VALIDADE E CONSEQÜÊNCIAS SOCIAIS DAS AVALIAÇÕES EM CONTEXTOS DE ENSINO DE LÍNGUAS*

Matilde V. R. Scaramucci
matilde@unicamp.br
Universidade Estadual de Campinas (Brasil)

Resumo: O conceito de validade é tão central à avaliação que seria praticamente impossível pesquisá-la ou praticá-la sem considerá-lo. Motivado, em grande parte, pelos desenvolvimentos recentes sobre avaliação de desempenho, que pressupõe a avaliação do uso da linguagem em contextos sociais, entretanto, esse conceito passou, nos últimos anos, por uma extensa revisão tanto em sentido como abrangência, a ponto de tornar-se um dos conceitos mais polêmicos e discutidos na área de avaliação no momento ou, para Chapelle (1999), uma das áreas mais interessantes e importantes dentro da Linguística Aplicada. Discuto, neste trabalho, o conceito moderno ou expandido de validade/validação, que considera não apenas as bases evidenciais, mas também as consequências sociais/efeitos retroativos no processo (Messick 1989).

Palavras-chave: validade, validação, testes de línguas, consequências sociais, efeitos retroativos

Abstract: The concept of validity is so central to testing and assessment that it would be practically impossible to research it or practice it without taking this concept into consideration. Motivated largely by the recent developments of performance assessment, which assesses language use in social contexts, however, this concept has been through extensive review in meaning and comprehensiveness to the point of becoming one of the most discussed and controversial concepts in the field of language assessment nowadays, or, as pointed out by Chapelle (1999), one of the most interesting and important aspects in Applied Linguistics. My aim in this article is to discuss the modern or expanded concept of validity/validation (as opposed to the traditional concept), which takes into consideration not only the evidencial basis but also the social consequences/washback in validating a test (Messick 1989).

Keywords: validity, validation, language tests, social consequences, washback.

* As reflexões sobre o tema em questão foram conduzidas durante o estágio de pós-doutorado realizado na Universidade de Melbourne, Austrália. Meus agradecimentos à FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) pelos recursos para a realização do estágio.

1 - Introdução

Grandes desenvolvimentos têm caracterizado a pesquisa em avaliação de línguas nos últimos anos, com contribuições importantes para a área de Linguística Aplicada. Uma das questões que tem mobilizado grande parte dos teóricos da avaliação é a busca de um entendimento mais preciso do que se convencionou chamar de testes ou avaliação de desempenho. As discussões motivadas pela complexidade dos fatores envolvidos nesse conceito foram tão intensas a ponto de determinarem reformulações não apenas das visões de linguagem e de proficiência subjacentes às avaliações tradicionais, mas, principalmente, dos conceitos utilizados como parâmetros para avaliar a qualidade/aceitabilidade desses instrumentos.

Como consequência, pudemos presenciar, a partir dos anos 90, uma verdadeira revolução no conceito de validade/validação que, entretanto, não ocorreu repentinamente, uma vez que suas bases foram estabelecidas de forma gradativa, a partir de contribuições de teóricos distintos durante os anos 80. De acordo com Chapelle (1999), esses elementos foram a noção de *afeto* (a medida em que um teste causa ansiedade indevida, levantado por Madsen (1983); efeito retroativo (efeito que o teste exerce nos processos de ensino e aprendizagem, levantado por Hughes (1989/1994)[1] como uma quarta qualidade de um teste, além das três validades; e ética, tratada nas questões "por que avaliar", juntamente com "o quê" e "como avaliar", levantada por Canale (1987).

Algumas publicações, ocorridas a partir de 1985, no entanto, foram decisivas para o que se poderia chamar uma nova teoria de validade: o código oficial de prática profissional nos Estados Unidos (*AERA/APA/NCME*[2] *Standards for Educacional and Psychological Testing*); e os trabalhos de Cherryholmes (1988), Moss (1992) e Wiggins (1993) por questionarem as bases filosóficas da visão tradicional, retomadas no artigo seminal de Messick (1989) intitulado "Validade", publicado na terceira edição do *Handbook of Educational Measurement*.

2 - Alguns conceitos básicos

Não poderia apresentar o que se convencionou chamar na literatura de conceito de validade moderno ou expandido sem antes abordar o conceito de validade tradicional. Importantes também se fazem, para o entendimento dessa problemática, de pelo menos mais dois outros conceitos: o de avaliação de desempenho e o de efeito retroativo/impacto das avaliações no ensino/aprendizagem. Abordo, em primeiro lugar, o conceito tradicional de validade, em seguida o de avaliação de desempenho, seguido pelo de impacto/efeito retroativo; por fim, retomo o conceito de validade, já expandido pelas contribuições dessas duas áreas.

[1] Menção a esse conceito já pode ser encontrada em Morrow (1986).

[2] *American Educational Research Association, American Psychological Association e The National Council on Measurement in Education, respectivamente.*

2.1 - Validade/validação: visão tradicional

Tradicionalmente, validade tem sido definida como uma característica ou qualidade de um teste, um critério para sua aceitabilidade. Se examinarmos as definições em livros que trazem noções básicas de avaliação publicadas nos anos 80, encontraremos definições como a de Hughes (1989/1994): um teste é válido se mede precisamente aquilo que deve medir. Essa visão, entretanto, também pode ser observada naqueles publicados nos anos 90, como Alderson, Clapham & Wall (1995), que retoma a definição de Henning (1987:96):

> Validade em geral refere-se à adequação de um teste ou de algum de seus componentes como uma medida do que esse teste deve medir. Um teste é válido na medida em que mede o que deve medir. Assim, o termo válido, quando usado para descrever um teste, deve geralmente vir acompanhado pela preposição "para". Qualquer teste, dessa forma, pode ser válido para alguns propósitos, mas não para outros. (Henning 1987 apud Alderson et al. pg 89)[3].

Além de enfatizar a validade como uma característica do teste, Henning (1987) reconhece o fato de não ser um atributo de tudo ou nada; mais adequado para explicar o conceito seria um contínuo com graus de validade que devem ser julgados em relação ao propósito do teste.

