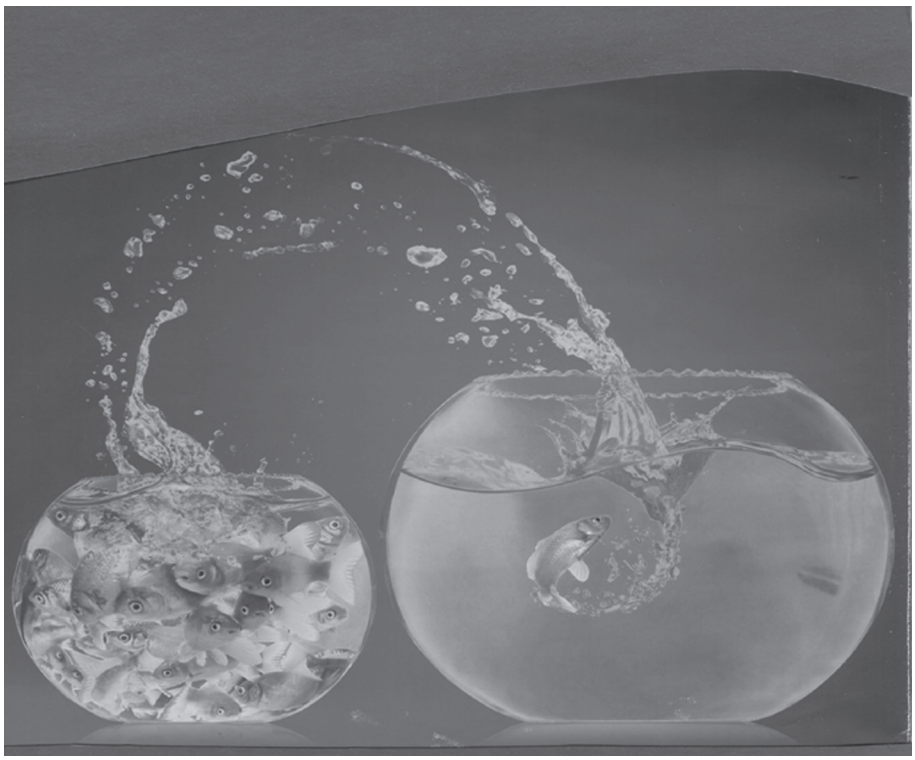
Figura 2 O peixe.



Fonte: Revista AutoData. São Paulo, n. 184, dez. 2004.

2 As tendências na área de avaliação, o efeito retroativo e o conceito de validade

What does washback look like? What brings washback about?
Why does washback exist?
J. Charles Alderson[1]

The extensive use of examination scores for various educational
and social purposes in society nowadays has made
the washback effect a distinct educational phenomenon.
Cheng, Watanabe; Curtis[2]

O objetivo deste capítulo é primeiramente apresentar a relação entre as concepções de linguagem, as abordagens de ensino de línguas estrangeiras e as avaliações de línguas, para então explorar conceitos como validade, confiabilidade e praticidade de testes.[3] Em seguida, discorreremos sobre modelos de leitura. Logo após esse tópico, falaremos sobre os conceitos de validade. Por último, um histórico do conceito de efeito retroativo será delineado, bem como sua evolução até os dias atuais.

1 Alderson, C. Foreword. In: CHENG, L.; WATANABE, Y. **Washback in Language Testing**. London: Lawrence Erlbaum Associates, Publishers, 2004. p. ix.

2 Ibid., p. xiii.

3 A palavra 'teste' neste trabalho será usada como sinônimo de prova, ou seja, um processo pelo qual se avalia algo. O teste, ou prova, é um dos vários instrumentos de avaliação que pode ser tanto oral como escrito. Quero distanciar-me do significado mais restrito que a palavra 'teste' tem no Português, geralmente conhecida somente por testes objetivos de múltipla-escolha, *cloze*, preenchimento de lacunas etc.

2.1 CONCEPÇÕES DE LINGUAGEM, ABORDAGENS DE ENSINO E DE AVALIAÇÃO EM LÍNGUA ESTRANGEIRA: UMA RESTROSPECTIVA HISTÓRICA

Testes de línguas estrangeiras evoluíram muito mais nos últimos cinquenta anos do que nas décadas e séculos anteriores. O desenvolvimento de testes se deu concomitante e dependentemente dos conceitos de linguagem e, consequentemente, das abordagens de ensino de língua estrangeira. Segundo Baker (1989, p. 2), sempre houve uma certa defasagem entre as concepções de avaliações com as de linguagem e abordagens de ensino/aprendizagem de línguas estrangeiras. Ele nos mostra, entretanto, a íntima relação existente entre abordagens de ensino e testes:

> Mudanças de ênfase no ensino de línguas estrangeiras provocaram [e provocam] inevitavelmente mudanças nos testes de LE. No entanto, métodos e teorias de testes têm sido mais relutantes às mudanças do que teorias de metodologias de ensino e desenvolvimento de cursos. Isto se deve principalmente porque testes de línguas estrangeiras modernas são baseados em princípios que, como a velha concepção estruturalista, se pautam na descrição da língua independente de qualquer uso particular que se faça dela. (tradução minha)[i] [4]

Skehan (1988, p. 3) também argumenta que a defasagem entre teorias de ensino de LE e teorias de avaliação só recentemente estreitou-se. Segundo o autor, "enquanto os testes de língua têm tido progresso em algumas áreas, no todo tem havido relativamente pouco progresso em testes de línguas até recentemente." (tradução minha)[ii]

Portanto, para delinear a evolução das visões da avaliação de línguas, é imprescindível traçar paralelos entre as diferentes concepções de linguagem e abordagens de ensino/aprendizagem de língua estrangeirs e as concepções de avaliação.

Spolsky (1976), em seu artigo seminal Language testing: art or science, identifica três estágios pelos quais a história de testes passou: o período pré-científico, o psicométrico-estruturalista e o psicolinguístico-sociolinguístico.

---

[4] Todas as citações que foram traduzidas para o português foram incluídas, na versão original do inglês, no final desta tese.

Morrow (1979), em seu artigo Communicative language testing: revolution or evolution, retoma os estágios de Spolsky classificando-os em: Garden of Eden (Jardim do Éden), Vale of Tears (Vale das Lágrimas) e The Promised Land (A Terra Prometida).

A seguir, irei tratar de cada um desses períodos, traçando paralelos com concepções de linguagem e também com as abordagens de ensino/aprendizagem de língua estrangeira.

2.1.1 Avaliação anterior aos anos 40: o período pré-científico ou *The Garden of Eden* (*O Jardim do Éden*)

A história do ensino de línguas estrangeiras e da avaliação não passou por mudanças profundas até meados do século XX. Sobre o ensino de LE, Bezerra da Maia et al. (2000) afirmam que:

> Até o século XIX, o fundamento teórico de novos métodos era praticamente inexistente e os métodos eram criados a partir das observações impressionistas de seus criadores. A abordagem gramatical, predominante até meados do século XIX [e ainda existente em alguns países], é composta de diversos métodos isolados, tendo eles em comum apenas a base estruturalista no ensino de línguas estrangeiras.

A avaliação sempre surgiu a partir das concepções de ensino de LE, como afirma Baker (1989, p. 2). "Mudanças nas abordagens para ensino de língua inevitavelmente resultaram em tentativas de desenvolvimento de métodos de avaliação apropriados para a nova pedagogia".

O primeiro período da avaliação, chamado por Spolsky (1976) de pré-científico, e denominada por Morrow de *Garden of Eden*, não era embasado em conhecimentos científicos, e, por esse motivo, a ideia de testes de línguas como uma atividade distinta não existia. Os testes eram considerados, segundo Spolsky, uma *arte*. Como o fundamento teórico de novos métodos era praticamente inexistente e os métodos eram criados a partir de observações impressionistas de seus criadores, os testes eram elaborados a partir dos mesmos meios de atividades utilizadas para o ensino da língua estrangeira, ou seja, tradução, ditado, composição etc. Nessa época, as atividades utilizadas em sala de aula correspondiam ao Método Clássico, que mais tarde seria chamado de Método de Tradução e Gramática. As atividades eram adaptadas para avaliar as

estruturas linguísticas e traduções de textos, geralmente literários, da língua estrangeira. A distinção entre ensino e testes não era bem delimitada.

Um exemplo de atividade dessa época pode ser observado no livro *Techniques and Principles in Language Teaching* de Larsen-Freeman (1986, p. 4-11). A autora sugere um trabalho de tradução do capítulo 4 – The Boys' Ambition – do livro Life on the Mississippi de Mark Twain. O primeiro parágrafo está transcrito abaixo.

> **When I** was a boy, there was but one permanent ambition among my comrades in our village on the west bank of the Mississippi River. That was, to be a steamboatman. We had transient ambitions of other sorts, but they were only transient. When a circus came and went, it left us all burning to become clowns; the first negro minstrel show that came to our section left us all suffering to try that kind of life; now and then we had a hope that if we lived and were good, God would permit us to be pirates. These ambitions faded out, each in its turn; but the ambition to be a steamboatman always remained.

Cada aluno é chamado para ler em voz alta uma oração e depois traduzi-la para a língua alvo. Ao termino da tradução do capítulo, a professora pergunta, na língua materna, se os alunos têm dúvidas. As dúvidas são sanadas na língua materna. Após essa atividade, os alunos são instruídos a responder as perguntas que se encontram ao final do texto na língua alvo. Logo a seguir, a professora apresenta alguns exercícios para eles trabalharem a língua.

> Exercise 2A
>
> These words are taken from the passage you have just read. Some of them are review words and others are new. Give the Portuguese translation for each of them. You may refer back to the reading passage.
>
> ambition gorgeous
>
> career loathe
>
> wharf envy
>
> tranquil humbly
>
> Exercise 2B
>
> These words all have antonyms in the reading passage. Find the antonym for each:
>
> love ugly
>
> noisy proudly

A seguir, a professora trabalha com palavras cognatas, prefixos e sufixos. Por exemplo: a terminação –ty em inglês corresponde à –dade em português. Retira, então, exemplos do texto. Nas aulas seguintes outros aspectos gramaticais e de tradução são trabalhados.

Como podemos observar, os exercícios são de tradução do inglês para o português e de gramática. Vemos pelo exemplo que, no ensino de língua estrangeira, as competências sociolinguística e ilocucionária (ver item 2.1.5) eram inexistentes.

Atividades como as supracitadas eram também utilizadas para avaliação, uma vez que não havia uma clara distinção entre ensino e avaliação.

A era pré-científica durou muitos anos, mas sua hegemonia foi abalada pelos novos acontecimentos mundiais que iriam mudar o rumo do ensino/aprendizagem e testes de língua estrangeira.

