Calidoscópio
Vol. 7, n. 1, p. 30-48, jan/abr 2009

Matilde V.R. Scaramucci
matilde@unicamp.br

Avaliação da leitura em inglês como língua estrangeira e validade de construto

EFL reading assessment and construct validity

RESUMO – Neste artigo, a análise conduzida por Tumolo e Tomich (2007) e Tumolo (2005) em itens extraídos da prova de inglês do vestibular da Unicamp é questionada. A análise, conduzida como parte de um estudo que investiga a defensibilidade de itens de três instrumentos de avaliação de leitura em inglês como língua estrangeira – exames internacionais de proficiência (TOEFL e IELTS), exames vestibulares de duas universidades brasileiras (Unicamp e UFSC) e exames elaborados pelo professor em sala de aula – tem a validade de construto como seu critério principal. Os itens são classificados em defensáveis e não defensáveis. Itens defensáveis, para os autores, seriam aqueles que "permitem a demonstração da habilidade como definida no construto, enquanto os não defensáveis não permitem coletar evidências para uma interpretação válida da habilidade relevante" (Tumolo e Tomich, 2007, p. 67). O argumento principal de nosso questionamento é o fato de o construto em que se baseia a análise ser distinto daquele que fundamenta a prova da Unicamp.

Palavras-chave: avaliação da leitura, inglês, língua estrangeira, validade de construto, exames vestibulares.

ABSTRACT – In this article, the analysis conducted by Tumolo and Tomich (2007) and Tumolo (2005) on items from the English subtest of the University of Campinas entrance examination has been questioned. The analysis, carried out as part of a study which investigates the defensibility of items from three different EFL assessment instruments -- international proficiency tests (TOEF and IELTS), entrance examinations of two Brazilian universities (Unicamp e UFSC) and classroom tests developed by the teacher -- has construct validity as its main criterion. The items were divided into defensible and non-defensible. According to the authors, defensible items are those which "allow for the demonstration of the reading ability as described in the construct, whereas non-defensible items do not allow for collecting evidence for a more valid interpretation of the relevant ability" (Tumolo and Tomich, 2007, p. 67). Our main criticism is that the reading construct used for the analysis is different from the one which underlies the Unicamp examination.

Key words: reading assessment, EFL, construct validity, entrance examination.

Introdução

Apesar de a avaliação em contextos de ensino/aprendizagem de línguas vir merecendo uma atenção bastante grande nos últimos anos no cenário internacional, constituindo-se uma subárea com desenvolvimentos importantes para a área de Linguística Aplicada, podemos dizer que ainda são escassos os estudos sobre avaliação em contextos de línguas no Brasil[1] quando comparados com outras temáticas. Quando se trata da avaliação da leitura em inglês como língua estrangeira, as contribuições brasileiras são ainda mais escassas, apesar do interesse pelo ensino dessa habilidade ter-se renovado com a publicação dos PCN nos anos 1990. E, principalmente, apesar da importância que a avaliação dessa habilidade tem em muitos vestibulares, considerados exames de alta relevância no contexto brasileiro, em vista de seus impactos ou consequências sociais e educacionais (vide Scaramucci, s.d., para uma discussão mais aprofundada sobre avaliação e impactos sociais[2]).

Nesse sentido, *Avaliando a leitura em inglês: uma reflexão sobre itens de testes,* de autoria de Tumolo e Tomich (2007, p. 67-88), publicado na *Revista Brasileira de Linguística Aplicada*, baseado em Tumolo (2005)[3] é uma

[1] Há muitos anos tenho me dedicado à área de ensino e pesquisa em avaliação em contextos de ensino/aprendizagem de línguas, estrangeira e materna. Teses e dissertações sobre esse tema têm sido desenvolvidas sob minha orientação no Programa de Pós-graduação em Linguística Aplicada da Unicamp.
[2] Tema de projeto de pós-doutorado, por mim realizado na Universidade de Melbourne, Austrália, sob a supervisão de Tim McNamara.
[3] O artigo baseia-se na tese de doutorado do primeiro autor (Tumolo, 2005), orientada pelo segundo, intitulada *Assessment of reading in English as a foreign language: investigating the defensibility of test items.*

dessas exceções. No artigo, os autores discutem, à luz do conceito de validade de construto, fundamental quando falamos de avaliação, itens selecionados de exames vestibulares, de proficiência e elaborados pelo professor para uso em sala de aula, classificando-os em "defensáveis" e "não defensáveis". Nas palavras dos autores, "os defensáveis permitem a demonstração da habilidade como definida no construto, enquanto os não defensáveis não permitem coletar evidências para uma interpretação válida da habilidade relevante" (Tumolo e Tomich, 2007, p. 67) e, portanto, seriam itens que levariam a inferências não válidas a respeito da habilidade em questão.

Embora relevante e esclarecedor em muitos aspectos, o artigo traz algumas questões que, a meu ver, mereceriam um tratamento mais qualificado para que pudessem (e retomo mais uma vez as palavras dos autores) "colaborar com o desenvolvimento de testes que possibilitem coletar evidências para uma interpretação mais válida da habilidade de leitura em inglês" (Tumolo e Tomich, 2007, p. 70) e "contribuir no sentido de trazer uma reflexão sobre o desenvolvimento de testes de tal forma a possibilitar o incremento das atividades de avaliação" (Tumolo e Tomich, 2007, p. 86) no Brasil.

Ao classificar itens de três modalidades de exames em defensáveis e não defensáveis e, sobretudo, ao identificar a origem desses itens – principalmente vestibulares nacionais e exames de proficiência internacionais importantes, tais como TOEFL e IELTS[4] – os autores sugerem que os resultados das provas que contêm esses itens não sejam válidos, o que é bastante comprometedor para esses exames. Entretanto, quando se trata de processos de validação, sabemos que o método de "julgamento de especialista" utilizado pelos autores é limitado, o que faz com que as análises nele baseadas sejam provisórias e parciais, até que argumentos contrários a refutem. Além disso, como é amplamente reconhecido na área e também pelos próprios autores, "validade é uma questão de grau, não é uma coisa de tudo ou nada. A validade da interpretação e a ação baseada nos itens serão relativos à força da argumentação, sendo mais válido ou menos válido em relação a todos os argumentos apresentados" (Tumolo, 2005, p. 236).

Neste artigo, portanto, organizado em seis partes, além desta introdução e das considerações finais, retomamos alguns dos exemplos analisados no referido artigo, questionamos aspectos da análise conduzida e os rediscutimos, também à luz do conceito de validade de construto, que apresentamos a seguir. Nossa análise focaliza itens citados pelos autores como "não defensáveis" extraídos da prova de inglês do exame vestibular da Universidade Estadual de Campinas (Unicamp).

O conceito de validade: visão tradicional

Uma das questões que tem mobilizado grande parte dos teóricos da avaliação nos últimos anos é a busca de um entendimento mais preciso do que se convencionou chamar de testes ou avaliação[5] 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 1990, 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. Algumas publicações, ocorridas a partir de 1985, 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[6] *Standards for Educacional and Psychological Testing* ou AERA/APA/NCME, 1985); 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* (Chapelle, 1999).

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, publicados nos anos 80, encontraremos definições como a de Hughes (1989): um teste é válido se mede precisamente aquilo que deve medir. Essa visão, entretanto, também pode ser observada naqueles publicados nos anos 1990, como Alderson *et al.* (1995), que retoma a definição de Henning (1987, p. 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 no grau em que mede o que deve medir. Assim, o termo válido, quando usado para descrever um teste, deve geralmente vir acompanhado pela

[4] Os exemplos do IELTS são discutidos apenas em Tumolo (2005) e não em Tumolo e Tomich (2007).

[5] Testes são instrumentos formais de avaliação; apesar de incluir testes, a avaliação é um processo mais amplo e não está limitada a eles. Neste trabalho, exames e testes são considerados sinônimos.

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

> preposição "para". Qualquer teste, dessa forma, pode ser válido para alguns propósitos, mas não para outros (Henning, 1987 *in* Alderson *et al.*, p. 89)[7].

Frequentemente, a validade tem sido abordada em relação à confiabilidade, mas, algumas vezes, também à praticidade, ambas também vistas como qualidades de um teste. Este 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, que resultado devemos tomar como evidência daquilo que pretendemos avaliar? 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) 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, tem por objetivo a redação de um texto pode ser 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.

Embora vários tipos de validade têm sido, tradicionalmente, reconhecidos na literatura, na medida em que ela pode ser estabelecida através de diferentes métodos, não parece haver concordância quanto aos termos e definições. Hughes (1989), por exemplo, refere-se à validade *de 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 contribuem as validades externa e interna.

Uma definição breve de validade de construto fornecida por Gronlund (1985, p. 58 *in* Alderson *et al.* 1995, p. 183) seria "o grau 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, p. 108)[8]:

> 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 dos escores ou notas do teste para garantir que eles signifiquem o que o especialista esperava que significassem.

O processo de validação é visto como uma atividade de pesquisa, através da qual teorias são testadas e confirmadas, modificadas ou abandonadas.

O conceito de validade: visão moderna ou expandida

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 a ação, as implicações de valor do significado dos resultados ou escores, 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.

---

[7] "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" (Henning, 1987 *in* Alderson *et al.*, p. 89).

[8] "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" (Ebel e Frisbie, 1991, p. 108).

Para Messick (1989), responsável pela revolução do conceito, 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 (Mislevy *et al*., 2003; Bachman, 2005), que tem como objetivo coletar informações a favor ou contra uma determinada interpretação dos escores do teste. O que é validado, portanto, não é o teste, mas as inferências derivadas dos resultados ou de outros indicadores, assim como as implicações para a ação determinadas pela interpretação.

Messick identifica duas fontes de ameaça à validade dos instrumentos de avaliação: *sub-representação do construto* e *variância irrelevante ao construto*. Enquanto a primeira compromete a autenticidade, a segunda compromete a característica do teste de "ser direto" (*directness*). As avaliações autênticas – testes de desempenho são autênticos –, portanto, implicam tarefas envolventes, que valem a pena realizar, 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 que nada seja deixado de lado, essas tarefas satisfazem o padrão de "mínima sub-representação do construto" (vide Araújo 2007, para uma discussão sobre o conceito de autenticidade nas avaliações em 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[9], 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" (Scaramucci, 2007).

No primeiro caso, teríamos um exame que, por ser definido de forma estreita, deixaria de incluir dimensões importantes do construto focal (*sub-representação do construto*). Como exemplo, poderíamos citar um teste de leitura que contempla itens que avaliam 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, além da *sub-representação da competência comunicativa*, teríamos também dificuldades relativas ao método de múltipla escolha, implicando possivelmente macetes para resolvê-las (*variância irrelevante ao construto*). Os dois testes teriam menores chances de exercer impactos positivos[10]. No processo de validação, portanto, o que necessitamos é coletar evidências que permitam balancear essas duas ameaças à validade de construto desse teste.