Frequentemente, a validade tem sido abordada em relação à confiabilidade, mas, algumas vezes, também à praticidade, ambas vistas como outras qualidades de um teste. Um teste não pode ser válido sem antes ser confiável (consistente e estável), uma vez que para ser válido necessita avaliar com precisão e de forma consistente. Se uma prova é corrigida por dois corretores e as notas obtidas são completamente distintas (10 e 0), por exemplo, qual dos resultados devemos tomar como evidência daquilo que pretendemos avaliar?

Por outro lado, um teste confiável pode não ser válido, entretanto. Um teste de produção escrita em língua estrangeira que solicita aos candidatos escreverem a tradução de 500 palavras em sua língua materna (Hughes 1989/1994) pode ser considerado um teste confiável, mas está longe de ser válido, na medida em que escrever em língua estrangeira é muito mais do que apenas traduzir palavras. Um teste de desempenho de produção escrita que, por outro lado, solicita aos candidatos redigirem um texto pode ser um teste válido, embora não necessariamente confiável, se não forem estabelecidos critérios claros para a correção, se os corretores não forem treinados para a tarefa, e assim por diante. Dessa forma, um aumento de validade geralmente leva a uma diminuição de confiabilidade e vice-versa, revelando a tensão existente entre os dois parâmetros.

[3] Validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure. A test is said to be valid to the extent that it measures what is supposed to measure. It follows that the term valid when used to describe a test should usually be accompanied by the preposition for. Any test then may be valid for some purposes, but not for others (citação original).

A praticidade é ainda outra característica discutida juntamente com validade e confiabilidade: diminui com um aumento de validade e aumenta com uma diminuição de validade. O inverso ocorre em relação à confiabilidade. Assim, podemos dizer que é inversamente proporcional à validade e diretamente proporcional à confiabilidade, uma vez que validade e confiabilidade estão em tensão. Retomando o exemplo de Hughes (1989/1994) acima, podemos dizer que o teste de tradução de palavras é muito mais prático para ser aplicado e corrigido do que o teste em que o candidato é solicitado a escrever um texto.

Embora vários tipos de validade têm sido, tradicionalmente, reconhecidos na literatura, na medida em que essa característica pode ser estabelecida através de diferentes métodos, não parece haver concordância quanto aos nomes e definições dados a esses tipos. Hughes (1989/1994), por exemplo, refere-se à validade d*e construto, de conteúdo, relacionada a critério* (que pode ser *preditiva* ou *paralela*) e *de face.*

Alderson *et al.* (1995), por sua vez, embora, seguindo Thorndike e Hagen (1986), saliente três tipos principais de validade – *racional, empírica e de construto* – prefere usar os termos *interna*, *externa* e *de construto*. Como o nome diz, *interna* tem a ver com os estudos que analisam o conteúdo do teste e seu efeito percebido, enquanto externa relaciona-se aos estudos que comparam as notas obtidas com medidas externas da competência/capacidade avaliada e, portanto, são correlacionais. Dentro do conceito de validade interna são discutidas as *validade de face, de conteúdo e de resposta* e, no de validade *externa*, os conceitos *de paralela* e *preditiva*. Um terceiro tipo ainda seria a *validade de construto*, que Alderson *et al.* (1995) tratam separadamente, por ser, segundo os autores, a mais complexa de explicar e uma espécie de termo superordenado, para o qual as validades externa e interna contribuem.

Uma definição breve de validade de construto fornecida por Gronlund (1985:58 apud Alderson *et al.* (1995:183) seria "a medida em que o desempenho em um teste pode ser interpretado como uma medida significativa de uma determinada característica ou qualidade". Uma definição mais completa é fornecida por Ebel e Frisbie (1991:108)[4]:

> O termo *construto* refere-se a um construto psicológico, uma conceitualização teórica sobre um aspecto do comportamento humano que não pode ser medida ou observada diretamente. Exemplos de construtos são inteligência, motivação para o rendimento, ansiedade, rendimento, atitude, dominância e compreensão em leitura. Validação de construto é o processo de coleta de evidência para dar apoio ao argumento de que um teste realmente mede o construto psicológico que os elaboradores querem que meça. O objetivo, nesse caso, é determinar o significado os escores ou notas do teste para garantir que eles signifiquem o que o especialista esperava que significassem.

[4] The term *construct* refers to a psychological construct, a theoretical conceptualisation about an aspect of human behaviour that cannot be measured or observed directly. Examples of constructs are intelligence, achievement motivation, anxiety, achievement, attitude, dominance, and reading comprehension. Construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend it to measure. The goal is to determine the meaning of scores from the test, to assure that the scores mean what we expect them to mean (citação original).

Dessa forma, a validação de construto pode ser vista como uma atividade de pesquisa, através da qual teorias são testadas e confirmadas, modificadas ou abandonadas.

2.2 - Avaliação de desempenho

A busca de um entendimento mais preciso do que se convencionou chamar de *avaliação de desempenho (performance assessment)* tem mobilizado grande parte dos teóricos da avaliação nos últimos anos. Fornecer concepções ou construtos que descrevam adequadamente a linguagem e o que significa dominá-la, ou ser proficiente tem sido um grande desafio (vide Shohamy 1995, para revisão e contribuições desses construtos).

O termo, que não teve origem na área de Linguística Aplicada, mas em contextos de educação geral e vocacional, tem-se referido à avaliação que pressupõe a demonstração direta da proficiência[5] almejada ou das capacidades adquiridas, em vez de limitar-se a avaliar indiretamente essa proficiência através de instrumentos que focalizam itens isolados de gramática. Testes de desempenho oferecem ao avaliando a oportunidade de produção espontânea e desempenho real para operar a língua em situações comunicativas autênticas (Morrow 1977), envolvendo-o em um ato de comunicação, seja oral ou escrito. O que se pretende avaliar, nesse caso, é a capacidade do avaliando de agir no mundo através da linguagem, num ato que, embora conjunto, exige a coordenação de ações individuais (Clark 2002). Nesse sentido mais amplo, essa capacidade é caracterizada por apresentar um requisito de desempenho que pressupõe o uso integrado de habilidades linguísticas e pragmáticas. Assim, não se restringe apenas à comunicação oral, mas envolve também a compreensão oral e escrita, e a produção escrita.

A avaliação de desempenho pressupõe, portanto, que a melhor maneira de avaliarmos se alguém é proficiente é colocá-lo em situação em que ele possa demonstrar diretamente essa proficiência. Em outras palavras, se o que desejamos é saber se sabe escrever, a melhor maneira é solicitarmos que escreva um texto; se é capaz de interagir em situações reais, simularmos situações reais de interação e fazê-lo desempenhar-se nessas situações.