2.1.2 Avaliação dos anos 40 aos 60: o período do psicométrico-estruturalista ou *The Vale of Tears (O Vale de Lágrimas)*

A avaliação começou a ser um objeto da ciência somente a partir da II Guerra Mundial, quando surgiu no cenário do ensino de língua estrangeira "*The Army Specialized Training Program*" (ASTP), ou mais popularmente conhecido como Método do Exército que evoluiu para Método Direto e posteriormente para Método Audiolingual. A evolução das abordagens de ensino de língua estrangeira, a partir do desenvolvimento da linguística estruturalista, juntamente com o progresso da psicologia behaviorista, abriu novos caminhos. Baker (1989, p. 29) ressalta a importância dessa época para o desenvolvimento do ensino/aprendizagem e avaliação em língua estrangeira:

> Um número de fatores contribuiu para o desenvolvimento de interesse em testes de línguas de modo sistemático e científico depois da guerra. Os programas desenvolvidos durante o período de guerra nos Estados Unidos e em outros lugares, e o crescimento de agências internacionais fizeram aumentar a importância (e fundos) para projetos de ensino de línguas estrangeiras. Métodos para avaliar a eficiência destes projetos foram requeridos e o trabalho feito nos Estados Unidos nesse período rapidamente tornou-se a ortodoxia prevalecente no campo de testes de línguas. Seria difícil exagerar a extensão com que ideias correntes sobre testes de línguas foram influenciadas por essa abordagem. (tradução minha)[iii]

A partir das novas concepções de língua e de ensino de língua estrangeira surge também uma nova época de avaliação de LE: a **psicométrico-estruturalista**. Sua origem pode ser traçada a partir de duas tradições acadêmicas distintas: a dos testes psicométricos da psicologia e dos linguistas estruturalistas.

A área da psicologia behaviorista nos ofereceu um grande número de testes que serviram para dar subsídios aos estudos da mente. Baker (1989, p. 29) comenta:

> Nas décadas de 20 e 30 vimos entrar em voga os testes psicológicos. Um grande número de testes investigava todo aspecto da psique, desde inteligência até aptidão profissional eram produzidos e predições milenares às vezes eram feitas sobre benefícios sociais que testes de grande escala desse tipo poderiam trazer. No entanto, poucos testes realmente retribuíram com soluções miraculosas que haviam sido prometidas. Eles sobrevivem hoje em forma de testes de inteligência e testes menos sérios do tipo testes de revistas femininas "Você é um bom marido?" etc. (tradução minha)[iv].

Dois aspectos dos testes psicológicos são incorporados: as questões de respostas fechadas – múltipla-escolha, que prometiam objetividade na correção, levando a uma maior confiabilidade dos testes (tais testes eram chamados de descrete-point objective tests (testes objetivos de itens isolados[5]) – e o sistema de procedimentos estatísticos para desenvolver e avaliar esse tipo de teste (análise de item, por exemplo). Os métodos e a terminologia dos testes psicológicos foram adotados no campo dos testes de línguas, e o termo mais

[5] Testes com itens isolados. Esses testes eram baseados em itens únicos e independentes, como conjugação de verbos ou identificação de elementos lexicais. Para cada item, o candidato tinha que preencher uma lacuna com uma palavra ou expressão ou escolher a melhor alternativa dentre 3 ou 4. Uma questão não tinha relação com a outra.

significativo dessa época é *'psicométrico'*, derivado da palavra *psicometria* que quer dizer registro e medida de fenômenos psíquicos por meio de métodos experimentais padronizados.

A segunda tradição acadêmica, que ajudou a delimitar os campos de testes de línguas dessa época, foi a da linguística estruturalista. Baker (1989, p. 30) faz um excelente relato dessa influência:

> A tradição psicométrica na psicologia proporcionou as ferramentas necessárias para a produção e desenvolvimento dos testes. O que foi requerido foi a base para o conteúdo dos testes que eram produzidos. Que tipo de coisa deveria ser testado em testes de línguas? Aqui, naturalmente, a estrutura usada na avaliação derivava da estrutura empregada nos programas de ensino: a descrição da língua era amplamente baseada no trabalho de linguistas estruturalistas americanos. Em termos gerais, a análise usada envolvia a quebra do sistema linguístico em pequenas unidades, e depois, descrito de maneira na qual essas unidades poderiam ser reagrupadas outra vez para formar trechos de fala. A descrição era hierárquica tendo na base da pirâmide os fonemas que eram combinados para produzir morfemas, que eram combinados. etc. (tradução minha)[v]

Portanto, os testes de línguas foram baseados na análise hierárquica da língua como pregava a linguística estruturalista e nos novos métodos de avaliação criados pela psicologia behaviorista. Um exemplo de exame dessa era pode ser observado na prova de inglês do vestibular da UFPR de 1977:

The items numbered from 40 to 50 are to be answered by filling in the blanks

40 1. Mary tasted the soup.
The soup _____ salty.
2. John smelled the flowers.
The flowers _____ nice.
3. Peter heard the noise.
The noise _____ terrible.
4. Paul saw the wounded man.
The wounded man _____ terrible.
5. Sheila touched the material.
The material _____ smooth.

a) taste, smell, hear, see, touch.
b) taste, smell, sound, look, feel.
c) felt, looked, sounded, smelled, tasted.
d) tasted, smelled, heard, saw, touched.
e) tasted, smelled, sounded, looked, felt.

41 "I have not seen John _____ three or four days, but his brother has been here _____ last Saturday".
a) for, since
b) since, since
c) for, for
d) since, for
e) _____, since

42 A: "Your speech was very good".
B: "I could have done better if I _____ more time".
a) have had
b) had had
c) will have had
d) would have
e) had

Nos três exemplos acima, notamos o uso de itens isolados, ou seja, a língua fragmentada, sem contextualização. O aluno/candidato somente necessitava da competência gramatical, desprezando-se, totalmente, a competência sociolinguística: saber o que falar, com quem, quando e como. Esses itens estavam longe de serem atividades que os alunos/candidatos fossem aplicar na futura vida acadêmica.

As limitações geradas pelo trabalho com itens isolados induziram alguns estudiosos a levantar problemas em relação a essa abordagem para testes em língua estrangeira. Oller (1979, p. 212) descreve as desvantagens desse tipo de testes:

> A análise de itens isolados necessariamente quebra os elementos da língua em pedaços e tenta ensiná-los ou testá-los separadamente com pouca ou nenhuma atenção para a maneira pela qual esses elementos interagem num contexto de comunicação maior. O que os torna ineficientes como base para o ensino ou testes de línguas é que as propriedades cruciais da língua são perdidas quando seus elementos são separados. O fato é que em qualquer sistema onde as partes interagem para produzir propriedades e qualidades organizacionais se restringem e se tornam propriedades cruciais do sistema que não podem simplesmente ser achadas em partes separadas. (tradução minha)[vi]

Morrow (1979, p. 145) argumenta que "conhecimento dos elementos da língua de fato não conta nada a não ser que o usuário seja capaz de combiná-los numa maneira nova e apropriada para satisfazer as demandas linguísticas da situação na qual ele deseja usar a língua".

Hoje, não obstante as muitas críticas feitas aos testes psicométrico-estruturalistas, Weir (1990, p. 2) reconhece-lhes algumas vantagens:

1) são facilmente quantificáveis;
2) permitem uma vasta cobertura de itens;
3) notas baseadas em testes de itens isolados são confiáveis, pois são objetivas.

Porém, não são necessariamente válidas. (ver item 2.3)

Portanto, apesar das poucas vantagens que esse tipo de teste oferece, a decadência do psicométrico-estruturalista ou, segundo Morrow, O Vale das Lágrimas, começou quando se percebeu que algumas características da língua em uso para comunicação foram negligenciadas. A autenticidade da língua em uso, o contexto em que ela é usada, para que propósito, são alguns aspectos da língua que começaram a ser repensados. Não se acreditava mais que a língua poderia ser fragmentada, ainda mais para ser ensinada ou avaliada. A competência linguística passou a ser vista como um fenômeno mais complexo no qual a língua era indivisível, e por isso não podia ser mais fragmentada. A língua seria um todo e o todo era muito mais do simplesmente a somatória de suas partes.

Os métodos estruturalistas desse período não tiveram vida longa, mas na área de avaliação, a abordagem não teve muitos opositores e muitas de suas categorias e termos ainda são correntemente utilizados – geralmente mal empregados nas abordagens que sucederam o Método Audiolingual. *O Vale de Lágrimas* foi uma tentativa de dar à avaliação em língua estrangeira um caráter científico e sério.

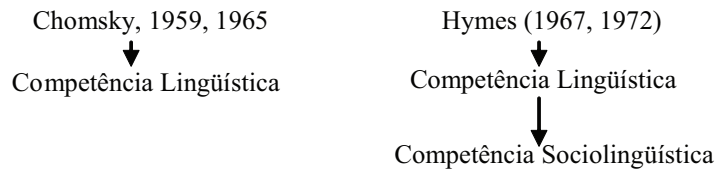
2.1.3 Avaliação nos anos 70: o período psicolinguístico-sociolinguístico ou *The Promised Land (A Terra Prometida)*

A inadequação dos testes de itens isolados para avaliar proficiência em língua estrangeira, juntamente com a queda do Método Audiolingual, fizeram surgir novos caminhos tanto para testes em língua estrangeira como tendências de ensino/aprendizagem de língua estrangeira. Novas concepções de linguagem foram a força motriz de mudanças nessas áreas.

Nas novas concepções, o termo novo mais debatido foi, e ainda é, a de competência. Para Chomsky (1966), que foi o primeiro a empregar o termo, competência significa conhecimento da língua, isto é, das suas estruturas e regras, e desempenho, o uso real da língua em situações concretas, numa construção marcadamente dicotômica, sem qualquer preocupação com a função social da língua.

Paralelo a era chomskiana, Hymes (1972) incorporou a dimensão social ao conceito de competência. Ao acrescentar "comunicativo" ao termo "competência", demonstrou claramente estar preocupado com o uso da língua. Assim, para Hymes, não é o bastante que o indivíduo saiba e use a fonologia, a sintaxe e o léxico da língua para caracterizá-lo como competente em termos comunicativos. É preciso que, além disso, esse indivíduo saiba e use as regras do discurso específico da comunidade na qual se insere. O indivíduo demonstra possuir competência se sabe quando falar ou não, a quem, com quem, onde e de que maneira se expressar. Deve-se a Hymes (1972) igualmente, a ampliação do conceito de competência por incluir a ideia de "capacidade para usar", unindo dessa forma as noções de competência e desempenho que estavam bem distintas na dicotomia proposta por Chomsky, em 1965. A partir de Hymes, e aparentemente inspirados por ele, vários autores enfrentaram a difícil tarefa de conceituar competência comunicativa. Cumpre, portanto, resumir o entendimento de alguns desses autores e firmar nossa posição para este trabalho.

Quadro 1 Emergência da competência comunicativa.

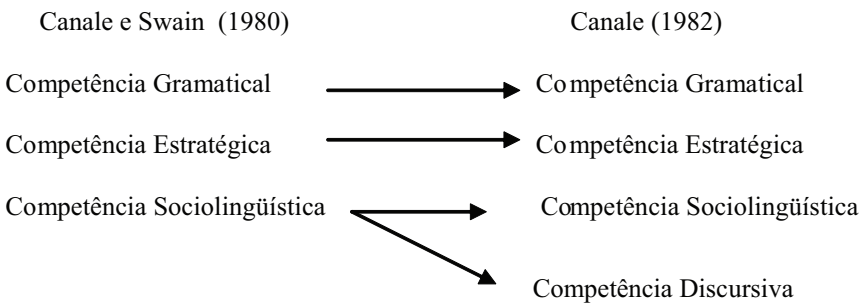


Fonte: Celce-Murcia, 1995, p. 700.

Nos anos 80, Canale e Swain (1980) e Canale (1983 a, b) levaram para a Linguística Aplicada o conceito de Hymes e formularam um modelo, ou melhor, uma concepção teórica, expandindo assim o conceito de competência comunicativa. Eles dividiram competência comunicativa em três subcompetências: a gramatical, a estratégica e a sociolinguística.

