individual. Pasquali (2003) diz que seu modo de ser e de atuar pode afetar bastante os resultados do teste. Sobre o assunto até este momento não existem pesquisas que permitam conclusões decisivas sobre o grau de influência que estas variáveis do examinador têm sobre os resultados dos testes. Considera-se que o psicólogo é um ser humano como todos os outros, com seus problemas inclusive, mas também é um técnico ou perito que deve ter desenvolvido algumas habilidades próprias da profissão, das quais obviamente ele deve fazer uso em situações como a testagem psicológica.

Em relação ao aplicador de testes psicológicos, caracteristicamente, deve ser um psicólogo e atender a alguns requisitos, tais como:

- *Conhecimento:* O aplicador deve conhecer intimamente o material utilizado, para que possa transmitir segurança e responder a alguma dúvida que possa surgir durante a testagem, bem como realizar análises fidedignas;

- *Aparência:* O aplicador deve causar boa impressão, usando roupas adequadas e limpas, evitando exageros até mesmo em perfumes;
- *Comportamento durante a testagem:* O papel do aplicador é o de conduzir a testagem, mantendo a ordem, o respeito, a orientação;
- *Gravação de sessões:* Somente com o consentimento do examinando.

É importante ainda ficar atento a não ceder a pressões quanto à utilização de determinado teste, mesmo que haja interesse em reduzir custos da avaliação que interfiram na qualidade do trabalho. É necessário prevalecer o princípio da isonomia, que consiste em tratar todos do mesmo modo, com condições de avaliação iguais (Pasquali, 2003).

Porém, o requisito mais importante para procedimento de teste é o preparo prévio do aplicador. Anastasi (1977) comenta que, durante os testes, não pode haver emergências, o que é a única forma de garantir a uniformidade de procedimento.

Lei e testes psicológicos

Esse tema é fundamental, pois as pessoas têm direitos garantidos em normas da Constituição dos países e das Nações Unidas. Por lei, os peritos devem prestar serviços de qualidade à sociedade, e esta qualidade pode ser judicialmente procurada por meio das leis pertinentes. O psicólogo responde por sua conduta nesta área

de testes. A lei considera o psicólogo como perito e, portanto, legalmente responsável em sua atuação profissional.

O princípio essencial estabelecido pelo Código de Nuremberg (1949) é o de garantir que fossem respeitadas as pessoas humanas que viessem a participar de experimentos médicos ou científicos. O princípio fundamental estabelecido por este Código é o de que toda experimentação com seres humanos requer o prévio consentimento livre e esclarecido do sujeito participante. Já na Declaração de Helsinque é reafirmado o princípio do consentimento livre e esclarecido e colocado o bem-estar do sujeito como prioritário. De acordo com Ambroselli (1987), a pesquisa médica aos interesses da ciência e aquelas da sociedade não devem jamais prevalecer sobre o bem-estar dos sujeitos.

Embasados nestes princípios, os comitês de ética em Psicologia, inclusive no Brasil, vêm elaborando normas que devem ser seguidas na aplicação de testes. De um modo geral, estas normas podem ser resumidas segundo as Normas para a testagem educacional e psicológica da *American Psychological Association* (APA, 1985, apud Cronbach, 1996, p. 97).

Sigilo e divulgação dos resultados

Devem seguir-se as normas do sigilo profissional contidos no Código de Ética do Psicólogo no Brasil (CFP, 1987). Nele constam informações sobre o que é vedado ao psicólogo e quais seus deveres.

A pessoa que se submete ao teste tem o direito de receber informações sobre os resultados da testagem. Também tem direito aos resultados o solicitante da avaliação, como empresas que solicitem

no psicólogo a avaliação psicológica no processo de recrutamento e de seleção e, outros exemplos, os juízes em casos de perícia judicial e, em casos de crianças, adolescentes, o responsável legal, que nem sempre é o solicitante. Eles apenas têm direito às informações estritamente necessárias à resposta da solicitação.

É preciso seguir as normas de sigilo entre profissional e paciente; assim, toda e qualquer informação sobre o sujeito, os arquivos e outras anotações provenientes do processo psicológico deve ser mantida em local seguro, de forma que ninguém possa ter acesso a ela. Pasquali (2003) menciona que os arquivos devem ser seguros, de modo que ninguém possa ter acesso a uma informação restrita sem autorização específica do profissional responsável. As identidades dos indivíduos devem ser codificadas de tal forma que somente o profissional responsável seja capaz de identificá-las. Em processos judiciais, o juiz pode solicitar abertura de registros sigilosos. É preciso ter em mente que o indivíduo não pode sair indevidamente prejudicado com a exposição de informações sigilosas.

Normatização

A normatização pressupõe que um teste necessita ser contextualizado para poder ser interpretado. Tal conceito diz respeito a padrões de como se deve interpretar um resultado que a pessoa atingiu em um teste. Segundo Urbina (2007), os resultados brutos não são muito úteis numa avaliação psicológica, representando um grupo de números que não transmitem nenhum sentido, mesmo

depois de mais de um exame aprofundado. Por meio da estatística descritiva (distribuições de frequência, gráficos, percentis, variabilidade, etc.), pode-se relacionar ou dar sentido aos dados de modo a facilitar a sua compreensão e utilização.

Exemplificando, um indivíduo que apresentou quarenta pontos num teste de inteligência não verbal e vinte pontos num teste de memória visual pouco significa dentro de uma avaliação psicológica. Outra forma de apresentação dos resultados pouco eficiente seria dizer que este sujeito acertou 70% das questões, pois comparado com um teste em que os indivíduos da amostra de padronização acertaram muitas questões (com médias altas), ou seja, um teste considerado fácil, é diferente de 70% de acerto nas questões em um teste considerado difícil, com médias de padronização baixas.

Assim, qualquer resultado bruto deve ser referido a alguma norma ou a algum padrão para que tenha algum sentido. A norma permite posicionar o resultado de um sujeito, possibilitando inferências:

- A posição em que a pessoa se localiza no traço medido pelo teste que produziu o resultado medido;
- A comparação da pontuação deste sujeito com os resultados de outras pessoas com características similares.

No processo de criação de normas, um teste deve ser aplicado a uma amostra representativa do tipo de pessoa para o qual foi planejado. Esse grupo, denominado de amostra de padronização, possibilita a composição de tabelas que serão estabelecidas como normas, indicando não somente as médias, mas também os

diferentes graus de desvios, acima ou abaixo da média. Isso possibilitará avaliar diferentes graus de superioridade ou inferioridade naquele determinado aspecto ou faceta que o teste se propôs a medir (Cronbach, 1996).

Segundo Pasquali (2001), o critério de referência ou a norma de interpretação é normalmente definido por dois padrões, sendo eles o nível de desenvolvimento do indivíduo humano, isto é, as normas de desenvolvimento, e um grupo padrão composto pela população típica para a qual o teste foi construído, também chamado de normas intragrupo.

