Validade – o Conceito, a Pesquisa, os Problemas de Provas Geradas pelo Computador

NICIA M. BESSA
Ph.D. em Educação, University of Pittsburgh
M.A. em Psicologia, State University of Iowa
nbessa@gbl.com.br

Resumo
Por muitos anos, o conceito de validade e a metodologia de validação da interpretação dos resultados, obtidos pelos examinandos em provas educacionais, evoluíram no sentido de uma incorporação de princípios e de métodos da psicologia cognitiva. Neste artigo, essa evolução é abordada como uma contínua discussão de temas, e como uma sucessão de linhas de investigação relevantes para a consolidação de uma teoria da validade e de uma metodologia de caráter científico. Ao focalizar desdobramentos mais recentes desse processo, trata-se com especial atenção o trabalho de pesquisadores que procuram estabelecer uma fundamentação adequada para uma nova metodologia de construção de testes, na qual os estudos sobre a validade de constructo são introduzidos no planejamento e no desenvolvimento da prova. Nessa perspectiva, são examinados problemas concernentes à validação, tanto nas provas educacionais produzidas artesanalmente, como nas geradas por programas de computador.
Palavras-chave: psicologia cognitiva, metodologia de testes, validade, validade de constructo, processo de validação.

Resumen
Durante muchos años, el concepto de validez y la metodología de validación de la interpretación de resultados, obtenidos por los alumnos en pruebas educativas, evolucionaron al incorporar algunos principios y metodologías de la psicología cognitiva. En este artículo, esta evolución es abordada como una continua discusión de temas y como una sucesión de líneas de investigación relevantes para la consolidación de una teoría de la validez y de una metodología de carácter científico. Al enfocar los alcances más recientes de este proceso, se trata con especial atención el trabajo de investigadores que procuran establecer fundamentos adecuados para una nueva metodología en la construcción de tests, al introducir los estudios sobre la validez de constructo en el planeamiento y desarrollo de la prueba. En esta perspectiva, se examinan problemas concernientes a la validación, tanto en las pruebas educativas producidas artesanalmente como en las generadas por programas de computación.
Palabras-clave: psicología cognitiva, metodología de tests, validez, validez de constructo, proceso de validación.

Abstract
For many years the concept of validity and a methodology of validating the interpretation of results obtained by examinees in educational tests has evolved in the direction of incorporating principles and methods of cognitive psychology. In this paper, this historical development is presented as a continuous discussion of topics, and as a succession of lines of investigation which are relevant for the consolidation of a theory of validity, and of a scientific methodology. With a focus on the most recent developments of this process, special attention is given to the work of researchers who try to establish an adequate foundation for a new methodology of test construction, in which construct validity studies are introduced in the design and development phases. Following this line, the paper focuses on problems of construct validation both in educational tests produced by educators, as well as test items generated by computers.
Key-words: cognitive psychology, testing methodology, validity, construct validity, validation processes.

INTRODUÇÃO

Este artigo focaliza os principais temas que têm sido discutidos a respeito do conceito e da pesquisa de validade das interpretações de resultados obtidos por aqueles que se submetem a provas educacionais. Dada a importância dessas provas, não somente em relação às ações que se baseiam na observação de diferenças individuais como também na avaliação de programas ou de projetos educacionais, a validade das interpretações dos resultados observados é crucial no que concerne à responsabilidade, envolvida em seus desdobramentos, perante a sociedade.

Na literatura das últimas três décadas sobre validade, são tratados como "provas educacionais" os vários procedimentos de coleta de dados – desde testes objetivos ou discursivos de conhecimentos até protocolos de observações do comportamento de indivíduos ou de grupos (*American Educational Research Association*, 1999; Cronbach, 1971; Messick, 1993).

De modo geral, os "resultados" são uma descrição e uma avaliação do comportamento observado nas provas educacionais, sejam expressos verbalmente de forma resumida, sejam quantitativamente em escores ou notas. Esses "resultados" são interpretados de várias maneiras. Pode-se, por exemplo, comparar o escore obtido por um indivíduo com as normas de uma população, ou localizar sua posição em uma distribuição de notas, ou concluir que suas respostas a determinadas questões mostram dificuldade em compreensão de leitura, ou classificá-lo como apto a passar a um curso de nível mais alto. O problema está em saber até que ponto tais interpretações são válidas – na investigação da validade dessas interpretações procura-se verificar qual o fundamento teórico e em que grau os dados empíricos lhes dão suporte.

Apesar da teoria psicométrica ser comum às provas educacionais e psicológicas, estas últimas não são focalizadas no presente artigo, que procura ater-se às provas educacionais por serem as que mais de perto interessam ao ensino, à pesquisa e à avaliação educacional.

Os principais temas versados na discussão sobre o conceito de validade são apresentados em um breve histórico, que destaca os diversos matizes que seu significado assume diante dos diferentes ângulos pelos quais se estudam os problemas da validade. A seguir, abordam-se as características principais das linhas de investigação do estudo da validade, em uma perspectiva de consolidação de uma metodologia que procura apropriar-se da teoria e de processos de pesquisa da psicologia cognitiva, e que enfrenta os novos problemas criados ao serem absorvidas as contribuições das ciências da computação.

UMA QUESTÃO ESSENCIAL

A perspectiva dos psicometristas remete aos demais especialistas em medidas educacionais a questão fundamental acerca dos resultados observados, obtidos pelos examinandos em provas educacionais. As teorias estatísticas dos escores de medidas educacionais tratam de modelos matemáticos cujos parâmetros não são definidos em termos de comportamentos observáveis. Na teoria clássica o "escore verdadeiro", representado por T, é uma abstração matemática – nas palavras de Lord (1980, p.5), um modelo estatístico é proposto, e é expresso em termos matemáticos, que não são definidos no "mundo real". Assim, também, no que concerne ao "escore verdadeiro platônico" (Lord, Novick, 1968, p.19), ou ao "escore verdadeiro" da teoria da generalização (Cronbach et al., 1972). Nas teorias que propõem "características latentes" para explicar o desempenho nas provas educacionais, essas variáveis latentes não são observáveis, não são mensuráveis diretamente – especificamente nas teorias da resposta ao item (TRI), a característica latente, representada por θ, assume valores conforme os pressupostos do modelo estatístico, mas não tem conteúdo substantivo (Hambleton, 1993).

Ao construir uma prova, cabe ao especialista definir detalhadamente o constructo focalizado, em termos da teoria cognitiva com que pretende explicar o desempenho dos examinandos. Uma vez coletados os resultados apresentados pelos examinandos, analisados os valores assumidos empregando-se o modelo psicométrico apropriado, e expostas as interpretações dos especialistas, questiona-se: até que ponto essas interpretações são adequadas, em face da definição do constructo proposto? Essa é a questão fundamental, a questão da validade, cuja resposta confere um sentido aos resultados observados.

A conceituação de validade e a concepção dos processos incluídos na validação refletem facetas importantes tanto do conhecimento científico como de questões sociais de diversas épocas: nos anos 50, a influência do behaviorismo e de certas correntes da filosofia da ciência; a partir dos anos 60, a preocupação com diferenças entre grupos populacionais – segundo o gênero, os níveis socioeconômicos, ou os conceitos pré-definidos de etnia; a partir dos anos 70 e 80, a influência dos avanços da psicologia cognitiva; e, sobretudo, a partir dos anos 80, o processo de validação sofre, também, o impacto das ciências da computação.

São correntes de pensamento sobre a conceituação e a investigação da validade que se sucedem, mas que também se superpõem no tempo e, em alguns casos, assumem novas nuances. Assim é que a maior parte das provas educacionais em uso atualmente reflete a base do pensamento

behaviorista, e coexiste com estudos inspirados nas teorias da cognição que procuram novas formas de avaliar o conhecimento do examinando; teorias da validade refletem a preocupação com diferenças entre grupos populacionais, enquanto o conceito de eqüidade continua reconhecidamente controverso (Cole, Zieky, 2001; Messick 1993; Zieky, 2002); os processos de validação continuam contemplando interpretações de resultados expressos em notas ou escores baseados em um conjunto de itens, ou de tarefas componentes da prova, enquanto as teorias psicométricas e as ciências da computação impulsionam a investigação para o estudo da validade em relação a cada item, a cada questão, desde a fase de planejamento e de construção da prova.

Ao longo dos últimos 55 anos, a conceituação da validade torna-se mais precisa, mas ainda é alvo de controvérsias. Há os que concebem a definição do constructo como um elemento isolado, ao qual o uso a que se destina a prova vem se somar; e há os que concebem a definição do constructo como uma construção em que todo um contexto – desde o uso, a população alvo até as condições de aplicação da prova – tem papel importante (Bennett, Bejar, 1997; Cole, Moss, 1993). Os processos de investigação também se tornam mais apurados: passam a dar atenção maior aos estudos de correlação entre escores e critérios diversos para uma série de evidências empíricas e para o suporte teórico das inferências sobre os resultados observados nas provas; ainda mais, vão até a exigência de técnicas do emprego de análise de cada questão antes de ser incluída na prova – ou seja, passam do exame da validade feito *a posteriori*, sobre os resultados de uma prova desenvolvida e aplicada, para a validação de cada tarefa ainda na fase de planejamento e de construção da prova, de modo que se verifique quais são os processos cognitivos envolvidos no desempenho do examinando e se a questão funciona adequadamente em relação aos fins propostos.

De uma forma ou de outra, a conceituação da validade se refere sempre à questão fundamental, expressa por Messick (1994, p.7): até que ponto, tendo em vista o constructo proposto, a teoria e as evidências empíricas dão suporte à interpretação do desempenho dos examinandos nas tarefas componentes da prova?

O CONCEITO DE VALIDADE

A análise a que se tem submetido o conceito de validade, nos últimos cinqüenta anos, levou a um refinamento para unificá-lo, a par de um maior detalhamento na identificação de fontes que podem invalidar a

interpretação dos resultados de provas educacionais. Numa reanálise do sentido dessa unificação, Kane (2001, 2006) propõe ângulos diferentes para a avaliação da validade e para o processo de validação.

AS DÉCADAS DE 50 E 60 – OS TIPOS DE VALIDADE

Na discussão sobre o conceito e os problemas da validade, nos anos 50 e 60, certos temas se destacam: a definição de conceitos abstratos que a prova pretende focalizar; a especificação e a generalização dos resultados da pesquisa de validação; a definição e a medida do que se considera como critério.

Até a década de 50, a preocupação com a validação das provas psicológicas e educacionais se revela nas pesquisas e na conceituação de três "tipos" de validade: validade de conteúdo, validade concorrente e validade preditiva. Na concepção da época, na validação do conteúdo procura-se verificar se a prova é constituída por uma amostra aceitável de situações (por exemplo, questões apresentadas ao examinando, momentos de observação, operações a executar) que permitam a observação de comportamentos dos quais se pretende extrair conclusões. No caso das provas educacionais, é comum serem constituídas de uma amostra aceitável de situações que representem programas curriculares e seus objetivos. Nos processos de validação preditiva e concorrente, procura-se comparar os resultados da prova a comportamentos exibidos em outras situações, tomando-se tais comportamentos como definição do que a prova pretende avaliar – situações e comportamentos que formam o que se denomina de "critério". São concepções que relacionam a validade ao uso que se pretende fazer dos resultados observados na prova.