Avaliações de desempenho contrapõem-se às de conhecimento, na medida em que, no primeiro caso, a proficiência é determinada com base no conhecimento sobre a língua sem a necessidade de prová-la em situações de uso, enquanto que, no segundo, é imprescindível a demonstração desses conhecimentos. Enquanto testes de conhecimento especificam, em termos exatos, os elementos que deverão fazer parte da proficiência, os testes de desempenho definem proficiência a partir do comportamento esperado em situações de uso. Isso quer dizer que testes de conhecimento focalizam a precisão da gramática e do vocabulário, ao passo que testes de desempenho levam em consideração o uso desses elementos dentro de uma atividade comunicativa e, portanto, consideram como critério "dizer a coisa

[5] Para uma discussão mais aprofundada do conceito de proficiência, vide Scaramucci 1999a.

certa na hora certa para a pessoa certa". Enquanto testes de conhecimento são geralmente vistos como "confiáveis", testes de desempenho são frequentemente vistos como subjetivos, não confiáveis e não precisos, principalmente quando se trata da oralidade (Shohamy 1995).

Essa modalidade de avaliação tem sido usada não apenas em situações de avaliação de rendimento, internas aos processos de ensino/aprendizagem, mas também naquelas de proficiência, que pressupõem uso futuro da língua. A elaboração de projetos, produção de textos, *portfolios*, simulações de problemas reais, entrevistas, debates, atividades de simulação em geral são alguns exemplos de avaliações de desempenho, alguns mais usados em avaliações de rendimento e outros de proficiência.

Vale lembrar que avaliações em geral são baseadas em inferências sobre um determinado critério, visto como o conjunto de comportamentos que se deseja avaliar. Esses comportamentos são subseqüentes a um teste e, portanto, não observáveis. A única maneira de torná-los observáveis é caracterizá-los para que possam ser simulados ou representados, sempre de forma amostral, na elaboração do instrumento. Os dados de desempenho observados a partir da aplicação do teste serão usados para fazermos inferências sobre o critério, permitindo observar o que antes não era observável (McNamara 1996). É importante, portanto, uma distinção clara entre o critério, ou seja, o comportamento comunicativo na situação alvo que se quer avaliar e o teste ou instrumento para avaliá-lo. Nos testes de desempenho, diferentemente dos testes tradicionais, a situação que serve como critério é simulada em um grau muito maior. Apesar de realistas, entretanto, as situações de avaliação não serão reais, mas sempre situações de avaliação, que somente poderão ser consideradas reais pelo fato de ocorrerem na vida real. É importante ressaltar, entretanto, que mesmo quando testes simulam comportamentos do mundo real – leitura de um jornal, simulação de uma conversa com um paciente, assistir uma palestra – salienta McNamara (2000), não são os desempenhos em si que são importantes, mas sim, as informações que fornecem em relação ao desempenho do avaliado em tarefas semelhantes ou relacionadas àquelas da vida real.

Avaliações de desempenho, portanto, procuram mostrar que, pelo fato de a linguagem ser essencialmente social, essa dimensão não pode ser ignorada na avaliação, mesmo que sua consideração leve a uma complexidade maior dos processos avaliativos (vide Scaramucci 2005 e Schlatter, Garcez & Scaramucci 2004, para variáveis relacionadas à correção, grades e corretor; vide também Sakamori 2006, para questões relativas ao comportamento dos entrevistadores em exames de desempenho).

2.3 - Efeito retroativo e impactos/conseqüências sociais

Muito popular em Educação e mais especificamente em Linguística Aplicada há anos, o conceito de *efeito retroativo (washback ou backwash)* tem sido usado para

referir-se ao impacto ou influência[6] que exames externos – principalmente aqueles de alta-relevância, tais como vestibulares e alguns testes de proficiência –, assim como a avaliação de rendimento, que ocorre internamente nos processos de ensino/aprendizagem[7] podem exercer, potencialmente, no ensino, na aprendizagem, no currículo, na elaboração de materiais didáticos e nas atitudes das pessoas envolvidas – alunos, professores, escola (vide especialmente Alderson & Wall 1993; Watanabe 1996; Alderson & Hamp-Lyons 1996; Bailey 1996; Cheng 1999; Wall 2000; Cheng, Watanabe & Curtis 2004. No Brasil, Scaramucci, 1998a, 1998b, 1999a, 1999b, 1999c, 1999d, 2000/2001, 2002, 2003, 2004; Avelar 2001; Pessoa 2002; Bartholomeu 2002; Souza 2002; Correia 2003; Lanzoni 2004; Retorta 2007).

Mal compreendido e pouco pesquisado, visto até os anos 90 de forma determinista – um bom exame terá um efeito benéfico, um mau exame um efeito maléfico[8] – o fenômeno[9] teve sua compreensão ampliada pelos estudos recentes, que mostraram não apenas sua complexidade, mas, sobretudo, a ingenuidade da visão determinista predominante. Outras forças presentes na sociedade e na escola – entre as quais destaca-se a formação do professor "ao de – interagem" com as características do exame na determinação de seu impacto. Um mesmo exame pode influenciar pessoas de formas distintas, dependendo de características pessoais, da natureza das inovações propostas pelos exames e das estratégias usadas para gerenciar as mudanças em contextos particulares (Markee 1997).

O efeito ou impacto pode manifestar-se de várias maneiras: de forma positiva, fazendo com que os alunos estudem mais, preparem-se melhor para as aulas e façam as lições de casa; ou de maneira maléfica ou negativa, na medida em que são responsáveis por um aumento de pressão e, consequentemente, de ansiedade, fazendo com que os alunos se sintam ansiosos por terem que agir sob pressão, com resultados negativos em seu desempenho e naquele do professor, que passaria a ensinar os conteúdos do exame, ou a "ensinar para o exame", causando um indesejável estreitamento do currículo (Alderson & Wall 1993).