Normas de desenvolvimento

Este tipo de normas se fundamenta em variáveis que podem ser expressas no desenvolvimento progressivo de aspectos psicológicos, tais como maturação psicomotora, maturação psíquica, idade mental, série escolar, entre outras. Tais características informam sobre aquilo que o indivíduo passa ao longo de sua vida. Neste sentido, são utilizados, como critério de norma, três principais fatores: a idade mental, a série escolar e o estágio de desenvolvimento.

A idade mental

Segundo Anastasi (1977), o conceito de idade mental foi introduzido por Binet e Simon na revisão de 1908 das escalas de Binet-Simon. Nestes casos, os itens individuais são agrupados em níveis de idade. Exemplificando, itens solucionados na amostra de

padronização pela maioria das crianças com nove anos de idade são atribuídos para ser aplicados e avaliados em crianças no nível de nove anos; os itens respondidos pela maioria das crianças com dez anos serão colocados ao nível de dez anos, e assim sucessivamente. Dessa forma, espera-se que o resultado de uma criança no teste corresponda ao nível mais alto dentro de sua idade; seguindo ainda o exemplo citado, crianças de oito anos devem ser capazes de responder às questões dentro deste nível de idade; se elas acertarem itens classificados como para dez anos, sua idade mental (IM) será dez, embora sua idade cronológica (IC) seja oito anos.

A autora explica que, na adaptação norte-americana da escala de Binet-Simon, a Stanford-Binet (Terman & Merrill, 1960), a idade mental (IM) foi expressa em termos da idade cronológica (IC), resultando no quociente intelectual (QI) por meio da fórmula:

$$QI = 100 \times \frac{IM}{IC}$$

Dessa forma, o QI é comparável em diferentes idades, na medida em que a interpretação de um determinado QI é sempre a mesma, qualquer que seja a idade do sujeito. Logo, se um sujeito responde a todas as questões relacionadas ao seu nível de idade cronológica, representará um QI de 100; por exemplo, uma criança de oito anos:

$$QI = 100 \times (8/8) = 100$$

Todavia, vale ressaltar que, apesar da aparente simplicidade lógica, o QI não é aplicável na maioria dos testes psicológicos, principalmente nos testes para adultos. O uso do QI deve ser precedido de uma verificação da variabilidade em diferentes idades, a fim de assegurar que foi satisfeita a condição de variabilidade uniforme do QI ou uma variabilidade proporcional crescente da idade mental. Vários testes de Inteligência que apresentam normas por idade não satisfazem as condições para a constância de QI (Anastasi, 1977).

Normas educacionais - Série escolar

É comum a interpretação de resultados de testes educacionais de aproveitamento em termos de normas de série. O conceito de série escolar como norma é empregado para testes de desempenho acadêmico e pode ser utilizado quando se trata de disciplinas que são oferecidas numa sequência de várias séries escolares. As normas são construídas por meio da pontuação bruta média obtida por alunos em cada série, resultando numa pontuação típica para cada série. Assim, a criança que apresentar uma pontuação típica da 4ª série obterá um escore padronizado de quatro (Anastasi, 1977; Pasquali, 2003).

Estágio de desenvolvimento

Piaget e seus colaboradores examinaram o desenvolvimento cognitivo e estabeleceram uma sequência de estágios consecutivos do desenvolvimento, denominados: sensório-motor, pré-operacional,

operacional concreto, operacional formal. Normas divididas por estágios de desenvolvimento são utilizadas por profissionais da psicologia infantil que estudam os desenvolvimentos mental e psicomotor em termos de idades consecutivas de desenvolvimento, como Gesell e Piaget.

Gesell e colaboradores (Ames, 1937; Gesell & Amatruda, 1947; Halverson, 1933; Knoblock & Pasamanick, 1974) estabeleceram normas para oito idades típicas (de quatro semanas a 36 meses) de desenvolvimento das crianças no que tange ao comportamento motor, adaptativo, linguístico e social. Pesquisadores e estudiosos da escola piagetiana (Laurendeau & Pinard, 1962, 1970; Pinard & Laurendeau, 1964) desenvolvem testes empregando estes estágios como método de interpretação dos resultados (Pasquali, 2001).

Normas intragrupo

Como os resultados brutos dos testes normalmente se apresentam em diferentes unidades, torna-se impossível a comparação direta de resultados. O nível de dificuldade de cada teste também pode influenciar nessa comparação entre resultados brutos. Assim, normas representadas por meio de transformações normativas permitem expressá-las em unidades que possibilitam comparações.

Anastasi (1977) elucida que existem várias maneiras por meio das quais os resultados brutos podem ser transformados; contudo, os resultados dos testes normalmente são expressos por três tipos: resultados por idade, já descritos anteriormente, percentis e resultado padrão. Assim, nas normas intragrupo, o critério de referência dos resultados são a população ou o grupo para o qual

o teste foi desenvolvido. A pontuação que o sujeito apresentou em um teste toma sentido em relação aos resultados dos demais sujeitos da população.

Percentis

O percentil indica a posição relativa do sujeito na amostra de padronização. Assim, a localização do sujeito, do ponto de vista percentílico, indica quanto por cento de pessoas desta população (amostra) apresentaram resultados inferiores ao dele. Por exemplo, se 40% dos sujeitos obtiveram um escore bruto menor do que vinte, este valor será expresso como percentil quarenta, o que indica que 40% dos sujeitos têm escore menor que vinte e 60% têm escore maior. Um percentil de cinquenta indica que o sujeito se situa na mediana dos escores da amostra (Pasquali, 2001).

Apesar de o percentil apresentar uma compreensão mais simples e ser comumente empregado na testagem psicológica, sua grande dificuldade situa-se no fato de que as distâncias entre escores percentílicos sucessivos não são constantes, mas variam segundo a posição do escore no início/fim da escala ou no meio dela. Portanto, os percentis não devem ser confundidos com "resultados de porcentagem", pois isso significaria porcentagem de itens respondidos corretamente nos testes, mas o percentil representa resultados transformados, apresentados em termos de porcentagem de pessoas que participaram da amostra de padronização (Cronbach, 1996).

Percentis são úteis também para comparar a realização do indivíduo em diferentes testes, não servindo apenas para mostrar a

posição do indivíduo na amostra normativa. Exemplificando, se uma criança obtém um resultado bruto de trinta, num teste aritmético, e de 58, num teste de leitura, não se podem comparar diretamente os resultados, porque suas unidades de medida têm características diferentes. Contudo, se a referência aos resultados percentílicos indicar que uma pontuação de trinta no teste de aritmética significa uma localização de percentil 65, enquanto um resultado de 58 num teste de leitura corresponda a um posto de percentil quarenta, então, pode inferir-se que a criança apresentou um resultado melhor no teste aritmético do que no de leitura (Anastasi, 1977).

Entre as vantagens da utilização dos resultados em percentil, verifica-se como primordial sua universalidade para interpretação e comparação de resultados. Pode ser usado tanto em crianças como em adultos, e é adequado para tipos de testes variados, seja para medir variáveis de personalidade, seja para medir atitudes, capacidades cognitivas, etc. Sua principal desvantagem encontra-se vinculada à distribuição dos sujeitos na amostra, pois, ao se aproximar da curva normal, são concentradas grandes quantidades de sujeitos representativos na mediana ou próximos ao centro da distribuição, enquanto os extremos são muito comprimidos, conforme observado na Figura 1.