A noção de "validade aparente" (*face validity*) – que corresponde ao que a prova, pelo tipo de questões ou de situações apresentadas, aparenta avaliar – já fora amplamente rejeitada, desde as primeiras análises sobre o assunto, por sua falta de fundamentação como processo científico (Cattell, 1964; Cureton, 1951; Mosier, 1947). A chamada "validade aparente" – tão cara a autores de questões de provas educacionais – não serve de suporte à interpretação dos resultados observados em relação ao que se pretende avaliar. Por exemplo: no caso das questões de provas educacionais em que se usa um parágrafo introdutório sobre o tema focalizado, antes de formular cada pergunta, supõe-se que se está apresentando a questão em um contexto temático. Na realidade (seja em questões de matemática, de física, ou de história, por exemplo), as diferenças observadas entre as notas obtidas pelos examinandos na prova podem ser devidas tanto ao

conhecimento do assunto em foco como a diferenças em compreensão de leitura de textos mais ou menos complexos, em rapidez de leitura, ou em familiaridade com esse tipo de questão, ou seja, introduz-se um elemento potencialmente associado ao que se conceitua como "variância irrelevante" em relação ao constructo que se pretende avaliar.

Algumas publicações da década de 50 e 60 constituem marcos importantes na conceituação da validade, por expressarem o pensamento de grupos de especialistas em medidas psicológicas e educacionais e por darem início a uma série de obras de referência que também servem de material instrucional: *Educational Measurement* (Lindquist, 1951), publicada pela *American Council on Education*; *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, preparada pela *American Psychological Association*, em 1954, logo seguida das *Technical Recommendations for Achievement Tests*, publicada pela *National Education Association*, ambas reformuladas na edição de 1966 dos *Standards for Educational and Psychological Tests and Manuals* (*American Educational Research Association*, apud Jackson, Messick, 1967).

Na primeira publicação do *Educational Measurement* (Lindquist, 1951), no capítulo de Cureton (1951), há clara prevalência a favor da validação em relação a cada uso particular dos resultados da prova. Com referência à validade preditiva, é uma concepção que suscita problemas de interpretação e de generalização das conclusões sobre as evidências empíricas coletadas. Ao examinar os muitos coeficientes de correlação entre os escores obtidos em uma prova e o desempenho escolar em certa disciplina, o problema do usuário é extrapolar para outras populações, outras situações, outros critérios, as conclusões a respeito das associações medidas no contexto particular de cada grupo de alunos. A relação entre a validade e o uso a que se destina a prova, e a especificidade dos estudos preditivos em face do critério são problemas recorrentes que continuam a ser abordados nos anos seguintes pelos especialistas.

De outro lado, a definição e a medida do critério suscitam problemas. A definição do critério em termos do comportamento observado não é simples. Cureton (1951) refere-se a problemas "lógicos" – termos como "habilidade", "proficiência", "perícia" representam conceitos abstratos, cujas definições operacionais permitem que instrumentos de avaliação e de medida sejam construídos e tenham sua validade investigada empiricamente. E há problemas metodológicos que exigem análise de aspectos que podem deturpar os resultados da investigação da validade: a fidedignidade da medida do critério, a escolha de um critério adequado, a possibilidade de fontes de tendenciosidade, a amostragem de comportamentos a serem observados (Cronbach, Meehl, 1955; Cureton,

1951). Acrescente-se que, na prática, é difícil para o investigador realizar uma pesquisa sobre a fidedignidade, ou a tendenciosidade do critério.

Em estudo de 1955 – época em que o behaviorismo procurava assentar a fundamentação filosófica para suas teorias – Cronbach e Meehl (1955) introduzem a concepção de "validade de constructo", entendendo por constructo a representação de algum atributo pessoal, supostamente refletido nas respostas dos examinandos, e que explica a variância em seu desempenho na prova. Na perspectiva de Cronbach e Meehl (1955), a validação de constructo é especialmente aplicável quando não se tem uma definicão operacional do constructo focalizado na prova. A primeira publicação da *Technical Recommentations*, pela *American Psychological Association,* em 1954, espelha esta concepção de Cronbach e Meehl de validade de constructo. Concepção essa que sofreu críticas, entre as quais a de Bechtold (1959, *apud* Jackson, Messick, 1967), que reflete a posição de boa parte dos psicólogos experimentais da época ao argumentar que, se um teste pretende avaliar certa característica dos indivíduos designada por um conceito abstrato (por exemplo, habilidade verbal), esse conceito deve fazer parte de uma teoria na qual uma cadeia conceitual inclua, no nível empírico, definições operacionais. A crítica de Bechtold provocou análises dos fundamentos filosóficos da conceituação da validade que se estenderam por vários anos (Cattell, 1964; Messick, 1993); e também um posicionamento menos extremado em edições posteriores dos *Standards for Educational and Psychological Tests and Manuals* – inclusive na edição de 1999 (*American Educational Research Association*, 1999).

Em sucessivas redefinições, o conceito de validade, que em anos anteriores se referia à prova, passou a referir-se aos resultados observados na prova e, a seguir, às interpretações desses resultados. E, apesar das referências generalizadas aos quatro "tipos" de validade, percebe-se a tendência emergente do conceito de validade de constructo abranger os demais "tipos" – por exemplo, na concepção de que a investigação da validade de constructo se nutre de diferentes estudos empíricos, sejam relativos ao conteúdo da prova, ou a correlações entre os resultados observados na prova e outras variáveis.

DA DÉCADA DE 70 À DÉCADA DE 90 – A UNIFICAÇÃO DO CONCEITO DE VALIDADE

Nos últimos trinta anos do século XX, além dos temas que já vinham permeando as concepções sobre a validade, tomam vulto a

unificação do conceito de validade, a conceituação de tendenciosidade (*bias*) e a investigação sobre suas possíveis fontes e conseqüências.

A relação entre o uso a que se destina a prova e a investigação da validade continua suscitando diferentes posicionamentos. Cronbach (1971) condiciona a validade preditiva a particularidades do critério e do contexto em que os dados são colhidos; de outro lado, reconhece que a tomada de decisões implica fazer generalizações e extrapolações, e propõe estudos de validação de constructo para propiciar uma base plausível para tais generalizações. Com o mesmo propósito, Messick (1993) alia à validação de constructo o exame da relevância e da utilidade da prova quanto ao uso a que se destina.

Deve-se observar que, na prática, é difícil dissociar o uso pretendido da prova do constructo focalizado. Em estudo sobre a validade de formas automáticas de atribuição de escores, Bennett e Bejar (1997) mostram que a definição do constructo focalizado está, na prática, interligada ao modelo de prova e de tarefas que a compõem; que, na construção de uma prova, há uma série de elementos interligados: a interface com o examinando, os instrumentos disponíveis para a criação das tarefas componentes da prova, o sistema de atribuição de escores, o sistema de interpretação dos resultados, e o sistema de comunicação da avaliação final aos interessados. Nessa perspectiva, a investigação da validade considera um constructo que se define no contexto de uma teoria cognitiva, de modelos psicométricos e de condições práticas.

Na mesma linha de pensamento, numa concepção alinhada com a prática de construção de provas educacionais, Cole e Moss (1993) sustentam que a avaliação da validade diz respeito exclusivamente ao grau em que as interpretações dos resultados obtidos na prova refletem o constructo visado; e que a definição do constructo está intimamente ligada à finalidade com que a prova é utilizada. O propósito da utilização da prova influencia o sentido, a interpretação dos escores, ou resultados em geral. Assim, a validade de constructo está essencialmente ligada ao contexto em que se usa a prova – um contexto que inclui uma prova com determinado conteúdo e objetivo, aplicada a certo grupo populacional, cujos resultados obtidos são utilizados para certos fins. Posicionamento semelhante é expresso nos *Standards for Educational and Psychological Tests and Manuals* (*American Educational Research Association*, 1999) – o enquadramento conceitual inclui uma descrição detalhada dos conhecimentos, das habilidades, das estratégias, dos processos e das características focalizadas; e essa descrição não só depende, como dela faz parte, a forma pela qual os resultados obtidos pelos examinandos serão utilizados.

Observe-se que, na prática, o contexto e o propósito da utilização da prova influenciam a sua construção e a escolha do tipo a ser usado, a interpretação que se pretende dar aos escores ou aos resultados observados, o tipo de informações a coletar para validar tais interpretações e, finalmente, o que se deve considerar como tendenciosidade. São exemplos: com o fim de distinguir melhor as diferenças entre indivíduos que compõem certo subgrupo de uma população é possível construir provas cujos escores observados tenham distribuição assimétrica positiva, ou assimétrica negativa; ou, usando a *Teoria da Resposta ao Item* (TRI), podem-se escolher itens de prova que melhor separem os examinandos em grupos de classes de diferentes níveis de habilidade (Hambleton, 1993). O problema é que a validação de constructo adquire um caráter particular, nessa concepção, dependendo de cada contexto, de cada utilização proposta para os resultados observados na prova.

A unificação dos chamados "tipos de validade" em uma só classe – a de "validade de constructo" – é acentuada por Messick (1993). Nessa vertente, validação de constructo compreende a análise teórica e a pesquisa de todo tipo de evidência empírica – inclusive da validade relativa ao critério e da relevância e da representatividade do conteúdo – que sirva de suporte à interpretação dos resultados obtidos pelos indivíduos na prova, em termos dos conceitos com que se procura explicar esse desempenho e sua relação com outras variáveis. Na concepção de Messick, os estudos do conteúdo e da relação com o critério têm importância como suporte e como parte da validação de constructo. A investigação sobre o conteúdo vai além de comparações com programas curriculares ou com um conjunto de situações que definem um universo que a prova deve representar; trata-se de estudos empíricos que sustentem a relação entre os resultados observados e a especificação do domínio de abrangência do constructo. Note-se que, ao enfatizar o estudo do conteúdo como referente à especificação do domínio abrangido pelo constructo, Messick (1993) vincula a investigação à teoria em que se fundamenta a prova. Nesse sentido, a análise do conteúdo é fundamental para que se verifique a possibilidade de sub-representação, ou de fontes de variância irrelevante.

A contribuição mais original da análise de Messick (1993) é a proposta de integração de facetas do conceito de validade, através da validade de constructo. No conceito de validade Messick distingue dois vetores: a) o da interpretação dos resultados obtidos pelos examinandos na prova, seja com base na análise teórica e nas evidências empíricas, seja com base no exame dos valores embutidos nessa interpretação e nas respectivas conseqüências; b) o da interpretação dos resultados e suas implicações com relação ao uso da prova, seja tomando por base sua relevância e utilidade,

seja com base na avaliação das conseqüências sociais de sua utilização. O cruzamento desses dois vetores revela o papel integrador da validação de constructo (Messick, 1994). A concepção de Messick amplia significativamente o domínio do conceito de validade – nessa vertente, a validade passa a depender da relação entre as interpretações dos resultados observados e o constructo, da referência ao uso, dos valores envolvidos na construção da prova e das conseqüências de sua utilização.

Com a noção de validação de implicações – sobre como responderão os indivíduos em situações estranhas à prova – que podem se aliar às descrições que procuram explicar os resultados na prova, Ebel (1963) e Cronbach (1971) já haviam introduzido o que mais tarde Messick (1993) chama de exame das "conseqüências" da interpretação e da utilização dos resultados. Percebe-se nas propostas de Cronbach (1971) e de Messick (1993, 1994), a preocupação com aspectos relacionados às possíveis fontes de tendenciosidade em provas psicológicas e educacionais e respectivas conseqüências individuais e sociais que, desde os anos 60, suscitava estudos dos especialistas em medidas.

Messick (1994) deixa bem claro que o exame dos valores não significa uma discussão de caráter sectário, ou uma justificativa com base em opiniões; diferentemente, propõe que se investigue se as interpretações dos resultados observados na prova e respectivas implicações refletem valores que não são parte do sentido do constructo focalizado na prova e da teoria em que o constructo se insere. Quanto à avaliação das conseqüências sociais da utilização da prova, Messick (1994) esclarece que não se trata de julgar se são positivas ou adversas; a proposta é de investigar se as conseqüências observadas são fruto de alguma fonte que possa invalidar o uso da prova. Seria o caso, por exemplo, de uma prova de compreensão de leitura, para a 4ª série do ensino fundamental, formulada em termos da linguagem e da cultura, típicas de grandes capitais do Sudeste brasileiro, aplicada a crianças de pequenos povoados do interior do Norte a fim de verificar a eficácia dos programas de ensino dessa região. Na perspectiva de Messick (1993), faria parte do processo de validação a investigação sobre valores associados ao conceito de compreensão de leitura que poderiam estar na base da escolha dessa prova, e sobre as conseqüências sociais – potencialmente negativas – de sua utilização. Para outros especialistas, as conseqüências sociais não fazem parte do processo de validação; nesse caso particular, bastaria constatar que a diferença entre grupos das duas regiões são devidas à intromissão de variável estranha ao constructo visado – um caso de tendenciosidade, que invalidaria a prova para o uso pretendido.