Entretanto, além de um efeito que independe das características do exame, também são reconhecidos efeitos positivos ou negativos relacionados às características propriamente ditas do exame, que teria a força e o poder de (re)direcionar o ensino, (re)definindo objetivos, conteúdos e habilidades/capacidades/competências desejáveis, e, portanto, considerado mecanismo potente para implementação de políticas públicas e educacionais. Entretanto, a forma como a avaliação tem sido usada, nesse caso, apenas como mecanismo de exclusão e manutenção do *status quo* tem recebido críticas contundentes por contemplar

[6] Neste texto, efeito retroativo e impacto são considerados sinônimos, seguindo Cheng, Watanabe & Curtis (2004). Para Alderson & Wall (1993), dentre outros, entretanto, impacto refere-se aos efeitos de um exame na sociedade em geral, enquanto efeito retroativo refere-se àqueles que ocorrem na sala de aula. Essa distinção também é feita por Messick (1996).

[7] A avaliação de rendimento também pode ser externa. Vide, por exemplo, o Exame Nacional do Ensino Médio (ENEM).

[8] Essa constatação não elimina a necessidade de se implementarem inovações e mudanças nas práticas avaliativas: elaborar um "bom" exame seria uma condição necessária, embora não suficiente, para desencadear mudanças.

[9] Convém ressaltar que efeitos retroativos ocorrem, potencialmente, com mais internsidade em exames/avaliações de alta relevância *(high-stakes)*, exames a partir dos quais decisões importantes são tomadas.

apenas os valores de grupos dominantes, não contribuindo necessariamente para a implementação de mudanças positivas. No Brasil, a implementação de mudanças educacionais através de exames já pode ser observada. Vide, por exemplo, exames tais como o Exame Nacional do Ensino Médio (ENEM), Exame Nacional de Desempenho de Estudantes (ENADE), Prova Brasil, dentre outros, e, mais recentemente, com o novo Plano de Desenvolvimento da Educação (PDE), a proposta de criação de um índice de desempenho baseado em avaliações, além da Provinha Brasil, para avaliar a eficácia das escolas na alfabetização de crianças de 6 a 8 anos. Observa-se, entretanto, que essas implementações,

> muito menos que no cenário internacional têm sido acompanhadas de evidências empíricas. Constata-se nesse caso, muito frequentemente, que exames são inseridos e descartados sem que uma análise e avaliação criteriosa de seus efeitos sejam realizadas, o que é de se lamentar, considerando-se os recursos de natureza diversa mobilizados para seu desenvolvimento e implementação (Scaramucci 2004:204).

A dimensão social da avaliação de línguas, salientada acima nas considerações sobre avaliação de desempenho, é retomada nos estudos sobre efeito retroativo e impacto sob outra perspectiva – aquela do papel e efeito da avaliação de línguas na determinação e reforço de determinadas políticas, na manutenção de identidades e de relações de poder na sociedade (McNamara & Rover 2006).

2.4 - Validade / validação: novo conceito

Retomo, nesta seção, o conceito de validade, já expandido pelas contribuições decorrentes das teorizações sobre avaliação de desempenho e efeito retroativo acima apresentadas.

A crítica principal ao conceito tradicional de validade era o fato de ser fragmentado e incompleto, elaborado exclusivamente por especialistas em medidas, seguindo uma visão essencialmente psicométrica e, como tal, não levar em conta, como base para ação, as implicações de valor do significado dos resultados ou escores e nem as consequências sociais do uso desses resultados, ou seja, a dimensão social e política que devia estar presente na avaliação de línguas, pelo fato de ser uma prática social. O novo conceito, embora contemplando múltiplas facetas, unifica-se em torno da validade de construto, passando a considerar não apenas bases evidenciais, mas também as consequenciais, como veremos mais adiante.

Para Messick (1989), validade pressupõe um julgamento que considera o grau em que explicações teóricas e evidências empíricas confirmam a adequação das interpretações e ações baseadas nos escores dos testes ou de outras formas de avaliação. Validade, portanto, não é uma propriedade do teste ou da avaliação, mas do significado dos seus resultados. Importante nesse caso é o argumento de validade, que tem como objetivo coletar informações a favor ou contra uma determinada interpretação dos escores do teste. O que é validado, portanto, são

as inferências derivadas dos resultados ou outros indicadores, assim como as implicações para ação determinadas pela interpretação.

No novo conceito de validade, Messick (1989) identifica seis diferentes tipos de evidência ou métodos para investigar hipóteses, que passam a substituir as três validades (*interna, externa e de construto*) do conceito tradicional. São elas: *validade de conteúdo, substantiva, estrutural, passível de generalização, externa e consequencial*.[10]

A Tabela 1, extraída de Chapelle (1999:258), apresentada a seguir, é elucidativa das mudanças ocorridas no conceito de validade.

Tabela 1 - Resumo dos contrastes entre as concepções tradicionais e modernas de validade (Chapelle 1999:258).

Passado	**Hoje**
Validade era considerada uma *característica de um teste:* a medida em que mede aquilo que pretende medir.	Validade é considerada um *argumento* relativo à interpretação e uso: a medida em que as interpretações e usos de um teste podem ser justificados.
Confiabilidade era vista como distinta e uma *condição necessária para validade.*	Confiabilidade pode ser vista como *um tipo de evidência de validade.*
A validade era frequentemente estabelecida através de *correlações* de um teste com outros.	Validade é argumentada com base em um número de tipos de *justificativas* e evidências, incluindo as consequências da avaliação.
Validade de construto era vista como um dos *três tipos de validade* (conteúdo, relacionada a critério e construto).	Validade é um *conceito unitário*, em que a validade de construto ocupa uma posição central (validade de conteúdo e relativa a critério podem ser usadas como evidência da validade de construto).
O estabelecimento da validade era uma tarefa de responsabilidade de *pesquisadores da avaliação*, responsáveis pelo desenvolvimento de testes de grande escala e alta relevância.	A justificativa de validade de um teste é de responsabilidade de *todos os usuários de um teste.*

[10] Foge do escopo deste trabalho um aprofundamento dessas noções.

Para caracterizar melhor a proposta moderna de validade, verdadeiro paradigma para a discussão da pesquisa e prática em medidas educacionais e psicológicas, apresento o que tem sido denominada como a "matriz progressiva de Messick".

Figura 1 - Matriz progressiva de Messick (1989:20).