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 por professores que ler significa apenas localizar informações explícitas, incentivando práticas em que as habilidades sub-representadas estarão ausentes[11]. No caso do teste de proficiência geral, que enfatiza o conhecimento da gramática através de itens de múltipla escolha, haveria 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.

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: evidências relativas ao conteúdo, substantiva, estrutural, passível de generalização, externa e consequencial.

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

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

Essa matriz oferece diretrizes no sentido de orientar como evidências podem ser produzidas ou o que constituem métodos para validação. Além disso, permite avaliar não apenas os testes, mas também suas consequências ou impactos sociais. 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

[9] 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.

[10] As questões sobre impacto/efeito retroativo não serão tratadas neste texto.

[11] 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.

Tabela 1. Resumo dos contrastes entre as concepções de validação tradicional e contemporânea (Chapelle, 1999, p. 258).
Table 1. Summary of the contrasts between the traditional and modern views of validity (Chapelle, 1999, p. 258).

Tradicional	Contemporânea
Considerada uma característica de um teste: o grau em que um teste mede aquilo que pretende medir.	Validade é um argumento relativo à interpretação e uso: o grau 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 sobre 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.

Tabela 2. Matriz progressiva de Messick (1989, p. 20).
Table 2. Messick's progressive matrix (1989, p. 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+ Consequências sociais

conta as consequências sociais relativas ao uso prático dos testes[12].

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. Por outro lado, há ainda outros (McNamara e Roever, 2006, por exemplo) que consideram "tímida" a forma como Messick incorpora as consequências e valores sociais/culturais em seu conceito de validade. Na opinião do autor, a consideração das consequências 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 nunca parece ter resolvido 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 (vide Scaramucci, s.d., para uma discussão dessa questão).

[12] Essa matriz, por exemplo, foi utilizada nos estudos de Kunnan (1999) e Hamp-Lyons e 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 1990, ainda se concentravam na *Interpretação do teste com base na categoria de base evidencial*, enquanto pouquíssima atenção tinha sido dada às outras categorias (Chapelle, 1999).

O construto "leitura" e sua avaliação

Tendo apresentado a visão expandida ou moderna de validade, voltamos nossa atenção neste momento para a introdução do artigo de Tumolo e Tomich (2007, p. 69), onde é apresentada a pergunta de pesquisa que fundamenta sua análise para a classificação de itens em defensáveis e não defensáveis: "os itens usados possibilitam coletar evidências para interpretação sobre habilidade em leitura em inglês como língua estrangeira?" Logo a seguir, na seção seguinte, apresentam "a definição do construto leitura a ser usado para nossa análise de itens de testes que objetivam avaliar a habilidade de leitura em inglês*"*. O construto é apresentado em termos de componentes -- seguindo Gagné *et al.* (1993) e Hutchins (1987) – "a compreensão da microestrutura, a identificação do esquema global, a aplicação de macro-regras para generalizar e condensar a uma representação da macroestrutura, e, por fim, a expressão dessa macroestrutura em um texto coerente" (Tumolo e Tomich, 2007, p. 73). Apesar de listarem esses componentes, não apresentam uma definição operacional de leitura: afinal, o que entendem como o ato de ler? Como esses componentes se interrelacionam? Para que esse construto possa ser utilizado, ele teria de ser mais bem explicitado. Um construto pressupõe uma concepção, nesse caso, de leitura.

Entretanto, esse não é o aspecto que mais chama a atenção no artigo e nem o foco de nossos comentários. O seu maior equívoco está, a nosso ver, no fato de os autores estabelecerem um construto para avaliar itens de testes que não foram elaborados com base nesse construto, mas sim em construtos distintos, que não são discutidos ou sequer apresentados. O que os autores parecem sugerir é que o construto "leitura" é universal e independente da situação de avaliação. Pode-se observar uma contradição no seu próprio texto na medida em que, ao discutirem o processo de elaboração de um teste, afirmam:

> O primeiro estágio, planejamento, envolve a definição do que se quer medir, as especificações norteadoras do desenvolvimento de um teste. Faz-se, assim, necessária a definição do construto, entendido como o conjunto de habilidades e/ou conhecimentos que pode ser *plausivelmente argumentado e/ou teoricamente justificado* [ênfase nossa] como esperado (Tumolo e Tomich, 2007, p. 68).

Ainda na mesma página:

> As perguntas a serem feitas em um processo de investigação de validade de construto devem se basear na seguinte pergunta geral: em que medida os itens do teste permitem demonstrar habilidades e/ou conhecimentos representativos e/ou relevantes em relação ao construto? [...] será que os itens não estão associados a um construto que não seja aquele *definido durante os estágios de elaboração do teste* [ênfase nossa] mas sim, a um construto alternativo? (Tumolo e Tomich, 2007, p. 68).

Como os próprios autores reconhecem, o construto que deve servir de base para avaliar os itens é o construto definido durante os estágios de elaboração do teste, e não um construto externo. Nesse caso, como já salientei antes retomando Ebel e Frisbie (1991, p. 108), construto "refere-se [...] a uma conceitualização teórica sobre um aspecto do comportamento humano que não pode ser medida ou observada diretamente". Construto, portanto, não é universalmente compartilhado, mas definido localmente, situado, uma construção, uma teorização que, como os próprios autores sugerem, deve ser "argumentado e teoricamente justificado".

O entendimento do que é um construto está na base do ato de avaliar. 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 subsequentes 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, 2000). É importante, portanto, uma distinção clara entre o critério, ou seja, o comportamento comunicativo na situação alvo que se quer avaliar (leitura, por exemplo) e o teste ou instrumento para avaliá-lo.

Apesar de realistas, entretanto, as situações de avaliação não são reais, porque sempre serão 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 artigo de jornal, simulação de uma conversa com um paciente, assistir uma palestra – salienta McNamara (2000, p. 7-8), 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.

Aparentemente o que é usado pelos autores para justificarem sua análise é o fato de os três tipos de testes analisados (testes de proficiência, de entrada ou vestibulares e testes de rendimento preparados pelo professor para a sala de aula) avaliarem a leitura em situações semelhantes, ou seja, "leitura de textos expositivos com o propósito de estudos, que é a situação de uso da língua alvo para as três situações de avaliação, ou seja, o critério". Entretanto, esse pressuposto – explicitado em Tumolo (2005) – é equivocado, porque mesmo se tratando de situações semelhantes, não necessariamente o construto é o mesmo.

Uma prova de proficiência em leitura para a pós-graduação, por exemplo, não tem necessariamente que ser configurada de forma semelhante a uma prova de proficiência em leitura em exames de proficiência geral como TOEFL e IELTS, que por sua vez também é distinta de uma avaliação de leitura em um teste de entrada, como é o

caso dos vestibulares. Cada uma das diferentes funções da avaliação ou tipos de exames (entrada, proficiência geral, proficiência específica, testes de sala de aula) pode ter conteúdos distintos com configurações também distintas de sub/habilidades e/ou conhecimentos e procedimentos distintos de elaboração: o construto leitura, portanto, teria que ser definido de acordo com a função do exame e da situação a que se propõe. Além disso, o construto envolve valores culturais e até mesmo o método de avaliar, como mostrarei mais adiante.

Outro aspecto que fragiliza a análise de Tumolo e Tomich (2007, p. 73) é reconhecido pelos próprios autores:

> As análises aqui feitas não são baseadas em itens testados em situações reais ou simulações, mas, sim em balanço de argumentos, procedimento de investigação de validade de natureza mais interpretativa, defendido por autores como Kane (1992) e Chapelle (1994, 1999). Elas têm por objetivo propiciar uma reflexão que auxilie o processo de desenvolvimento de itens que permitem interpretações e ações mais válidas em relação ao construto.

Embora concordemos que a validação de um exame é um processo de coleta de evidências relevantes às inferências e hipóteses do argumento interpretativo, não podemos deixar de salientar que esse processo é em geral conduzido durante a elaboração de um teste pelo próprio elaborador, a partir de suas especificações. Embora uma análise crítica de pontos fracos ou problemáticos seja parte importante das evidências em um processo de validação, ela nunca é final ou conclusiva, porque argumentos contrários para justificar esses aspectos considerados problemáticos podem ser levantados. Nesse caso em especial, observamos que a análise dos autores é feita com dados escassos; faltam informações importantes sobre o desempenho dos candidatos, critérios e procedimentos de correção, entre outros.

Ainda outro aspecto que merece destaque é o fato de o artigo discutir itens de maneira descontextualizada, sem analisá-los no conjunto de itens que compõem uma prova[13]. Para que esta possa ser uma amostra representativa do que se deseja avaliar, ela é configurada de forma a contemplar os vários aspectos importantes do construto. Descontextualizá-los significa perder informações valiosas.

Feitas essas colocações, passamos a seguir à discussão do que os autores consideram como itens defensáveis e não defensáveis.

Itens defensáveis e não defensáveis

Como mencionado antes, um construto único, proposto pelos autores, foi utilizado para classificar os itens de três exames distintos em defensáveis e não defensáveis. Novamente, defensáveis são aqueles "que permitem coletar evidências para a interpretação de habilidade com maior grau de validade de construto". Os autores passam então a julgar itens geralmente utilizados para avaliar a leitura em língua estrangeira:

> Testes que avaliam a habilidade de leitura em língua estrangeira têm tradicionalmente incorporado itens que focam a compreensão da ideia principal do texto e algumas vezes de parágrafos, compreensão do significado de palavras em contexto, como também na identificação de informações localizadas (datas, números, lugares, nomes, etc.) e identificação de referência (pronomes relativos, pronomes objetivos, pronomes possessivos, etc.) (Tumolo e Tomich, 2007, p. 73).

O que se observa aqui é que o construto aparentemente foi esquecido. O que deve orientar a escolha dos itens (de compreensão geral, ou de detalhes do texto, ou ainda de sumarização) é o objetivo da avaliação e seu construto, e não o que é mais simples ou menos problemático de se avaliar. Os próprios autores citam Hutchins (1987) a respeito disso, conforme segue:

> Itens que focam a compreensão de ideia principal do texto têm o argumento favorável que avaliam todo o trabalho de leitura envolvendo processos inferiores, como estabelecimento da relação sintática e atribuição de significado de palavras, e processos superiores de inferências, como também a progressão temática e a sumarização. Apesar dos argumentos favoráveis, existem aqueles que consideramos desfavoráveis. A construção da ideia principal depende do processo de sumarização. Esse processo envolve o desenvolvimento de uma macroestrutura textual construída, envolvendo o conhecimento de mundo de cada indivíduo, levando a interpretações distintas de um mesmo texto (Hutchins, 1987, *in* Tumolo e Tomich, 2007, p. 73).