Não obstante a considerável repercussão das concepções de Messick sobre a validade, não houve nem há um consenso a respeito da inclusão do exame de valores e de conseqüências no processo de validação. Observa-se uma aceitação generalizada sobre a necessidade de se proceder a tal exame, mesmo entre aqueles que classificam a investigação sobre valores e sobre conseqüências como matéria de interesse de política educacional; mas ainda é controvertida a tese da inclusão da investigação acerca dos valores e das conseqüências no processo de validação (Cole, Moss, 1993; Kane, 2001, 2006).

Ao amarrar o propósito do uso e o contexto em que a prova é utilizada à conceituação de validade, Cole e Moss (1993) prescindem, no processo de validação, do exame da base de conseqüências e de implicações dos valores associados ao constructo – na validação, importa investigar indicações empíricas de que se apresentam, ou não, quaisquer fontes de invalidação, sejam fontes de variância irrelevante em relação ao constructo visado, seja um sistema inadequado de atribuição de escores, por exemplo. Para Cole e Moss (1993) a análise dos valores e o exame das conseqüências da utilização da prova – importantes sem dúvida – são parte das discussões que interessam aos responsáveis pela tomada de decisões, ou são de interesse da política educacional; mas não são parte do processo de validação das interpretações dos resultados obtidos nas provas. Nesse ponto, o processo de validação defendido nos *Standards for Educational and Psychological Tests and Manuals* (*American Educational Research Association*, 1999) afasta-se da perspectiva de Cole e Moss (1993) ao incluir, de um lado, a investigação sobre as conseqüências da utilização da prova e, de outro, os valores – que são contemplados com referência ao exame dos chamados "benefícios" (*American Educational Research Association*, 1999) – que o uso da prova pode trazer aos indivíduos e às instituições.

A partir do final dos anos 60, tomou vulto a discussão sobre o estudo da tendenciosidade das interpretações dos resultados observados. Para Cole e Moss (1993), ela é definida tecnicamente como a validade diferenciada de uma certa interpretação dos escores ou notas de subgrupos de examinandos – uma interpretação é tendenciosa quando não é igualmente válida para diferentes grupos de examinandos.

Particularmente em relação a procedimentos de seleção de pessoal, diversos modelos e processos estatísticos foram propostos para verificar objetivamente a tendenciosidade com base nos resultados observados nas provas, conforme a proporção de candidatos aceitos e rejeitados (Cole, Zieky, 2001; Messick, 1993). Do ponto de vista de Cole e Zieky (2001), por serem modelos referentes ao desempenho futuro e não a inferências em face do constructo focalizado na prova, não são considerados dentro da

questão da validade, e sim como concernentes a questões de política social. Quanto aos métodos estatísticos propostos para investigar a tendenciosidade de itens de provas educacionais, a medida de funcionamento diferencial (DIF) tem sido utilizada na seleção de questões na fase de construção desses testes. Além disso, a análise qualitativa de fontes que possam explicar as diferenças observadas entre grupos de examinandos é essencial para se concluir sobre a tendenciosidade dos itens (Sireci, Patsula, Hambleton, 2005).

A questão da eqüidade no desenvolvimento e no uso das provas educacionais está ligada à questão da tendenciosidade. Nos *Standards* (*American Educational Research Association*, 1999) reconhece-se que o termo é empregado com sentidos diversos, entre os quais o de ausência de tendenciosidade. A conceituação de eqüidade, baseada simplesmente na diferença entre resultados obtidos por grupos populacionais diversos, é rejeitada pelos mesmos motivos que é rejeitada na definição de tendenciosidade: a existência desse tipo de diferenciação tanto pode ser verdadeira como ser devida a fatores que invalidam a interpretação dos resultados observados (Cook, Schmidt-Castallar, Brown, 2005); e o julgamento baseado simplesmente na observação de tais diferenças está sujeito a juízos de valor que contaminam a investigação de sua validade (*American Educational Research Association*, 1999; Cole, Moss, 1993). Note-se que os problemas relativos à eqüidade envolvem questões de política educacional que refletem tensões sociais e interesses de natureza diversa. Aos especialistas responsáveis pelos estudos, pelo desenvolvimento, e pelo uso de provas educacionais cabe aprofundar a reflexão sobre conceitos, e a investigação sobre procedimentos que melhor reflitam o respeito aos princípios de eqüidade em relação a indivíduos e grupos sociais, dentro dos limites da área da avaliação educacional (Camilli, 2006; Cole, Zieky, 2001).

Principalmente nos últimos dez anos nota-se uma preocupação em caracterizar o processo de avaliação da validade como uma investigação de caráter científico que serve de base para uma argumentação sobre o grau de validade das interpretações dos resultados observados (*American Educational Research Association*, 1999; Kane, 2001). Messick (1993) observa que, na evolução do pensamento sobre a teoria da validade, acabaram por se tornar flexíveis as exigências de que essa teoria fosse fundamentada na vertente da filosofia da ciência refletida no pensamento de Cronbach e Meehl (1955). A opção defendida por Messick, então, é conceber a validação como um processo de investigação científica cuja função é colher evidências e ordenar argumentos que sirvam de suporte, ou que

contrariem as interpretações e o uso dos resultados das provas educacionais.

Na teoria da validade dos últimos anos do século XX, fica exposta a necessidade de distinguir com clareza os problemas referentes à validade e os problemas concernentes a políticas educacionais. O chamado "argumento de validade" (*American Educational Research Association*, 1999; Kane, 2006) é uma justificação, fundamentada numa investigação de cunho científico, do grau em que a teoria e as evidências empíricas dão suporte à interpretação dos resultados obtidos pelos examinandos. Nessa argumentação, é essencial que possíveis fontes de tendenciosidade sejam investigadas. Além disso, interessa aos responsáveis por decisões, baseadas nos resultados da prova, que as implicações e possíveis conseqüências do uso da prova sejam analisadas com base na teoria e em estudos empíricos – e esse não é um problema referente à validade, é um problema de política educacional. Caso as diferenças entre subgrupos populacionais sejam válidas – ou seja, caso não se constatem fontes de tendenciosidade que invalidem os resultados observados – investigam-se os possíveis fatores que contribuem para tais diferenças; evidentemente, diferenças reais, assim reveladas, são de interesse da política educacional.

TENDÊNCIAS ATUAIS – UNIFICAÇÃO OU UNIFORMIZAÇÃO?

Nestes primeiros anos do século XXI, a reflexão a respeito dos problemas e dos conceitos da teoria da validade continua com as concordâncias e discordâncias presentes no início da década de 90. Na realidade, notam-se mais diferenças no modo de conduzir a pesquisa de validação do que na conceituação da validade.

A abrangência do conceito de validade de constructo tem sido submetida a uma análise que tem resultado numa abertura maior em relação à investigação da validade de interpretações de resultados de provas que se referem apenas às relações entre o desempenho dos examinandos em situações bem delimitadas e o comportamento observado em condições também bem especificadas.

Ao focalizar o processo de validação, Kane (2001, 2006) distingue duas acepções diferentes do termo: a) no sentido de pesquisa de evidências que sirvam de suporte ao uso da prova e às interpretações dos resultados observados – que fundamenta o que denomina de *argumento de interpretação*; b) no sentido de avaliação do uso e das interpretações dos resultados observados nas provas, segundo critérios propostos – que constitui o *argumento de validade*. É nesse segundo sentido que Kane (2006)

analisa de modo detalhado a validação, que concebe como uma argumentação na qual são avaliados os fundamentos, a coerência, as inferências, os pressupostos, as evidências, as extrapolações e generalizações do *argumento de interpretação*. Para Kane (2001), é indesejável que a unificação do conceito de validade – como validade de constructo – seja entendida como uma uniformização do processo de validação, ou seja, não se deve admitir que toda e qualquer interpretação dos resultados de provas educacionais deva ser em termos de constructos teóricos. No caso de interpretações de respostas a estímulos específicos, obtidas em condições bem especificadas, o argumento de validade deve reportar-se à fundamentação do sistema de atribuição de escores, das generalizações em relação ao conjunto de respostas possíveis, das implicações e das extrapolações extraídas dos resultados observados; mas não se trata de validação de constructo (Kane, 2006). Ao propor que a validação de constructo não seja estendida uniformemente a todas as formas de interpretação de resultados observados em qualquer prova educacional, Kane (2001, 2006) sistematiza uma prática da investigação da validade que se distancia da concepção de Messick (1993).

Kane (2006) analisa também as interpretações qualitativas das observações do comportamento em diferentes ocasiões e contextos, das quais são exemplo as que os professores fazem nas escolas em relação a seus alunos. O *argumento de validação* das interpretações qualitativas inclui a avaliação de sua coerência, de sua amplitude e da fundamentação do enquadramento conceitual em que se desenvolvem. Com essas concepções, Kane (2006) avança na direção de alguns problemas que, na prática, continuam sendo enfrentados na investigação da validade das interpretações dos resultados obtidos pelos examinandos em provas educacionais

Nas três últimas décadas, tem ocorrido uma transformação na construção de provas educacionais que diz respeito à sua fundamentação na teoria psicológica. Em sua maior parte, as provas atualmente em uso se fundamentam nas teorias da psicologia diferencial e na perspectiva behaviorista; procuram refletir diferenças relativamente estáveis entre os indivíduos, ou entre grupos, em relação a habilidades, ou a conhecimentos que podem ser medidos. A análise de domínios de conhecimento leva à seleção de uma amostra de tarefas, ou questões, para estruturar a prova; infere-se a competência do examinando em relação a todo o domínio. A avaliação do desempenho no conjunto de questões da prova se expressa em relatórios ou em escores totais. Neste caso, o processo de validação refere-se ao constructo que a prova, em seu conjunto, pretende refletir, no contexto em que é utilizada. Ao investigar as relações entre esses escores

totais e outras variáveis, a validade de constructo acaba por ser concernente ao que as questões da prova – uma vez construída e aplicada – medem, ou seja, a definição do constructo focalizado passa a depender das relações empiricamente observadas entre os resultados obtidos pelos examinandos e outras variáveis (Embretson, Gorin, 2001; Kane, 2001). Essa transformação é descrita resumidamente por Pellegrino e Glaser (1980), que se reportam a investigações que procuram explicar diferenças individuais, medidas por testes de conhecimentos ou de aptidões, em termos de estruturas e de processos cognitivos – nessa perspectiva, deve-se procurar explicar as características psicométricas da questão de prova nos termos das explicações propostas nos estudos da cognição e do desenvolvimento cognitivo.

Seguindo essa mesma linhagem de estudos, Embretson posiciona-se entre os que orientam a evolução para os estudos do século XXI, ao propor que a elaboração das questões que devem constituir a prova siga processos semelhantes àqueles empregados em pesquisas experimentais da cognição nas quais as características de cada tarefa proposta aos sujeitos são sistematicamente manipuladas para testar hipóteses derivadas da teoria cognitiva, ou seja, as questões de provas educacionais devem ser derivadas da teoria cognitiva particular em que se fundamenta a construção da prova, e submetidas ao mesmo processo de experimentação sistemática – tal como nos estudos da cognição. Esse processo é concernente ao que Embretson designa *representação do constructo* (Embretson, Gorin, 2001; Embretson, 2005), isto é, refere-se aos processos cognitivos, às estratégias, aos conhecimentos diretamente envolvidos no desempenho da questão da prova. Embretson propõe ainda que o estudo das relações das notas, atribuídas ao desempenho nas provas com medidas de outras variáveis, seja elemento importante como indicação da utilidade da prova como medida de diferenças individuais. São, de modo geral, estudos baseados em correlações, dos quais, no contexto dessa proposta, a definição do constructo não depende – uma vez que a validade de constructo é investigada preliminarmente em relação a cada questão –, e que seguem hipóteses derivadas da *representação do constructo*.