A *competência gramatical* refere-se ao domínio do sistema abstrato da língua-alvo e ense estritamente ligada ao nível da sentença, implicando conhecimento da sintaxe, da morfologia e da fonologia da LE. A *competência sociolinguística* implica o domínio das regras socioculturais da língua e do discurso, requerendo, portanto, o conhecimento do contexto social no qual a língua é usada e o da cultura dos falantes daquela língua. A *competência estratégica* engloba as estratégias linguísticas utilizadas pelos falantes para compensar a falta ou o não domínio pleno do código linguístico em questão. Canale subdividiu a competência sociolinguística em: *competência discursiva* e *competência sociolinguística.* A primeira relaciona-se com a correta organização de textos que segue regras de coesão e coerência determinadas pelo texto em si e pelo seu gênero em particular; a segunda implica o domínio das regras socioculturais da língua e do discurso, requerendo, portanto, o conhecimento do contexto social no qual a língua é usada e da cultura dos falantes daquela língua.

Quadro 2 Elaboração da competência comunicativa.



Fonte: Celce-Murcia, 1995, p. 701.

Canale e Swain (1980, p. 6) sugeriram a inclusão de tais competências, pois:

> Se a abordagem comunicativa para o ensino de segunda língua for adotada, então princípios de desenvolvimento de programas têm que integrar aspectos de ambas competências: gramatical e sociolinguística. Mais ainda, metodologia de ensino e instrumentos de avaliação têm que ser desenvolvidos para dar conta não somente da competência comunicativa, mas também do desempenho comunicativo, isto é, a demonstração real desse conhecimento em situações reais de segunda língua e para propósitos autênticos de comunicação. Também é importante manter em mente que uma pessoa não pode mensurar diretamente a competência: somente o desempenho é observável. (tradução minha)[vii]

Essas visões mais amplas de competência comunicativa fizeram com que tanto profissionais na área de ensino como na de testes repensassem suas metodologias e procedimentos.

Na área de avaliação em língua estrangeira, um novo indicador de proficiência ganha apoio com os **testes integradores**. Justamente nas décadas de 70 e começo dos 80, com inovações nas concepções de linguagem, a controvérsia entre testes de itens isolados e testes integradores ganha força. Testes de itens isolados foram criticados assim que emergiu a *era psicolinguística-sociolinguística* ou *integrative-sociolinguistics,* termos usados por Brown (1994, p. 262).

Precursor dos testes integradores, Oller (1979, p. 37) via-os como um instrumento que mensurava a capacidade de unir habilidades díspares de maneiras mais próximas do processo do uso linguístico real. Ele afirma:

> O conceito de teste integrador nasceu em oposição à definição de teste de itens isolados. Se testes de itens isolados fragmentam a língua em pedaços, testes integradores juntam as partes de volta. Testes de itens isolados tentam avaliar conhecimento linguístico um pouquinho de cada vez; testes integradores tentam avaliar a capacidade do aprendiz de usar todos os pedacinhos juntos ao mesmo tempo, e possivelmente enquanto exercitando vários componentes, tradicionalmente reconhecidos, do sistema gramatical, e talvez mais do que uma habilidade ou aspectos dessas habilidades. (tradução minha)[viii]

Os testes integradores foram baseados na *'Unitary trait hypothesis'* ou *'Unitary Competence Hypothesis'* – UCH (hipótese de competência unitária) proposta por Oller (1979), que sugeria que a proficiência linguística era mais 'unitária' (no sentido *uno*, singular, único e indivisível) que os testes de itens isolados sustentavam, isto é, vocabulário, gramática, fonologia, as quatro habilidades (o ensino de fala, compreensão auditiva, leitura e escrita, separadamente) e qualquer outro fragmento da língua não podem ser distinguidos uns dos outros. Tais tipos de testes requeriam do examinando que ele demonstrasse mais de um nível linguístico (micro e macro) ao mesmo tempo, às vezes, no caso do ditado, usando até duas habilidades como produção escrita e compreensão auditiva. A *hipótese de competência unitária* afirmava que existe um fator geral da proficiência da língua que todos os itens isolados não conseguem somar num todo.

Avaliações com procedimentos holísticos tais como testes *cloze* e ditados eram considerados por Oller (1979, p. 37) como sendo testes integradores devido ao fato de irem além da mensuração limitada de partes da competência linguística. Portanto, o significado do termo integrador aqui tem um sentido *sui generis*: os procedimentos de avaliação defendidos por Oller sugerem que tais instrumentos são capazes de fornecer amostras da língua de forma unitária, no seu todo. Baker (1989, p. 66) bem definia o que Oller cunhou de UCH e testes integradores:

> O que Oller disse foi que a proficiência linguística é indivisível, que testes somente diferem na sua eficácia da mensuração desse único fator, e que o elaborado aparato de dimensões, e testes usados por psicometricistas podiam ser repostas por um teste que iria diretamente avaliar o fator de indivisibilidade singular de proficiência linguístico. Testes que eram capazes de fazer isso, Oller cunhou de integradores; eles incluíam testes 'cloze'onde os candidatos tinham que restaurar lacunas de palavras, num intervalo regular, e ditados, nas quais os candidatos tinham que escrever palavras de um texto que era lido em voz alta. (tradução minha)[ix]

Um exemplo de método integrador, considerado por Oller, é o *cloze*:

The guitar has a long history. The Ancient Egyptians _____ (1) simple stringed instruments, and the GreeK-Ns and Romans also made music _____ (2) plucK-Ning strings by their fingers. The first true guitar music came during the 15th _____ (3) in Spain. At first it was an _____ (4) for poor people and travelling musicians, but soon rich people all _____ (5) Europe were learning to play the guitar.

The guitar travelled far and fast. When Cortes reached Mexico in the 16th century he had a guitar player _____ (6) his soldiers. A century later, the guitar was _____ (7) played all over South America. The Spanish Americans made some changes to the instrument and developed their _____ (8) style of playing. In North America new _____ (9) of music, jazz and popular music especially, led _____ (10) new K-Ninds of guitar. In the modern world there are 4 main _____ (11) of guitar: the classical, the flamenco, the steel stringed and electric guitars.

Ao preencher as lacunas, o examinando tem que utilizar informações que são inferidas através de fatos, eventos, ideias, relacionamentos, contexto social que são mapeadas de modo pragmático por sequências linguísticas contidas no texto.

Apesar de Oller afirmar que testes de *cloze* e ditados avaliam a competência unitária, até hoje não há resultados de nenhuma pesquisa que concluam quais são exatamente os construtos que tais métodos abordam. Além disso, vários teóricos criticaram os testes integradores. Alderson (1980, p. 59) questiona a confiabilidade dos testes *cloze*. Baseado em resultados de sua pesquisa, ele mostrou que as notas dos examinandos num teste *cloze* foram afetadas pela alteração dos pontos onde as lacunas começaram, ou pelo uso de diferentes intervalos de lacunas.

Davies (1981, p. 182) também aponta deficiências dos testes integradores:

> Apesar de Oller ter afirmado que seus testes integradores representassem uma proficiência linguística melhor que qualquer outro tipo de teste ou combinação de testes, isto não é, em si, um argumento a favor da hipótese de competência unitária, nem que testes integradores de *cloze* e ditados avaliam todas ou a maior parte das habilidades. Alta correlação entre *cloze* e outros instrumentos de mensuração pode somente refletir que eles estão mensurando diferentes habilidades que são altamente

> correlacionadas entre indivíduos; entretanto, pode haver indivíduos cujos desempenhos em várias habilidades diferem consideravelmente. (tradução minha)[x]

Weir (1990, p. 5-6) sustenta a ideia que testes *cloze* e ditados *não* provocam a produção espontânea dos examinandos, ou seja, na vida real eles não utilizariam o cloze em uma situação do cotidiano, e as normas linguísticas usadas não são as dos examinandos, mas sim as do examinador. Ele também afirma que ambos os tipos de testes mensuram conhecimento do sistema linguístico e não a capacidade de operá-lo em situações reais da vida, ou seja, esses tipos de testes nos dizem algo a respeito da competência linguística do examinando e não sobre a capacidade de uso.

As controvérsias suscitadas pelos testes de itens isolados e dos testes integradores abriram novos caminhos para uma evolução do período psicolinguístico-sociolinguístico.

2.1.4 Avaliação no final do século XX

Morrow (1979, p. 149) critica os testes integradores dizendo:

> [...] nem testes de *cloze* nem ditados oferecem qualquer prova convincente da capacidade do candidato para, de fato, usar a língua, para traduzir a competência (ou falta dela) que ele está demonstrando no desempenho real 'em situações do dia a dia', ou seja, de estar de fato usando a língua para ler, escrever, falar ou ouvir de maneiras e contextos que correspondem à vida real.

Adotar o 'uso' como critério de bons instrumentos de avaliação, segundo o autor, pode nos levar a considerar precisamente o porquê nem os testes de itens isolados nem os integradores são instrumentos satisfatórios de avaliação.

Para Morrow (op. cit., p. 150/51), 'A Terra Prometida' preconiza as seguintes características, para que um teste possa ser considerado um teste comunicativo:

1) Os testes deverão ser referenciados em critério para um desempenho operacional de um conjunto de tarefas com linguagem autêntica. Em outras palavras, tais testes mostrarão se (ou o quão bem) o candidato pode desempenhar um conjunto de atividades específicas.
2) Será crucial estabelecer sua própria validade como uma mensuração das operações que tais testes afirmam avaliar. Assim, validades de

conteúdo, de construto e o efeito retroativo serão importantes, mas a validade concomitante com testes existentes não será necessariamente significativa.

3) Os testes terão uma ênfase na avaliação qualitativa e não diretamente quantitativa. Poderá ser necessário converter os resultados em escores numéricos, mas o foco na análise qualitativa será sempre evidenciado.
4) Confiabilidade, claramente importante, será subordinada à validade.

Morrow (1979, p. 150/51) afirma que elaborações de testes com tais características oferecem numerosas vantagens. Primeiramente, tais aspectos suscitam testes com tarefas baseadas em desempenho, e apesar do desempenho ser, por natureza, um fenômeno integrador, qualquer tentativa de quantificá-lo poderá consequentemente provocar problemas de confiabilidade. Porém, para escolher e elaborar um teste com tarefas de desempenho, primeiramente, o profissional deverá se perguntar o objetivo daquela avaliação: se é de proficiência, diagnóstico ou de rendimento. Segundo, o conteúdo a ser avaliado é uma questão a ser considerada. Quando testes com tarefas de desempenho são desenvolvidos, pode-se perguntar: que desempenho? E Morrow (op. cit., p. 155) responde dizendo que:

> um dos aspectos característicos da abordagem comunicativa para o ensino de língua é que tal abordagem nos força ou nos capacita a fazer suposições sobre os tipos de comunicação que nossos aprendizes necessitarão. Isso também se aplica aos testes comunicativos. (tradução minha)[xi]

A Terra Prometida de Morrow (1979) está sintonizada com as características das abordagens comunicativas postuladas por Canale e Swain (1980), e Canale (1983 a, b), porque há uma preocupação em transformar exames em instrumentos de avaliação que contemplem características essenciais e que envolvam elementos das competências gramaticais, sociolinguísticas, discursivas e estratégicas. Os testes comunicativos, segundo Morrow, devem simular situações onde o aluno possa usar a língua que empregaria na vida real, avaliando a competência (ou falta dela) em contextos que ocorrem em seu dia-a-dia.