Nesse caso, a defesa de um item de compreensão da ideia principal do texto parece ser feita com base em outros critérios. Nesse caso específico, arriscaríamos dizer que esta sempre seria defensável, na medida em que é a "entrada" para o texto, essencial para uma compreensão mais detalhada. Mas essa decisão somente poderá ser tomada com base no construto que fundamenta o teste. Portanto, não se trata de uma opção, mas de uma necessidade, dependendo do construto.

Além disso, percebe-se a dificuldade dos autores em reconhecerem o papel desempenhado pelos conhecimentos de mundo do leitor em uma visão de leitura como construção de sentidos. Em nossa opinião, a possibilidade de interpretarmos um texto de formas distintas – que pode ser mais "aberto" ou mais "fechado", dependendo do gênero e de outras características – pela variação do conjunto de conhecimentos do leitor é um dos grandes desafios e a grande riqueza dos processos de avaliação

[13] Em Tumolo (2005), embora uma prova completa da Unicamp seja discutida, não são feitos comentários em relação à prova como um todo e nem ao seu construto. As especificações da prova aparecem em um dos anexos da tese.

da compreensão em leitura. Lidar com essas questões tem sido uma experiência extremamente interessante para os que trabalhamos com correção de questões de itens de respostas abertas em testes de grande escala. Os autores reconhecem que

> [esses itens] podem levar à dificuldade, em situações de teste, de identificar o que pode ser considerado como resposta correta ou incorreta. Assim, no balanço dos argumentos para validade de construto, esses itens parecem ser defensáveis com a ressalva que as distinções sejam consideradas no momento da interpretação (e pontuação) das respostas (Tumolo e Tomich, 2007, p. 73-74).

Apesar de reconhecer essas possibilidades, os autores não mencionaram aspectos relativos à correção da prova da Unicamp quando classificaram esses itens em não defensáveis.

Por outro lado, itens não defensáveis, para os autores, são aqueles que apresentam

> problemas no momento da coleta de evidências para interpretação da habilidade sendo medida, o que pode ser identificado em uma investigação de validade de construto. Estaremos focando três tipos de problemas que podem ser recorrentes no processo de desenvolvimento e uso de testes: (a) itens independentes do texto, (b) itens que fornecem pistas para respostas, e (c) itens de vocabulário (Tumolo e Tomich, 2007, p. 78).

Os autores discutem, então, itens não defensáveis extraídos da prova da Unicamp de 1998. Para ilustrar nossas observações, retomamos os itens considerados não defensáveis, e os recolocamos em seu contexto, trazendo outros elementos para justificá-los. Apresentamos, portanto, os contra-argumentos e as justificativas para esses itens com base em seu construto, que explicitamos a seguir, conforme apresentado no Manual do Candidato à época em que a prova foi elaborada.

A prova de língua estrangeira (inglês) da Unicamp

O formato do atual vestibular da Unicamp[14] foi implantado em 1987. Até essa época, o exame, elaborado pela Fundação Carlos Chagas, era distinto, mais próximo dos demais vestibulares no país. O objetivo de sua reformulação foi selecionar um perfil de aluno considerado desejável pela universidade e promover, ao mesmo tempo, maior interação com o ensino de nível médio; o que se desejava, portanto, era um exame que pudesse ser, pelo menos de maneira potencial, educacionalmente benéfico. O impacto que exames de alta-relevância têm no ensino e na aprendizagem é inegável e foge da alçada dos elaboradores controlarem esse impacto, que, muitas vezes, não é positivo mesmo quando a proposta do exame é boa (vide Scaramucci, 2004).

O novo teste propunha-se a selecionar candidatos que, além de dominarem os conteúdos adquiridos no ensino médio, considerados necessários para o ensino superior, fossem também capazes de organizar suas ideias e exprimir-se com clareza, estabelecer relações, interpretar dados e fatos e elaborar hipóteses. Pela sua própria natureza, minimizaria o incentivo a práticas mecanicistas, de treinamentos e memorização de fórmulas.

É, portanto, dentro dessa filosofia que a prova de inglês foi pensada. Ela focaliza a compreensão em leitura e, como tal, é composta de doze questões, redigidas em português (que, na época, valiam cinco pontos cada uma), elaboradas a partir de textos autênticos de gêneros, assuntos, extensões, fontes e níveis de dificuldade variados. As respostas devem ser redigidas também em *português*, uma vez que seu objetivo é avaliar a compreensão em leitura do candidato e não sua capacidade de redigir textos em inglês. Embora a justificativa para um foco na compreensão seja a necessidade de os candidatos lerem bibliografia em inglês durante sua vida universitária, não se trata aqui da compreensão de textos acadêmicos específicos das áreas em questão. Não se espera de um candidato que não tenha sido exposto aos conceitos de sua área (o que deverá ocorrer na universidade) ser capaz de ler textos acadêmicos específicos em um nível de compreensão adequado (Scaramucci, 1999).

Os textos, portanto, de natureza diversa, oferecem aos candidatos de áreas diferentes possibilidades variadas de construírem sentidos. Os critérios utilizados para a seleção de textos são: diversidade temática, com temas de interesse dos candidatos, sem se restringirem a um único domínio do conhecimento; diversidade de gênero, através de poemas, artigos de jornais e de revistas, artigos científicos e de vulgarização científica dentre outros; autenticidade, isto é, textos não extraídos de livros para o ensino de inglês nem muito menos elaborados especialmente para a prova. Poemas, editoriais e textos de outros gêneros também são incluídos uma vez que é parte do construto que os candidatos sejam capazes de ler e interpretar textos diversos. Em outras palavras, ler esses gêneros é parte do construto avaliado.

O método utilizado para avaliar, como já mencionado, são perguntas de respostas abertas ou disser-

[14] O exame vestibular da Unicamp é realizado em duas fases: a primeira, obrigatória para todos os candidatos, é constituída de uma única prova composta de uma redação e de um conjunto de doze questões gerais sobre o conteúdo das disciplinas do núcleo comum do ensino médio, ou seja, Matemática, Química, Física, História, Geografia e Biologia. A segunda fase, realizada em quatro dias consecutivos, é constituída de oito provas, também de respostas discursivas, das disciplinas Língua e Literaturas de Língua Portuguesa, Biologia, Química, Física, Geografia, História, Matemática e Língua Estrangeira (Inglês ou Francês). A prova de francês foi eliminada na versão atual. A revisão desse formato está atualmente em discussão.

tativas, que permitem uma gama variada de respostas, julgadas como *mais adequadas* ou *menos adequadas* aos objetivos de leitura determinados pelas perguntas, e não apenas *uma resposta correta*, como no caso das questões de múltipla escolha. A escolha desse método é fundamental para garantir a validade do construto avaliado, no caso, a leitura como um processo de construção de sentidos, que reconhece a possibilidade de leituras distintas para um mesmo texto, dependendo dos conhecimentos que cada leitor traz para a tarefa, assim como de leituras distintas para um mesmo leitor, dependendo de sua motivação, interesses, envolvimento e/ou objetivos que tem para a leitura. A substituição de itens de múltipla escolha pelos de resposta aberta minimizaria aspectos irrelevantes ao construto – dificuldades relativas ao método, implicando possivelmente macetes para resolvê-las (Scaramucci, 1999).

A função das perguntas em uma prova dessa natureza é exatamente a de determinar objetivos comuns, que vão delimitar o leque de respostas possíveis, o que é essencial na avaliação, principalmente quando o número de candidatos é muito grande, como é o caso do vestibular da Unicamp, que conta com aproximadamente 13.000 candidatos nessa fase (para uma análise mais detalhada do programa com as especificações da prova, e das questões comentadas, vide Scaramucci, 1998 e Scaramucci e Oliveira, 1999). Podemos dizer, portanto, que o método de avaliar, neste caso, envolvendo itens de respostas abertas, é parte do construto: ser capaz de expressar-se por escrito, *em português*, a respeito do que foi lido é considerado elemento importante da capacidade de ler em inglês, conforme o construto definido para a prova em questão[15]. Reconhecer alternativas em uma prova de múltipla escolha seria incompatível em vários aspectos com a visão de leitura que fundamenta essa prova, na medida em que há apenas manipulação dos elementos lidos e o risco de uma preparação equivocada para esse tipo de item. Em outras palavras, o que estou querendo dizer é que a compreensão somente será confirmada a partir do momento em que o leitor é capaz de expressar-se a respeito do que leu. Isso faz com que o exame seja menos excludente, na medida em que é sensível a níveis mínimos de proficiência.

Para que esse conceito de leitura subjacente não seja, pois, distorcido, a banca, ao elaborar a prova, também elabora, para cada questão, uma grade de correção, que é o conjunto de expectativas para a resposta considerada *mais adequada* (e não correta) em vista dos objetivos propostos. Essa grade (escala de 0 a 5[16], em que cada faixa tem seus descritores) é completada pela banca de correção a partir de uma amostra real de provas, que passa a incluir outras respostas não esperadas. A correção só é iniciada quando todos os níveis são descritos adequadamente.

O programa, que contém as especificações do exame, tem como objetivo explicitar o conceito ou visão de leitura e de linguagem a ele subjacentes ou seu construto. A leitura é vista como um processo ativo (e não mais uma tarefa passiva de decodificação ou de tradução) "que resulta na produção de um texto novo pelo leitor" [...] "diferentes leitores podem produzir diferentes leituras do mesmo texto, o que não significa, em outro extremo, que qualquer leitura possa ser feita" [...] "A leitura pode ser definida como o resultado de uma operação de atribuição de sentido que atua sobre o texto em sua globalidade, recuperando seu funcionamento" (Vestibular Nacional Unicamp, 1996, p. 25). Ela é concebida, portanto, como uma prática que requer capacidades de uso de linguagem e não apenas de conhecimentos sobre ela. Por sua vez, "o texto não é uma soma de frases, mas um todo que se articula" (Vestibular Nacional Unicamp, 1996, p. 26). O leitor é ativo e participante, usando seu conhecimento prévio para interagir com o texto.

A dificuldade das questões não está somente relacionada à dificuldade ou extensão dos textos. Pode-se observar um texto fácil com perguntas difíceis e um texto difícil com questões fáceis; da mesma forma, um texto longo pode ser fácil e um curto, difícil. Assim, a prova tem de ser analisada como um conjunto de questões/textos/respostas/correção.

As questões são de natureza diversa, buscando "mobilizar diferentes aspectos [da competência de leitura do candidato] e diferentes procedimentos frente ao texto. Para responder a essas questões [o candidato] estará trabalhando ora com informações veiculadas no texto, ora com a argumentação que o constitui" (Vestibular Nacional Unicamp, 1996, p. 25). O programa ainda apresenta uma lista de objetivos ou propósitos possíveis para a leitura, que são aqueles determinados pelos itens. Esses objetivos ou propósitos envolvem níveis variados de compreensão, pressupondo sub-habilidades de leitura diversas, que vão desde a simples localização de informações pontuais explícitas e inferência de itens lexicais até a reconstrução da cadeia argumentativa, reconhecimento de relações e contradições entre textos, envolvendo aspectos mais implícitos e inferenciais.