	Inferências	**Usos**
Base evidencial	Validade de construto	Validade de construto + Relevância/utilidade
Base consequencial	Validade de construto + Implicações de valor	Validade de construto + Implicações de valor + Relevância /utilidade + Conseqüências sociais

Essa matriz oferece diretrizes no sentido de orientar como evidências podem ser produzidas, ou o que constituem métodos para validação. Ela permite avaliar não apenas a avaliação de línguas, mas também os efeitos retroativos de um exame, como mostrarei mais adiante. Podemos observar que a dimensão social da avaliação está representada em duas das células da matriz: naquela que considera as implicações de valor, focalizando o caráter social e cultural dos significados atribuídos aos escores do teste; e naquela que leva em conta as consequências sociais relativas ao uso prático dos testes.[11]

3 - Validade e consequências sociais

A partir dos fios condutores da discussão sobre os conceitos teóricos abordados na seção anterior busco tecer a relação entre "validade" e "consequências sociais" assim como justificar sua importância na avaliação em contextos de língua, considerando que tanto a língua como a avaliação são práticas sociais.

Um grande argumento em favor de avaliações de desempenho, em contraposição a avaliações de conhecimento, conduzidas, em geral, através de itens de múltipla escolha tem sido seu efeito potencial positivo na melhoria do ensino e aprendizagem de línguas. Alguns autores chegam até mesmo a afirmar que, por serem autênticas e diretas, as avaliações de desempenho têm *validade sistêmica (systemic validity)* (Frederiksen & Collins 1989), capazes de induzirem mudanças

[11] Essa matriz, por exemplo, foi utilizada nos estudos de Kunnan (1999) e Hamp-Lyons & Lynch (1998) para mostrar – a partir de uma análise dos processos de validação conduzidos nos últimos anos em práticas de língua estrangeira e segunda língua – que, praticamente dez anos após o novo conceito de validade ter sido proposto, os métodos de validação, pelo menos até o final dos anos 90, ainda se concentravam na *Interpretação do teste com base na categoria de base evidencial*, enquanto muita pouca atenção tinha sido dada às outras.

que levariam ao desenvolvimento das capacidades/habilidades/conhecimentos/competências que pretende avaliar. Conceito semelhante *(validade retroativa ou washback validity)* é proposto por Morrow (1986): um teste tem validade retroativa se influenciar as pessoas a fazerem coisas que não necessariamente fariam sem o exame e, portanto, deve ser avaliado pelo grau de influência positiva no ensino.

Retomando a matriz acima, podemos observar que Messick (1996) considera as consequências de uso de um exame – *validade de consequência (consequential validity)* – um dos aspectos importantes da validade de construto, e efeito retroativo[12] um tipo de consequência. Nesse tipo de validade, são consideradas as implicações de valor decorrentes da interpretação dos resultados do teste como base para a ação ou para as decisões a serem tomadas (reprovar o aluno, fazê-lo rever conteúdos, entre outras), assim como as reais e potenciais conseqüências de uso do teste na sociedade, especialmente decorrentes das fontes de invalidade envolvendo questões como viés *(bias)*, justiça *(fairness)*, além do efeito retroativo.

Entretanto, da mesma forma que Alderson & Wall (1993), o autor discorda do uso dos conceitos de *validade sistêmica e validade retroativa* propostos por Frederikson & Collins (1989) e Morrow (1986), mostrando que esse tipo de evidência não pode ser considerado de forma isolada, uma vez que os valores sociais dos resultados pretendidos e não pretendidos decorrentes da interpretação e do uso derivam e contribuem para o significado dos resultados.

Para construir seu argumento em defesa de um conceito unificado de validade, o autor salienta que o efeito retroativo de um teste somente pode ser relacionado à sua validade se pudermos ter evidência de que o efeito foi realmente do teste e não de outras forças existentes na sociedade. Assim, bons escores ou notas em um dado exame podem não ser necessariamente decorrentes de sua qualidade, mas de um bom ensino, de professores bem formados, entre outros fatores. Nesse caso, efeito retroativo seria apenas um dos aspectos a serem considerados na validação de construto, nos levando a concluir que não seria desejável, portanto, que as consequências de um teste em geral e nem muito menos os efeitos retroativos em particular fossem utilizados, de forma independente, para estabelecimento de sua validade. Os aspectos consequenciais deveriam ser considerados em conjunto com as outras evidências propostas por Messick em sua matriz; um teste válido seria aquele que apresentaria um conjunto de evidências convergentes (Messick 1996).

Qual seria a explicação, de acordo com Messick (1996), para um efeito retroativo potencial maior das avaliações de desempenho? Avaliações autênticas, pontua o autor, implicam tarefas que valem a pena e são envolventes, aplicadas em cenários realistas ou contendo simulações próximas às tarefas que acontecem na vida real, tanto em termos de tempo como de recursos. Como a maior preocupação, nesse caso, é de que nada seja deixado de lado, essas tarefas satisfazem o padrão de "mínima sub-representação do construto" *(minimal construct underrepresentation)* (vide Araújo 2007, para uma discussão sobre o conceito de autenticidade nas

[12] Vide nota 9.

avaliações de língua estrangeira). Da mesma forma, o fato de serem diretas ou abertas, uma vez que o avaliado pode responder sem estar contido pelo tipo de formato ou de método[13], que poderia introduzir fatores contaminantes, faz com que as avaliações de desempenho satisfaçam outro padrão de validade, ou seja, "não conter variância irrelevante ao construto" *(construct irrelevant variance).* Nesse caso, portanto, tanto a sub-representação do construto – que compromete a autenticidade – como a variância ou dificuldade que é irrelevante ao construto, que compromete o fato de ser direto – são ameaças à validade.

No primeiro caso, teríamos um exame que, por ser definido de forma estreita, deixa de incluir dimensões importantes do construto focal. Poderíamos, como exemplo, citar um teste de leitura que contempla itens de localização de informações, deixando de lado aqueles que avaliam a capacidade de fazer inferências; ou, ainda, um teste de proficiência geral que focaliza apenas a competência linguística a partir de itens de múltipla escolha, deixando de avaliar a competência comunicativa na interação face a face. Nesse segundo caso, o teste, por incluir aspectos irrelevantes ao construto – dificuldades relativas ao método de múltipla escolha, implicando possivelmente macetes para resolvê-las –, além da sub-representação da competência comunicativa, teria menores chances de exercer efeitos retroativos positivos. No processo de validação, portanto, o que necessitamos é coletar evidências que permitam balancear essas duas ameaças à validade de construto desse teste.