2.1.5 Avaliação no novo século: o paradigma comunicativo

Nos anos 90, novos modelos de competência surgiram para complementar e ampliar as concepções de Canale e Swain (1981) e Canale (1983). Bachman

(1991) e Bachman e Palmer (1996) retomam as concepções de competência dos anos 70 e 80 e avançam na teorização do conceito, afirmando, por exemplo, que:

> apesar de especialistas de testes em língua estrangeira terem sempre, provavelmente, reconhecido a necessidade de basear o desenvolvimento e uso de testes em língua estrangeira em uma teoria de proficiência linguística, recentemente, eles nos têm chamado a atenção para a incorporação de um quadro teórico do que é proficiência linguística com os métodos e tecnologia envolvidos na sua mensuração (tradução minha)[xii]. (Bachman 1991, p. 81)

Bachman propõe, portanto, um quadro teórico de competência de linguagem focalizando seu trabalho em contribuições para a área de testes em línguas. Ele advoga que a capacidade para usar uma língua de maneira comunicativa envolve tanto o conhecimento da língua quanto a capacidade de implementar ou usar esse conhecimento. Competência abarca conhecimentos específicos que são usados na comunicação. O modelo que concebeu inicialmente compreendia os seguintes conhecimentos:

a) *competência linguística*, subdividida em organizacional e pragmática;
b) *competência estratégica*, que é a capacidade mental de implementar os componentes da competência de linguagem em um uso comunicativo da linguagem contextualizado;
c) *mecanismos psicofisiológicos*, os quais dizem respeito aos processos neurológicos e psicológicos na produção real da língua como um fenômeno físico.

Mais recentemente, no entanto, Bachman (1991b, p. 683) reviu seu modelo e operou nele algumas alterações. Primeiramente, o que chamava de competência passou a se denominar "conhecimento". Em nota de rodapé argumenta que o termo competência traz consigo uma grande e desnecessária bagagem semântica, e por isso não é mais tão útil como conceito. Assim sendo, saber usar uma língua tem a ver com "a capacidade de utilizar o conhecimento da língua em sintonia com as características do contexto para criar e interpretar significados".

O esquema teórico da competência de linguagem inclui diferentes componentes. Competência de linguagem pode ser classificada em dois tipos: *conhecimento organizacional* e *conhecimento pragmático*.

O *conhecimento organizacional* compreende as habilidades envolvidas no controle da estrutura formal da língua para produzir ou reconhecer

orações corretas gramaticalmente, compreendendo seu conteúdo proposicional e ordenando-as para formar um texto. Tais habilidades são de dois tipos: *conhecimento gramatical* e *conhecimento textual*. O *conhecimento gramatical* inclui competências envolvidas no uso da língua. Elas consistem em um número relativamente independente de competências tais como *conhecimento de vocabulário*, *morfológico*, *sintático* e *fonológico* e *gráfico*. O *conhecimento textual* inclui o conhecimento de convenções de junções de expressões para formarem um texto, escrito ou falado, que é estruturado a partir de regras de coesão e organização retórica.

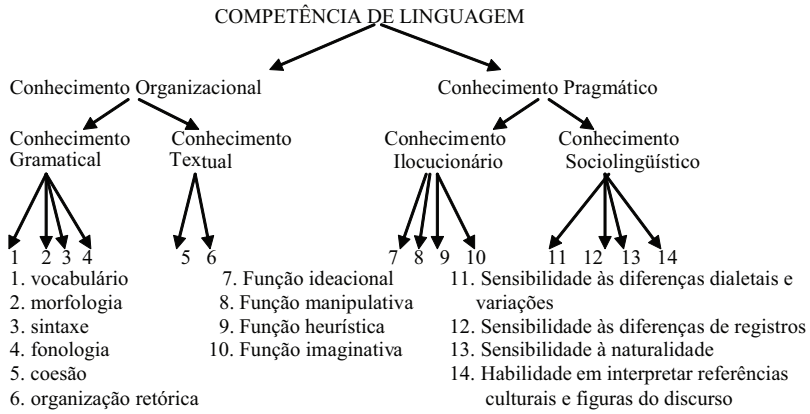
O *conhecimento pragmático* faz a ponte entre os componentes do *conhecimento organizacional* com o usuário da língua e seu contexto de comunicação. Assim, a pragmática se preocupa com a relação entre expressões e os atos ou funções que o falante (escritor) pretende desempenhar através dessas expressões num contexto de uso de língua determinado pela adequação de uso dessas expressões. O *conhecimento pragmático* pode ser subclassificado em *conhecimento ilocucionário* e *conhecimento sociolinguístico*.

O *conhecimento ilocucionário* foi introduzido a partir da teoria dos atos de fala e pela descrição de funções linguísticas descritas por Halliday (1973, 1976). Ele foi subdivido em *funções ideacionais* que expressam significados em termos da nossa experiência do mundo real, e incluem o uso da língua para expressar proposições ou para trocar informações sobre nosso conhecimento ou sentimento; *funções manipulativas* que são também chamadas de *funções instrumentais*, pois é com elas que se fazem coisas e se faz com que coisas sejam feitas (dar sugestões, perguntar, pedir, ordenar etc); as *funções heurísticas* pertencem ao uso da língua para entender o conhecimento do mundo e ocorre concomitantemente com atos de ensino, de aprendizagem, de solucionar problemas etc; e, por último, as *funções imaginativas* que nos capacitam a criar ou a entender o próprio ambiente para humor ou propósitos estéticos em que o valor do ato deriva da maneira pela qual a língua é usada.

O *conhecimento sociolinguístico* (CS) tem a preocupação com a adequação das funções e como elas são desempenhadas, pois variam de um contexto linguístico para outro, de acordo com aspectos socioculturais e do discurso. O CS é sensível a convenções de uso linguístico que são determinadas por contextos de uso linguísticos – tal conhecimento possibilita desempenhar funções linguísticas dentro de contextos apropriados. O CS preocupa-se com a sensibilidade quanto às diferenças dialetais e variantes, às diferenças de registros, à naturalidade e com a habilidade em interpretar referências culturais e figuras do discurso.

O quadro 3 mostra um esquema da competência de linguagem de Bachman.

Quadro 3 Competência de linguagem.



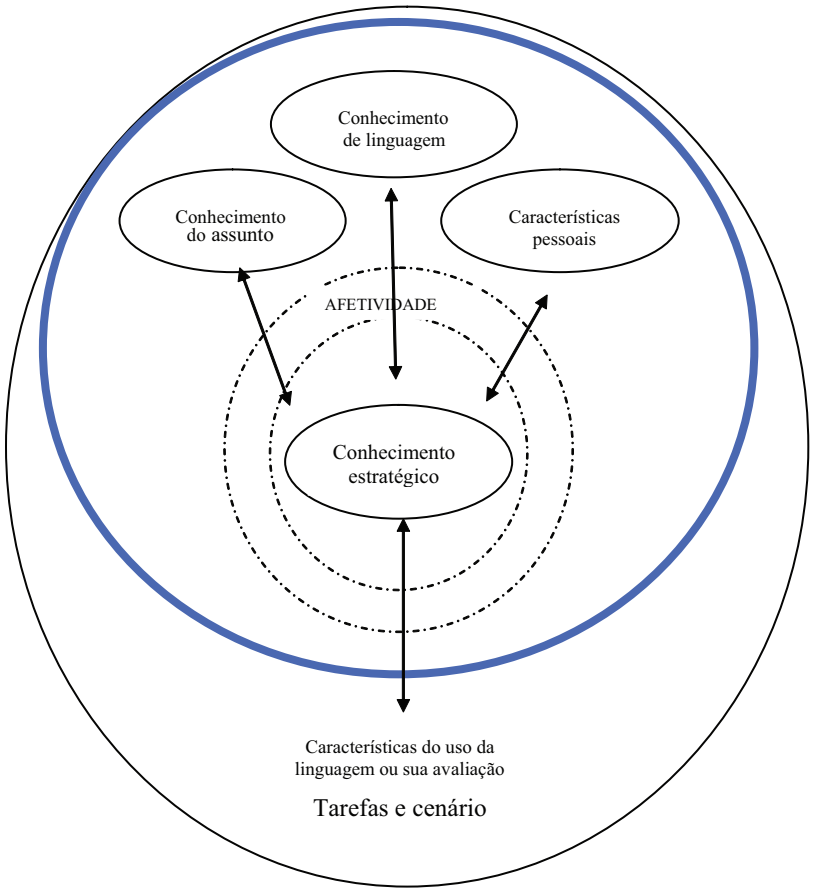
Fonte: Bachman, 1990, p. 87.

Em 1996, Bachman e Palmer especificam a descrição da habilidade linguística para o desenvolvimento de testes em língua estrangeira. Para eles (1996, p. 61),

> o uso linguístico pode ser definido como criações ou interpretações de significados pretendidos num discurso por um indivíduo, ou as dinâmicas e negociações interacionais de significados pretendidos entre dois ou mais indivíduos numa situação particular. Em usar a língua para expressar, interpretar, ou negociar significados pretendidos, o usuário da língua cria discursos.

Nessa perspectiva, a natureza interacional do uso linguístico é enfatizada. O uso da língua envolve interações complexas e múltiplas entre várias características individuais do usuário da língua, e, entre tais características, as características do usuário ou situação de teste. Portanto, o modelo de competência de linguagem de Bachman e Palmer enfatiza a interação entre as áreas de habilidade linguística (conhecimento linguístico e competência estratégica ou metacognitiva), conhecimento do assunto e do esquema afetivo, e como elas interagem com as características da situação do uso linguístico ou tarefa de teste. O quadro 4 nos mostra alguns componentes do uso linguístico e desempenho de linguagem de um teste.

Quadro 4 Alguns componentes do uso linguístico e do desempenho de linguagem de teste.



Fonte: Bachman; Palmer, 1996, p. 63.