São salientados ainda, no programa, outros elementos que podem ser utilizados na "busca de um significado para o texto", mas que também podem ser avaliados diretamente pelas questões, tais como o autor e o público a que se destina o texto, o contexto sócio-histórico em que

[15] Para McNamara (2000, p. 26), "há duas grandes abordagens para se entender a relação entre o método de avaliar e o conteúdo da avaliação. A primeira vê o método como um aspecto do conteúdo, e levanta questões de autenticidade; a segunda, mais tradicional, trata método independentemente do conteúdo, e permite formatos de resposta mais obviamente não autênticos".

[16] No exame atual, as notas para cada questão são de 0 a 4, com dois itens em cada uma.

foi escrito, sua finalidade, o veículo em que foi publicado, sua configuração gráfica: fotos, ilustrações, gráficos, títulos, dentre outros.

Por ser um vestibular de altíssima demanda e pelo fato de a prova de inglês ser uma das mais discriminatórias entre alunos de diferentes níveis socioeconômicos, são incluídas perguntas de nível variado de dificuldade de forma a permitir distinções finas não apenas entre candidatos de níveis altos e baixos de proficiência, mas, principalmente, entre os candidatos bons e excelentes (em geral, alunos de Medicina, dada a relação demanda/oferta de vagas) e entre os candidatos de níveis baixos de proficiência. Aliás, esse é um aspecto reconhecido por Tumolo e Tomich (2007, p. 70) como positivo:

> A ideia é que um teste, através de seus itens, deve permitir que a pontuação conseguida por cada pessoa reflita seu nível da habilidade, neste caso, a habilidade de leitura em inglês como língua estrangeira. Assim, um teste que permite interpretações mais válidas é um teste que possibilita a diferenciação entre um indivíduo com maior nível da habilidade em questão daquele com menor nível da habilidade.

Itens não defensáveis da prova de 1998

Tendo caracterizado o construto que fundamenta a prova da Unicamp, passamos a seguir a analisar os itens considerados não defensáveis, entretanto, agora com base no construto que a fundamenta. Antes, porém, gostaríamos de observar que, pela amostra de itens considerados defensáveis pelos autores, fica evidente tratar-se de uma concepção distinta de leitura quando comparada àquela que fundamenta o exame da Unicamp, na medida em que, aparentemente, são valorizadas questões pontuais, em pequenos trechos de textos de gêneros semelhantes, sem título e sem contextualização suficiente, todos avaliados por meio de itens de perguntas de múltipla escolha que avaliam mais proficiência linguística e menos proficiência em leitura. Não seria coincidência, portanto, que todos os itens das provas do vestibular da UFSC, com exceção de um, analisados em Tumolo (2005) tenham sido considerados defensáveis.

(a) Itens independentes do texto

Um dos itens apresentados em Tumolo e Tomich (2007) como não defensável foi extraído da prova de inglês do vestibular Unicamp de 1998 (veja Anexo A). Esse item é o segundo de um conjunto de três, elaborado a partir de um trecho de um livro de Philip Ridley intitulado *Mercedes Ice*, que inicia a prova de 1998. Os três itens focalizam os aspectos mais relevantes do texto, sua ideia principal e detalhes, configurando um conjunto de sub-habilidades consideradas importantes dentro do construto que fundamenta a prova. Julgamos, portanto, importante retomar esse conjunto antes de considerar o item não defensável.

O primeiro item (*Quem é quem nessa história*) – vale ressaltar, considerado defensável pelos autores – cumpre uma função procedimental, ou seja, chamar a atenção do candidato para informações relativas aos personagens, oferecendo-lhe condições mais adequadas para responder às perguntas subsequentes. Embora não seja o item mais simples da prova, é importante na medida em que leva o candidato a refletir sobre o funcionamento do texto, auxiliando-o na reconstrução de sua estrutura narrativa. O objetivo da questão é a identificação dos três personagens a partir dos diálogos, relacionando nomes e respectivos papéis na família. A identificação dos nomes é extremamente fácil. Já as relações de parentesco exigirão mais do candidato, que terá de inferi-las através do contexto/palavras e expressões-chave tais como "*her mum*", "*you're as foolish as your father*" e "*Listen, young lady*". O enunciado mais vago "quem é quem", que não explicita ao candidato a existência de relações de parentesco entre os personagens, é exatamente o que torna a pergunta interessante. Para respondê-la adequadamente, portanto, o candidato deveria dizer que Harold é o pai, Doll é a mãe; Rosie é a filha (Scaramucci e Oliveira, 1999, p. 116).

Já o item 2 (*A que se refere "Shadow Point"? Por que recebeu esse nome?*) é criticado sob a alegação de poder ser respondido sem a necessidade de leitura do texto. Em outras palavras, é considerado um *item independente do texto* e pouco confiável, uma vez que "para responder, o leitor pode fazer uso da figura e de processo de inferência, baseado em seu conhecimento de mundo e, portanto, não defensável. Dessa forma, estaria respondendo ao item sem referência ao texto escrito, podendo, assim, fornecer apenas evidências fracas para a interpretação da habilidade em leitura em inglês como língua estrangeira" (Tumolo e Tomich, 2007, p. 79).

Concordamos com os autores – e igualmente com Alderson *et al.* (1995) e Nuttall (*in* Tumolo e Tomich, 2007) – que itens que *apenas* avaliam conhecimento de mundo não sejam defensáveis em uma prova de vestibular. Com a devida argumentação, no entanto, sua inclusão até mesmo poderia ser justificada em um teste de sala de aula, em que o foco do ensino seja, por exemplo, ativação e uso de conhecimento prévio. Entretanto, o item em questão não é "independente do texto" porque, embora a informação visual complemente ou ofereça subsídios para a interpretação, não a esgota.

O item em questão, subdividido em duas partes, visa, na primeira, a recuperação do referente de *Shadow Point* (um prédio, edifício em construção ou outros termos que expressassem a ideia de construção alta e capaz de gerar sombra) e, na segunda, a explicação do motivo pelo qual lhe foi dado esse nome (o fato de estar fazendo sombra na região ao seu redor).

O leitor poderá observar, pelos elementos visuais que acompanham o texto em que se baseia o item em

questão (Anexo A), um prédio que faz sombra em algumas casas. Entretanto, somente conseguirá relacionar *Shadow Point* a esse prédio se souber o que é *Shadow*, o que é *Point*, conhecer a estrutura de um sintagma nominal, relacionando-o ao que está sendo discutido no texto. Se se tratasse de uma simples questão de interpretação dos elementos visuais, todos os candidatos deveriam acertar esse item, o que não ocorreu. Daí a importância de dados sobre o uso do teste, ou seja, sobre as respostas dos candidatos em um processo de validação. Foram comuns, por exemplo, respostas em que houve a identificação do referente com um elemento da realidade, não possibilitado nem pelo texto nem pela figura, tais como Big Ben, o Muro de Berlim, entre outros; ou ainda respostas que, inequivocamente, procederam à identificação do referente com o processo, como, por exemplo: "*Shadow Point* refere-se ao *Big Ben*. Ele recebeu esse nome por ser o maior e mais alto relógio do mundo, tendo grande importância para os ingleses"; "*Shadow Point* se refere a [sic] construção do edifício. Recebeu esse nome por se tornar dia a dia mais alto"[17] (Scaramucci e Oliveira, 1999, p. 117).

Além disso, a capacidade de interpretar informações visuais – que, a propósito, podem ter funções distintas dentro de um texto – e relacioná-las com texto escrito no processo de construção de sentidos é cada vez mais necessária em um mundo em que os letramentos devem ser múltiplos e é, portanto, parte do construto avaliado na prova de inglês da Unicamp. Dessa forma, a conclusão de que o item era pouco confiável por ser independente do texto não procede e, aparentemente, é resultado de uma análise baseada em uma concepção distinta de leitura.

Outro aspecto que não podemos deixar de salientar é que a pergunta em questão tem uma segunda parte (*Por que recebeu esse nome?*), que avalia o entendimento da justificativa ou explicação do referente, que não é tão simples. Nesse caso, podemos dizer que a função da primeira parte seria apenas estabelecer o contexto para que essa pergunta pudesse ser respondida.

(b) Itens que fornecem pistas para as respostas

O item 3 (*O texto menciona mudanças. Que mudanças são essas?*) por sua vez, complementa a anterior, na medida em que focaliza a identificação das mudanças ocorridas no local descrito pelo texto como consequência da sombra causada pelo prédio. Há, nesse caso, um encadeamento de mudanças, que se inicia com a construção do prédio, considerada a primeira mudança ("*the Point got taller and taller*") e o crescimento da sombra ("*the shadow got longer and longer*"), sua consequência. Todas as outras mudanças que se seguiram são, por sua vez, consequências dessas duas primeiras ("*All around flowers died, grass turned brown and rooms became dark and cold. Old people had to turn on heaters, even in the middle of summer*").

A questão pode ser considerada fácil porque não pressupõe uma interpretação mais elaborada, mas sim a mera localização do trecho que menciona as mudanças. Os elementos solicitados envolvem itens lexicais básicos, tais como *trees, flowers*, *got, became, turned, once, now, taller*, *longer,* etc.

Para o obtenção da *nota 5* era, portanto, necessário fazer referência a todos os elementos da resposta (de forma genérica ou com exemplos) deixando clara a ideia de encadeamento desses elementos. São exemplos de *notas 5*: "Com o crescimento desse grande edifício, as flores ao seu redor morreram, a grama ficou amarronzada e os quartos tornaram-se escuros e frios. Pessoas idosas tinham que ligar os aquecedores até em pleno verão. Isto ocorreu devido à enorme sombra projetada por esse edifício."; "A medida que o *Shadow Point* crescia, ia mudando o cotidiano das pessoas. Assim, a sombra que o *Shadow Point* gerava fez com que os mais velhos tivessem de ligar o aquecedor mesmo no meio do verão. As flores atingidas pela sombra morreram, a grama tornou-se marrom e toda a área se tornou fria e escura" (Scaramucci e Oliveira, 1999, p. 117).

Esse item também é criticado como não defensável sob a alegação de que "fornece, na língua mãe, informações para sua resposta. Essas informações sobre mudanças, juntamente com a figura fornecida com o texto, já fornecem subsídios para a resposta, permitindo, assim, que o item seja respondido, em grande parte, com base nas informações do próprio item dadas na língua mãe, não na língua-alvo" (Tumolo e Tomich, 2007, p. 80).