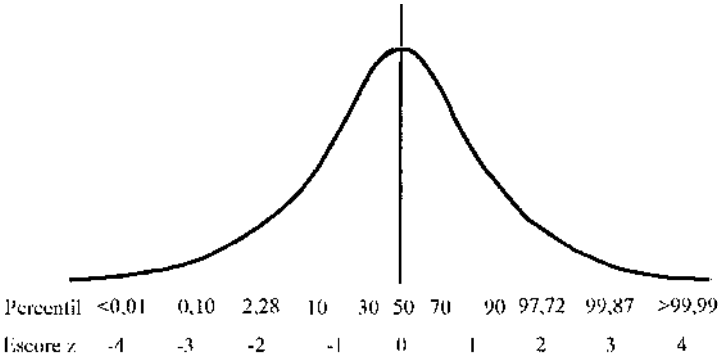


Figura 1 - Distâncias intervalares de escores percentílicos e escores *z* (Pasquali, 2001).

Escore Padrão

Com os avanços no desenvolvimento da testagem psicológica, observa-se cada vez mais o crescente uso de resultados em formato de *escore padrão*, também conhecidos como *resultados padrão* ou simplesmente *escore z*. O escore padrão revela a distância do sujeito em relação à média, em termos do desvio padrão da distribuição (Anastasi & Urbina, 2000).

Normas fundamentadas no escore padrão baseiam-se no cálculo de um escore z que está relacionado ao resultado bruto do sujeito, podendo ser calculado de duas formas distintas, que resultarão ou em um escore padrão ou em um escore padrão normalizado. O primeiro resultado pode ser determinado por uma transformação

linear, enquanto o outro se dá por meio de uma transformação não linear, ambas a partir dos resultados brutos originais.

Escore padrão linear

Derivado do resultado bruto, o escore padrão linear pode ser expresso em uma escala padrão sem afetar a posição relativa dos indivíduos no grupo e nem mudar a forma da distribuição original. Os escores padrão são úteis para expressar os escores brutos de formas paralelas de um mesmo teste, sobretudo se as formas possuem dificuldades diferentes. Os escores padrão também facilitam a comparação e a interpretação dos resultados (Pasquali, 2001).

Em sua maioria, os resultados padrão linearmente derivados são denominados apenas de "escore padrão" ou "escore z" e representam a relação entre o desvio do escore bruto em comparação à média e ao desvio padrão dos escores do mesmo grupo. Para calcular um escore z, basta encontrar a diferença entre o resultado bruto do indivíduo e a média do grupo normativo, e depois dividir essa diferença pelo desvio padrão (DP) do grupo normativo. A fórmula usada para cálculo deste procedimento é:

$$Z = \frac{X - M}{DP}$$

Onde:
X = escore bruto
M = média do grupo
DP= desvio padrão

Exemplificando:

Imagine-se o cálculo do escore z de dois sujeitos em um grupo no qual a média é igual a 75 e o desvio padrão igual a cinco.

Resultado bruto de André: $X_1 = 85$

$$Z_1 = \frac{85 - 75}{5}$$

$$Z_1 = +2{,}00$$

Resultado bruto de Roberto: $X_2 = 68$

$$Z_2 = \frac{68 - 75}{5}$$

$$Z_2 = -1{,}4$$

Escore padrão normalizado (EPN)

Assim como explicado anteriormente, o objetivo da transformação dos resultados brutos em outro tipo de escala derivada é tornar compatíveis os resultados obtidos em diferentes testes. Contudo, existem casos em que os resultados não se apresentam na mesma forma do ponto de vista da distribuição normativa dos

resultados; assim, por exemplo, comparar o resultado de um teste com distribuição assimétrica com o resultado de outro cuja distribuição não é assimétrica, aproximando-se da curva normal.

Testes com resultados de formas diferentes precisarão de transformações não lineares para ajustar os resultados a qualquer tipo especificado de curva de distribuição. Tal transformação não linear é calculada por meio das tabelas da curva normal, e incide basicamente em transformar os percentis em escores z (Anastasi, 1977).

A idade mental e os resultados de percentil, descritos anteriormente, representam transformações não lineares, mas estão limitadas a desvantagens já discutidas em cada tópico. Apesar da possibilidade de ajustar os resultados a algum outro tipo de distribuição, a curva normal é comumente utilizada para este fim. O motivo relaciona-se ao fato de as distribuições de resultados brutos frequentemente se aproximarem mais da curva normal do que de qualquer outro tipo de curva, assim como nas medidas físicas, tais como peso e altura, que usam escalas de unidades iguais, obtidas por meio de operações físicas, produzindo geralmente distribuições normais. Outra conveniência importante da curva normal é apresentar muitas propriedades matemáticas proveitosas, facilitando outros cálculos e comparações.

Segundo Anastasi (1977), os resultados padrão normalizados apresentam-se em termos de uma distribuição que foi transformada a fim de se adequar a uma curva normal, podendo ser calculados por meio de referência a tabelas que deem a porcentagem de casos colocados em diferentes distâncias DP da média de uma curva normal. Para isso, inicialmente, verifica-se a porcentagem de sujeitos da amostra de padronização que estejam no nível de cada resultado bruto ou acima dele. Em seguida, essa porcentagem

deverá ser localizada na tabela de frequência da curva normal, obtendo-se assim o correspondente escore padrão normalizado. Estes escores são apresentados de forma similar ao escore padrão linear, ou seja, com uma média de zero e um DP de um.

Assim, um EPN de zero indica que o sujeito está na média de uma curva normal, superando 50% do grupo. Já um resultado de -1,00 significa que o respondente supera aproximadamente 16% do grupo, enquanto um resultado de + 1,00 indica uma superação de 84%. Tais porcentagens correspondem, respectivamente, a uma distância de 1 DP abaixo e 1 DP acima da média de uma curva normal.

Simplificando, o escore padrão normalizado (EPN) são escores padrão forçados a apresentar uma distribuição normal pela conversão dos equivalentes percentis dos escores brutos em correspondentes escores padrão ao longo da curva normal, independentemente da forma original da distribuição. Os escores padrão normalizados geralmente apresentam média cinquenta e desvio padrão dez.

Transformações do escore padrão

Nos exemplos anteriormente apresentados, observou-se a presença de casas decimais e valores negativos (pois o z vai de menos infinito a mais infinito, na prática, de -5 a +5), o que tende a produzir números que podem tornar-se difíceis e confusos de utilizar tanto para cálculos quanto para descrições. Assim, costuma-se empregar outra transformação linear visando a deixar o escore padrão em um formato mais conveniente (Anastasi, 1977; Pasquali, 2001).