O quadro 4 mostra algumas das mais importantes interações envolvidas no uso da linguagem como base conceitual para organizar o pensamento e linguagem e também como base conceitual para o desenvolvimento de testes. Os componentes dentro do círculo azul, *conhecimento do assunto* (*topic knowledge*), *conhecimento de linguagem* (*language knowledge*), *características pessoais* (*personal characteristics*), *conhecimento estratégico* (*strategic compentence*) e *afetividade* (*affect*) fazem parte das características individuais de um usuário de uma língua. Os componentes dentro do círculo maior incluem características de tarefas ou cenários com as quais o usuário da língua interage. Os círculos pontilhados indicam interações. A figura indica que a *competência estratégica* é o componente que integra os outros componentes

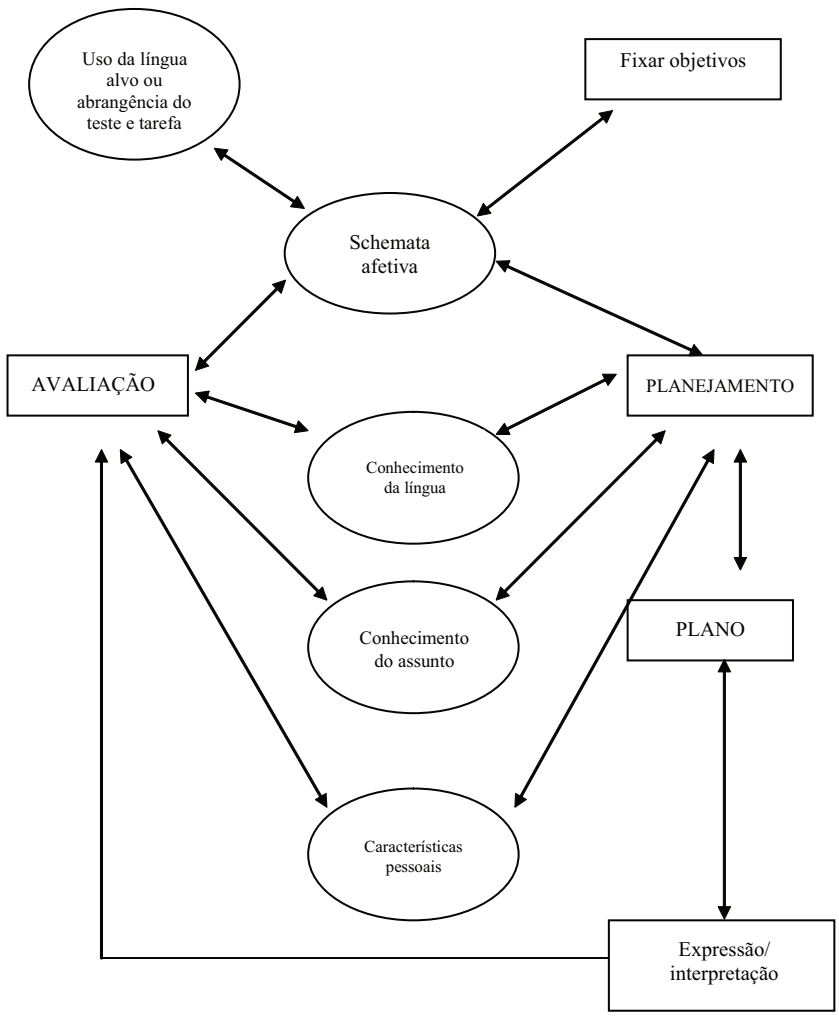
ao individual, assim como prevê um elo cognitivo com as características do uso linguístico da tarefa e cenário. Para Bachman e Palmer, tal figura serve para nos manter em contato e relembrar as habilidades importantes ou atributos que influenciam a utilidade de testes em língua estrangeira.

As *características individuais* são atributos individuais que não fazem parte da habilidade linguística que um examinando utiliza num teste, mas que ainda assim influenciam os seus desempenhos nos testes. Dentre as *características individuais* podemos ressaltar idade, sexo, nacionalidade, *status* da residência, língua nativa, nível e tipo de educação geral, tipo e quantidade de preparação ou experiência anterior com o tipo de teste. O *conhecimento do assunto* também é chamado de conhecimento de *esquemata* ou conhecimento do mundo real, e são as estruturas de conhecimento que temos armazenados na memória de longo prazo. A *esquemata afetiva* está *emocional* ou *afetivamente correlacionada* com o conhecimento do assunto. Ela provê a base para que o usuário da língua acesse, consciente ou inconscientemente, as características do uso linguístico da tarefa e seu cenário em termos de experiências emocionais passadas em contextos similares. A *resposta afetiva* do usuário linguístico a uma tarefa em particular, ou seja, a facilidade ou a limitação das respostas em um dado contexto de teste, é determinada pela *esquemata afetiva* em combinação com as características particulares da tarefa. O *conhecimento de linguagem* já foi descrito no parágrafo anterior.

A *competência estratégica* é um conjunto de componentes metacognitivos ou estratégicos que são processos executivos de nível mais alto (*higher order executive processes*) e que provê uma função gerencial cognitiva no uso linguístico, assim como em outras atividades cognitivas. Os componentes metacognitivos operam em três áreas gerais: determinação de objetivos, (*goal-setting*), avaliação (*assessment*) e planejamento (*planning*). Para determinar objetivos é preciso que alguém decida o que vai fazer. Isso envolve identificação: da **tarefa** do uso linguístico ou da tarefa de testes; da **escolha** de uma ou mais tarefas dentre um conjunto de tarefas possíveis, quando houver escolha; da **decisão** de tentar ou não cumprir a tarefa. A avaliação requer fazer um balanço do que é necessário, do que o usuário da língua tem para trabalhar e como ele tem feito a tarefa. Tal avaliação inclui a das características do uso linguístico ou da tarefa do teste, a do conhecimento de assunto do indivíduo e linguístico, e da correção e adequação das repostas da tarefa do teste. O planejamento abrange a utilização do *conhecimento linguístico*, do *conhecimento do assunto* e da *esquemata afetiva* para indicar como completar a tarefa do

teste com sucesso. Apresenta três aspectos: a da seleção de um conjunto específico de elementos linguísticos (conceitos, palavras, estruturas, funções) e de *conhecimento do assunto* que serão usados no planejamento; a da formulação de um ou mais planos cujas realizações serão as respostas à tarefa; e a da seleção de um plano para a implementação das respostas às tarefas. O quadro 5 mostra as estratégias metacognitivas do uso linguístico e do desempenho da linguagem do teste.

Quadro 5 Estratégias metacognitivas no uso linguístico e no desempenho em testes de línguas.



Fonte: Bachman e Palmer, 1996, p. 72.

Todos os modelos que descrevem o que é competência de linguagem possuem limitações devido ao caráter complexo da linguagem.

O modelo de Hymes se limita a conhecimento de língua e capacidade para uso dentro da competência comunicativa. Canale e Swain, apesar de desenvolverem e acrescentarem mais componentes a seu modelo, tais como competência gramatical, sociolinguística e estratégica e Canale posteriormente acrescentar competência de discurso, não contemplam fatores não linguísticos tais como afetividade e personalidade. Um outro problema com o modelo de Canale é que a competência do discurso que deveria estar contida dentro de conhecimento de linguagem é uma competência independentemente dela (coesão e coerência são os elementos que compõem tal competência).

Bachman (1990) e Bachman e Palmer (1996) sofisticam o modelo ainda mais e são os que mais se aproximam de uma descrição mais detalhada do que é ter competência comunicativa. Isto porque primeiro eles retomaram a capacidade de uso de Hymes dentro da competência estratégica que não faz parte da competência de linguagem, e, segundo, incluíram outros fatores que não são somente cognitivos, tais como fatores afetivos e pessoais. Porém, o maior problema desse modelo é que seus criadores não mostram como os componentes interagem em uma situação real de comunicação ou avaliação.

2.1.5.1 Evolução das avaliações no paradigma comunicativo

Até a era psicométrico-estruturalista, testes e exames centravam-se em avaliar fragmentos da língua, ou seja, somente as estruturas eram focadas. A partir do paradigma comunicativista, a noção de capacidade de uso tomou força, e muitos testes começaram a avaliar leitura, escrita, compreensão auditiva e fala, porém, separadamente. Métodos de múltipla-escolha, *cloze* e verdadeiro-falso, típicos da era psicométrico-estruturalista, continuaram a ser utilizados, e por esse motivo, a avaliação, nesses casos, permanecia indireta.[6]

A partir da avaliação das habilidades separadamente, os testes evoluíram para a avaliação integrada das habilidades, utilizando-se, para isso, as tarefas.

---

[6] O teste indireto avalia o conhecimento de língua do aluno indiretamente por métodos (que não são usados em situações corriqueiras na vida real) tais como múltipla-escolha, *cloze* ou verdadeiro/falso, ao contrário do teste direto que avalia o aluno diretamente através de situações reais de desempenho, geralmente, por tarefas.

Os exames de proficiência TOEFL[7] e IELTS,[8] que dividiam o exame em seções separadas como leitura, escrita e compreensão auditiva, recentemente passaram a se preocupar com a integração das habilidades[9] em forma de tarefas (ver Anexo C).

O Certificado de Língua Portuguesa para Estrangeiros do Ministério da Educação – Celpe-Bras (ver Anexo B) é um outro exemplo da nova era de testes baseados em tarefas e nos quais as habilidades são integradas para simularem situações que aconteceriam no dia-a-dia do candidato. Scaramucci (1999a) apresenta as principais características desse exame:

1. Ênfase na comunicação/interação – o saber expressar-se e interagir com outras pessoas através da língua alvo.
2. Habilidades integradas – ler e escrever, ouvir e anotar, ver, ouvir e falar; é, portanto, uma avaliação que integra, geralmente, duas habilidades.
3. Tarefas – o exame todo é feito por tarefas que têm um propósito comunicativo, especificando para a linguagem usos que se assemelham àqueles que se têm na vida real.
4. Conteúdos autênticos ou contextualizados – tais como jornais, revistas, usados por falantes nativos na sua comunicação.
5. A correção é predominantemente qualitativa e holística.

As características do Celpe-Bras, apontadas acima, mostram como os exames de proficiência evoluíram nos últimos anos. O que se tenta fazer hoje é simular tarefas,[10] o mais próximo possível da vida real tanto para exames de proficiência, como para outros tipos de avaliações, sejam elas de rendimento, de classificação ou diagnóstica. No Brasil, no entanto, o exame Celpe-Bras é uma exceção, e testes em língua estrangeira, de modo geral, ainda têm muito a evoluir.

---

7 Informações sobre o TOEFL podem ser encontradas em <http://toeflpractice.ets.org>. Acesso em: 22 fev.2006.

8 Informações sobre o IELTS podem ser encontradas em <http://www.ielts.org>. Acesso em: 22 fev. 2006.

9 Integrar habilidade é quando um aluno/candidato desenvolve uma tarefa utilizando uma habilidade – a leitura, por exemplo – e, a partir das informações adquiridas nessa tarefa, desenvolve uma outra utilizando uma outra habilidade – a escrita, por exemplo.

10 Segundo Scaramucci, tarefa é um convite para agir no mundo, um convite para o uso da linguagem com um propósito social. Uma tarefa envolve basicamente uma ação, com um propósito, direcionada a um ou mais interlocutores.

Como bem coloca Scaramucci (1999a, p. 106),

> A coerência entre ensino e avaliação é fundamental não apenas quando se está considerando a questão sob o ponto de vista da sala de aula e do professor, mas também sob o ponto de vista mais externo, ou de uma política educacional.
>
> Entretanto, essa coerência nem sempre se observa na realidade. Embora uma tendência mais inovadora possa ser observada no ensino de português –LE no Brasil [e eu estendo tal observação para o ensino de LE em geral] tanto no Brasil como nos países onde o português é ensinado, observa-se, infelizmente, que em muitos desses contextos, a avaliação ainda é conduzida nos moldes tradicionais, o que acaba por comprometer o ensino.

Scaramucci (1999a, p. 108) comenta o atraso dos testes em relação aos avanços na área de ensino de LE dizendo que "apesar dos avanços recentes com relação ao desenvolvimento de diretrizes e princípios comunicativos, e, mais especificamente, materiais comunicativos, a área de avaliação da competência comunicativa pode ser considerada ainda em estágio embrionário."