Entretanto, "fornecer pistas para as respostas" é parte do construto da prova da Unicamp. A inclusão dessas pistas é deliberada, na medida em que só levarão à construção de sentidos se forem entendidas pelos leitores como pistas e puderem ser "decodificadas". O que queremos dizer é que o "contexto" para inferência desses elementos não "está no texto", mas é construído a partir dos conhecimentos diversos que têm de ser mobilizados de forma ativa pelo leitor. Além do mais, estão à disposição de todos os candidatos. Não seria, portanto, um exemplo do que salienta Popham (*in* Tumolo e Tomich 2007, p. 79) como "pistas não intencionadas" (*unintended clues*), mas de "pistas intencionadas". Não seriam, como ressalta Messick (1989) (*in* Tumolo e Tomich 2007, p. 79), pistas externas que trazem, para a

[17] Também foi aceita outra variante dessa resposta, permitida pela pergunta, em que o nome *Shadow Point* é uma decisão tomada por Doll em função do papel da sombra e das consequências que provoca.

resposta, facilidade irrelevante ao construto, permitindo que alguns indivíduos respondessem corretamente. Não se trata de uma facilidade irrelevante ao construto, mas de uma facilidade relevante ao construto.

Mesmo com essas pistas, muitos, muitos candidatos erraram a resposta. Portanto, o item mostrou-se capaz de discriminar entre os que conseguem usar as pistas e aqueles que não conseguem usá-las, e, portanto, cumpre com seu papel na configuração das sub-habilidades avaliadas na prova de acordo com seu construto. Portanto, a crítica não se justifica quando o construto correto é utilizado.

Ainda dessa mesma prova de 1998, também o item 7 é considerado não defensável sob a mesma alegação, ou seja, oferecer pistas para as respostas. Nas palavras dos autores, "a pergunta (*De que maneira a violência urbana pode estar afetando a saúde de pessoas idosas*?) contém a proposição de que a violência urbana pode estar afetando a saúde das pessoas idosas... além das pistas fornecidas na língua-mãe para a resposta, por um processo de inferência baseado no conhecimento de mundo, pode-se chegar à resposta satisfatória" (Tumolo e Tomich, 2007, p. 80).

Esse item é o único elaborado a propósito de um texto extraído da revista *New Scientist* que, embora curto, é interessante pela tese que apresenta: a saúde de pessoas idosas está sendo prejudicada, de maneira indireta, pela violência urbana. A pergunta busca recuperar exatamente os elementos que dão sustentação a essa tese, que deverão estar encadeados na mesma ordem em que são apresentados no texto. Para isso, o candidato teria de entender o funcionamento do texto como um todo. Embora este seja relativamente redundante e contenha muitas palavras cognatas, é necessário que o candidato compreenda a lógica da argumentação, não bastando entender trechos isolados. Apesar de o texto mencionar exemplos, na resposta adequada há necessidade de se fazer uma generalização a partir desses exemplos. Para a atribuição da *nota 5*, é necessário incluir os seguintes elementos, que devem estar encadeados de forma adequada: (a) medo de sair de casa por causa das gangues de rua (ou violência urbana). Esse elemento foi muitas vezes inferido quando o candidato dizia que havia um impedimento/inibição/etc. das caminhadas devido à violência urbana ou ao medo; (b) interrupção da caminhada. Esse elemento foi inferido quando a resposta mencionava interrupção de uma atividade física. Algumas vezes, esse elemento não apareceu de modo explícito, mas estava claramente subentendido como o elo entre medo (ou medo de caminhar) e controle de peso; (c) controle de peso. Esse elemento pode aparecer exatamente como expresso no texto: a interrupção do exercício físico provoca aumento de peso, ou reescrito: exercício físico (caminhada) proporciona controle/diminuição de peso ou manutenção da forma (física); (d) agravamento das doenças. As seguintes respostas são exemplos de nota 5: "Para pessoas hipertensas e diabéticas é fundamental fazer caminhadas para evitar o ganho de peso. As gangues de rua amedrontam as pessoas, inclusive os idosos citados no texto consequentemente, estas deixam de fazer caminhadas, ganhando peso e agravando seu quadro clínico. Assim, com as gangues de rua, aumenta-se o índice de doenças, em pessoas idosas, como a diabetes e doenças cardiovasculares"; "Pessoas idosas que precisam fazer caminhadas regulares têm medo de fazê-las, por causa da violência urbana, principalmente das gangues. Parando de fazer exercícios elas agravam problemas de saúde como hipertensão, diabetes e ainda ganham peso (engordam)" (Scaramucci e Oliveira, 1999, p. 121-122).

O foco desse item, portanto, não é a identificação, no texto, de que a violência urbana afeta a saúde das pessoas, porque essa informação já é dada, de forma proposital, pela pergunta. O que se procura resgatar, nesse caso, é exatamente a explicação para esse fato, que é muito mais interessante. A primeira parte do item tem uma função procedimental, estabelecendo uma base comum e um caminho de entrada para o texto, definindo claramente o objetivo de leitura, necessário para que a resposta possa ser mais bem avaliada e considerada adequada em testes de respostas abertas. Segundo Tumolo e Tomich (2007), essa "falha", ou seja, fornecer pistas, explicaria por que o item foi considerado o mais fácil nas estatísticas baseadas nas pontuações, e mais fácil ainda para os futuros estudantes de Medicina[18]. Itens dessa natureza têm sua função na prova. Como salientamos antes, cumprem o papel de discriminar de forma mais "fina" entre níveis mais baixos de proficiência. Apesar de fácil, não foi respondido por todos. Mais uma vez, as pistas não podem ser consideradas ameaças à validade de construto, conforme salientam os autores citando Messick (1989), pois as evidências ou pistas em português não são irrelevantes ao construto mas sim, parte dele.

(c) Itens de vocabulário

Uma terceira categoria de itens não defensáveis seriam itens "que medem conhecimento de vocabulário, *independentemente do contexto em que está inserido e de possíveis processos de inferência*" [ênfase nossa] (Tumolo e Tomich, 2007, p. 80).

[18] Tumolo (2005) explica a maior facilidade da questão entre os alunos da área de Biológicas (a partir de um comentário em Scaramucci e Oliveira, 1999, p. 122) pelo fato de o texto versar sobre um assunto da área médica o que, na opinião do autor, privilegiaria candidatos dessa área. Discordamos dessa análise, visto tratar-se de um texto de vulgarização científica, que não pressupõe conhecimentos específicos de área. O fato de ter sido mais fácil para os alunos da área de Biológicas já era esperado e significa que o item realmente discriminou, uma vez que esses alunos são, em geral, excelentes candidatos, com altos níveis de proficiência em leitura em inglês.

Também concordo com os autores que avaliar vocabulário independentemente do contexto e de possíveis processos de inferência não é avaliação de leitura, mas de vocabulário. Sem dúvida alguma, "conhecimento de vocabulário é essencial e tem uma correlação positiva com a habilidade de leitura (Tumolo e Tomich, 2007, p. 81). "Medir conhecimento de vocabulário para a interpretação da habilidade de leitura não é, porém, um procedimento adequado, já que conhecer determinadas palavras não garante a leitura, como também não conhecer determinadas palavras não impede a leitura" (Tumolo e Tomich, 2007, p. 85). "Deve, porém, ser considerado distinto do construto de leitura, por ter características próprias, podendo, até mesmo, ser considerado um construto pré-requisito" (Tumolo e Tomich, 2007, p. 81). Essa distinção, todavia, requer uma análise mais precisa e cuidadosa (para considerações mais aprofundadas sobre o papel do vocabulário na leitura, vide Scaramucci, 1995).

O item citado no artigo em questão para ilustrar essa categoria foi extraído de uma prova da Unicamp de 2000. Em vez de discutir esse exemplo, entretanto, prefiro usar outro, extraído da prova de 1998 (item 6), utilizado em Tumolo (2005), uma vez que o leitor poderá avaliá-lo melhor dentro do contexto da prova, que se encontra no Anexo A. O item em questão era o terceiro de um conjunto de três, elaborados a propósito de um trecho de um texto mais longo sobre geofagia, extraído da página da revista *Nature,* intitulado *The soil-eaters.* O vocabulário do texto pode ser considerado fácil, rico em palavras cognatas, prefixos e sufixos.

O item (*Dê um significado para a palavra "but" no trecho "[...] on the whole [soil eaters] are regarded as quite 'normal' to most but outsiders")* focalizava o sentido menos comum de *but* (*exceto*, *exceção* ou *exclusão*). Mesmo desconhecendo o significado da palavra, o leitor poderia responder adequadamente à questão, pois teria a chance de inferi-la pelo contexto. Embora focalize apenas um item lexical, a pergunta não deixa de ser uma questão de leitura, uma vez que para respondê-la o leitor terá que interpretar adequadamente tanto o trecho que a contém como o que vem em seguida – focalizado na pergunta anterior (item 5), comentada mais adiante. Portanto, o entendimento correto de *but* é fundamental para que o leitor possa relacionar que a geofagia é comum em seu contexto mas "desconhecida, pouco relatada, mal compreendida ou ignorada pela maioria das pessoas" fora do seu ambiente, que são exatamente as pessoas no mundo desenvolvido, como mostra o trecho a seguir:

> The reasons for soil consumption are many and often misunderstood, say the researchers Peter Abrahams and Julia Parsons. *But geophagists – as soil-eaters are known – on the whole are regarded as quite 'normal' to most but outsiders* [*ênfase nossa*]. "Despite the widespread distribution of geophagy, both today and in the past, it is largely unknown, under-reported, misunderstood or ignored by most people in the developed world", say Abrahams and Parsons. [This is why] "the adjectives 'eccentric', 'perverted', 'odd', and 'bizarre' have all been applied to geophagy" [...].

Portanto, não se trata apenas de um item de avaliação de vocabulário, mas sim de um item de leitura que, mais uma vez, procura chamar a atenção do leitor para algo que lhe poderia passar despercebido por julgar já conhecê-lo (sentido mais frequente de *but*). Trata-se, portanto, de mais um item procedimental, que faz com que o candidato analise o trecho e possa, com isso, relacioná-lo ao que vem a seguir. Portanto, a função desse item na prova está inteiramente de acordo com a afirmação de Tumolo e Tomich:

> o processo de inferência de significado de palavras desconhecidas pode ter uma contribuição positiva fundamental e palavras desconhecidas podem ter seu significado inferido em contexto com auxílio de outras ao redor (Tumolo e Tomich, 2007, p. 82).