O argumento principal de Messick (1989) é que, minimizando as fontes de invalidade na concepção do teste, suas deficiências e aspectos contaminadores, estimuladores potenciais de efeitos negativos, aumentaríamos a probabilidade de efeitos positivos ocorrerem. Se o exame contiver fontes de invalidade, os escores altos obtidos não poderão ser atribuídos ao exame, mas às boas práticas de ensino, bons professores, etc. Essas invalidades, por outro lado, poderão exercer efeitos negativos, levando a práticas ruins.

Retomando os exemplos acima, podemos dizer, portanto, que o fato de o teste de leitura acima mencionado não incluir itens que avaliam a inferência de sentidos a partir do contexto, mas apenas habilidades de localização das informações explícitas poderia ser interpretado pelos professores que ler significa apenas localizar informações explícitas, incentivando práticas em que as habilidades sub-representadas estarão ausentes[14]. No caso do teste de proficiência geral, que enfatiza o conhecimento da gramática através de itens de múltipla escolha há grandes chances de os professores passarem a dar mais atenção às dificuldades envolvidas na resolução de itens de múltipla escolha de gramática do que ao desenvolvimento da competência comunicativa propriamente dita.

Nessa linha de argumentação, podemos dizer que os testes que conseguem minimizar as sub-representações do construto e suas dificuldades irrelevantes e,

[13] Métodos são meios de se avaliar. A múltipla escolha e o resumo, por exemplo, são métodos de avaliação de compreensão em leitura. Todo método tem efeitos contaminantes.

[14] Não podemos nos esquecer que a prática do professor tem-se fundamentado, em geral, em uma visão de leitura como decodificação e, portanto, o exame, nesse caso, poderia reforçar essa prática.

portanto, são válidos, também podem apresentar escores ruins que, nesse caso, não seriam atribuídos ao uso do teste. De acordo com Messick (1989), entretanto, tais consequências adversas associadas ao uso de um instrumento válido não seriam de responsabilidade do elaborador, mas resultado de uma política educacional e social.

As implicações do argumento de Messick (1989) para a consideração sobre efeito retroativo na validação de instrumentos é clara: em vez de buscarmos efeitos retroativos positivos como um sinal de validade do teste (o que é pressuposto nos conceitos de *validade sistêmica e retroativa*), deveríamos buscar validade no projeto de elaboração do teste como base para alcançarmos esses efeitos (o que tanto Messick (1989) como Alderson (2004) denominam de *washback by design)*, ou seja, para melhorá-lo nos aspectos que podem, potencialmente, apresentar problemas ou levar a consequências ruins ou que estão realmente causando consequências ruins. Assim, testes válidos, que, baseados nos seis aspectos da validade de construto, buscam minimizar a sub-representação do construto e suas irrelevâncias, devem aumentar a probabilidade de efeito retroativo potencial e ajudar a separar o que é apenas efeito retroativo do teste de práticas educacionais boas e ruins, apesar da qualidade do teste. (Messick 1996).

4 - O grande debate

Torna-se desnecessário dizer que a visão de Messick não é consensual, embora endossada por muitos, dentre os quais podemos destacar Kane (2001), Linn (1997) e Shepard (1997). Há, entretanto outros, dentre os quais Mehrens (1997) e Popham (1997) que, apesar de reconhecerem a importância das evidências obtidas a partir de estudos sobre efeito retroativo/impactos sociais, não as consideram parte de um conceito de validade.

Nesse debate, posições distintas podem ser identificadas assim como um conflito existente entre o ponto de vista do teórico em avaliação e o do prático, aquele responsável pelo desenvolvimento de exames. Embora a maioria dos autores concorde que as conseqüências do uso de um instrumento de avaliação têm que ser consideradas, não há unanimidade em que a investigação de conseqüências intencionais e não intencionais faça parte de um plano de validação. Alguns autores que argumentam a favor dessa investigação salientam que exames que se propõem a promover mudanças no ensino e na aprendizagem, além do propósito de selecionar ou classificar, devem ser avaliados em relação à consecução dessas intenções. Convém lembrar que muitos exames, nos últimos tempos, tanto no cenário internacional como no Brasil, têm sido elaborados como instrumentos de política educacional e pública, tendo em vista o poder que exercem nas pessoas e na sociedade em geral. O que muitos autores questionam é se realmente esses instrumentos têm promovido mudanças benéficas na educação, principalmente quando acompanhados de punições e recompensas para os professores e escolas (conceito tradicionalmente conhecido como responsabilização *(accountability)*). Não estariam apenas levando as escolas a uma preocupação com resultados positivos

nas avaliações, em tentar apenas a aumentar os escores, sem resultados positivos na aprendizagem? Não estariam apenas levando as escolas a treinamentos para os exames, principalmente no caso em que lhe são atribuídas responsabilidades pelas mudanças intencionadas? Portanto, nesse caso, para esses autores, as conseqüências deveriam ser avaliadas como parte fundamental do processo de validação, mesmo que isso torne o processo muito mais complexo.

Os oponentes, por outro lado, argumentam que embora as conseqüências sejam importantes, elas não devem fazer parte de um plano de validação, na medida em que uma consideração de conseqüências pressupõe considerações políticas e sociais, que fogem do controle do elaborador de exames. Na maior parte das vezes, o elaborador não tem a autoridade para decidir avaliar as conseqüências sociais do exame que elabora uma vez que essa é uma atribuição dos responsáveis pelas políticas implementadas a partir de exames. Nesse caso, a responsabilidade pela coleta de evidências sobre as conseqüências tem sido atribuída a todos os protagonistas do contexto em questão, ou seja, também o usuário, além dos pais, professores, alunos, etc. Se a validação é um processo contínuo e sem fim, caberia a todos os envolvidos a coleta de dados sobre esse processo.

O exame e a documentação de consequências são, portanto, considerados questões bastante difíceis, em que obstáculos variados são encontrados. Um deles, talvez o principal é que, para muitos, consequências implicam uma relação de causa-efeito e os dados teriam que ser coletados de maneira a mostrar evidências dessa relação. Para outros, entretanto, consequências, não implicam uma relação determinista de causa-efeito, uma vez que esse impacto é mediado por outras forças (formação do professor, por exemplo).

Outra questão que merece destaque é a importância de estudos de base (*baseline studies*), entre outros, em que o contexto de implementação seja bem descrito, de forma que se possa explorar com mais profundidade a relação entre consequências desejadas ou intencionais e consequências reais.