Mesmo com tal atraso, pode-se notar uma certa evolução em relação às fases anteriores. De acordo com Bachman (1991), os anos 80 e 90 podem ser caracterizados pelas décadas de testes comunicativos. Existe uma mudança em ênfase de linguístico para uma dimensão comunicativa. A ênfase não está mais na precisão linguística, mas na habilidade de funcionar eficazmente através da língua num contexto ou situação particular. Essa nova visão surgiu a partir da evolução do conceito de competência comunicativa sugerido por Canale (1980), Canale e Swain (1982) Bachman (1991) e Bachman e Palmer (1996). Weir (1990, p. 9) afirma:

> Testes comunicativos têm a preocupação em descobrir o que o aprendiz sabe sobre a forma da língua e sobre como usá-la apropriadamente em contextos de uso (competência), e também têm que lidar como o aprendiz pode, de fato, demonstrar esse conhecimento numa situação comunicativa significativa (desempenho), isto é, o que ele possa fazer com a língua – sua habilidade para comunicar com facilidade e eficiência em cenários sociolinguísticos específicos. (tradução minha)[xiii]

Madsen (1983, p. 7) também define testes comunicativos como exames que combinam várias sub-habilidades. "Em particular, testes comunicativos necessitam mensurar mais do que habilidades linguísticas isoladas: eles devem indicar o quão bem uma pessoa possa funcionar na sua segunda língua."

Schlatter et al (2004, p. 366) refere-se à avaliação de desempenho como sendo uma demonstração direta da proficiência almejada ou das capacidades adquiridas, em vez de limitar-se a avaliar indiretamente essa proficiência através de instrumentos que focalizam itens isolados de gramática. A avaliação de desempenho pressupõe que a melhor maneira de avaliar se alguém é proficiente é colocá-lo em situação em que ele possa demonstrar essa proficiência diretamente. Um exemplo de um exame de desempenho é o Certificado de Proficiência em Língua Portuguesa - Celpe-Bras (ver Anexo B.1).

Para Hughes (1994, p. 19-20), testes comunicativos devem ocorrer quando há necessidade de mensurar a habilidade de participar em atos de comunicação (incluindo leitura e compreensão auditiva).

Baker (1989, p. 77) define testes comunicativos como um teste direto referenciado em desempenho. Para ele, esse tipo de teste envolve simulações de atividades futuras ou potenciais e os resultados dos testes podem ser usados para predizer a habilidade do candidato para desempenhar situações similares no futuro.

Porém, os elaboradores de testes enfrentam um problema grave quando constroem e aplicam testes comunicativos: a fragilidade da confiabilidade, quando a correção não é criteriosa (ver 2.3.5). Apesar do teste comunicativo ter alta validade (ver 2.3.1), a confiabilidade tem sido o componente que mais tira o sono dos elaboradores. Weir (1990, p. 15) observa que

> [...] além do sério problema da confiabilidade, associada com avaliação de desempenho, um outro aspecto que envolve a adoção da abordagem comunicativa para testes de línguas é a generalização dos resultados produzidos por tais testes. (tradução minha)[xiv]

Scaramucci (1999a, p. 108) relata a dificuldade encontrada na hora da correção de exames baseados em tarefas em geral: "Há dificuldades num sistema de notas qualitativo, assim como na padronização de um exame nesses moldes. Enfim, a falta de tradição na avaliação comunicativa faz de toda experiência uma tarefa árdua e desafiadora."

Scaramucci (op. cit., p. 111) explica como a equipe enfrentou o problema da confiabilidade:

> Entretanto, para que uma abordagem comunicativa possa ser operacionalizada sem distorções, é necessário tomar algumas precauções não apenas na elaboração do exame, mas também na implementação de pro-

> cedimentos de correção e, sobretudo, em sua validação. Muito frequentemente, o valor de uma avaliação direta, com todos os desafios que constitui, é perdido pelo reducionismo dos procedimentos de correção. Por isso, no caso do Celpe-Bras, foram adotados procedimentos predominantemente qualitativos e holísticos de correção. Digo predominantemente, já que tem sido impossível evitarem-se quantificações, uma vez que esse exame terá de passar por um processo de validação. Os resultados da avaliação são expressos em descritores de competência e desempenho, faixas ou ainda escalas que mostram o que cada candidato é capaz de fazer em termos comportamentais, não se restringindo aos números, como nos exames tradicionais.

Todos os testes que foram citados neste capítulo têm algum tipo de problema: podem carecer de validade (ver item 2.3.1), ou confiabilidade (ver item 2.3.5), ou ainda praticidade (ver item 2.3.6). Os testes da era psicométrico-estruturalista tendem a ser bastante confiáveis, pois testes objetivos são facilmente corrigidos e os escores dificilmente se alteram, se outro corretor os corrigir. Porém, a validade de construto desses tipos de testes é questionável. Por outro lado, os testes da era psicolinguístico-sociolinguística tendem a ter uma maior validade de construto por utilizar instrumentos de avaliação onde se tenta mostrar a habilidade linguística de funcionar eficazmente num contexto ou situação particular, mas tem uma frágil confiabilidade por serem testes subjetivos e sua correção vulnerável. Um bom teste deveria apresentar um equilíbrio entre as três características, porém esse equilíbrio é bastante difícil de se obter. Na seção 2.3 examinarei essas concepções e sua importância quando temos que optar entre a validade, confiabilidade ou praticidade.

2.2 OS MODELOS/VISÕES DE LEITURA

Até os anos 60, a leitura não era considerada um objeto independente para estudos científicos. A partir dessa década, o estudo da leitura emergiu no cenário acadêmico dando origem à basicamente três diferentes modelos teóricos: *o modelo de decodificação* ou *ascendente*, o *modelo psicolinguístico* ou *descendente* e o *modelo interativo*. Diferentes visões ou concepções de leitura/compreensão são pressupostas nos diferentes modelos, com consequentes implicações, não apenas para o seu ensino, mas também para sua avaliação.

Segundo Scaramucci (1995, p. 11/12)

> *No modelo ascendente ou de decodificação a leitura é vista como extração de significados na qual o fluxo da informação é ascendente, ou seja, se inicia com a percepção dos dados na página impressa, procedendo em uma sequência fixa, sempre das unidades menores (reconhecimento de letras e palavras) para as maiores (frases, orações, parágrafos), até chegar ao significado que está cristalizado no texto.*

No final dos anos 60 e começo dos 70, o modelo de leitura vigente era o *descendente* ou psicolinguístico e a ênfase do texto passa para o leitor. A teoria que embasava a visão de leitura descendente era advogada por Goodman (1967) e Smith (1978 a, 1978 b), ou seja, a leitura implicava no uso mínimo de sinais linguísticos e conhecimento de mundo para que o leitor pudesse confirmar, rejeitar ou redefinir hipóteses criadas por ele para que as mensagens pudessem ser reestruturadas. Kleiman (1989) afirma que, nesse modelo, a direção do fluxo principal da informação passa a ser descendente, processo iniciado a partir do leitor e procedendo em direção ao texto, que é visto como objeto indeterminado e incompleto, cabendo ao leitor impor-lhe uma estrutura, (re) criando um significado.

A partir da década de 80, a leitura começa a ser vista como um processo de construção de sentidos: o *modelo interativo*. Scaramucci (op. cit., p. 18), antes de caracterizar o modelo, salienta como o termo interativo tem sido usado com sentidos diferentes dentro da literatura, referindo-se, principalmente, a dois conjuntos independentes, mas relacionados de pesquisa. A pesquisadora nos alerta para a diferença entre os conceitos de *processo interativo* e *modelo interativo*. Scaramucci discute (op. cit., p. 18)

> *A visão de leitura como um processo interativo é discutida por Widdowson (1979), como um processo de combinação da informação textual (ascendente) com a informação que o leitor traz para o texto (descendente), ou da interação entre a mente do leitor e os elementos do texto. Durante a leitura ocorre a ativação dos vários tipos de conhecimento na mente do leitor, que, por sua vez, como resultado da informação nova fornecida pelo texto, são refinados e ampliados. A leitura caracteriza-se, assim, como um diálogo de negociação do sentido entre o leitor e o texto.*

> *Embora essa visão tenha trazido contribuições muito importantes, introduzindo aspectos pragmáticos ou interacionais não considerados nos outros modelos, o conceito assume semelhanças entre os processos de leitura em língua materna e língua estrangeira e a possibilidade de transferência automática linguística da LM para L2, deixando de lado, principalmente, a diferença entre a proficiência linguística em LE e LM, ou a limitação no vocabulário/sintaxe dos leitores em LE. Uma outra implicação dessa visão é a ideia de que os textos são informações, quando, na realidade, são apenas marcas em uma página, que os leitores devem converter em linguagem/informação.*

Uma outra versão para o termo *interativo* está relacionada ao seu uso em modelo interativo (Scaramucci, op. cit., 18-19). A leitura passou a ser vista como um processo mais complexo, como uma prática social: um processo de construção de sentidos. A polarização dos *modelos ascendentes* ou *descendentes de leitura* deu lugar *ao interativo*, no qual a leitura passou a ser vista como um processo cognitivo e, ao mesmo tempo, perceptivo, envolvendo uma combinação desses dois, além de outros níveis de conhecimento, inclusive o linguístico. Tal concepção caracteriza-se por apresentar uma bidirecionalidade de fluxo da informação, isto é, do texto para o leitor (ascendente) e também do leitor para o texto (descendente): ambos dependentes de certos tipos de conhecimento prévio e certos tipos de habilidades de processamento de informação (Eskey 1988, p. 96). Neste caso, a construção do significado durante o processo de leitura é feita através de uma interação leitor/texto, ou leitor/pistas indexicais para o leitor, mais os conhecimentos e habilidades descritos por Eskey. (Scaramucci, 1995).

As duas visões do termo interativo não são excludentes, pois os modelos interativos incorporam as implicações de leitura como um processo interativo. Portanto, o *modelo interativo* servirá como arcabouço teórico deste trabalho, uma vez que é mais abrangente e, portanto, mais condizente com as necessidades do leitor em seu processo de aquisição de leitura em LE.

2.2.1 Uma matriz de questões

Para que pudéssemos analisar as questões dos exames de inglês do vestibular da UFPR de 2000 a 2006 adotamos a taxonomia de questões desenvol-

vida por Cherem & Nery (1993) e Nery (2003). Tal taxonomia foi escolhida, basicamente, por duas razões: primeiramente, porque sua origem se deu a partir do trabalho das autoras de elaborarem e corrigirem os exames de francês do vestibular da Unicamp, e, num segundo momento, por se tratar de uma taxonomia que explora desde questões com predominância informativa até as com predominância argumentativas.