Em provas de entrada e proficiência, em que o foco geralmente é o produto da compreensão e não seu processo, itens que focalizam inferência de palavras pelo contexto, apesar de considerada uma sub-habilidade importante para a leitura, podem não se justificar porque o foco, nesse caso, não seria em *como* o leitor chegou a um sentido, mas *se* chegou a um sentido. Outro argumento muitas vezes citado para não incluí-los é que nunca teremos garantias do que o item está realmente avaliando (vide Alderson e Lukmani (1989), para uma discussão interessante a respeito de objetivos em perguntas de avaliação de leitura): ele pode funcionar como um item de leitura para o candidato que desconhece o sentido da palavra e usou as pistas para inferi-lo; ou como um item de conhecimento de vocabulário para aquele que já sabe o que a palavra significa. Entretanto, não seria esse o caso aqui, visto tratar-se de um item pouco frequente: as chances de os candidatos não conhecerem esse sentido são grandes, fazendo com que tenham que recorrer às pistas do contexto. Dada a correlação entre conhecimento de vocabulário e leitura (bons leitores geralmente têm um bom vocabulário) imaginamos que os poucos que conheçam esse sentido menos frequente tenham também grandes chances de serem bons leitores.

(d) Itens de natureza diversa

Havia mais dois itens (4 e 5) elaborados a propósito do mesmo texto. O primeiro deles – também considerado não defensável em Tumolo (2005) – (*O primeiro parágrafo se dirige a um público-leitor específico. Que público é esse? Justifique sua resposta*) chamava a atenção para um aspecto pouco trabalhado na escola, apesar de extremamente desejável, que é a relação entre as estratégias discursivas usadas pelo autor e o universo referencial do leitor presumido. O objetivo específico era identificar o público leitor que o autor tinha em mente, justificando a resposta através dos elementos de contextualização fornecidos no primeiro parágrafo do texto, a saber, familiaridade ou conhecimento de hábitos urbanos relacionados

à alimentação: restaurantes típicos, *fast food* e entrega a domicílio, conforme trecho a seguir:

> It's lunchtime somewhere in rural tropical Africa. You're hungry, but the nearest restaurant is too far to walk. There's no Italian, Chinese, Indian or fast food and the telephone pizza delivery company is a little reluctant to send its dispatch rider beyond the city walls.
> Moreover, you're on a tight budget. What are you to do? The answer, quite literally, may lie in the soil directly beneath your feet (Scaramucci e Oliveira, 1999, p. 118).

Para o autor, esse item não seria defensável porque

> O você hipotético (que se refere ao público que o escritor tem em mente, no Caderno de Questões Comentadas), não é geralmente usado em textos acadêmicos, e portanto, não necessariamente parte do critério de referência usada neste estudo. Para que ele pudesse ser identificado, o candidato teria que esquecer a informação normalmente usada para a identificação do leitor potencial, relevante para os estudos universitários, tais como o veículo usado para publicação, o registro, a função, o tópico. No caso, a conclusão seria que o público leitor seria alguém com acesso à internet, uma vez que o texto tem o formato de um texto publicado nesse meio, alguém que lê revistas de ciência, não alguém hipotético que está em uma área rural da África e é familiar com todas as comodidades da vida moderna, tais como restaurantes de *fast food* (Caderno de Questões Comentadas) [...].
> Esse é outro item cujo foco, evidentemente, não é trabalhado na escola secundária (Caderno de Questões Comentadas), que, juntamente com o fato de não ser a inferência usual para a identificação do público alvo, pode explicar o nível de dificuldade representada por 70% de notas zero. Se não trabalhado previamente, e não parte do critério, por que usar esse tipo de questão? O que está sendo determinado através de questões como estas? Há argumentos fortes para não usar esse item para a obtenção de evidências para a inferência da capacidade de leitura (Tumolo, 2005, p. 207-208).[19]

Vários são os aspectos que merecem ser questionados nessa longa argumentação, todos eles motivados pelo uso de um construto que é distinto daquele que fundamenta a prova da Unicamp. A afirmação de que "o você hipotético não é geralmente usado em textos acadêmicos, e, portanto, não necessariamente parte do critério de referência usado neste estudo", é equivocada, porque o que interessa é que seja parte do critério de referência do *exame em questão*, que não pressupõe a leitura de textos acadêmicos. Como já havíamos salientado anteriormente, o programa, que contém as especificações, é claro nesse sentido e "[inclui] outros elementos [...] tais como o autor e o público a que se destina o texto, o contexto sócio-histórico em que foi escrito, sua finalidade, o veículo em que foi publicado, sua configuração gráfica: fotos, ilustrações, gráficos, títulos, dentre outros" (Scaramucci, 1999, p.10). Portanto, a crítica não se justifica.

Além disso, discordamos da afirmação de que "o leitor teria que esquecer a informação normalmente usada para a identificação do leitor potencial" [...] tais como o veículo usado para publicação, o registro, a função, o tópico, o veículo em que foi publicado". Embora a informação, por exemplo, do veículo usado para publicação possa ser usada, ela é insuficiente, neste caso. Outros elementos, relacionados ao tópico – familiaridade ou conhecimento de hábitos urbanos relacionados à alimentação, restaurantes típicos, *fast food* e entrega a domicílio -- são aqui mais produtivos. Dessa maneira, a resposta sugerida pelo autor, ou seja "alguém com acesso à internet, uma vez que o texto tem o formato de um texto publicado nesse meio, alguém que lê revistas de ciência, [e] não alguém hipotético que está em uma área rural da África e é familiar com todas as comodidades da vida moderna, tais como restaurantes de *fast food* (Caderno de Questões Comentadas)" não seria adequada porque pressupõe uma interpretação equivocada (Tumolo, 2005, p. 206). Portanto, para a obtenção da nota 5, era necessário incluir a identificação do público leitor não apenas como "alguém com acesso à internet" (o que seria óbvio e incompleto) ou como "alguém hipotético que se encontra em uma área rural da África e é familiar com todas as comodidades da vida moderna, tais como restaurantes de *fast food*" (o que seria equivocado, porque esse público, na realidade, não está na área rural da África), mas como um público urbano, que desconhece a geofagia e precisa ser contextualizado através das coisas que lhe são familiares. Como o texto aborda um hábito alimentar desconhecido, usa, para contextualizar seu público leitor, hábitos alimentares conhecidos por esse público. Portanto, respostas que afirmaram tratar-se apenas de um público com acesso à internet foram consideradas inadequadas, como mostram os exemplos de nota 0 mais adiante.

São exemplos de *nota 5*: "O primeiro parágrafo se dirige a um leitor que vive nas grandes cidades porque restaurantes italianos, chineses, *fast-food* ou mesmo um disque-pizza são característicos de um meio urbano"; "Este parágrafo dirige-se a um público leitor do "mundo

[19] "The hypothetical you (referred to as the public that the writer has in mind, in the Review Book) is not usually used in academic texts, thus not necessarily in the criterion of reference used in this study. For it to be identified, the test taker must forget the usual information used for the identification of the target reader, rather relevant for university studies, such as the media used for publication, the register, the function, the topic. In this case, the conclusion would be that the target reader is someone with access to the Internet in that it has the format of a text published on it, someone who reads science magazine, not someone hypothetical who is in a rural area in Africa and is familiar with all commodities [sic] of modern life, such as fast food restaurants (Review Book). This is another item whose focus is, admittedly, not worked upon in the secondary school (Review Book), which, together with the fact that it is not the usual inference for the identification of the target reader, might account for the level of difficulty found represented by 70% of score zero (Review Book). If not worked upon previously, and not in the criterion, why to use this kind of question? What is being determined through the use of questions such as these? This item seems to have stronger argument against its use to provide evidence converging to the inference of reading ability" (Tumolo, 2005, p. 207-208).

civilizado", isto é, das cidades, pois são mencionados restaurantes de várias nacionalidades e também serviço de entrega, coisas típicas das cidades"; "Esse parágrafo se dirige a pessoas de países desenvolvidos que não conhecem a realidade de fome em determinadas regiões do planeta como na África tropical. Isso pode ser percebido porque ele se dirige a um público que tem facilidades para obter alimentos, podendo, por exemplo, pedir uma pizza pelo telefone e não conhece sobre aqueles que comem terra para sobreviver" (Scaramucci e Oliveira, 1999, p. 119).

Sem dúvida alguma, esse era um item considerado difícil na prova, elaborado para fornecer informações que possibilitassem uma discriminação mais fina entre candidatos de níveis de proficiência mais alta, como já dissemos, necessária quando a população de candidatos é muito heterogênea. Os exemplos de respostas nota 0 a seguir mostram essa heterogeneidade, quando comparadas com as respostas de nota 5 acima: "O público norte-americano que come *fast-food*"; "Pessoas ricas, de primeira classe"; "O parágrafo se dirige a pessoas que utilizam a *Internet*"; "O parágrafo se dirige a turistas que estão na zona rural da África, na hora do almoço, e não sabem ficar sem um hambúrguer ou uma pizza"; "O parágrafo se dirige aos habitantes da África que comem terra pois não tem restaurantes, *fast-food* e *disk-pizzas* por perto"; "As pessoas que estão engordando, pois ele sugere que elas não vá [sic] a Restaurantes Chinês, Italiano, Indiano e joguem fora os telefones das pizzarias e Serviços de entrega de comida a domicílio. Sugerindo que estas façam uma dieta saudável"; "Ao público Italiano, Chinês e Indiano. O público Italiano pede pizza pelo telefone, os chineses gostam de restaurantes e os Indianos gostam muito de Literatura"; "Aos literários. Porque na última frase o autor se dirige diretamente ao público chamando-os de "*quite literally*"; "São os turistas, porque além de o texto apresentar as comidas mais conhecidas do mundo e que não existem nas savanas africanas, o texto ainda menciona o aperto dentro de um veículo" (Scaramucci e Oliveira, 1999, p. 119).

O segundo item (*Qual é a explicação de Abrahams e Parsons para o uso de adjetivos como "eccentric", "perverted", "odd" e "bizarre" para caracterizar a geofagia*?), por sua vez, considerado defensável em Tumolo (2005), tinha por objetivo identificar uma relação de causa e efeito entre dois trechos do texto, sinalizada pelo articulador lógico *why* juntamente com o elemento anafórico *this*. A questão, que se mostrou de dificuldade média, pressupunha também o conhecimento dos prefixos negativos *un-, under-* e *miss-*, e de itens lexicais em sua maioria básicos ou cognatos. Apesar de ser muito localizada, não exigindo o processamento do texto como um todo, pressupunha o reconhecimento de vários elementos coesivos, que também é uma sub-habilidade importante e parte do construto avaliado na prova em questão (Scaramucci e Oliveira, 1999, p. 119).