Quando a validação, num processo de experimentação sistemática, desce ao estudo de cada tarefa, ou de cada item, desde o planejamento e a construção da prova, a investigação se aproxima do ideal de Cronbach (1957) de aliar a pesquisa experimental ao estudo de medidas de diferenças individuais. Na realidade, dentro dessa metodologia, os estudos são mais próximos do ideal de Cronbach (1971) – sobre a relação entre a psicologia experimental e os estudos da psicologia diferencial e das medidas

psicológicas – do que a teoria de validade desenvolvida pelo próprio autor (1971).

Atuando no mesmo sentido, outras vertentes importantes são as transformações operadas, principalmente nos últimos trinta anos, nas áreas da tecnologia da informação e das teorias psicométricas. Os modelos psicométricos orientam a interpretação dos resultados quantitativos da avaliação do desempenho dos examinandos; entretanto, não oferecem evidências que possam substituir os estudos empíricos para investigar a validade de constructo da interpretação de uma questão ou de uma prova. Contudo, os avanços da tecnologia da informação tiveram papel importante não só nas transformações quanto ao modo de apresentar provas educacionais ao examinando e de analisar os resultados observados, mas também na construção desses instrumentos. O impacto das novas tecnologias da informação tem sido de importância crucial no processo de integração da construção da prova com a pesquisa de validação, que desce ao nível dos itens, ou tarefas (Drasgow, Luecht, Bennett, 2006; Embretson, 2005; Pellegrino, Chudowsky, Glaser, 2001). Em todo esse processo observa-se que o conceito de validade permanece com suas raízes nas idéias dominantes desde a década de 90.

Quanto ao conceito de tendenciosidade, sua ligação com o conceito de validade marca uma possibilidade de maior entendimento entre os especialistas (Cole, Zieky, 2001). Ao examinar o caso de provas adaptadas a grupos culturais diferentes, Van de Vijner e Poortinga (2005) distinguem três tipos de tendenciosidade que podem ser diagnosticados por meio da investigação empírica e da aplicação de métodos estatísticos adequados: tendenciosidade de constructo, do método e do item da prova. Se houver *tendenciosidade de constructo* – caso em que a definição do constructo varia de um grupo populacional a outro – a comparação de resultados será inviável; mas se o caso é somente de tendenciosidade quanto ao método ou ao item, há a possibilidade de corrigir diferenças quanto à validade e de adaptar a prova aos grupos em questão (Van de Vijner, Poortinga, 2005).

No início do século XXI, no que se refere à área dos testes educacionais, não se conseguiu um consenso sobre a definição de eqüidade (Cole, Zieky, 2001). Possivelmente, isto ocorreu por tratar-se de um conceito cuja análise e definição se insere melhor nas reflexões da política educacional. Embora conceitos e métodos tenham sido desenvolvidos para tentar evitar a intromissão de fontes que possam invalidar diferenças entre resultados de provas aplicadas a diferentes grupos de examinandos, persistem problemas que refletem o contexto social em que as provas educacionais são utilizadas, e que são próprios dos estudos da área da política educacional.

Cole e Zieky (2001) observam que somente a partir dos anos 60 os especialistas em medidas educacionais expressam, em estudos teóricos e empíricos, uma preocupação nítida com a eqüidade em relação aos constructos, aos objetivos, ao processo de desenvolvimento, ao uso e aos resultados de provas educacionais para grupos culturais diferentes. É justo registrar, porém, que desde as primeiras edições, em 1954, da *Technical Recommendations* pela *American Psychological Associacion*, e do *Educational Measurement* (Lindquist, 1951), fica evidente que esses especialistas desejam contribuir para a qualidade das provas – em termos de conteúdo, arquitetura, aplicação e apuração dos resultados – colocadas à disposição da sociedade. E essa é outra faceta importante do seu sentido de responsabilidade social.

A PESQUISA DE VALIDAÇÃO

Embora o conceito de validade, em sua evolução, venha orientando o rumo das investigações, o uso da prova educacional continua sendo fator importante a motivar e a definir o escopo da pesquisa de validação. Ao emprego na seleção de pessoal e no acesso às universidades correspondem estudos cuja metodologia e cujos objetivos são adequados à pesquisa das relações com variáveis definidas como critérios. Os testes padronizados para acompanhamento do desempenho de alunos da escola fundamental e média têm suscitado indagações sobre o conteúdo das provas educacionais, sua relação com o currículo e com os objetivos do sistema educacional. Com a generalização do conceito de validade de constructo, as análises fatoriais, inicialmente mais comuns na área dos testes psicológicos, passaram a figurar nos estudos de validação das provas educacionais.

Esses são estudos que não só caracterizam uma fase do desenvolvimento da investigação da validade que toma como base os resultados gerais – ou escores – obtidos pelos examinandos na prova, mas que também continuam sendo fonte essencial de informação para dar suporte à sua interpretação, por meio das análises de alinhamento, dos processos correlacionais e das análises da variância. São esses estudos – classificados por Embretson (Embretson, Gorin, 2001; Embretson, 2005) como *tradicionais* – que fundamentam as interpretações, em termos de diferenças individuais, dos resultados que são observados no conjunto de questões da prova e são expressos de forma global.

O desenvolvimento das pesquisas nas áreas da psicologia cognitiva, dos modelos psicométricos apropriados à análise de cada item de prova, e das ciências da computação vem impulsionando a investigação da validade

de constructo no sentido de focalizar cada questão proposta, desde a fase de planejamento da prova. Além disso, o uso generalizado do computador tem concorrido para reforçar pressões sociais que levam os pesquisadores a encarar problemas inteiramente novos, tanto na construção das provas educacionais como na investigação da validade – trata-se agora de desenvolver metodologia adequada à geração de provas por programas computacionais específicos, e de métodos de validação para o caso de itens de provas produzidos pelo computador durante a respectiva aplicação ao examinando.

A PESQUISA TRADICIONAL DE VALIDAÇÃO

Tradicionalmente, os estudos de validade são baseados em correlações entre os resultados obtidos na prova e variáveis diversas. São, por exemplo, investigações em que se correlacionam os escores obtidos pelos examinandos na prova e critérios vários, ou são estudos que empregam análises fatoriais ao focalizar a validade de constructo. Esta metodologia dos estudos de validação, que reflete a concepção que representa a forte influência do pensamento de Cronbach (Cronbach, Meehl, 1955; Cronbach, 1971), dominou até o fim da década de 90 e ainda prevalesce na maior parte da literatura especializada: busca-se a definição do constructo na rede de relações entre os resultados observados na prova e outras variáveis selecionadas (Embretson, Gorin, 2001; Embretson, 2005).

Nas provas empregadas na seleção, na classificação de pessoal, na promoção de alunos, ou no acesso à universidade, a ênfase está em se obter uma ordenação dos resultados obtidos pelos examinandos, de modo a diferenciar níveis de desempenho, tão consistentemente quanto possível. É importante estabelecer diferenças entre os resultados obtidos pelos examinandos, de maneira a prever diferenças futuras quanto à sua atuação em áreas relacionadas ao respectivo desempenho na prova. São provas menos adequadas ao diagnóstico de dificuldades do aluno – de modo geral, podem ser mais adequadas como fonte de informação para políticas sociais; mas, por sua natureza, sua contribuição é pobre como base para o professor ajustar o processo instrucional a características individuais de seus alunos. Nesses casos, à pesquisa de validação interessa sobretudo estabelecer o grau de correlação entre os escores obtidos pelos examinandos na prova e a variável definida como critério.

São exemplos os estudos que focalizam a correlação entre os escores observados em provas de admissão a cursos superiores e resultados em medidas de critérios diversos. No Brasil, alguns estudos pioneiros (Bessa,

Mettel, 1965; Monteiro, 1964), tomando como critério as notas em exames vestibulares, ou em cursos pré-vestibulares, usaram processos de correlação, de análises de regressão univariada e multivariada, e de análises de discriminação para avaliar a relação com escores obtidos previamente nos testes do DAT – Formas A e B (*Differential Aptitude Tests*, adaptação do ISOP-FGV) – então usados na orientação educacional –, ou com os Testes de Desenvolvimento Educacional (Bessa, 1971), que refletiam o currículo de nível médio da época. Na mesma linha de interesse, a associação de notas do exame vestibular com o desempenho no curso de Engenharia foi estimada (Bessa, 1980). Silveira e Pinnent (2001) pesquisaram as correlações entre provas de admissão a duas universidades às quais um mesmo grupo de candidatos foi submetido à mesma época.

O problema da generalização das correlações entre escores em provas de acesso à universidade e o desempenho no curso superior tem sido focalizado em estudos metanalíticos ou no exame de dados acumulados por longos períodos. O estudo de Boldt (1986), por exemplo, focaliza o resultado de pesquisas de correlação entre escores no SAT (*Scholastic Aptitude Test*) com as notas médias obtidas no primeiro ano do curso superior, em 99 universidades. A hipótese testada de que as correlações do SAT-V e do SAT-M podem ser generalizadas por todas as instituições é parcialmente aceita, existindo entretanto uma substancial diferença entre as universidades. Já um relatório do Boars (2002) – Conselho da Universidade da Califórnia – tem como foco específico a comparação dos resultados do SAT-I e do SAT-II em relação ao desempenho dos alunos no *college* dessa universidade. Análises de regressão múltipla, usando dados de 77.800 alunos, mostraram que o SAT-II, ao ser incluído na equação, juntamente com as notas médias escolares, eleva de 15,4% para 22,2% a variância explicada das notas médias no primeiro ano do curso universitário.

Em relação a provas utilizadas na avaliação de programas educacionais, o interesse dos pesquisadores se volta freqüentemente para a validade de constructo e para a análise de conteúdo. Apesar de serem provas que objetivam estabelecer diferenças e servir de base para interpretações sobre níveis de desempenho acadêmico de grupos populacionais diversos, além das correlações com critérios apropriados interessa também pesquisar as evidências que sirvam de suporte às interpretações concernentes ao constructo visado. Na linha das avaliações da validade de constructo a metodologia de investigação varia bastante, sendo as análises fatoriais empregadas freqüentemente no Brasil. Num esforço de oferecer uma base cognitiva para explicar o desempenho dos examinandos, provas do Saeb e do ENC têm sido submetidas a análises

fatoriais – veja-se, por exemplo, a página do Departamento de Psicologia da Universidade de Brasília, de 2004, com resumos de dissertações com análises fatoriais de provas do Saeb e do ENC, além de estudos sobre a tendenciosidade de itens com relação a diferenças entre vários grupos, inclusive por regiões do país. Vale notar o interesse particular de estudos brasileiros por diferenças entre regiões, focalizadas também em outros trabalhos (Soares, Genovez, Galvão, 2005). Outros estudos fatoriais ampliam a área de provas focalizadas, ao pesquisar a validade de constructo em campos tão diversos como educação física (Balbinotti et al., 2004) e compreensão de leitura (Santos et al., 2002).