Segundo Pasquali (2001), para essa transformação, normalmente o *z* é multiplicado por um coeficiente e ao produto é agregada uma constante, utilizando a fórmula:

$$\text{Escore transformado} = a + b(z)$$

Assim, o coeficiente de multiplicação do *z* (isto é, o *b* da fórmula), tanto quanto a constante somada (o *a* da fórmula), é arbitrário, o que resulta em diversas formas de apresentação das normas quantas se desejar. Não obstante, alguns valores *a* e *b* são habitualmente mais empregados, o que permite estabelecer normas já tradicionalmente conhecidas, tais como o escore T, o QI de desvio, o escore CEEB (*College Entrance Examination Board*), entre outros.

A principal vantagem de se empregarem transformações já utilizadas universalmente é que tornam todas estas normas comparáveis entre si (Pasquali, 2001). As fórmulas de transformação para algumas destas normas são:

- Escore T = 50 + 10z
- QI de Desvio = 100 + 15z (Escalas de Wechsler) ou
- QI de Desvio = 100 + 16z (Stanford-Binet)
- CEEB = 500 + 100z.

Outra transformação bem conhecida é a escala de padrão nove, também chamada de classes normalizadas ou estanino, empregada pela Força Aérea dos Estados Unidos, durante a Segunda Guerra Mundial. Esta transformação fornece um sistema

de escores de apenas um dígito, com uma média de cinco e um DP de 1,96. A palavra *stanine* é derivada da expressão *standard nine point scale*, ou seja, o nome padrão nove baseia-se no fato de os resultados irem de um a nove. Realizar a transformação a resultados de apenas um dígito facilita a realização de cálculos, sobretudo com máquinas (Anastasi, 1977).

O estanino apresenta grande vantagem prática, pois é de fácil utilização para representar resultados de sujeitos submetidos à testagem, mas o seu cálculo é relativamente trabalhoso. Esta transformação consiste em dividir os z, que vão comumente de -3 a +3, em uma quantidade de classes. As divisões de classes mais usadas são cinco, sete, nove (estaninos) e onze. Como exemplo, ao dividir os z em cinco classes, obtém-se uma divisão facilmente compreensível e prática dos resultados dos sujeitos, tais como:

Superior / Média Superior / Média / Média Inferior / Inferior

As formas de transformação dos escores descritas produzem resultados análogos se as distribuições de frequência forem normais. Quanto mais as distribuições se espaçam da normalidade, menor é a recomendação e a utilização de transformações não linear dos escores (Pasquali, 2001).

Considerações finais

A Resolução nº 002/2003 do Conselho Federal de Psicologia (CFP, 2003) regulamenta a utilização, a elaboração e a comercialização de testes psicológicos, restringindo o uso por parte dos

psicólogos apenas a testes que tenham sido comprovados por estudos científicos e encaminhados para a avaliação da Comissão de Avaliação Psicológica do Sistema de Avaliação de Testes Psicológicos (SATEPSI). Entre as recomendações da resolução, são solicitados a apresentação da fundamentação teórica que embasa o teste, as evidências empíricas de validade e precisão, os dados informando as propriedades psicométricas dos itens do instrumento, as informações a respeito do sistema de correção e interpretação dos resultados, e também dos procedimentos padronizados de aplicação e interpretação.

No estabelecimento de normas de testes, englobando desde sua criação, desenvolvimento e utilização, deve-se dar grande atenção à amostra de padronização. A amostra em que se fundamentam as normas deve ser satisfatoriamente ampla, a fim de proporcionar valores estáveis. Outra amostra da mesma população escolhida de maneira semelhante não deve apresentar normas consideravelmente díspares das obtidas anteriormente. As normas com amplo "erro de amostragem" representariam um pequeno valor para a interpretação dos resultados de um teste (Anastasi & Urbina, 2000).

Do mesmo modo, é importante a exigência de que a amostra seja representativa da população considerada. Devem-se investigar, cuidadosamente, fatores relevantes durante a seleção dos sujeitos que irão compor a amostra de padronização, capazes de tornar a amostra não representativa. Entre estes fatores, devem ser considerados aspectos como idade, sexo, escolarização, nível socioeconômico, acesso à informação, entre outros.

Usualmente, considera-se a Padronização como todo o processo de estabelecer procedimentos padronizados e valores normativos

para a comparação e a avaliação do desempenho dos indivíduos ou de grupos. O processo de desenvolvimento de um teste padronizado exige fases, tais como pré-testagem de itens, análise de itens, estudos de validade e precisão, desenvolvimento de normas, etc.

Entende-se, portanto, a Normatização como o conjunto de valores típicos descritivos do desempenho, num determinado teste, de um grupo específico de indivíduos supostamente representativos de uma determinada população. As normas dão valores típicos para diferentes grupos homogêneos (segundo a idade, a escolaridade etc.), por meio da equivalência dos escores brutos, obtidos no teste, com alguma forma de escore derivado (desvio de QI, percentil, estanino, etc.). Vale ressaltar que as normas não devem ser consideradas exclusivamente como padrões ou níveis desejáveis de desempenho.

Por fim, destacou-se a importância do processo de padronização e de normatização como meios para aumentar a probabilidade de que as ferramentas para avaliação psicológica estejam cada vez mais adequadas ao trabalho desenvolvido por profissionais, e também como instrumentos para que avaliações não sejam invalidadas em virtude da má utilização dos testes, trazendo benefícios infindáveis principalmente para o indivíduo submetido à testagem.

Em síntese, a apreciação e a compreensão cada vez mais cuidadosa e consistente teoricamente das técnicas e ferramentas de avaliação psicológica, capazes de indicar, com maior precisão, os caminhos para tomada de decisão, surgem como uma necessidade prioritária nos cenários nacional e internacional.

A consolidação do campo da avaliação psicológica dentro da psicologia reveste-se de capacidade potencial de colaborar não

apenas para a melhoria da qualidade de vida das pessoas, mas também para que organizações e instituições disponham ainda mais de ferramentas competitivas no atual contexto globalizado, a partir da qualidade dos serviços oferecidos a seus clientes. Assim, a testagem provavelmente apresentará um melhor desempenho e, desta forma, cooperará mais eficazmente para o sucesso das avaliações psicológicas, o que poderá reverter-se em melhores produtos e serviços oferecidos.

Questões

1) A testagem psicológica diferencia-se de outras técnicas de avaliação, por se tratar de procedimentos referenciados a normas e a diretrizes interpretativas padronizadas. Para alguns autores há uma distinção clara entre padronização e normatização. Defina estes dois conceitos.
2) Suponha que, durante um processo de Recrutamento e Seleção, a utilização de testes psicológicos tenha sido feita de forma inadequada. Em uma situação como esta, o resultado do examinando pode ficar comprometido? Justifique.
3) O aplicador do teste é um elemento importante no processo de testagem. Explique os motivos pelos quais o aplicador pode afetar os resultados do teste.
4) A principal vantagem de se empregar transformações já utilizadas universalmente é que tornam os resultados comparáveis entre si. Assim sendo, seria o QI aplicável em todos os testes psicológicos?
5) Os escores de postos de percentil são o método mais direto e disseminado para transmitir resultados de testes referenciados em normas, contudo, os escores de percentil são muitas vezes confundidos com escores percentuais. Sendo estes dois tipos de escores distintos, justifique a diferença entre eles.