A prova, conforme cópia no Anexo A, envolvia ainda mais 5 itens, elaborados a propósito de dois textos, dos quais apenas um item havia sido considerado defensável em Tumolo (2005). Para não nos estender demasiadamente, não comentaremos esses itens. O que merece ser salientado, entretanto, é que enquanto as críticas aos itens iniciais foram relativas, em grande parte, às "facilidades" por eles introduzidas, as críticas aos itens finais foram relativas às "dificuldades", o que mostra, efetivamente, que a prova incluía itens de natureza e dificuldades diversas, conforme previsto pelo construto em que se fundamenta, e teria, portanto, de ser avaliada no conjunto de seus itens e textos.

Considerações finais

Nosso objetivo, neste artigo, foi mostrar que a classificação de um item em defensável e não defensável apenas com base no julgamento do especialista, conforme conduzida em Tumolo e Tomich (2007) e Tumolo (2005), não pode ser considerada definitiva, mas provisória e parcial até que outros argumentos a refutem. Procuramos mostrar, ainda, o que é reconhecido pelos próprios autores, que "validade é uma questão de grau, não é uma coisa de tudo ou nada. A validade da interpretação e a ação baseada nos itens serão relativos à força da argumentação, sendo mais válido ou menos válido em relação a todos os argumentos apresentados" (Tumolo, 2005, p. 236). A conclusão, portanto, de que a prova da Unicamp contém "muitas fontes de invalidade" (Tumolo, 2005, p. 217) é no mínimo precipitada.

Ao questionar a análise conduzida, nosso argumento principal foi mostrar que cada teste tem seu construto, definido na fase de elaboração do instrumento. E é com base nesse construto que a defensibilidade dos itens deve ser julgada: em outras palavras, em que medida o instrumento é a operacionalização desse construto. Portanto, concluímos que a análise dos autores é equivocada, na medida em que considera um construto externo para avaliar os itens da prova de inglês da Unicamp.

Para finalizar, não poderíamos deixar de ressaltar a importância de se levar em conta, em um processo de validação, a situação de uso do teste e as evidências de seus resultados, conforme proposto por Messick e seguido por muitos outros autores. Como salienta Bachman (1990, p. 279) "testes não são desenvolvidos e usados em um tubo de ensaio psicométrico, despido de valores; eles são sempre elaborados para servir às necessidades de um sistema educacional ou da sociedade como um todo"[20].

[20] "Tests are not developed and used in a value-free psychometric test-tube; they are virtually always intended to serve the needs of an educational system or society at large" (Bachman, 1990, p. 279).

Referências

ALDERSON, J.C.; CLAPHAM, C.; WALL, D. 1995. *Language test construction and evaluation*. Cambridge, Cambridge University Press, 310 p.

ALDERSON, J.C.; LUKMANI, Y. 1989. Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language*, **5**(2):253-270.

ARAÚJO, K. da S. 2007. *A perspectiva do examinando sobre a autenticidade de avaliações em leitura em língua estrangeira*. Campinas, SP. Dissertação de Mestrado. Universidade Estadual de Campinas – Unicamp, 153 p.

AERA/APA/NCME. 1985. *Standards for Educacional and Psychological Testing*. Washington, DC, American Psychological Association, 98 p.

BACHMAN, L. 1990. *Fundamental considerations in language testing*. Oxford, Oxford University Press, 408 p.

BACHMAN, L. 2005. Building and supporting a case for test use. *Language Assessment Quarterly*, **2**(1):1-43.

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

CHAPELLE, C.A. 1994. Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, **10**(2):157-187.

CHERRYHOLMES, C. 1988. *Power and criticism: poststructural investigations in education*. New York, Teachers College Press, 223 p.

EBEL R.L.; FRISBIE, D.A. 1991. *Essentials of Educational Measurement*. 5ª ed., Englewood Cliffs, Prentice-Hall, 622 p.

GAGNÉ, E.D.; YEKOVICH, C.W.; YEKOVICH, F.R. 1993. *The cognitive psychology of school learning*. New York, Harper Collins College Publishers, 512 p.

GRONLUND, N.E. 1985. *Measurement and Evaluation in Teaching*. 4ª ed., New York, Macmillan. Pub. Co.; London, Collier Macmillan, 597 p.

HAMP-LYONS, L.; LYNCH, B. 1998. Perspectives on validity: A historical analysis of language testing conferences. *In:* A. KUNNAN (ed.), *Validation in language assessment.* Mahwah, L. Erlbaum, p. 253-277.

HENNING, G. 1987. *A Guide to language testing*. Cambridge, Cambridge University Press, 198 p.

HUGHES, A. 1989. *Testing for language teachers*. Cambridge, Cambridge University Press, 172 p.

HUTCHINS, J. 1987. *Meaning: the frontier of informatics. Informatics 9. Proceedings of a conference jointly sponsored by Aslib, the Aslibi Informatics Group, and the Information Retrieval Specialist Group of the British Computer Society, King's College Cambridge*. London, Aslib, p. 151-173.

KANE, M.T. 2001. Current concerns in validity theory. *Journal of Educational Measurement*, **38**(4):319-342.

KANE, M.T. 1992. An argument-based approach to validity. *Psychological Bulletin*, **112**(3):527-535.

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

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

MCNAMARA, T.; ROEVER, C. 2006. *Language Testing: The social dimension.* Blackwell Publishing Limited, 284 p.

MCNAMARA, T. 2000. *Language Testing*. Oxford, Oxford University Press, 140 p.

MEHRENS, W.A. 1997. The consequences of consequencial validity. *Educational Measurement: Issues and Practice*, **16**(2):17-18.

MESSICK, S. 1989. Validity. *In:* R. L. LINN (ed.), *Educational measurement*. 3ª ed., New York, American Council on Education e Macmillan, p. 13-103.

MISLEVY, R.J.; STEINBERG, L.S.; ALMOND, R.G. 2003. On the structure of the educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, **1:**3-62.

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

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

SCARAMUCCI, M.V.R. [s.d.]. Validade e consequências sociais das avaliações em contextos de ensino/aprendizagem de línguas. [em preparação].

SCARAMUCCI, M.V.R. 2007. Validade e consequências sociais da avaliação de desempenho no contexto de línguas. Campinas, SP. Projeto de pós-doutorado, FAPESP, 19 p.

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. 1999. Vestibular e ensino de língua estrangeira (Inglês) em uma escola pública. *Trabalhos em Linguistica Aplicada,* **34:**7-29.

SCARAMUCCI, M.V.R. 1998. Língua Estrangeira (Inglês). *Caderno de Questões. A Unicamp comenta suas provas.* Comvest, Unicamp, p. 99-109. Disponível em: http://www.comvest.unicamp.br/vest_anteriores/1998/download/comentadas/LEstrangeira.pdf. Acesso em: 01/02/2009.

SCARAMUCCI, M.V.R. 1995. *O papel do léxico na compreensão em leitura em língua estrangeira: foco no produto e no processo*. Campinas, SP. Tese de Doutorado. Universidade Estadual de Campinas - Unicamp, 345 p.

SCARAMUCCI, M.V.R.; OLIVEIRA, P.S. 1999. Língua Estrangeira (Inglês). *Caderno de Questões. A Unicamp comenta suas provas*. Comvest, Unicamp, p. 112-128. Disponível em http://www.comvest.unicamp.br/vest_anteriores/1999/download/comentadas/LEstrangeira.pdf. Acesso em: 01/02/2009.

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

THORNDIKE, R.L.; HAGEN, E.P. 1986. *Measurement and Evaluation in Psychology and Education*. New York, Macmillan, 544 p.

TUMOLO, C.H.S.; TOMICH, L.M.B. 2007. Avaliando a leitura em inglês: uma reflexão sobre itens de testes. *Revista Brasileira de Linguística Aplicada*, 7(2):67-88.

TUMOLO, C.H.S. 2005. *Assessment of reading in English as a foreign language: investigating the defensibility of test items*. Florianópolis, SC. Tese de Doutorado. Universidade Federal de Santa Catarina – UFSC, 314 p.

VESTIBULAR NACIONAL UNICAMP. 1996. *Manual do candidato*. Comvest, Unicamp.

WIGGINS, G.P. 1993. *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, Jossey-Bass Publishers, 316 p.

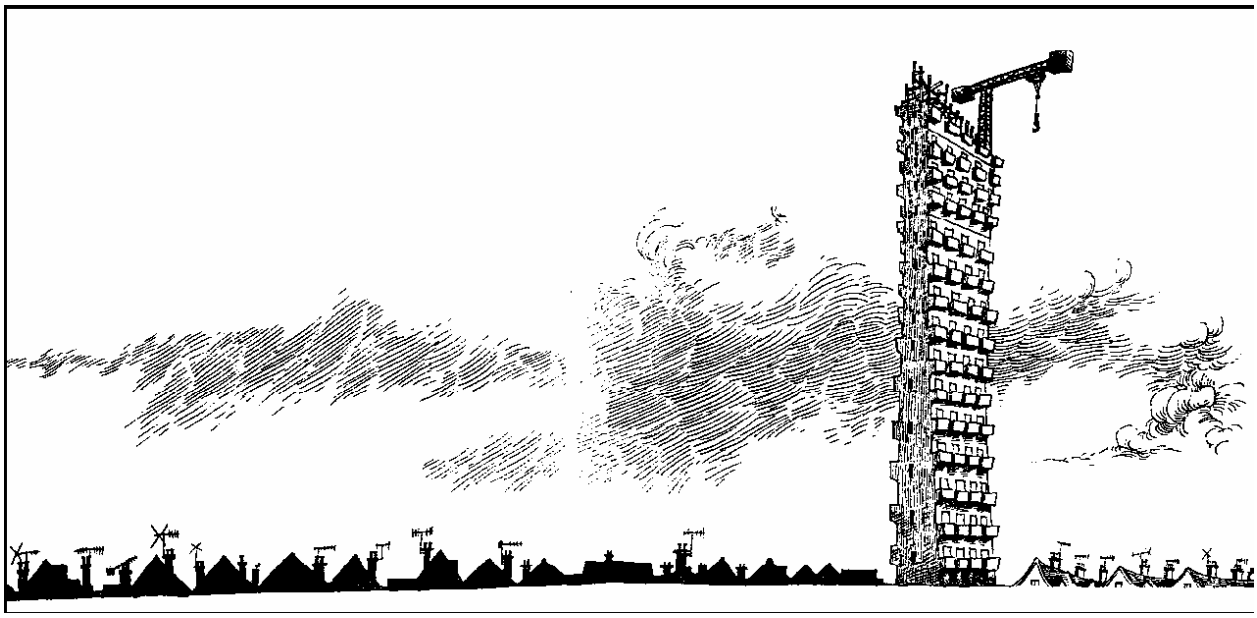
Submetido em: 12/03/2009
Aceito em: 30/03/2009

Matilde V. R. Scaramucci
Universidade Estadual de Campinas
Rua Sérgio Buarque de Holanda, 571
13083-859 - Campinas, SP, Brasil

Anexo A: Prova do exame vestibular de inglês da Unicamp de 1998

Todas as perguntas deverão ser respondidas em português.