Com o uso crescente de provas computadorizadas, alguns estudos investigam a validade de diferentes versões de provas, algumas impressas e outras aplicadas com o uso do computador. Um estudo de Lawrence e Feigembaum (1997), por exemplo, compara resultados da aplicação de uma versão experimental do SAT com outra computadorizada; embora os autores concluam que as correlações encontradas sugerem que ambas refletem os mesmos constructos, tanto na parte verbal como na parte de Matemática, mostram-se cautelosos na generalização desses achados. Bennett e Rock (1998) empregam diversos processos ao comparar o teste GRE CAT – *Computerized Adaptive GRE General Test* (versão computadorizada do *Graduate Record Examination*) – com uma forma experimental, também computadorizada, do *General Explanations Test* (GE), com o propósito de examinar a validade de constructo deste último. Resultados de correlações simples e de análises fatoriais mostram que o GE é fracamente relacionado ao GRE, e num processo de regressão linear múltipla hierarquizada, o GE não apresenta incremento significativo à explicação da variância das notas médias no primeiro ano universitário além daquela obtida pela inclusão, na equação, dos escores no GRE.

Numa outra vertente, Primi et al. (2001) analisam a definição de competências e de habilidades proposta no desenvolvimento do Enem. Além de contribuir para esclarecer conceitos como os de competência e de habilidade, a análise desses autores envereda por uma investigação de fundamentos teóricos que sustentem a validade de constructo das interpretações dos resultados obtidos pelos examinandos no Enem, e chama a atenção para um ponto crucial no desenvolvimento de provas educacionais: a necessidade de especificação do modelo teórico diante da natureza das questões apresentadas na prova.

Note-se que o avanço dos estudos da cognição tende a refletir-se no sentido de exigir maior clareza e objetividade na fundamentação das provas educacionais, inclusive com exigências quanto ao suporte em pesquisa empírica. É o caso, por exemplo, do estudo de validação no qual

Ayala et al. (2002) partem de uma teoria da multidimensionalidade do desempenho dos alunos em ciências e avaliam as intercorrelações entre os resultados apresentados pelos examinandos em três testes de *performance* e outros três de múltipla escolha, cada teste focalizando um dos constructos: conhecimento básico e raciocínio, raciocínio espacial-mecânico e raciocínio quantitativo em ciências. As conclusões sugerem a necessidade de outros estudos: as intercorrelações observadas indicam que os três testes de *performance* referem-se mais a medidas de conhecimento básico e de raciocínio quantitativo; e os protocolos, com as descrições feitas pelos examinandos das respectivas estratégias de resolução dos problemas dos testes de performance – processo de "pensar alto" – apontam que os procedimentos dos alunos variam de acordo com o conhecimento de que cada um dispõe.

O Saeb tem motivado estudos que focalizam o conteúdo das provas. No trabalho de Rodrigues (2006) faz-se uma avaliação das provas de Matemática de 1997 e 1999 do ponto de vista do conteúdo, em face das matrizes curriculares que foram associadas a categorias de competências cognitivas, conforme definidas no plano das provas; além disso, com análises qualitativas e quantitativas de cada prova e de cada item procura-se esclarecer a interpretação tanto do desempenho dos alunos como do desempenho dos itens.

É interessante observar que a crítica de Messick (1993) tornou mais claro o papel das análises de conteúdo da prova como contribuição ao argumento de validade. Num desdobramento importante, a metodologia e o escopo de certas análises de alinhamento ampliam os limites da avaliação de conteúdo. Trata-se de uma avaliação minuciosamente arquitetada de um sistema de ensino, do ponto de vista da congruência de todos os elementos que o compõem. Na concepção de Webb (1997), o alinhamento refere-se ao grau em que todos os elementos da política educacional de um sistema atuam em conjunto para guiar a instrução e, em última análise, a aprendizagem. Todo o sistema de avaliação da aprendizagem faz parte desse conjunto e é, como um sistema, incluído na análise do alinhamento – obviamente sendo parte importante o exame do conteúdo das provas educacionais. Entre várias metodologias, a proposta de Porter (2001) desenvolve medidas do conteúdo do currículo – indicadores curriculares – e de suas relações com medidas de avaliação e com padrões de expectativas pré-estabalecidos pelo sistema instrucional. Com base nesses indicadores, Porter propõe uma metodologia quantitativa para avaliação do alinhamento do conteúdo, denominada de *currículo proposto*, de *currículo posto em prática*, de *conteúdo curricular da avaliação* e de *conteúdo curricular dominado pelo aluno*. De modo geral, a metodologia de

alinhamento representa um passo à frente – em objetividade, rigor e possibilidade de quantificação – no que concerne à avaliação do sistema de ensino do ponto de vista do conteúdo das provas educacionais.

A pesquisa de Ferrara (2004) exemplifica a importância dos estudos de alinhamento entre a declaração do objetivo do item de um teste educacional e o comportamento dos examinandos ao procurarem respondê-lo. O trabalho procura identificar e explicar o alinhamento entre o objetivo de cada um dos itens da prova e os conhecimentos, habilidades e processos – detalhadamente definidos e codificados – identificados numa observação tão objetiva quanto possível do comportamento dos examinandos. O alinhamento, entre o que se pretende medir e as respostas do examinando que são observadas realmente, é, para Ferrara (2004), uma evidência que concorre com outras na argumentação sobre a validade de constructo das interpretações dos resultados observados na prova.

Na metodologia dessas várias linhas de investigação, em que se procura obter evidências que sirvam de suporte à validade de constructo, percebe-se que a própria definição do constructo depende da rede de relações entre resultados observados na prova e variáveis externas. Na análise de Embretson (Embretson, Gorin, 2001; Embretson, 2005), esta concepção tradicional da validação de constructo limita o papel da teoria cognitiva na elaboração da prova, pois pressupõe que relações sejam empiricamente observadas entre os resultados da prova e outras variáveis – ou seja, depois da prova pronta e aplicada – para que se possa conferir um sentido ao constructo que se pretende medir.

Pellegrino, Chudowsky e Glaser (2001) destacam três fontes que, nas últimas décadas, vêm contribuindo para estudos que preparam uma base para uma transformação na construção de provas educacionais, e que se refletem na investigação da validade: os avanços nas teorias da cognição e do desenvolvimento cognitivo, nas teorias psicométricas e na tecnologia da informação. Seria possível acrescentar a essa lista a pressão das preferências dos examinandos, que se soma ao interesse de instituições usuárias (organismos governamentais, centros de treinamento de pessoal, ou de desenvolvimento de testes educacionais) para promover maior facilidade de acesso de indivíduos e de grupos aos meios de avaliação, maior rapidez na apuração e na comunicação de resultados, a adaptação do teste ao indivíduo, e a diminuição de custos em todo o processo desde a produção da prova (Drasgow, Luecht, Bennett, 2006). Não são tendências atualmente observadas no Brasil, mas, certamente, no futuro, terão reflexos no país.

A INFLUÊNCIA DOS ESTUDOS DA COGNIÇÃO E DO AVANÇO DA TECNOLOGIA COMPUTACIONAL

...technological innovation in assessment should be grounded on the constructs we aim to measure rather than in the technology per se.
(Bejar, 2002, p. 202)

Com o avanço das teorias da cognição e das ciências da computação, a exigência de um modelo cognitivo estipulado no planejamento da prova é estendida à criação de cada item, ou tarefa; e os estudos empíricos em que se fundamenta o modelo cognitivo servem de suporte à validade de constructo. O progresso das ciências da computação estimulou não apenas o uso do computador como instrumento de aplicação de provas e de avaliação das respostas dos examinandos, mas também estudos sobre provas adaptáveis ao indivíduo e sobre questões geradas por *softwares* específicos. Aos estudos da psicologia cognitiva e ao desenvolvimento da tecnologia computacional somou-se a contribuição da evolução da teoria psicométrica. Como conseqüência, a investigação da validade de constructo desce ao nível de cada questão proposta para integrar a prova.

A integração do modelo cognitivo ao planejamento da prova vale, também, para aquelas que não são computadorizadas, como mostram os esquemas gerais de elaboração de provas propostos por Mislevy (2002) e por Pellegrino, Chudowsky e Glaser (2001). Ambos os esquemas, baseados em análises minuciosas da elaboração de provas educacionais, sejam ou não computadorizadas, deixam muito clara a complexidade desse processo, que exige equipes de especialistas com domínio das teorias cognitivas, do sistema de ensino e das teorias em que este se fundamenta, das teorias e da prática das medidas educacionais, das teorias psicométricas, da metodologia da pesquisa empírica na área da cognição – e, eventualmente, também, de problemas específicos que devem ser estudados quando se emprega qualquer tecnologia. Esses esquemas enfatizam a necessidade de se abordar os problemas da validação, desde a concepção e a produção de cada questão proposta para compor a prova, qualquer que seja a teoria cognitiva que oriente o projeto e a tecnologia adotada.

Ao focalizar cada questão proposta para integrar a prova, vários processos de validação de constructo têm sido empregados, inspirados na metodologia das pesquisas empíricas da psicologia cognitiva. Pellegrino,

Chudowsky e Glaser (2001) destacam o emprego de processos de análise cognitiva de itens ou tarefas, como, por exemplo, análise de erros, ou análise de protocolos de descrições feitas pelos examinandos das respectivas estratégias adotadas na resolução de problemas. São processos de pesquisa empírica valiosos principalmente na exploração de tarefas que possam representar o constructo focalizado, e que requerem que a investigação se limite a grupos relativamente pequenos – veja-se, por exemplo, a primeira fase dos estudos de Newstead et al. (2002) que exploram características de questões de raciocínio analítico e suas relações com os respectivos índices de dificuldade. Nessas pesquisas, além de verificar o tempo de resposta a cada questão, empregam-se processos em que o examinando resolve problemas "pensando alto".

O estudo de Ferrara (2004) exemplifica a análise prévia de cada questão de uma prova, de modo a especificar detalhadamente os requisitos para respondê-la, em termos de conhecimentos, de estratégias usadas pelo examinando, e de processos particulares de encarar o problema. O método empregado é o da gravação por áudio e vídeo enquanto o examinando "pensa alto" ao procurar a resposta para cada questão da prova. A comparação dos resultados dessa análise com testes educacionais atualmente usados – revistos por Ferrara, DeMauro (2006), e que são fruto de um processo intuitivo de desenvolvimento de questões de provas tradicionalmente empregado – dá a medida da importância da introdução da validação em relação à interpretação de cada questão, desde o planejamento, na metodologia da construção de provas educacionais.

Estudos preliminares têm abordado uma variedade de problemas que vão desde os efeitos de diferentes interfaces com que se apresentam as questões ao examinando até os problemas da análise psicométrica e da especificação do modelo cognitivo. Seja na Teoria Clássica, seja na TRI, os respectivos modelos psicométricos expressam a propensão dos examinandos a exibirem um desempenho de nível mais ou menos alto em determinadas situações – definidas por questões, ou tarefas, procedimentos de exame etc., isto é, ensejam interpretações quanto a diferenças entre o nível de desempenho dos indivíduos, mas não sobre processos cognitivos que possam explicar esse desempenho. Além disso, alterações têm sido introduzidas em modelos da TRI de modo que expressem o peso que variáveis cognitivas selecionadas podem assumir no desempenho, em certas situações definidas na prova (Embretson, 2005). Situações mais complexas, representadas em tarefas ou itens de uma prova, têm motivado a criação de modelos psicométricos também mais complexos, que vêm ao encontro das necessidade de se interpretar os resultados observados em termos não só de diferenças entre indivíduos em certo momento, mas

também em diferentes estágios de desenvolvimento, ou entre classificações de indivíduos e de itens, ou de múltiplos constructos (Mislevy, 2006; Pellegrino, Chudowsky, Glaser, 2001). Do ponto de vista da validade de constructo, faz parte do argumento de validade uma avaliação da integração do modelo psicométrico à natureza da prova, aos objetivos de sua utilização, e ao tipo de interpretação que se faz dos resultados obtidos pelos examinandos na prova.