Não podemos nos esquecer, ainda, da dificuldade decorrente do fato de que a questão envolve valores: o que é uma conseqüência positiva e o que é uma negativa depende de quem faz a pesquisa, suas crenças e valores.

5 – Considerações finais

Meu objetivo, neste artigo, foi discutir o novo conceito de validade, que passou por reformulações nos últimos anos, passando a incluir as conseqüências intencionais e não intencionais assim como as conseqüências reais das avaliações no ensino, na aprendizagem, na vida das pessoas em geral como parte das evidências de validade no processo de validação.

Apesar do reconhecimento praticamente unânime da importância dessas evidências, não podemos deixar de salientar que, na visão de Messick, seu principal proponente, no entanto, somente podemos contabilizar como consequências de uma avaliação aquelas que realmente puderem ser atribuídas ao teste e a nada mais. O que nos coloca um problema de difícil solução: se as características desse

instrumento interagem com outras forças na sociedade, como podemos separar o que é efeito de um teste de outras ações, tais como bom ensino, por exemplo?

O que podemos fazer – como elaboradores, usuários e elaboradores de políticas que se utilizam de exames – é, na opinião de Messick (1989) e também de Alderson (2004), analisar as consequências de um teste não para concluir se é válido ou não, mas apenas para melhorá-lo nos aspectos que podem, potencialmente, apresentar problemas ou levar a consequências maléficas ou que já estão apresentando consequências maléficas. Mesmo em testes considerados inovadores os impactos podem ser negativos dependendo de como as pessoas interpretam essas características.

Autores como McNamara & Roever (2006), no entanto, consideram "tímida" a forma como Messick (1989) incorpora as conseqüências e valores sociais/ culturais em seu conceito de validade. Na opinião do autor, a consideração das conseqüências sociais parece andar na contramão da teoria de validade, que ainda permanece calcada, em grande parte, pelos princípios decorrentes de suas origens no campo da psicologia, individualista e cognitivamente orientado. Para ele, um problema que Messick (1989) nunca resolveu é a relação entre as duas dimensões menos socialmente orientadas da linha superior da matriz e as duas dimensões da linha inferior, o que permanece como uma das questões fundamentais da área.

Recebido em junho de 2011; aceite em junho de 2011.

Referências

Alderson, J. C. 2004. Foreword. In: L. Cheng; Y. Watanabe; A. Curtis (Eds.) *Washback in Language Testing – Research Contexts and Methods.* New Jersey: Lawrence Erlbaum Associates.

Alderson, J.C.; Hamp-Lyons, L. 1996. TOEFL preparation courses: a study of washback. *Language Testing*, 13: 280-297.

Alderson, J. C. ; Clapham, C; Wall, D. 1995. Language test construction and evaluation. Cambridge: Cambridge University Press.

Alderson, J. C. ; Wall, D. 1993. Does washback exist? *Applied Linguistics.* 14 (2): 115- 129.

Araújo, K. da S. 2007. *A perspectiva do examinando sobre a autenticidade de avaliações em leitura em Língua Estrangeira.* Dissertação de mestrado. Programa de pós-graduação em LinguísticaAplicada, IEL, Unicamp.

Avelar, S.L.T. 2001. *Mudanças na concepção e prática da avaliação e seu efeito no ensino/aprendizagem de língua estrangeira (Inglês) em uma escola de ensino médio e técnico.* Dissertação de mestrado. Programa de Pós-graduação em Linguística Aplicada, IEL, Unicamp.

AERA/APA/NCME. 1985. *Standards for Educacional and Psychological Testing.* Washington DC: American Psychological Association.

Bailey, K. 1996. Working for washback: A review of the washback concept in language testing. *Language Testing*. 13: 257-279.

Bartholomeu, M. A. 2002. *Prova de língua estrangeira (Inglês) dos vestibulares e sua influência nas percepções, atitudes e motivações de alunos do terceiro ano do ensino médio.* Dissertação de mestrado, Programa de Pós-graduação em Linguística Aplicada, IEL, Unicamp.

Canale, M. 1987. The measurement of communicative competence. In R. B. Kaplan, et al. (eds) *Annual Review of Applied Linguistics*. 8: 67-84. New York: Cambridge University Press.

Chapelle, C. A. 1999. Validity in language assessment. *Annual Review of Applied Linguistics*. 19: 254-272.

Cheng, L. 1999. Changing assessment: washback on teacher perspectives and actions. *Teaching and teacher education*. 15 (3): 253-271.

Cheng, L.; Watanabe, Y.; Curtis, A. 2004. *Washback in Language Testing – Research Contexts and Methods.* New Jersey: Lawrence Erlbaum Associates.

Cherryholmes, C. 1988. *Power and criticism: Poststructural investigations in education*. New York: Teachers College Press.

Clark, H. 2002. O uso da linguagem. *Cadernos de Tradução*, 9:49-71. Porto Alegre: Instituto de Letras, UFRGS.

Correia, R. M. D. 2003. *O efeito retroativo da prova de inglês do vestibular da Unicamp na preparação de alunos em um curso preparatório comunitário.* Dissertação de mestrado. Programa de Pós-graduação em Linguística Aplicada, IEL, Unicamp.

Ebel R. L.; Frisbie, D. A. 1991. *Essentials of Educational Measurement*. 5th edition. Englewood Cliffs: Prentice-Hall.

Gronlund, N. E. 1985. *Measurement and Evaluation in Teaching*. New York: Macmillan.

Frederiksen, J. R.; Collins, A. 1989. *Essentials of Educational Measurement*. 5th edition. Englewood Cliffs: Prentice-Hall.

Hamp-Lyons, L.; Lynch, B. K., 1998. Perspectives on validity: a historical analysis of language testing conference abstracts. In: A.J. Kunnan, (Editor). Issues in Language Testing Research: Conventional Validity and Beyond, Mahwah: Lawrence Erlbaum, 253–277.

Henning, G. 1987. *A Guide to Language Testing*. Cambridge: Cambridge University Press.

Hughes, A. 1989/1994. *Testing for language teachers.* Cambridge: Cambridge University Press.

Kane, M.T. 2001. An argument-based approach to validity. *Psychological Bulletin*. 112: 527-535.