Leia o trecho abaixo e responda às questões 1, 2 e 3.



Day by day the Point got taller and taller. And day by day the shadow got longer and longer.

All around flowers died, grass turned brown and rooms became dark and cold.

Old people had to turn on heaters, even in the middle of summer.

'It's just so ugly,' said Doll to Harold as they ate dinner one night.

'Once I used to look out of the window and see trees and flowers, hear singing birds. Now all I see is that ugly grey thing. There're no flowers, no trees, no light, no grass, no birds, nothing.'

'Oh, it's not that bad,' said Harold.

'Don't give me that,' snapped Doll. 'You don't have to watch it. Day in and day out. Watch it getting bigger and bigger and bigger.'

Rosie sat at the table and ate her dinner. She thought her mum was being stupid, although she didn't say so. Instead, she just filled her mouth with a forkful of mashed potato and stared at her plate.

Later, though, while Doll was washing up, Rosie couldn't help saying, 'I don't think it's ugly.' 'Well, you're as foolish as your father, then.' 'I just think it's . . it's a gigantic finger pointing up to the sky. Or a tall flower. Or a wonderful steeple –´

'Listen, young lady,' interrupted Doll. 'It's not a finger and it's not a flower and it's not a steeple. It's just a shadow. Nothing else. It's just a point of shadow.'

And that was how the Point became known as Shadow Point.

(*Philip Ridley. Mercedes Ice. London, Puffin Books. 1996, p. 18-19*)

1. Quem é quem nessa história?

2. A que se refere "Shadow Point"? Por que recebeu esse nome?

3. O texto menciona mudanças. Que mudanças são essas?

As questões 4, 5 e 6 dizem respeito ao texto abaixo.

nature science update
[Update] [Next Article]
The soil-eaters
By Ehsan Masood

It's lunchtime somewhere in rural tropical Africa. You're hungry, but the nearest restaurant is too far to walk. There's no Italian, Chinese, Indian or fast food and the telephone pizza delivery company is a little reluctant to send its dispatch rider beyond the city walls.

Moreover, you're on a tight budget. What are you to do? The answer, quite literally, may lie in the soil directly beneath your feet.

According to two researchers from the University of Wales at Aberystwyth, UK, the tradition of soil consumption is still very much alive in the African tropics, India, Jamaica and it has also been reported in Saudi Arabia. Despite the advent of modern religions and the end of the slave trade, soil eating is not uncommon, though mostly confined to the poorer sections of society.

The reasons for soil consumption are many and often misunderstood, say the researchers Peter Abrahams and Julia Parsons. But geophagists – as soil-eaters are known – on the whole are regarded as quite 'normal' to most but outsiders.

"Despite the widespread distribution of geophagy, both today and in the past, it is largely unknown, underreported, misunderstood or ignored by most people in the developed world", say Abrahams and Parsons. [This is why] "the adjectives 'eccentric', 'perverted', 'odd', and 'bizarre' have all been applied to geophagy".[...]

(*Nature News Service, 1996*)

4. O primeiro parágrafo se dirige a um público-leitor específico. Que público é esse? Justifique sua resposta.

5. Qual é a explicação de Abrahams e Parsons para o uso de adjetivos como *"eccentric"*, *"perverted"*, *"odd"* e *"bizarre"* para caracterizar a geofagia?

6. Dê um significado para a palavra "*but*" no trecho "*…on the whole [soil eaters] are regarded as quite 'normal' to most but outsiders*".

Leia o texto abaixo e responda à questão 7.

A sidelight on urban violence in the US could also be showing up a similar situation in some parts of the UK. A doctor in Arkansas has pointed out that the rise of street gangs is affecting preventive medicine for elderly people. He mentioned two patients of his, both in their early 60s, one with hypertension and the other with diabetes. Both took regular walks of a mile or two several times a week, but they have become too frightened of street gangs to go out.

Their walks ceased several months ago. Consequently both had gained about 10 pounds in weight, not a good thing for either condition. So street gangs, apart from the obvious damage they can cause, might also be worsening cardiovascular disease and diabetes in the elderly. I do not know whether anyone has noticed gains in weight for the same reason among elderly patients in some parts of London, for example.

Bill Tidy

(*New Scientist,* 28 September 1991)

7. De que maneira a violência urbana pode estar afetando a saúde de pessoas idosas?

Leia os dois textos abaixo, da seção *Letters*, e responda às questões 8, 9, 10 e 11.

MURPHY WAS A PERFECTIONIST

As the son of the man whose name is attached to "Murphy's law," I want to thank you for accurately and respectfully identifying the origin of this "law" in your recent article ["The Science of Murphy's Law," by Robert A.J. Matthews, April]. My father was an avid reader of *Scientific American*, and I can assure you that were he still alive, he would have written to you himself, thanking you for a more serious discussion of Murphy's Law than the descriptions on the posters and calendars that treat it so lightly.

Yet as interesting as the article is, I suggest that the author may have missed the point of Murphy's Law. Matthews describes the law in terms of the probability of failure. I would suggest, however, that Murphy's law actually refers to the CERTAINTY of failure. It is a call for determining the likely causes of failure in advance and acting to prevent a problem before it occurs. In the example of flipping toast, my father would not have stood by and watched the slice fall onto its buttered side. Instead he would have figured out a way to prevent the fall or at least ensure that the toast would fall butter-side up.

Murphy and his fellows engineers spent years testing new designs of devices related to aircraft pilot safety or crash survival when there was no room for failure (for example, they worked on supersonic jets and Apollo landing craft). They were not content to rely on probabilities for their successes. Because they knew that things left to chance would definitely fail, they went to painstaking efforts to ensure success.

EDWARD A. MURPHY III, Sausalito, California

After receiving more than 362 intact issues of *Scientific American*, I received the April issue – with the article on Murphy's Law – that was not only assembled incorrectly by the printer but also damaged by the U.S. Post Office during delivery. My teenage daughter is taking this magazine into her science class to talk about Murphy's Law. The condition of this issue is an excellent example for her presentation.

BRAD WHITNEY, Anaheim, California

(*Scientific American*, August 1997)

8. O que deu origem a esses dois textos?

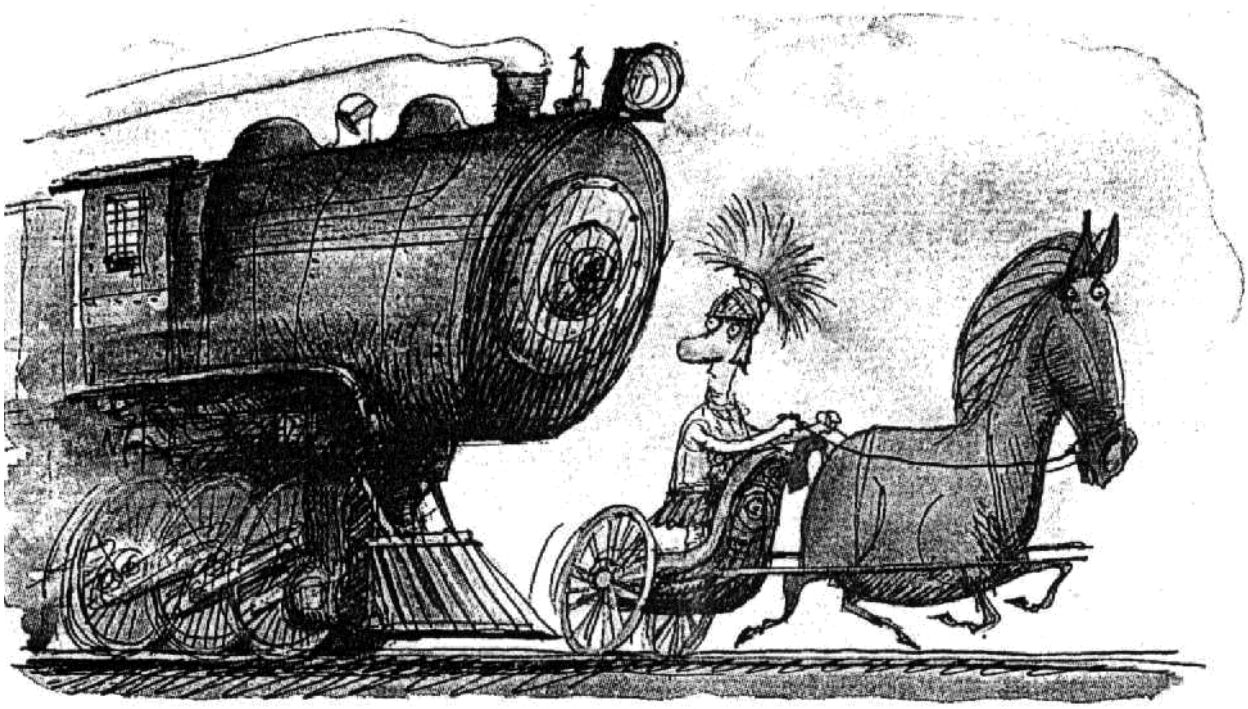
9. O primeiro texto destaca dois pontos positivos e faz uma ressalva. Transcreva o quadro abaixo para o seu caderno de respostas, preenchendo-o com as informações necessárias:

Pontos positivos	1.
	2.
Ressalva	

10. O segundo texto afirma: *"The condition of this issue is an excellent example for her presentation".* Explique por quê.

11. Explique por que Murphy pode ser considerado um perfeccionista.

Leia o texto abaixo e responda à questão 12.



Caesar´s Ghost

The real *reason why things never change*

The U.S. standard railroad gauge – the distance between the rails – is 4 feet, 8.5 inches. Why that exceedingly odd number? Because that's the way they built them in England, and the U.S. railroads were built by English expatriates. Why did the English people build them like that? Because the first rail lines were built by the same people who built the pre railroad tramways, and that's the gauge they used. Why? Because the people who built the tramways used the same jigs and tools for building wagons, which used that wheel spacing. OK! Why did the wagons use that odd wheel spacing? Well, if they tried to use any other spacing their wagons would break on some of the old long-distance roads, because that's the spacing of the old wheel ruts.

So who built the old rutted roads? The first long-distance roads in Europe were built by Imperial Rome for the benefit of their legions and have been used ever since. The initial ruts, which everyone else had to match for fear of destroying their wagons, were first made by Roman war chariots, which, because they were made for or by Imperial Rome, were all alike in the matter of wheel spacing. So, the U.S. standard railroad gauge of 4 feet, 8.5 inches derives from the original specifications for an Imperial Roman army war chariot. Specs and bureaucracies live forever.

From Kyoto Journal (#33). *Subscriptions: $40 for 4 issues from 31 Baud St., York, NY 10012*. RICHARD THOMSON *(UTNE READER,* July-August 97, p. 32).

12. Explique o título desse texto.