Algumas pesquisas ilustram o avanço no sentido da integração do modelo cognitivo, do modelo psicométrico e da investigação da validade desde a fase do planejamento da prova. Essa integração se observa, claramente, por exemplo, no *sistema cognitivo de planejamento* da prova proposto por Embretson (Embretson, Gorin, 2001; Embretson, 2005). A proposta de Embretson é particularmente importante porque sistematiza a aplicação da metodologia da pesquisa experimental na validação da interpretação dos resultados obtidos pelos examinandos nos itens da prova, seguindo a linha preconizada, entre outros, por Pellegrino e Glaser (1980). Além do esquema conceitual, no qual distingue dois aspectos da validade de constructo – a representação do constructo e a rede das relações dos resultados da prova com outras variáveis –, são estabelecidos estágios no procedimento da validação (Embretson, Gorin, 2001) que servem como guia para orientar a investigação. O *sistema cognitivo* proposto por Embretson (Embretson, Gorin, 2001; Embretson, 2002, 2005) baseia-se na teoria cognitiva do processamento da informação, e especifica processos envolvidos na solução dos problemas apresentados no item, seu impacto no desempenho do examinando e nas características do item sobre os processos cognitivos.

As pesquisas de Embretson (Embretson, Gorin, 2001; Embretson (2002, 2005), acerca dos itens de testes psicológicos que focalizam relações espaciais (completação de figuras), ilustram a especificação de um modelo cognitivo e a metodologia de validação dos itens. Embretson (2002, 2005) deriva um *modelo cognitivo* para os itens do tipo matrizes (empregados nos testes de Matrizes Progressivas, de Raven), fundamentado numa teoria cognitiva e respectivas pesquisas empíricas que indicam processos cognitivos – como *descobrimento de correspondências* e de *indução de relações* – na base da resolução desses problemas. Esses dois processos são incluídos em *modelos cognitivos* do item. São também incluídas características das figuras: fusão, distorção e sobreposição. As variáveis integrantes do *modelo cognitivo* são operacionalizadas e valores são atribuídos a cada uma. Com a aplicação dos itens a grupos de indivíduos, os parâmetros de dificuldade e de discriminação são estimados. O *modelo cognitivo* é avaliado conforme a estrutura cognitiva postulada; as características dos itens que

operacionalizam os processos cognitivos são tomadas como variáveis independentes, num modelo hierárquico de regressão múltipla em que as variáveis dependentes são a dificuldade, a discriminação e o tempo de resposta dos itens. Nesse estudo, duas das variáveis – *correspondência entre figuras* e *número de regras* –, que operacionalizam dois processos cognitivos, apresentam correlação positiva significativa com a dificuldade dos itens. Com o conjunto das variáveis cognitivas, as correlações múltiplas com as estimativas das dificuldades foram de 0,79 e de 0,81 conforme o *modelo cognitivo* respectivo. O mesmo efeito é observado em outro modelo cognitivo do item, quanto à variável que operacionaliza a carga de memória ativa, sendo a correlação obtida de 0,82. Nessa relação entre as variáveis cognitivas incluídas no modelo e as respostas dos examinandos, fundamenta-se a validade das interpretações dos resultados observados. Ou seja, no sistema proposto por Embretson, busca-se uma explicação para as respostas do examinando nas variáveis incluídas no modelo cognitivo – nos termos da teoria cognitiva específica da qual o modelo é derivado, a qual se enquadra na teoria geral do processamento da informação.

A mesma metodologia é aplicada na área das medidas educacionais: a pesquisa sobre itens de leitura de textos (Gorin, Embretson, 2006), empregados no *Graduate Record Examination – Verbal*, sugere a possibilidade de serem produzidos itens especificamente para medir o raciocínio verbal ou a proficiência em leitura. Os resultados indicam, entretanto, a necessidade de continuação dos estudos para esclarecer dúvidas remanescentes no argumento de validade.

O *sistema cognitivo* de planejamento de prova proposto por Embretson (Embretson, Gorin, 2001; Embretson, 2005) caracteriza-se por: a) ter em seu cerne a validação experimental da explicação das respostas dos examinandos a cada item; b) ter a validade das interpretações dos resultados obtidos pelos examinandos em uma prova avaliada previamente, durante a construção das questões a serem incluídas, em termos de uma explicação que tem raízes nas pesquisas da psicologia cognitiva. Tanto quanto o estoque atual de pesquisas na área da psicologia cognitiva permite, o *sistema cognitivo* poderá responder satisfatoriamente à advertência de Bejar (2002) de que as inovações tecnológicas nas medidas educacionais devem ter suporte no constructo que se deseja medir. O problema está na possibilidade de compatibilizar o desenvolvimento dos estudos da psicologia cognitiva com a vasta gama de necessidades e interesses das medidas educacionais.

VALIDADE DE ITENS GERADOS PELO COMPUTADOR

A pressão social para que sejam usadas novas tecnologias na avaliação educacional, principalmente em países europeus e na América do Norte, tem sido mais um estímulo para que os pesquisadores encarem uma variedade de problemas teóricos e práticos. Dada a comodidade oferecida pelo computador na aplicação de provas individuais, é natural – e até previsível – a preferência dos examinandos por essa modalidade, em vez da aplicação de uma prova única, a grande massa de indivíduos ao mesmo tempo, em um mesmo local. Em contrapartida, o interesse dos pesquisadores em testar os limites da potencialidade da tecnologia expressa-se, desde a década de 80, nos estudos a respeito da geração de itens pelo computador.

Como se trata de questões produzidas artesanalmente, a teoria e a metodologia da validação servem para provas aplicadas por meio de computador ou não. Problemas novos surgem quando se trata de questões geradas pelo computador para compor provas educacionais. Muito mais do ponto de vista das medidas educacionais do que do ponto de vista tecnológico, os problemas envolvidos na geração de questões de provas pelo computador ainda desafiam os especialistas (Bejar, 2002; Drasgow, Luecht, Bennet, 2006; Embretson, 2005).

No estado atual da arte, trata-se de desenvolver uma classe de itens – ou uma *estrutura* (Embretson, 2005), ou um *modelo de item* (Bejar et al., 2003) – com base na qual o computador deverá gerar variações dentro de regras especificadas no programa. Irvine (2002) e Kyllonen (2002) referem-se a características – chamadas de *radicais* – de questões de provas que controlam a dificuldade do item, e a outras características – denominadas *incidentais* – que não controlam a dificuldade. De modo geral, o objetivo é gerar itens cujas características psicométricas são automaticamente derivadas de princípios que orientam o planejamento da prova.

Quando os princípios que norteiam o desenvolvimento do item se fundamentam numa teoria cognitiva fortemente ancorada na pesquisa empírica, é possível – dentro de certos limites – propor um *modelo cognitivo* do item, predizer parâmetros de dificuldade, e explicar o desempenho dos examinandos em termos dos processos cognitivos postulados (Drasgow, Luecht, Bennet, 2006; Embretson, 2005). É o caso dos estudos, fortemente amparados na teoria e na pesquisa cognitiva, que Embretson (2002, 2005) vem realizando sobre itens de completação de figuras. Essa série de estudos parte da avaliação da validade de constructo em relação a cada item de completação de matrizes – do tipo utilizado nos testes de Matrizes Progressivas, de Raven – e exemplifica a geração de novos itens com

aplicação de um programa computacional específico. Para tanto, uma *estrutura formal* de cada item é definida – composta dos elementos que operacionalizam as variáveis cognitivas postuladas (no exemplo, são *indução de relações* e *descoberta de correspondência entre figuras*). Operacionalizadas, às variáveis que compõem o *modelo cognitivo* Embretson (2002) acrescenta as características das figuras apresentadas: sobreposição, fusão e distorção. Itens que têm a mesma *estrutura formal* são considerados equivalentes; e elementos da *estrutura* podem ser substituídos por outros, desde que a *estrutura* seja mantida. Uma vez quantificadas as variáveis representativas do *modelo cognitivo*, foram gerados automaticamente cinco itens para cada uma das 30 *estruturas* definidas, num total de 150 itens. Embretson (2005) aplicou, a uma amostra de adultos, 90 desses itens gerados pelo computador. Para estimar o impacto de cada variável do *modelo cognitivo* sobre a resposta do examinando, Embretson (2005) alterou um modelo da TRI, com dois parâmetros, introduzindo pesos correspondentes a cada uma dessas variáveis. Os que avaliam a dificuldade e a discriminação do item foram substituídos pela soma dos produtos desses pesos pelos valores atribuídos às respectivas variáveis. Depois de estimados os valores desses pesos, pode-se avaliar a dificuldade e a discriminação do item. Com os resultados da experiência, Embretson (2005) verifica que: a) os itens gerados pelo programa computacional refletem o impacto das mesmas variáveis introduzidas no *modelo cognitivo* que foram testadas em estudos anteriores (Embretson, 2002); b) o modelo psicométrico correspondente ao *modelo cognitivo* permite a previsão da dificuldade do item sem que este precise ser testado empiricamente.

Outras linhas de pesquisa perseguem o objetivo de construir modelos de itens com base teórica e empírica para preservar a medida do constructo ao serem gerados automaticamente novos itens – veja-se, por exemplo, Dennis et al. (2002), Kyllonen (2002) e Newstead et al. (2002). Entretanto, nem sempre é possível contar com uma sólida base de pesquisas cognitivas em face das necessidades dos sistemas de avaliação educacional que cobrem uma área enorme e diversificada de conhecimentos, e que se aplicam a objetivos variados. Nesses casos, alguns pesquisadores experimentam a viabilidade de processos de produção automática de itens de provas, mesmo sem contar com a base da teoria e da pesquisa cognitiva. É nesse sentido, por exemplo, que Bejar et al. (2003) sugerem a criação do que chamam de *modelos de itens*, que não são baseados em uma teoria cognitiva, mas desenvolvidos com o apoio de uma variedade de processos usados comumente na construção artesanal das provas educacionais.

Segundo Bejar (Bejar, 2002; Bejar et al., 2003) um *modelo de item* é definido como uma "classe", ou um tipo de item do qual possam ser geradas variações (*variantes*), que sejam itens equivalentes, ou seja, que todos mantenham as características psicométricas do modelo original. Na descrição do desenvolvimento das questões de um teste destinado ao exame de licenciamento profissional de arquitetos (Bejar, 2002), em que a validade de constructo é enfatizada, percebe-se que a lógica da aplicação do *modelo de item* para geração automática de questões de prova tem raízes na lógica da produção artesanal dessas questões. Nesse estudo, trata-se de questões complexas, que requerem conhecimentos e habilidades específicas, cuja análise demandou dos pesquisadores, além da base teórica propiciada pela literatura concernente, uma análise do trabalho do arquiteto. Os *modelos de itens* foram definidos, as definições e limitações para gerar as *variantes* de cada um deles foram estabelecidas, mas essas *variantes* foram produzidas artesanalmente. Já em outro estudo, sobre questões utilizadas no teste de GRE (*Graduate Record Examination General Test*), o objetivo é produzir itens, por um programa computacional específico, durante a aplicação da prova.