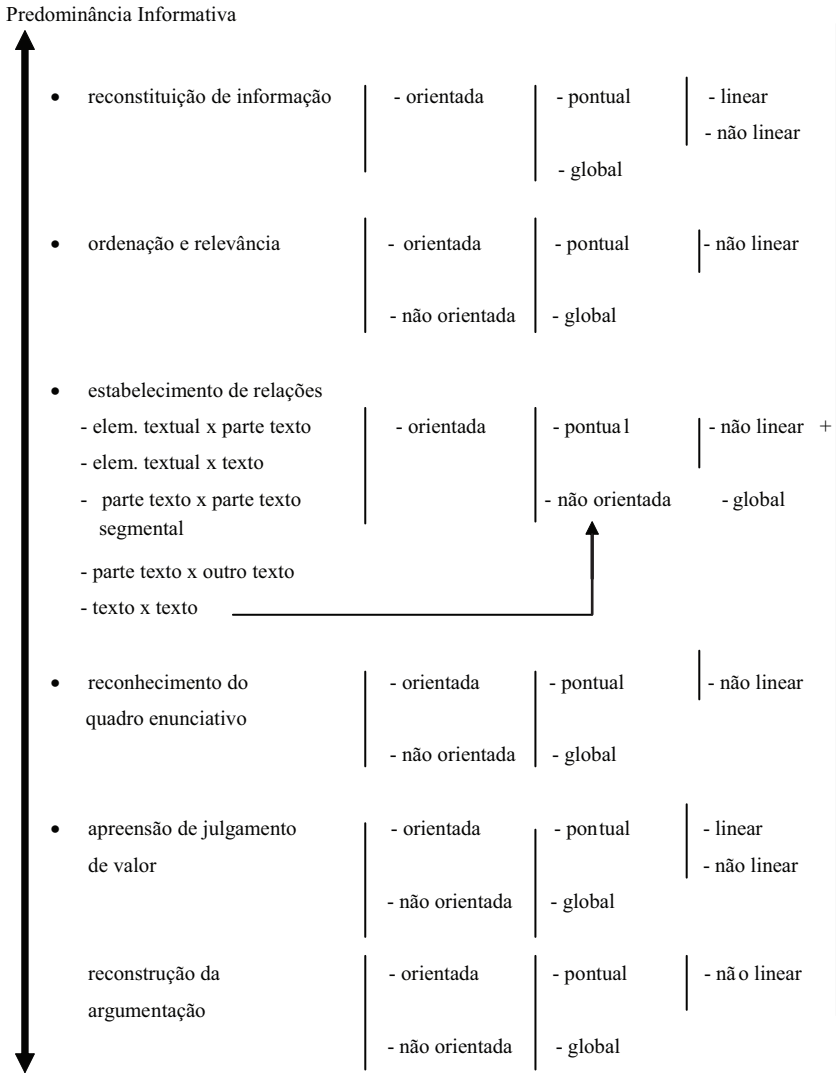
Essa tipologia de questões elaborada para respostas abertas[11] surgiu a partir da experiência dessas autoras como elaboradoras e corretoras do exame de francês do vestibular da Unicamp.[12] Essa tipologia se configura em um eixo que compreende dois pólos: *informativo versus argumentativo*, sendo que o grau de complexidade de leitura de um texto pode ser marcado por esses pólos. Segundo Cherem & Nery (1993, p. 99), o pólo *informativo* determina, por definição, um maior grau de simplicidade de leitura, enquanto que o pólo *argumentativo* marca um maior grau de complexidade. Os diversos tipos de questões de respostas abertas de um exame de leitura se distribuem nesse eixo, caminhando do *informativo* ao *argumentativo*, ou seja, do mais simples ao mais complexo. Dentro do eixo *informativo/argumentativo*, constam as seguintes categorias: *reconstituição da informação, ordenação e relevância, estabelecimento de relações, reconhecimento do quadro enunciativo, apreensão de julgamento de valor, reconstrução da argumentação*.

[11] Optei traduzir *open-ended questions* por questões de respostas abertas ou dissertativas.

[12] Apesar da taxonomia ter sido elaborada para questões abertas, percebi que ela poderia ser usada também para respostas fechadas, como a múltipla-escolha.

O esquema abaixo sintetiza a taxonomia de questões:

Figura 3 Esquema geral matriz de questões.



Fonte: NERY, R. *Questões sobre Questões de Leitura*. Campinas, SP: Alínea Editora, 2003, p. 56.

Segundo Cherem e Nery (1993) e Nery (2003),

> As questões de *reconstituição da informação* exigem apenas que o leitor/candidato identifique e extraia as informações solicitadas, tais como elas aparecem no texto.
>
> As *de ordenação e relevância* incidem sobre informações que se articulam no texto. Assim, a tarefa que se impõe ao leitor/candidato é reconstituir sua ordenação, a partir do grau de relevância das informações.
>
> As questões em que há *estabelecimento de relações* requerem que o leitor/candidato apreenda a relação existente entre dois ou mais elementos do texto. Essa relação pode ser entre:
>
> - um elemento textual e uma parte do texto,
> - um elemento textual e o texto todo,
> - uma parte do texto e outra parte do mesmo texto,
> - um texto e outro texto,
> - uma parte do texto e outro texto.
>
> As questões de *reconhecimento do quadro enunciativo* demandam que o leitor/candidato reconheça os componentes da cena enunciativa (como o sujeito que enuncia e como o sujeito a quem enuncia se constituem na configuração discursiva) e que compreenda as estratégias discursivas a que eles se prestam.
>
> As de *apreensão de julgamento de valor* exigem que o leitor/candidato apreenda segmentos do texto que veiculam um julgamento de valor - que se apresenta enquanto tal – sobre informações fornecidas no texto.
>
> As de *reconstrução da argumentação* requerem que o leitor/candidato reconstrua a linha de argumentação que filtra e articula a informação.
>
> Além das categorias do contínuo *informativo versus argumentativo*, subcategorias que a elas se combinam, foram determinadas, sobre a forma de "pares": *pontual x global, linear x não linear e orientada x não orientada.*
>
> Uma questão é *pontual* quando sua resposta exige uma apreensão localizada, quando as informações solicitadas (reconstituição da informação), os elementos a serem selecionados e ordenados (ordenação e relevância), as relações (estabelecimento de relações), o julgamento de valor apresentado enquanto tal (apreensão de julgamento de valor) incidem sobre um aspecto particular do texto; ou quando a questão incide apenas sobre um componente da:

- rede enunciativa (reconhecimento do quadro enunciativo),
- cadeia argumentativa (reconstrução da argumentação) do texto.

E é *global* quando sua resposta exige uma operação de atribuição de sentido que atua sobre o texto em sua globalidade, ou seja, a questão incide sobre:

- informações difusas *(reconstituição da informação*),
- elementos a serem selecionados e ordenados (*ordenação e relevância*)
- um julgamento de valor que se apresenta enquanto tal (*apreensão de julgamento de valor*), em todo o texto; ou quando sua resposta obriga a um movimento de:
- apreensão da rede enunciativa (*reconhecimento do quadro enunciativo*),
- reconstrução da cadeia argumentativa *(reconstrução da argumentação*) do texto em sua globalidade.

Uma questão *pontual* pode ser *linear* – quando o aspecto sobre o qual ela incide se localiza num ponto específico do texto enquanto materialidade, e somente em um – ou *não linear*, quando este aspecto se localiza em mais de um ponto do texto enquanto materialidade. Devido a sua natureza, as questões do tipo *ordenação e relevância, estabelecimento de relações, reconhecimento do quadro enunciativo e reconstrução da argumentação* são sempre *não lineares.*

Uma questão é *orientada* quando a própria formulação da questão contém orientações para um determinado percurso de leitura, ou seja, quando há orientações sobre:

- número e/ou natureza dos elementos a serem selecionados e ordenados *(ordenação e relevância*),
- elementos e natureza das relações a serem estabelecidas (*estabelecimento de relações*),
- a configuração da rede enunciativa do texto (*reconhecimento do quadro enunciativo*),
- a presença de um julgamento de valor apresentado enquanto tal (*apreensão de julgamento de valor),*
- a configuração da cadeia argumentativa do texto (*reconstrução da argumentação).*

Há, portanto, submissão do leitor-candidato ao universo discursivo do leitor-elaborador-da-questão, ou seja, do "primeiro leitor".

> Vale ressaltar, como já observado, que as questões do tipo reconstituição da informação, devido à sua própria natureza, são sempre orientadas, constituindo um "olhar" sobre o texto. Esse tipo de questão contém, na sua própria formulação, elementos da estrutura "informacional" do texto: diz-se ao leitor-candidato para que aspectos do texto ele deve "olhar", em detrimento de outros. Ela é, portanto, por si só e em si mesma, orientada.
>
> Será não orientada quando, na formulação da questão, não há nenhuma orientação sobre o percurso de leitura a ser feito para se chegar à resolução da questão. O que a formulação da questão apresenta é apenas um "foco" sobre o texto, e não um "olhar" sobre ele.

Segundo Nery (2003, p. 70), a Matriz de Questões aqui apresentada, resultante de uma prática de ensino e avaliação de leitura, deixa transparecer que ela não é o resultado da aplicação direta e mecânica de uma teoria, não sendo, tampouco, um produto acabado. Ela não pode ser, portanto, aplicada de maneira mecânica. Trata-se de um instrumental para ensino e avaliação de leitura que envolve categorias operacionais a serem manuseadas com flexibilidade.

Segundo a autora, o grau de complexidade de leitura exigida por uma questão, pode resultar, dentre outros fatores, da combinação das diferentes categorias:

Simplicidade	Complexidade
Informativo	Argumentativo
Pontual Linear Orientada	Global Não linear Não orientada
	Fonte: NERY (2003, p. 50).

Uma categoria só, por si mesma, não pode determinar o grau de complexidade de uma questão. Por exemplo, uma questão do tipo *reconstituição de informação global* pode ser mais complexa do que uma do tipo *reconstrução da argumentação orientada pontual*. Outros fatores externos ao texto também podem contribuir para o grau de complexidade, como o "conhecimento prévio do assunto", nível de conhecimento da língua estrangeira, dentre outros.

Para o propósito de análise dos tipos de questões do exame de inglês do vestibular da UFPR, a Matriz nos servirá de condutor para nos levar a uma compreensão mais minuciosa dessas provas.

Para que um exame de leitura seja considerado bem elaborado e inserido dentro de uma visão de leitura como construção do significado, é desejável que

tal instrumento de avaliação inclua não somente questões com predominância informativa, mas também questões com predominância argumentativa, pois, geralmente tais questões (se bem elaboradas e com um grau de complexidade adequado) requerem do leitor um nível de capacidade de leitura muito maior do que se ele tivesse que *apenas reconstituir informações orientadas pontuais e lineares* do texto.

2.3 O CONCEITO DE VALIDADE

A definição do que é validade tem mudado bastante desde sua introdução no mundo da avaliação em língua estrangeira nos anos 60 e 70. Até os anos 90, a validação de um exame passava pelo crivo de três conceitos distintos: validade, confiabilidade e praticidade. O conceito de validade, por sua vez, era subdividido, basicamente, em quatro subcategorias: 1) validade de conteúdo, 2) validade relacionada a critério, 3) validade de construto e 4) validade de face. A validade, nessa época, era frequentemente estabelecida através de correlações com outros testes.

2.3.1 Validade de conteúdo

De acordo com Hughes (1994, p. 22), um teste tem validade de conteúdo se seu conteúdo constitui uma amostra representativa das habilidades, sub-habilidades, estruturas etc., que tal teste pretende avaliar. Anastasi (1961, p. 135) provê um conjunto de diretrizes úteis para estabelecer validade de conteúdo:

> 1) O domínio do comportamento a ser testado deve ser sistematicamente analisado para nos certificarmos de que todos os aspectos principais são cobertos no teste, e proporcionalmente corretos;
> 2) O domínio em consideração deve ser prévia e totalmente descrito, ao invés de fazê-lo somente após o teste ter sido elaborado;
> 3) Validade de conteúdo depende da relevância das respostas individuais dos testes em relação à área de comportamento que está sendo considerado, ao invés da aparente relevância do conteúdo do item. (tradução minha)[xv]

O ajuste direto e a adequação da amostra do teste é, assim, dependente da qualidade da descrição do comportamento da língua alvo que está sendo testado. Portanto, a melhor salvaguarda contra testes que podem minar a vali-

dade de conteúdo, é escrever especificações[13] detalhadas do teste, e assegurar que o conteúdo do teste se torne um reflexo justo dessas especificações.