Kunnan, A. J. 1999. Recent developments in language testing. *Annual Review of Applied Linguistics*. 19: 235-253.

Lanzoni, H. 2004. *Exame de proficiência em leitura de textos acadêmicos em*

Inglês: um estudo sobre efeito retroativo. Tese de doutorado inédita. Programa de Pós-graduação em LinguísticaAplicada, IEL, Unicamp.

Linn, R. L. 1997. Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice.* 16(2):14-16.

Madsen, H. S. 1983. *Techniques in testing.* Oxford: Oxford University Press.

Markee, N. 1997. The diffusion of innovation in language teaching. *Annual Review of Applied Linguistics.* 13: 229-243.

McNamara, T. 2000. *Language testing*. Oxford: Oxford University Press.

McNamara, T. 1996. *Measuring second language performance.* London: Longman.

McNamara, T.; Roever, C. 2006. *Language Testing: The social dimension.* Blackwell Publishing Limited.

Mehrens, W. A. 1997. The consequences of consequential validity. *Educational Measurement: Issues and Practice.* 16(2): 16-18.

Messick, S. 1996. Validity and washback in language testing. *Language Testing.* 13(3): 241-256.

Messick. 1989. Validity. In R. L. Linn (Ed.), *Educational measurement* (3th edition). New York: American Council on Education & Macmillan, 13-103.

Morrow, K. 1986. The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing*. London: NFER/Nelson.

Morrow, K. 1977. *Techniques of evaluation for a notional syllabus*. London: Royal Society of Arts.

Moss, P. A. 1992 Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research.* 62: 229-258.

Pessoa, A. R. 2002. *O efeito retroativo do Programa de Avaliação Seriada da Universidade de Brasília no Ensino de língua estrangeira do Distrito Federal.* Dissertação de mestrado. Programa de Pós-graduação em Linguística Aplicada, Universidade de Brasilia, DF.

Popham, W. J. 1997. Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice.* 16(2): 9-13.

Retorta, M. S. 2007. *O efeito retroativo do vestibular da Universidade Federal do Paraná: uma investigação em escolas públicas, particulares e cursos pré-vestibulares.* Tese de doutorado inédita. Programa de Pós-graduação em Linguística Aplicada, IEL, Unicamp.

Sakamori, L. 2006. *A atuação do entrevistador na interação face a face no exame Celpe-Bras*. Dissertação de mestrado. Programa de Pós-graduação em Linguística Aplicada, IEL, Unicamp.

Scaramucci, M. V. R. 2005. Prova de redação nos vestibulares: educacionalmente benéfica para o ensino/aprendizagem da escrita? Em Flores, V. do N. et al. (orgs.), *A redação no contexto do Vestibular 2005 – a avaliação em perspectiva.* Editora UFRGS, 37-57.

Scaramucci, M. V. R. 2004. Efeito retroativo da avaliação no ensino/aprendizagem de línguas: o estado da arte. *Trabalhos em Linguística Aplicada*. 43 (2): 203-226.

Scaramucci, M. V. R. 2003. Evaluación de "proficiencia" em lengua extranjera: relaciones com la enseñanza y evaluación de rendimiento. Palestra apresentada no Sexto Ciclo Internacional de Enseñanza de Lenguas Extranjeras, Buenos Aires, Argentina.

Scaramucci, M. V. R. 2002. Entrance examinations and TEFL in Brazil: a case study. *Revista Brasileira de LinguísticaAplicada*. 2(1): 61-81.

Scaramucci, M. V. R. 2000/2001. Propostas curriculares e exames vestibulares: potencializando o efeito retroativo benéfico no ensino de LE (Inglês). *Revistas Contexturas.* 5:97-109, APLIESP, São José do Rio Preto, SP.

Scaramucci, M. V. R. 1999a. Proficiência em LE: Considerações terminológicas e conceituais. *Trabalhos de Linguística Aplicada.* 36:11-22.

Scaramucci, M. V. R. 1999b. Celpe-Bras: um exame comunicativo. In: M. J. Cunha; P. Santos (orgs.) *Ensino e Pesquisa em Português para Estrangeiros.* Editora da Universidade de Brasília, Brasilia, DF. 75-81.

Scaramucci, M. V. R. 1999c. Vestibular e ensino de língua estrangeira (inglês) em uma escola pública. *Trabalhos em Linguística Aplicada.* (34): 7-20.

Scaramucci, M. V. R. 1999d. University entrance examinations and EFL teaching: a study of washback in a Brazilian context. Trabalho apresentado no 12th World Congress of Applied Linguistics (AILA99), em Tóquio, Japão.

Scaramucci, M. V. R. 1998a. Vestibular: instrumento direcionador do ensino de segundo grau? Trabalho apresentado no V Congresso Brasileiro de Linguística Aplicada (V CBLA), em Porto Alegre, Rio Grande do Sul.

Scaramucci, M. V. R. 1998b. O efeito retroativo dos vestibulares de língua inglesa da Unicamp no ensino de segundo grau de escolas públicas e particulares de Campinas. Relatório final de pesquisa, FAPESP 95/06551-9.

Schlatter, M. Garcez, P. M.; Scaramucci, M. V. R. 2004. O papel da interação na pesquisa sobre aquisição e uso da língua estrangeira: implicações para o ensino e para a avaliação. *Letras de Hoje.* 39 (3): 345-378.

Shepard, L. A. 1997. The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice.* 16(2):5-13.

Shohamy, E. 1995. Performance assessment in language testing. *Annual Review of Applied Linguistics.* 15: 188-211.

Souza, L. G. 2002. *Ensino da produção escrita em língua estrangeira (inglês) em um curso de línguas: influência da avaliação ou da concepção de escrita do professor?* Dissertação de mestrado. Programa de Pós-graduação em LinguísticaAplicada, IEL, Unicamp.

Thorndike, R. L.; Hagen, E. P. 1986. *Measurement and Evaluation in Psychology and Education.* New York:Macmillan.

Wall, D. 2000. The impact of high-stakes testing on teaching and learning: can this be predicted or controlled? *System*. 28(4):499-509.

Watanabe, Y. 1996. Does grammar translation come from entrance examination? Preliminary findings from classroom-based research. *Language Testing.* 13(3): 318-333.

Wiggins, G. P. 1993. Assessing student performance: *Exploring the purpose and limits of testing*. San Francisco: Jossey-Bass Publishers.