Referências

Alchieri, J. C. & Cruz, R. M. (2003). *Avaliação psicológica: conceito, métodos e instrumentos*. São Paulo: Casa do Psicólogo.

Ambroselli, C. (org.) (1987). *Comitês d'Éthique a Travers Le Monde: recherches em cours 1986*. Obra coletiva publicada por L'Inserm (pp. 5-12). Paris: Éditions Tierce.

Ames, L. B. (1937). The sequential patterning of prone progression in the human infant. *Genetic Psychology Monographs, 19*, 409-460.

Anastasi, A. (1977). *Testes Psicológicos*. São Paulo: EPU.

Anastasi, A. & Urbina, S. (2000). *Testagem psicológica*. Porto Alegre: Artes Médicas.

Binet, A. & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11*, 191-244.

Conselho Federal de Psicologia (1987). *Resolução 002 de 15 de agosto de 1987*. Brasília: CFP.

Conselho Federal de Psicologia (2003). *Resolução 002 de 24 de março de 2003*. Brasília: CFP.

Cronbach, L. J. (1996). *Fundamentos da testagem psicológica*. Porto Alegre: Artes Médicas.

Freud, S. (1960). The psychopathology of everyday life. In *Standard Edition*, Vol. 6. Londres: The Hogarth Press. (Originalmente publicado em 1901).

Gesell, A. & Amatruda, C. S. (1947). *Development diagnosis* (2a ed.). Nova Iorque: Hoeber-Harper.

Halverson, H. M. (1933). The acquisition of skill in infancy. *Journal of Genetic Psychology, 43*, 3-48.

Knoblock, H. & Pasamanick, B. (orgs.) (1974). *Gesell and Amatruda's developmental diagnosis* (3a ed.). Nova Iorque: Harper & Row.

Laurendeau, M. & Pinard, A, (1962). *Causal thinking in the child: A genetic and experimental approach*. Nova Iorque: International Universities Press.

Laurendeau, M. & Pinard, A. (1970). *The development of the concept of space in the child*. Nova Iorque: International Universities Press.

Pasquali, L. (2001). *Técnicas de Exames Psicológicos - TEP: Manual*. São Paulo: Casa do Psicologo.

Pasquali, L. (2003). *Psicometria: teoria dos testes na psicologia e na educação*. Petrópolis: Vozes.

Pasquali, L. (org.) (2010). *Instrumentação Psicológica: fundamentos e práticas*. Porto Alegre: Artmed.

Pinard, A. & Laurendeau, M. (1964). A scale of mental development based on the theory of Piaget. Description of a project. *Journal of Research in Science Teaching, 2*, 253-260.

Terman, L. M. & Merrill, M. A. (1960). *Stanford-Binet Intelligence Scale: Manual for the third revision, Form L-M*. Boston: Houghton Mifflin.

Treffinger, D. J. (1987). Research on creativity assessment. In S. G. Isaksen (Org.), *Frontiers of creativity research: Beyond the basics* (pp. 103-119). Buffalo: Bearly.

Urbina, S. (2007). *Fundamentos da testagem psicológica*. Porto Alegre: Artmed.

Werlang, B. S. G., Villemor-Amaral, A. E. & Nascimento, R. S. G. F. (2010). Avaliação psicológica, testes e possibilidades de uso. In Conselho Federal de Psicologia, *Avaliação psicológica: diretrizes na regulamentação da profissão* (pp. 87-100). Brasília: CFP.

Capítulo 7

A ética no uso de testes no processo de Avaliação Psicológica

Maria Cristina Barros Maciel Pellini
Irene F. Almeida de Sá Leme

A Avaliação Psicológica, tarefa prevista em lei como privativa do psicólogo[1], nos últimos anos vem difundindo-se, trazendo muitas contribuições em diversas áreas do conhecimento da psicologia. Pode definir-se a Avaliação Psicológica como um processo técnico e científico de coleta de dados e interpretações, com pessoas ou grupos de pessoas, por meio de informações obtidas em questionários, métodos, instrumentos psicológicos, entrevistas, entre outros (Noronha & Alchieri, 2002; Primi, Flores-Mendoza & Castilho, 1998; Wechsler, 1999).

Enquanto a Avaliação Psicológica refere-se a um processo amplo que envolve a integração de informações provenientes de

[1] Lei 4119/62, artigo 13º, parágrafo 1º.

diversas fontes, como testes, entrevistas, observações, análises de documentos, entre outras, a testagem psicológica deve ser considerada como uma das etapas da avaliação, por meio da utilização de testes psicológicos de diferentes tipos.

Pasquali e Alchieri (2001) definem testes psicológicos como um procedimento sistemático para observar um comportamento e descrevê-lo com a ajuda de escalas numéricas. Tradicionalmente, são encontrados testes com o objetivo de mensurar áreas tais como inteligência, cognição, psicomotricidade, atenção, memória, percepção, emoção, afeto, motivação, personalidade, dentre outras, nas suas mais diversas formas de expressão, segundo padrões definidos pela construção dos instrumentos.

Pellini, Rosa e Vilarinho (2002) apontam um ponto importante que merece reflexão quanto à qualidade dos instrumentos, mais especificamente, quanto à qualidade científica destes, com validação e normas atualizadas e adequadas à população que irá utilizá-los. Os *Princípios Éticos e Código de Conduta da American Psychological Association* (1992) dizem, em seu Artigo 2.07:

1. Os psicólogos não baseiam sua avaliação ou decisões de intervenção ou recomendações sobre dados ou resultados de testes que estejam desatualizados para a atual finalidade.
2. De modo semelhante, os psicólogos não baseiam tais decisões ou recomendações em testes e medidas obsoletas e não úteis para a atual finalidade.

A respeito de trabalhos e pesquisas com os instrumentos de avaliação psicológica, Jacquemin (1997) comenta:

> Os trabalhos sérios e confiáveis que permitem conhecer a situação real a respeito dos testes, sua utilização e edição no Brasil, são bastante escassos. A publicação de Hutz e Bandeira (1993) apresenta um levantamento da literatura realizado a partir dos resumos publicados em cerca de 1300 periódicos indexados na *Psychological Abstracts* (1974-1992) e a análise manual dos periódicos brasileiros (1987-1992), para conhecer as tendências contemporâneas dos testes. Os resultados apontam uma situação bastante precária no Brasil, tornando o trabalho do psicólogo brasileiro, em psicodiagnóstico, bastante difícil e problemático (inclusive do ponto de vista ético) (Hutz, 1989) (p. 59).

Alves (1998) comenta sobre a qualidade psicométrica dos testes como fundamental para a sua utilização, pois o emprego de instrumentos não padronizados para a realidade brasileira "leva muitas vezes ao uso de testes totalmente inadequados, o que também invalida todas as conclusões tiradas a partir dessas avaliações" (p. 22). A autora levanta, nessas considerações, a questão da atualização das normas dos testes e a necessidade de pesquisas periódicas para o estabelecimento dessas normas (Pellini, Rosa & Vilarinho, 2002).