O estudo de Bejar et al. (2003) oferece oportunidade para se avaliar o problema da validade de constructo em relação aos itens gerados pelo computador, no caso específico de não se partir de uma teoria cognitiva fortemente ancorada na pesquisa empírica. Ao estudar a viabilidade do desenvolvimento de uma prova adaptada ao indivíduo, composta de questões produzidas pelo computador durante a aplicação do instrumento ao examinando, Bejar et al. (2003) experimentam itens cuja modelagem não tem apoio consistente numa teoria cognitiva. A proposta de um *modelo de item*, nesse caso, refere-se a uma classe de questão de prova da qual são geradas *variantes*; os parâmetros de dificuldade, de discriminação e de acerto casual devem ser iguais por todas as *variantes* – denominadas de itens *isomorfos* – de um mesmo *modelo de item*. Para esse estudo, foram selecionados 147 itens da parte quantitativa do GRE *General Test* dos quais foram derivados outros tantos *modelos de itens*; segundo regras pré-estabelecidas, partes de cada item original (por exemplo, certos números, ou palavras) podem ser substituídas pelo programa computacional para formar *variantes* do *modelo* derivado daquele item. Nesse estudo foram usadas as estimativas, obtidas em estudos anteriores, dos três parâmetros de um modelo da TRI, para os 147 itens originais que serviram de base para desenvolver os respectivos *modelos de itens*. Essas estimativas foram submetidas a um processo estatístico para compensação de possíveis desvios (Bejar et al., 2003), sendo os valores obtidos impostos a todas as *variantes* do respectivo *modelo de item*. Deste modo, durante a aplicação da

prova adaptada a cada indivíduo, o programa computacional usava esses valores para o cálculo da probabilidade de se obter uma resposta correta para determinado item, assim como para estimar o nível de habilidade θ do examinando ao qual se aplicava a prova. Ou seja, para cada valor de θ, da habilidade do examinando, o programa selecionava uma *variante* do *modelo do item*, calculava a probabilidade de uma resposta correta para essa *variante*, dados os respectivos valores de dificuldade, de discriminação e de acerto casual; tendo em vista a resposta do examinando, o programa a classificava como correta ou incorreta e renovava a estimativa do nível de habilidade θ; este ciclo recomeçava e repetia-se até o limite de 28 questões de prova.

Para efeito de estudo de validação, Bejar et al. (2003) comparam os resultados da prova gerada pelo programa computacional com aqueles obtidos no GRE, usado igualmente com adaptação ao examinando, e aplicado juntamente, na mesma ocasião, aos mesmos indivíduos. A correlação obtida entre os escores das duas provas foi de 0,87, que é semelhante à encontrada entre o teste e reteste do GRE, segundo os autores. Os resultados são considerados promissores, tanto do ponto de vista da viabilidade do processo de geração automática de itens durante a aplicação do teste como da correlação com o critério focalizado na pesquisa. Entretanto, como os *modelos de itens* não se apoiam especificamente numa teoria e em pesquisas cognitivas, fica prejudicada a proposta de manipulação de características do modelo de item e respectivas relações com processos cognitivos e com as características psicométricas dos itens gerados automaticamente. Em contrapartida, correlações com outras variáveis podem dar suporte a interpretações sobre diferenças individuais e, particularmente no caso de testes cujo objetivo é apenas servir à previsão do desempenho em relação a um critério bem delimitado, podem constituir uma base para o argumento de validade. De qualquer forma, estudos que focalizam possíveis desvios do *isomorfismo* continuam tentando explorar formas de corrigir seus efeitos sobre as características psicométricas dos itens gerados automaticamente (Graf et al., 2005).

Em resumo, na prática, as condições de construção e de aplicação de provas educacionais em grande escala somam-se ao desenvolvimento científico e tecnológico, e impulsionam o sistema para o uso do computador desde o planejamento da prova e da geração das questões. O *sistema cognitivo* proposto por Embretson (Embretson, Gorin, 2001; Embretson, 2005) é uma forma de responder ao problema da validade de constructo, no caso do emprego dos itens gerados pelo computador – modelam-se variáveis cognitivas que explicam o desempenho do examinando em cada item da prova, e planeja-se cada questão de prova de

modo a representar as variáveis cognitivas que explicam a dificuldade respectiva. Contudo, persiste a necessidade de pesquisas, na área da cognição, que cubram a enorme gama de constructos focalizados, principalmente nas provas que se relacionam aos currículos escolares. As soluções propostas no estudo de Bejar et al. (2003) representam passos importantes, especialmente do ponto de vista tecnológico. Quanto à validade, o suporte que vêm oferecendo, atualmente, baseia-se nos estudos correlacionais entre os escores totais da prova gerada pelo computador e medidas de critérios selecionados. Nesses casos, em que na elaboração de cada questão de prova não se conta com apoio sólido na teoria e nas pesquisas cognitivas, não se tem a mesma possibilidade de explicação para as respostas ao item, em termos dos processos cognitivos envolvidos. Fica até certo ponto prejudicado o uso das questões de prova no diagnóstico com a finalidade de adaptar os processos instrucionais ao aluno.

COMENTÁRIOS FINAIS

A teoria da validade evoluiu paulatinamente, e a validação da interpretação dos resultados obtidos pelos examinandos numa prova educacional tomou a direção de uma concepção análoga à de uma teoria científica. A interpretação desses resultados é tratada como uma hipótese que tem raízes na teoria cognitiva, e que depende, para sua validação, das evidências coletadas dentro do contexto em que a prova é desenvolvida e usada. A validação é encarada como uma investigação de caráter científico, tanto no que concerne à pesquisa de processos cognitivos que expliquem o comportamento dos examinandos diante das questões de prova como no que diz respeito ao estudo das relações entre os resultados obtidos na prova e outras variáveis. Nesse amadurecimento da concepção de validade, o processo de validação passou a ser concebido em relação à metodologia das pesquisas da psicologia da cognição; tende-se a conceber o processo de validação em termos mais próximos da metodologia da pesquisa experimental. Um amadurecimento que, todavia, pode não ter atingido sua plenitude, mas que transparece nos métodos de algumas pesquisas de validação das últimas décadas.

Embora essa seja uma conceituação de validade que se firma entre os especialistas, não tem uma penetração generalizada entre os usuários, e coexiste com uma idéia de validação limitada a processos correlacionais. Em que pesem essas incongruências, o argumento com que se avalia o grau de validade de uma interpretação dos resultados de uma prova pode tomar várias formas e usar vários critérios de julgamento, mas de modo geral

focaliza a lógica dessa interpretação e o rigor dos processos com que se buscam as evidências empíricas em que se fundamenta.

Se o panorama é animador do ponto de vista conceitual e das pesquisas de validação, na prática são comuns os desvios flagrantes quanto às exigências mínimas feitas pelos especialistas no que concerne às normas para garantir ao usuário a qualidade das provas educacionais. Aparentemente, trata-se de um fenômeno de ordem generalizada, não adstrito a alguns países ou culturas; é possível que a falta de divulgação de informação a respeito das medidas educacionais seja um dos fatores que contribuem para ainda predominar, em certos meios, a conceituação popular de validade como *validade aparente*.

Um dos aspectos da evolução da pesquisa de validação tem especial relevância não só em relação à construção das provas educacionais, mas também a aplicações no ensino e na aprendizagem: o objeto do estudo de validação passa a ser cada questão de prova, em lugar de serem tomados como base os resultados no conjunto da prova. Não se trata apenas de estudar as relações de cada questão com as demais componentes da prova. Trata-se de estudar cada questão no contexto de uma teoria cognitiva, numa pesquisa que assume caráter experimental, com fins de relacionar o comportamento do examinando, em face da questão da prova, a processos cognitivos que o expliquem. Esta é uma concepção que leva a uma maior aproximação entre as medidas educacionais e a individualização dos processos instrucionais, visto que oferece uma informação mais detalhada sobre as diferenças entre os examinandos quanto a processos cognitivos envolvidos nas respostas à questão da prova.

A integração da metodologia da pesquisa experimental, dos estudos da psicologia cognitiva, e das ciências da computação aponta para um futuro promissor, tanto em relação à investigação da validade de constructo como à construção das provas educacionais. Na prática, porém, a sedução que as novas tecnologias exercem não pode ser subestimada; os estudos sobre as vantagens e desvantagens de diferentes abordagens – principalmente os que se referem ao emprego do computador na geração de questões de provas educacionais – mostram que há diferenças entre os tipos de informação obtida, conforme seja a metodologia empregada ao se construir a prova. A argumentação em favor da validação de constructo, introduzida desde o planejamento e a elaboração de cada questão a ser incluída na prova, parece ter o apoio da comunidade de especialistas; entretanto, este é um caminho reconhecidamente mais longo, mais difícil, e que envolve maiores custos. A tradição de construção artesanal de provas leva usuários e instituições financiadoras de projetos a encararem a elaboração desses instrumentos de medida como um processo

relativamente rápido – comumente, o critério da *validade aparente* ainda domina em meios não especializados. Isso dificulta a passagem para a concepção da construção de provas educacionais como uma tecnologia baseada na teoria e na pesquisa científica. O problema não se restringe, porém, ao estudo de características técnicas de um instrumento de medida; a prova educacional é um produto elaborado e entregue por especialistas ao usuário. Como tantos outros, é um produto cujo uso pode implicar tomada de decisões sobre indivíduos, ou sobre grupos de pessoas, ou sobre políticas educacionais. A validação de questões de uma prova envolve, portanto, problemas concernentes a relações entre especialistas e usuários, entre especialistas e a sociedade em geral. Não se trata de um problema técnico apenas, mas de um problema que envolve a responsabilidade social dos que trabalham na construção e no uso da prova educacional.

REFERÊNCIAS BIBLIOGRÁFICAS

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION. American Psychological Association. National Council on Measurement in Education. Standards for Educational and Psychological Tests and Manuals,1966. In: JACKSON, D. N.; MESSICK, S. (eds.) *Problems in Human Assessment*. N.Y.: McGraw-Hill,1967. p.169-189.

_____. American Psychological Association. National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, D.C.: AERA,1999.

AYALA, C. C. et al. On Science achievement from the perspective of different types of tests: a multidimensional approach to achievement validation. *CSE Technical Report 572*, Los Angeles: University of California, July 2002. Disponível em: <www.cse.ucla.edu/CRESST/Reports/Report 5722002.pdf > Acesso em: 23 mar. 2004.

BALBINOTTI, M. A. A. et al. Proposição e validação de um instrumento para avaliação do treino técnico-desportivo de jovens tenistas. *Revista Brasileira de Educação Física e Esporte*, v. 18, n.3, p.213-226, jul./set. 2004. Disponível em: <www.usp.br/eef/rbefe/v18n32004/v18p213.pdf> Acesso em: 5 maio 2006.

BECHTOLD, Harold P. Construct validity: a critique. American Psychologist, 1959, n. 14, p.619-629. In: JACKSON, D. N.; MESSICK, S. (eds.) *Problems in Human Assessment*. N.Y.: McGraw-Hill, 1967, p.133-146.

BEJAR, Isaac I. Generative testing: from conception to implementation. In: IRVINE, S. H.; KYLLONEN, P. C. (eds.) *Item generation for test development*. Mahwah, N. J.: Lawrence Erlbaum, 2002, p.199-217.

BEJAR, I. I. et al. A Feasibility study of on-the-fly item generation in adaptive testing. *Journal of technology, learning and assessment*, 2003, v.2, n. 3. Disponível em: <http://www.jtla.org> Acesso em: 24 jan. 2007.

BENNETT, Randy E.; ROCK, Donald A. Examining the validity of a computer-based generating–explanations test in an operational setting. *ETS Research Report*. Princeton, N. J.: Educational Testing Service, July, 1998.

BENNETT, Randy E.; BEJAR, Isaac I. Validity and automated scoring: it´s not only the scoring. *ETS Research Report*. Princeton, N. J.: Educational Testing Service, 1997.

BESSA, Nicia M. *Teste de desenvolvimento educacional*; Relatório Técnico. Rio de Janeiro: Fundação Getúlio Vargas; ISOP; CETPP, 1971.

_____. Aspectos metodológicos do processo de seleção para o ingresso nas universidades. *Educação e Seleção*, n. 2, p. 39-56, dez. 1980.

BESSA, Nicia M.; METTEL, Thereza L. Validade de três testes do DAT (Forma B). *Arquivos Brasileiros de Psicotécnica*, v. 14, n. 3, p. 5-15, jul./set. 1965.

BOARS – Board of Admission and Relations with Schools of the University of California´s Academic Senate. The Use of Admissions Tests by the University of California. *Research Report*, California, 2002. Disponível em: <www.universitycalifornia.edu/senate/committees/boars/ar/boars01-02 ar.pdf> Acesso em: 6 maio 2006.

BOLDT, Robert F. Generalization of SAT Validity Across Colleges. *College Board Report*, n. 86-3/ETS N.Y.: College Entrance Examination Board, 1986.

CAMILLI, Gregory. Test Fairness. In: BRENNAN, R. L. (ed.) Educational Measurement. Connecticut: Praeger Publishers, 2006. p. 221-256.

CATTELL, Raymond B. Validity and reliability: a proposed more basic set of concepts. Journal of Educational. Psychology, 1964, n. 55, p. 1-2. In: MEHRENS, W. A.; EBEL, R. L. (eds.) *Principles of Educational and Psychological Measurement*. Chicago, Ill.: Rand McNally, 1967, p. 337-365.

COLE, Nancy S.; MOSS, Pamela A. Bias in test use. In: LINN, R. L. (ed.) *Educational Measurement*. 3rd.ed. Phoenix, AZ: Orix Press, 1993, p. 201-219.

COLE, Nancy S.; ZIEKY, Michael J. The New Faces of Fairness. *Journal of Educational Measurement*, v. 38, n. 4, p. 369-382, 2001.

COOK, Linda L.; SCHMIDT-CASCALLAR, Alicia P.; BROWN, Catherine. Adaptive achievement and aptitude tests: a review of methodological issues. In: HAMBLETON, R. K.; MERENDA, P. F.; SPIELBERGER, C. D. (eds.) *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, N.J.: Lawrence Erlbaum, 2005. p.171-192.

CRONBACH, Lee J. The Two disciplines of scientific psychology. *American Psychologist*, n. 12, p. 671-684, 1957.

_____. Test validation. In: THORNDIKE, R. L. (ed.) *Educational Measurement.* 2nd.ed. Washington, D.C.: American Council on Education, 1971. p. 443-507.

CRONBACH, Lee J.; MEEHL, Paul E. Construct validity in psychological tests. Psychological Bulletin, 1955, n. 52, p. 281-302. In: JACKSON, D. N.; MESSICK, S. (eds.) *Problems in Human Assessment*. N.Y.: McGraw-Hill, 1967. p.57-77.

CRONBACH, Lee J. et al. *The Dependability of Behavioral Measurements*: theory of generalizability for scores and profiles. N.Y.: John Wiley, 1972.

CURETON, Edward E. Validity. In: LINDQUIST, E. F. (ed.) *Educational Measurement*. Washington, D.C.: American Council on Education, 1951. p. 621-684.

DENNIS, I. et al. Approaches to modeling item: generative tests. In: IRVINE, S. H.; KYLLONEN, P. C. (eds.) *Item Generation for Test Development*. Mahwah, NJ: Lawrence Erlbaum, 2002. p.53-71.

DRASGOW, F.; LUECHT, R. M.; BENNETT, R. E. Technology and Testing. In: BRENNAN, R. L. (ed.) *Educational Measurement*. Westport, CT.: American Council on Education/Praeger, 2006. p. 471-515.

EBEL, Robert L. The Social consequences of educational testing. ETS Invitational Conference, 1963. In: ANASTASI, A. (ed.) *Testing Problems in Perspective*. Washington, D.C.: American Council on Education, 1948-1966. p. 18-28.

EMBRETSON, Susan E. Generating abstract reasoning items with cognitive theory. In: IRVINE, S. H.; KYLLONEN, P. C. (eds.) *Item Generation for Test Development*. Mahwah, N. J.: Lawrence Erlbaum, 2002. p. 219-260.

_____. Measuring human intelligence with artificial intelligence. In: STERNBERG, R. J.; PRETZ, J. E. (eds.) *Cognition and Intelligence*. Cambridge, UK: Cambridge University Press, 2005. p.251-267.

EMBRETSON, Susan E.; GORIN, Johanna S. Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*. Winter 2001, v. 38, n. 4, p. 343-368.

FERRARA, S. *Examining test score validity by examining item construct validity:* preliminary analysis of evidence of the alignment of targeted and observed content, skills and cognitive processes in a middle school science assessment. Paper presented at the Annual Meeting of the American Educational Research Association, 2004. Disponível em: <www.air.org/News/default.aspex> Acesso em: 28 ago. 2006.

FERRARA, S.; DeMAURO, G. E. Standardized assessment of individual achievement in K-12. In: BRENNAN, R. L. (ed.) *Educational Measurement*. Westport, CT.: American Council on Education/Praeger, 2006. p. 579-621.

GORIN, Joanna S.; EMBRETSON, Susan E. Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, v. 30, n.5, p. 394-411, September 2006.

GRAF, E. A. et al. Psychometric and cognitive analysis as a basis for the design and revision of quantitative item models. *ETS Research Report* (RR-05-25). Princeton, NJ: Educational Testing Service, 2005. Disponível em: <http://www.ets.org/Research/researcher/RR-05-25.html> Acesso em: 17 ago. 2006.

HAMBLETON, Ronald K. Principles and selected applications of item response theory. In: LINN, R. L. (ed.) *Educational Measurement*. Phoenix, AZ: American Council on Education/Orix Press, 1993. p. 147-200.

HAMBLETON, Ronald K. Issues, designs and technical guidelines for adapting tests into multiple languages and cultures. In: HAMBLETON, R. K.; MERENDA, P. F.; SPIELBERGER, C. D. (eds.) *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, NJ: Lawrence Erlbaum, 2005. p.3-38.

IRVINE, Sidney H. The Foundations of item generation for mass testing. In: IRVINE, S. H.; KYLLONEN, P. C. (eds.) *Item Generation for Test Development*. Mahwah, NJ: Lawrence Erlbaum, 2002. p.3-4.

KANE, Michael T. Current concerns in validity theory. *Journal of Educational Measurement*. v. 38, n. 4, p.319-342, winter 2001.

_____. Validation. In: BRENNAN, R. L. (ed.) *Educational Measurement*. Westport, CT.: American Council on Education/Praeger, 2006. p.17-64.

KYLLONEN, Patrick C. Item generation for repeated testing of human performance. In: IRVINE, S. H.; KYLLONEN, P. C. (eds.) *Item Generation for test Development*. Mahwah, NJ: Lawrence Erlbaum, 2002. p. 251-275.

LAWRENCE, Ida; FEIGENBAUM, Miriam. Linking scores for computer-adaptive and paper-and-pencil administration of the SAT. *Research Report*. Princeton, N. J.: Educational Testing Service,1997.

LINDQUIST, E. F. Preliminary considerations in objective test construction. In: _____. (ed.) *Educational Measurement*. Washington, DC: American Council on Education,1951. p.119-158.

LORD, Frederic M. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum, 1980.

LORD, Frederic M.; NOVICK, Melvin R. *Statistical Theories of Mental Test Scores*. Reading, Mass: Addison-Wesley,1968.

MESSICK, Samuel. Validity. In: LINN, R. L. (ed.) *Educational Measurement*. 3rd.ed. Phoenix, AZ: American Council on Education/Orix Press, 1993. p.13-103.

_____. *Foundations of Validity*: meaning and consequences in psychological assessment. *European Journal of Psychological Assessment*, v. 10, n.1, p.1-9, 1994.

MISLEVY, Robert J. Cognitive psychology and educational assessment. In: BRENNAN, R. L. (ed.) *Educational Measurement*. Westport, CT.: American Council on Education/Praeger, 2006. p. 257-305.

MISLEVY, R. J.; STEINBERG, L. S.; ALMOND, R. J. On the roles of task model variables in assessment design. In: IRVINE, S.; KYLLONEN, P. C. (eds.) *Item Generation for Test Development*. Mahwah, NJ: Lawrence Erlbaum, 2002. p. 97-128.

MONTEIRO, Kilda. Estudo com o DAT (Forma A). *Arquivos Brasileiros de Psicotécnica*, v. 16, n. 4, p. 47-54, out./dez. 1964.

MOSIER, Charles I. A Critical examination of the concepts of face validity. educational and psychological measurement, n.7, p.191-205, 1947. In: MEHRENS, W. A.; EBEL, R. L. (eds.) *Principles of Educational and Psychological Measurement*. Chicago, ILL: Rand McNally, 1967. p. 207-218.

NEWSTEAD, S. et al. Using the psychology of reasoning to predict the difficulty of analytical reasoning problems. In: IRVINE, S. H.; KYLLONEN, P. C. (eds.) *Item Generation for Test Development*. Mahwah, N. J.: Lawrence Erlbaum, 2002. p. 35-51.

PELLEGRINO, James W.; GLASER, Robert. Components of inductive reasoning. In: SNOW, R. E.; FEDERICO, P. A.; MONTAGUE, W. E. (eds.) *Aptitude, Learning and Instruction*: cognitive process analyses of aptitude. v.1. Hillsdale, N. J.: Lawrence Erlbaum, 1980. p.177-217.

PELLEGRINO, J. W.; CHUDOWSKY, N.; GLASER, R. (eds.) *Knowing what Students Know*. Committee on the Foundations of Assessment, National Research Council. Washington, DC: National Academy Press, 2001.

PORTER, Andrew C.; SMITHSON, John L. Defining, developing, and using curriculum indicators. *Research Report Series*. Consortium for policy research in education, University of Pennsylvania, December, 2001. Disponível em: <www.cpre.org/Publications/rr48.pdf> Acesso em: 17 maio 2006.

PRIMI, R. et al. Competências e habilidades cognitivas: diferentes definições dos mesmos constructos. *Psicologia*: teoria e pesquisa, v.17, n.2, 2001. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext &pid=S0102-37722001000200007&Ing=en&nrm=iso> Acesso em: 6 maio 2007.

RODRIGUES, Margarida, M. M. Proposta de análise de itens das provas do Saeb sob a perspectiva pedagógica e a psicométrica. *Estudos em Avaliação Educacional*, v. 17, n. 34, p. 43-77, maio/ago. 2006.

SANTOS, A. A. A. et al. O Teste de CLOZE na avaliação da compreensão em leitura. *Psicologia*: reflexão e crítica, v.15, n.3, p.549-560, 2002.

SILVEIRA, Fernando L.; PINNENT, Carlos E. A Questão da redação no concurso vestibular à universidade: validade e poder decisório. *Estudos em Avaliação Educacional*, n. 24, p.147-164, jul./dez. 2001.

SIRECI, S. G.; PATSULA, L.; HAMBLETON, R. K. Statistical methods for identifying flaws in the test adaptation process. In: HAMBLETON, R. K.; MERENDA, P. F.; SPIELBERGER, C. D. *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, NJ: Lawrence Erlbaum, 2005, p. 93-115.

SOARES, T. M.; GENOVEZ, S. F. de M.; GALVÃO, A. F. Análise do comportamento diferencial dos itens de Geografia: estudo da 4ª série avaliada no Proeb/Simave 2001. *Estudos em Avaliação Educacional*, v.16, n. 32, p. 81-102, jul./dez. 2005.

VAN DE VIJVER, Fons J. R.; POORTINGA, Ype H. Conceptual and methodological issues in adapting tests. In: HAMBLETON, R. K.; MERENDA, P. F.; SPIELBERGER, C. D. *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, NJ: Lawrence Erlbaum, 2005. p. 39-63.

WEBB, Norman L. Criteria for alignment of expectations and assessments in mathematics and science education. *Research Monograph nº 8*. Council of Chief State Officers. Washington, DC, 1997. Disponível em: <www.wcer.wisc.edu/addingvalue/Related%20Bibliography/Articles/WebbAll.doc > Acesso em: 18 maio 2006.

ZIEKY, Michael J. Ensuring the fairness of licensing tests. *CLEAR Exam Review*, v. 12, n.1, p. 20-26, winter 2002. Disponível em: <http://www.ets.org/Media/Research/pdf/FAIRNESS-LICENSING.pdf.> Acesso em: 12 mar. 2006.

Recebido em: maio 2007

Aprovado para publicação em: julho 2007