O Conselho Federal de Psicologia (CFP), órgão que analisa e avalia os instrumentos de uso restrito dos psicólogos, orienta os profissionais a observarem os estudos realizados com cada teste, principalmente no que se refere aos estudos de validade, de precisão e de padronização. Assim, os requisitos básicos para uma

determinada utilização são os resultados favoráveis desses estudos, orientados para os problemas específicos relacionados às exigências de cada área. Conforme a Resolução do CFP nº 002/2003, que regulamenta a produção e a utilização de testes psicológicos, "as condições de uso dos instrumentos devem ser consideradas apenas para os contextos e propósitos para os quais os estudos empíricos indicaram resultados favoráveis". De acordo com os Artigos 10 e 16 da Resolução CFP n.º 002/2003 (transcritos abaixo), só será permitida a utilização dos testes psicológicos que obtiverem o parecer favorável pelo CFP e será considerada falta ética a utilização de instrumento que não esteja em condição de uso.

> Art. 10 - Será considerado teste psicológico em condições de uso, seja ele comercializado ou disponibilizado por outros meios, aquele que, após receber Parecer da Comissão Consultiva em Avaliação Psicológica, for aprovado pelo CFP.
>
> Parágrafo único - Para o disposto no *caput* deste artigo, o Conselho Federal de Psicologia considerará os parâmetros de construção e princípios reconhecidos pela comunidade científica, especialmente os desenvolvidos pela Psicometria.
>
> Art. 16[2] - Será considerada falta ética, conforme disposto na alínea "c" do Art. 1º e na alínea "g" e "h" do Art. 2º do Código de Ética Profissional do Psicólogo, a utilização de testes psicológicos que não

[2] Alterado pela Resolução CFP 023/2007; artigo 9.º.

> constam na relação de testes aprovados pelo CFP, salvo os casos de pesquisa.
>
> Parágrafo Único - O psicólogo que utiliza testes psicológicos como instrumento de trabalho, além do disposto no *caput* deste artigo, deve observar as informações contidas nos respectivos manuais e buscar informações adicionais para maior qualificação no aspecto técnico operacional do uso do instrumento, sobre a fundamentação teórica referente ao constructo avaliado, sobre pesquisas recentes realizadas com o teste, além de conhecimentos de Psicometria e Estatística.

É função do Psicólogo a avaliação e a escolha dos métodos e técnicas a serem utilizados em sua prática profissional. No caso dos testes, é importante, primeiramente, a consulta ao Sistema de Avaliação de Testes Psicológicos (SATEPSI), disponível no site do Conselho Federal de Psicologia (www.pol.org.br), buscando verificar se o instrumento escolhido consta na listagem com parecer favorável para uso profissional. Após a confirmação do parecer favorável do instrumento, é igualmente importante consultar o manual do referido teste, de modo a obter informações adicionais acerca do construto psicológico que ele pretende medir, bem como sobre os contextos e propósitos para os quais sua utilização se mostra apropriada.

Para a utilização de alguns instrumentos que exigem uma ou mais habilidades específicas (teóricas e de interpretação) por parte do aplicador, deve-se verificar, também, se não existem dificuldades tanto por parte do psicólogo (conhecimento para a interpretação

conforme a teoria e constructo em que o teste foi criado), ou, ainda, dificuldades físicas ou psíquicas do examinando. Ressalta-se a obrigatoriedade de utilizar-se o teste dentro dos padrões referidos por seu manual e cuidar da adequação do ambiente, do espaço físico, do vestuário dos aplicadores e de outros estímulos que possam interferir na aplicação.

Pellini e Pereira (2008), em matéria publicada no *Jornal Psi – CRPSP*, destacam que o uso de instrumentos de forma imprópria, com parecer desfavorável, ou mesmo uma interpretação errônea, pode prejudicar os examinados, além de implicar falta ética por parte do profissional. As autoras mencionam ainda que os instrumentos utilizados devem estar de acordo com as normas para evitar prejuízos à população usuária. Este importante alerta se deve à disseminação do uso dos testes psicológicos em processos seletivos e em várias outras circunstâncias, de maneira irresponsável.

O uso de instrumentos não favoráveis pode causar prejuízos aos usuários e resultar numa avaliação inadequada, como: candidatos "não recomendados" para assumirem cargos/funções em processos seletivos para uma vaga em uma empresa ou concurso público (na área de recursos humanos); pacientes para realização de cirurgia bariátrica se submeterem a processos avaliativos em que o recurso utilizado não afira as verdadeiras condições psíquicas desses pacientes (na área clínica/hospitalar); riscos de envolvimento em acidentes por candidatos que receberam carteira nacional de habilitação (CNH) sem terem a aptidão necessária; recomendação indevida de "apto" a candidatos ao registro e porte de arma de fogo, sem ter a condição necessária, trazendo riscos para o próprio candidato ou para as demais pessoas da sociedade, entre outros. Quando realizadas as avaliações psicológicas nestes contextos,

deve garantir-se que os instrumentos utilizados atendam aos critérios de aplicação, correção e interpretação, definidos em seus manuais, previamente analisados e aprovados pelo CFP, a fim de evitar danos e consequências, muitas vezes, irreversíveis.

As autoras alertam, nessa mesma matéria do *Jornal Psi - CRPSP* (2008), para outra questão que também requer atenção e refere-se ao fato de que muitos testes estrangeiros são trazidos ao Brasil, colocados em uso, inclusive por não psicólogos, e utilizados como parâmetros para comparações de sujeitos que a eles se submetem.

Isso traz prejuízo ao usuário, que pensa que está adquirindo serviços profissionais, quando, na verdade, está sendo avaliado por pessoas sem formação nem qualificação requeridas para a realização da avaliação, com instrumentos que, ainda que tenham excelente reputação em seu país de origem, não estão adaptados à população brasileira, o que pode implicar desvios significativos de resultados.

Muitos instrumentos jamais passaram por estudos de validação e, mesmo que tais estudos tenham sido realizados em outros países, é imprescindível a adaptação à nossa realidade.

Assim, os testes de qualquer natureza importados de outros países devem ser traduzidos para a língua portuguesa e padronizados a partir de estudos realizados com amostras nacionais, considerando a relação de contingência entre as evidências de validade, precisão e dados normativos com o ambiente cultural onde foram realizados os estudos para sua aplicação prática profissional (Pellini & Pereira, 2008).

Estima-se que as pesquisas com testes demoram em torno de quatro anos para o cumprimento das exigências mínimas de estudos

para sua padronização, e esse é outro componente dificultador no Brasil. Entre os diversos fatores que envolvem e demandam muito tempo de realização, encontram-se as dificuldades em aplicar os instrumentos em pesquisa[3]. Diante dessa informação, ressalta-se a pertinência e a necessidade de parcerias para a aplicação dos instrumentos nos mais variados contextos: clínicas, escolas, organizações, instituições de ensino etc.

Acredita-se que, consolidando essas parcerias entre instituições, pesquisadores e população, ocorra uma coleta de dados mais efetiva e representativa, atentando-se sempre aos cuidados para a aplicação de forma correta e sistemática, além de se observar a regulamentação relacionada à pesquisa com seres humanos.

É importante salientar que esse é um requisito orientado pelo Conselho Federal de Psicologia, conforme seu artigo 1º da Resolução nº 006/2004, que altera o artigo 14 da Resolução nº 002/2003:

> Art. 1º - Altera o art. 14 da Resolução CFP n.º 002/2003, que passa a ter a seguinte redação:
>
> Art. 14 - Os dados empíricos das propriedades de um teste psicológico devem ser revisados periodicamente, não podendo o intervalo entre um estudo e outro ultrapassar: 15 (quinze) anos, para os dados referentes à padronização, e 20 (vinte) anos, para os dados referentes à validade e precisão.

[3] Leme, I. & Rabelo, I., matéria publicada na *newsletter* DIPSI, Blumenau, 2007.

Enfim, para que um profissional atue de forma ética quanto ao uso de instrumentos, deve procurar manter contínuo aprimoramento profissional; utilizar, no contexto profissional, apenas testes psicológicos com parecer favorável, que se encontram listados no site do SATEPSI; realizar a avaliação psicológica em condições ambientais adequadas, de modo a assegurar a qualidade e o sigilo das informações obtidas; guardar os documentos produzidos decorrentes de Avaliação Psicológica em arquivos seguros e de acesso controlado; proteger a integridade dos instrumentos, não os comercializando, publicando ou ensinando àqueles que não são psicólogos.

A devolutiva no processo de Avaliação Psicológica

É importante mencionar outro aspecto fundamental envolvido no processo de avaliação, que se refere à entrevista devolutiva.

Conforme a Resolução do CFP n.° 01/2002, que regulamenta a Avaliação Psicológica em Concursos Públicos e em processos seletivos da mesma natureza, a devolutiva é direito de todo candidato sujeito a processos de avaliação psicológica:

> Art. 6º - . . .
>
> § 1º - O sigilo sobre os resultados obtidos na avaliação psicológica deverá ser mantido pelo psicólogo, na forma prevista pelo código de ética da categoria profissional.

> § 2º - Será facultado ao candidato, e somente a este, conhecer o resultado da avaliação por meio de entrevista devolutiva.

Deste modo, o candidato deverá ser informado sobre os serviços prestados e orientado quanto aos resultados obtidos, conforme artigo 1º, alíneas "g" e "h" do Código de Ética, que diz ser responsabilidade do psicólogo "informar a quem de direito os resultados decorrentes da prestação de serviços psicológicos, transmitindo somente o que for necessário, para a tomada de decisões, que afetam o usuário ou beneficiário" e "orientar a quem de direito sobre os encaminhamentos apropriados, a partir da prestação de serviços psicológicos, e fornecer, sempre que solicitado, os documentos pertinentes ao bom termo do trabalho".

A entrevista devolutiva, em sua maioria, é resultante de um processo de avaliação psicológica, sendo esta entendida como um processo técnico-científico de coleta de dados, estudos e interpretação de informações a respeito dos fenômenos psicológicos que são resultantes da relação do indivíduo com a sociedade, utilizando-se de estratégias psicológicas: métodos, técnicas e instrumentos.

Os resultados dessas avaliações ou devolutivas devem considerar e analisar os condicionantes históricos e sociais e seus efeitos no psiquismo, com a finalidade de servirem como instrumentos para atuar não somente sobre o indivíduo, mas na modificação desses condicionantes que operam desde a formulação da demanda até a conclusão do processo, segundo a Resolução nº 07/2003 do Conselho Federal de Psicologia.

A devolutiva não se constitui apenas em transmitir os resultados de um processo de avaliação psicológica, mas é, também, o

fruto de um trabalho realizado a partir de uma solicitação externa (Pellini, 2006).

Essas solicitações são oriundas de diversas áreas de atuação profissional: clínica, organizacional, educacional, judiciária, hospitalar, entre outras, e se configuram em solicitações que exigem cuidados e responsabilidades. O psicólogo deve ter sempre claro em seu trabalho o objetivo de estar realizando uma avaliação. Dependendo do motivo da solicitação, ele pode mudar radicalmente, por exemplo, o destino de uma pessoa, família, o desenvolvimento de uma criança ou uma decisão judicial.

É importante destacar o tipo de linguagem a ser empregada na devolutiva. No caso de trabalhar a devolutiva entre colegas psicólogos, o comunicado pode ser feito em termos técnicos, constando as referências aos recursos utilizados e discutindo-se os detalhes dos aspectos mais primitivos às defesas mais regressivas e mais maduras do cliente. Já em relação a outros profissionais, o psicólogo deve compartilhar somente as informações relevantes, resguardando o caráter confidencial e preservando o sigilo.

Algumas categorias profissionais têm características distintas a serem observadas. Uma solicitação feita por um juiz, por exemplo, que nomeia um psicólogo como perito no sistema judiciário, deve resultar em um laudo ou em um parecer, sendo que esses tipos de documentos escritos devem ser formulados com os devidos cuidados de redação e transmitindo somente o que for necessário para a tomada de decisões e para que os operadores do Direito possam compreendê-los.

Para uma devolutiva solicitada por escola, o psicólogo deve referir-se exclusivamente às questões levantadas na demanda

inicial, em linguagem acessível a quem vai receber o resultado, tomando as devidas precauções no sentido de não invadir a intimidade do caso por questões que não se relacionem ao campo pedagógico.

Nas situações de recrutamento e seleção, alerta-se para a importância de ter claro o perfil do cargo para selecionar as técnicas que serão utilizadas e os procedimentos, de forma a não causar danos aos candidatos. No momento da devolutiva, o psicólogo deve comunicar claramente ao solicitante se suas características estão ou não contemplando os anseios da empresa.

Nesses casos, é necessário ter o cuidado de não utilizar expressões como "você não passou no teste" ou "você não passou na avaliação psicológica", porque o candidato poderá considerar-se incapaz e portador de alguma dificuldade ou "problema". Isso pode não corresponder à realidade, mas apenas ao fato de que ele não apresenta as características exigidas para o desempenho da função (Pellini, 2006).

A importância da devolutiva nos processos de avaliação psicológica para obtenção da CNH é outro aspecto que devemos considerar. Existe a obrigatoriedade do cumprimento do Código de Ética do Psicólogo, no art. 1°, alínea g, e da Resolução nº 007/2009 do Conselho Federal de Psicologia que "institui normas e procedimentos para a avaliação psicológica no contexto do Trânsito" e revoga a Resolução CFP nº 012/2000.

Esta recente resolução destaca, em seu art. 1º, as normas e os procedimentos para avaliação psicológica de candidatos à Carteira Nacional de Habilitação e condutores de veículos automotores, no item *III - Dos instrumentos de avaliação psicológica*, alínea

"a"- *Entrevista Psicológica*, em que obriga o psicólogo a realizar a entrevista devolutiva, apresentando de forma clara e objetiva a todos os candidatos o resultado de sua avaliação psicológica.

A partir dos contextos acima, ressalta-se, ainda, que a devolutiva no processo de avaliação psicológica, assim como em qualquer área em que o psicólogo estiver atuando, deve sempre ser realizada de forma a promover o crescimento do indivíduo, e não o contrário.

Quanto à guarda do material produzido que fundamentou a avaliação psicológica, este deve ser guardado pelo prazo mínimo de cinco anos, e o psicólogo e/ou a instituição em que foi realizada a avaliação psicológica é responsável pelos materiais relativos à avaliação.

Para referência e orientação quanto à elaboração desse documento (de acordo com os princípios técnicos e éticos necessários para elaboração qualificada da comunicação escrita), devem ser seguidos os parâmetros da Resolução CFP nº 007/2003, que institui o *Manual de Elaboração de Documentos Escritos* produzidos pelo psicólogo, decorrentes de avaliação psicológica.

Quanto à responsabilidade técnica, o psicólogo deve ser capaz de transmitir ao candidato informações que o esclareçam sobre sua condição psicológica atual, e, se necessário, encaminhá-lo a outro profissional ou serviço especializado, conforme previsto no art. 1º alínea "g" e "h" do Código de Ética Profissional (2005).

Tal contexto remete à reflexão sobre o tema da devolutiva, que, além de sua importância, teve alteração introduzida pelo Código de Ética - Resolução CFP nº 010/2005. A mudança introduzida pelo Código de Ética vigente é que este prevê como dever

do psicólogo que a devolutiva seja também fornecida por escrito à pessoa atendida, caso esta venha a solicitar que seja feito dessa forma.

> Art. 1º - São deveres fundamentais dos psicólogos: . . .
>
> h. Orientar a quem de direito sobre os encaminhamentos apropriados, a partir da prestação de serviços psicológicos, e fornecer, sempre que solicitado, os documentos pertinentes ao bom termo do trabalho.
>
> A lei 10.241/1999, que dispõe sobre os direitos dos usuários dos serviços e das ações de saúde no Estado, também especifica que a prestação das informações deve ser fornecida por escrito:
>
> Artigo 2º - São direitos dos usuários dos serviços de saúde no Estado de São Paulo:
>
> IX - receber por escrito o diagnóstico e o tratamento indicado, com a identificação do nome do profissional e o seu número de registro no órgão de regulamentação e controle da profissão.

Como mencionado, a devolutiva refere-se ao momento em que o psicólogo transmite à pessoa atendida o resultado do trabalho realizado, orientando-o e fazendo os encaminhamentos necessários. Isso pode ocorrer tanto durante o atendimento (por exemplo, no decorrer de um processo psicoterapêutico) ou na sua finalização (por exemplo, após a realização de uma avaliação psicológica).

Caberá ao psicólogo, no entanto, avaliar quais informações devem ser documentadas, considerando: a situação específica,

os objetivos propostos do trabalho para o qual foi contratado e a fundamentação teórica do seu trabalho.

Ressaltam-se ainda os cuidados e deveres do psicólogo nas suas relações com a pessoa atendida quanto ao sigilo profissional, às relações com a justiça e ao alcance das informações, identificando riscos e compromissos em relação à utilização das informações presentes nos documentos em sua dimensão de relações de poder, conforme dispõe o *Manual de Elaboração de Documentos Decorrentes de Avaliação Psicológica*, em seus princípios éticos.

Considerações finais

Conclui-se que a ética no uso de testes no processo de Avaliação Psicológica corresponde à elaboração ou à escolha adequada de instrumentos, considerando a correta condição de aplicação e análise de seus resultados. Implica definir o que aferir, como aferir, as consequências dessa aferição e o uso dos resultados obtidos, ou seja, significa um processo. Segundo Sass (2000), é um equívoco considerar a Avaliação Psicológica tão somente como geradora de um produto.

Este autor destaca que a avaliação psicológica é marcada fortemente pelo aspecto técnico, o que parece ocultar a sua principal determinação: o aspecto político, pois sua dimensão técnica, dotada de procedimentos que avaliam pessoas e tomam decisões por elas, incide diretamente sobre a ação ético-política que o psicólogo executa em relação àquele que é avaliado.

Portanto, parece que o profissional da psicologia somente pode atuar de forma crítica e ética.

Questões

1) Defina Avaliação Psicológica.
2) Ao utilizar um teste psicológico, quais aspectos devem ser observados?
3) Qual a conduta adequada para o uso de testes internacionais na população brasileira?
4) No que consiste a entrevista devolutiva?

Referências

Alves, I. C. B. (1998). Avaliação Psicológica: ética, situação no Brasil e na formação do psicólogo. In *IV Encontro mineiro – O uso dos testes psicológicos* (pp. 17-32). Belo Horizonte: Vetor.

American Psychological Association. (1992). *Ethical principles of psychologists and code of conduct*. Washington, DC: American Psychological Association.

Conselho Federal de Psicologia. (2002). *Resolução 001/2002*. Recuperado em 19 de abril de 201, de HTTP://www.pol.org.br.

Conselho Federal de Psicologia. (2003). *Resolução 002/2003*. Recuperado em 19 de abril de 2011, de HTTP://www.pol.org.br.

Conselho Federal de Psicologia. (2003). *Resolução 007/2003*. Recuperado em 19 de abril de 2011, de HTTP://www.pol.org.br.

Conselho Federal de Psicologia. (2004). *Resolução 006/2004*. Recuperado em 19 de abril de 2011, de HTTP://www.pol.org.br.

Conselho Federal de Psicologia. (2005). *Resolução nº 010/2005*. Recuperado em 19 de abril de 2011, de HTTP://www.pol.org.br.

Conselho Federal de Psicologia. (2007). Resolução nº 023/2007. Recuperado em 19 de abril de 2011, de HTTP://www.pol.org.br.

Conselho Federal de Psicologia (2011). Sistema de Avaliação de Testes Psicológicos – SATEPSI. Recuperado em 21 de março de 2011, de http://www2.pol.org.br/satepsi/sistema/admin.cfm.

Jacquemin, A. (1997). As técnicas de exame psicológico como instrumento na pesquisa e objeto de pesquisa. *Boletim de Psicologia, XLVII*(107), 57-68.

Jornal PSI CRP-SP. A inserção do psicólogo no Poder Judiciário e sua interface com o Direito. Questões éticas. (Janeiro/março, 2006). *Jornal Psi CRPSP, 146*. Recuperado em 21 de março de 2011 de: http://www.crpsp.org.br/portal/comunicacao/jornal_crp/146/frames/fr_q. aspx.

Noronha, A. P., & Alchieri, J. C. (2002). Reflexões sobre os Instrumentos de Avaliação Psicológica. In R. Primi. (Org.), *Temas em Avaliação Psicológica*. (pp. 7-16). Campinas: Impressão Digital do Brasil, IBAP.