2.3.2 Validade de construto

A principal pergunta que um elaborador de um teste se faz quando está verificando a validade de construto é: "O que este teste está avaliando?" Indubitavelmente, esta pergunta não é facilmente respondível uma vez que nem todas as habilidades linguísticas são diretamente observáveis.

Hughes (1994, p. 26) define validade de construto da seguinte maneira:

> Um teste, ou parte dele, ou uma técnica de teste, tem validade de construto se pudermos demonstrar que ele avalia somente a habilidade (não o conteúdo) que deveria estar avaliando. A palavra 'construto' refere-se a qualquer habilidade(ou característica) subjacente que foi descrita numa teoria de habilidade linguística. (tradução minha)[xvi]

Bachman (1991, p. 254) afirma que a "validade de construto preocupa-se em saber até que ponto o desempenho nos testes é consistente com as predições que fazemos baseados nas teorias de capacidades ou construtos."

Alderson (1995, p. 183) vê tal concepção da seguinte maneira:

> O termo construto refere-se ao construto psicológico, um conceito teórico sobre um aspecto do comportamento humano que não pode ser mensurado ou observado diretamente. Exemplos de construto são: inteligência, motivação, ansiedade, atitude, domínio da língua e compreensão em leitura. Validade de construto é um processo que une evidências que dão suporte ao argumento de que um dado teste realmente mensura o construto psicológico que o escritor do teste pretende mensurar. O objetivo é assegurar que os escores significam o que nós esperamos que eles signifiquem. (tradução minha)[xvii]

---

[13] Especificações de um teste são os conteúdos que foram trabalhados em aula em forma de lista a serem avaliadas. Podem ser embasadas em funções lingüísticas, tais como saber cumprimentar, pedir favor etc, ou em itens gramaticais, tais como presente simples, passado simples, ou ainda, serem mais detalhadas relacionando funções lingüísticas com gramática, léxico, gênero textual e assim sucessivamente. É desejável que as especificações usadas para nortear o planejamento das aulas sejam as mesmas para orientar a elaboração das avaliações.

A validade de construto deve ser uma preocupação central para qualquer tipo de avaliação, pois necessitamos responder à pergunta se nosso teste é realmente um bom instrumento para avaliar o que realmente queremos avaliar.

2.3.3 Validade referenciada em critério

De acordo com Weir (1990, p. 27), a validade referenciada em critério é predominantemente quantitativa e uma concepção *a posteriori*, que se preocupa com a extensão com que os escores dos testes se correlacionam com adequados critérios externos de desempenho. A validade relacionada em critério é dividida em dois tipos:

a) **validade comparativa** (*concurrent validity*) – "onde os escores dos testes são correlacionados com uma outra mensuração, geralmente um teste mais antigo e bem estabelecido, feitos ao mesmo tempo" (Weir 1990, p. 27). Hughes (1994, p. 23) nos dá um exemplo de validade comparativa: quarenta e cinco alunos têm que se submeter a um teste oral baseado em cinquenta funções que foram ensinadas durante um curso. Seria impraticável gastar mais de uma hora com cada aluno para nos certificarmos que eles foram proficientes nas cinquenta funções. Portanto, a validade comparativa é necessária. Um teste oral de dez minutos é aplicado aos alunos. Somente uma amostra representativa de funções será cobrada. Em seguida, uma amostra de alunos é escolhida aleatoriamente e submetida a um teste de uma hora cobrindo todas as cinquenta funções. São necessários quatro ou cinco avaliadores para assegurar a confiabilidade dos escores. Esse teste mais longo seria o de critério com o qual o teste original seria comparado e julgado. Os escores inteiros dos alunos seriam comparados com os escores que eles obtiveram no teste de dez minutos. Se a comparação dos dois testes mostrar um alto grau de concordância, então a versão mais curta do teste pode ser considerada um teste válido. Se a comparação mostrar um baixo nível de concordância, então a versão mais curta do teste tem que ser re-examinada.

b) **Validade preditiva** – de acordo com Weir (1990, p. 27), é quando "os escores de testes são correlacionados com um critério de desempenho futuro." Hughes (1994, p. 25) também dá um exemplo de validade preditiva quando um teste de proficiência, por exemplo, prediz a habilidade de um aluno lidar com um curso de graduação numa universidade britânica.

2.3.4 Validade de face

Este é um aspecto da validade bem mais subjetivo e menos científico do que os outros tipos de validade. De acordo com Hughes (1994, p. 27), considera-se a validade de face de um teste quando ele parece mensurar o que ele deveria estar mensurando, e isso, inevitavelmente, envolve o julgamento, tanto dos elaboradores de testes quanto dos examinandos. Quanto mais validade de face tiver um teste, mais motivados os examinandos ficam, não só por submeterem-se a ele, mas também por acreditarem serem os resultados um instrumento verdadeiro de avaliação. Se um teste carece desse tipo de validade, pode acontecer de ele não ser aceito pelos candidatos, professores e educadores como um teste válido, mesmo que ele tenha um alto grau de validade de conteúdo, de construto e de critério-relacional.

Anastasi (1961, p. 138) diz que a validade de face

> não é validade num sentido técnico; ela refere-se, não ao que o teste realmente mensura, mas ao que ele parece superficialmente estar mensurando. Validade de face tem a preocupação de 'parecer válido' para o examinando, quem o faz, para o pessoal administrativo que decide usá-lo, e outros observadores tecnicamente não treinados. Fundamentalmente, a pergunta da validade de face depende da opinião e da relação com o público. (tradução minha)[xviii]

Entretanto, problemas podem surgir. Um teste pode ter um alto grau de validade de face, mas carecer de validade de conteúdo e de construto. Um outro problema que pode surgir é quando professores e alunos consideram um teste com validade de face, mas o elaborador de testes não, ou vice-versa.

Além de se analisar a validade, os elaboradores também se preocupavam com a confiabilidade e praticidade de um exame.

2.3.5 O conceito de confiabilidade

Um teste tem que ser confiável para que possa ser válido. Quando consideramos confiabilidade, perguntamo-nos o quanto podemos confiar nos resultados dos procedimentos de um teste, ou ainda, se os mesmos resultados podem ser produzidos consistentemente? O objetivo da confiabilidade é produzir um teste que possa ter escores muito similares se o mesmo teste for aplicado

em populações similares (ou até a mesma), em momentos diferentes sem que tenha havido aprendizado entre eles. Quanto mais similares forem os escores entre os dois testes, mais confiável o teste é. Confiabilidade se preocupa com aspectos tais como:

1) o mesmo avaliador dará a mesma nota para o mesmo teste em momentos diferentes?
2) diferentes avaliadores darão a mesma nota para o mesmo teste?
3) o teste permitirá que os examinandos tirem a mesma nota em momentos diferentes?

Para estimar confiabilidade, são comumente usados o método do *teste-reteste*, o de *formas equivalentes* e o *split-half* (dividido pela metade). O método teste-reteste é estimado depois de se aplicar o mesmo teste duas vezes a um mesmo grupo de indivíduos. Calcula-se, então, o coeficiente de correlação entre os pares de escores das duas aplicações. O método de formas equivalentes consiste em aplicar dois testes que sejam equivalentes para um mesmo grupo de indivíduos e calcular o coeficiente de correlação entre os escores. O método *split-half c*onsiste em dar nota separadamente a itens de números ímpares e pares no mesmo teste e depois calcular a correlação entre estes dois subtestes. O resultado do coeficiente é, então, ajustado para uma confiabilidade de teste inteiro usando a fórmula *Spearman-Brown*. Este tipo de método é aplicado a testes de itens isolados porque os itens têm médias e variâncias iguais, isto é, os itens estão mensurando a mesma habilidade ou característica.

2.3.6 Praticidade

Além da validade e confiabilidade, um bom teste tem também que ser prático. Isso significa que ele deve ser empregado dentro de limites de tempo e orçamento disponíveis. Ele também deve ter um alto grau de custo-benefício. Por exemplo: um teste pode ser muito longo e, portanto, requerer muito tempo dos candidatos para responder e muito tempo para o professor corrigir. Um exemplo de teste não prático seria fazer uma entrevista de trinta minutos para avaliar desempenho oral de uma grande quantidade de candidatos, com somente poucos avaliadores disponíveis. Uma solução possível seria fazer uma entrevista em grupo, ou entrevistas individuais mais curtas.

2.3.7 O dilema da validade, confiabilidade e praticidade

O que está errado no teste[14] abaixo?

1) Elvis aprendeu a cantar
   a) em casa.
   b) na igreja.
   c) na escola.
2) Elvis frequentemente ia à igreja porque:
   a) seus pais eram pobres.
   b) ele aprendeu a cantar lá.
   c) seus pais eram religiosos.

Problema: de acordo com o gabarito, a resposta correta para a questão 1 é (b), e (c) para questão 2. Os alunos que respondem corretamente à primeira questão provavelmente transferem seu conhecimento prévio para a questão 2 e escolhem (b). Os itens são muito parecidos e confusos. É um exemplo de interdependência de itens.

3) O estilo de música cantado por Elvis era:
   a) música *country* branca.
   b) pop negro no sul.
   c) música *country* e *blues*.

Problema: A segunda alternativa não é uma continuação linguística lógica da questão.

Este teste, no entanto, pode ser considerado confiável pelo fato da resposta ser a mesma, independentemente do corretor, mas carece de validade, pois seus erros colocam em xeque o construto que está sendo avaliado.

Podemos concluir que um teste pode ser confiável sem ser válido; o contrário, entretanto, não é verdade. Para que um teste seja válido, ele tem que ser, necessariamente, confiável.[15] Isso nos conduz a um dilema: Weir (1990, p. 33) afirma que às vezes é essencial sacrificar um grau de confiabilidade para aumentar validade. Se, no entanto, a validade for sacrificada para aumentar a confiabilidade, acabaremos com um teste que mensurará confiavelmente algo

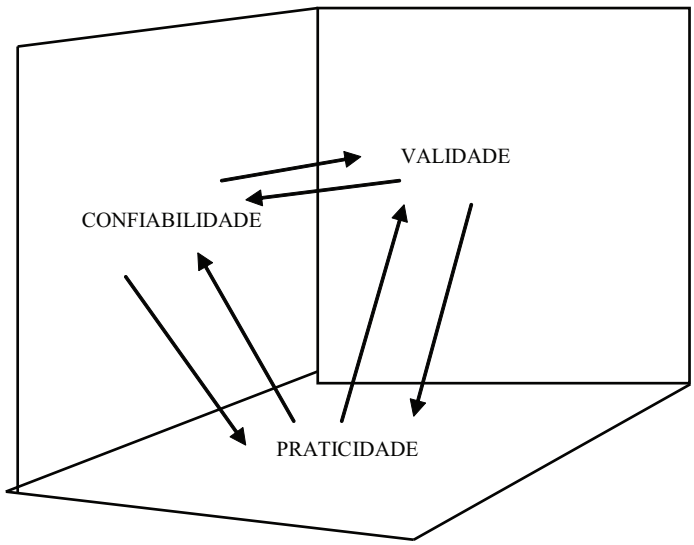
[14] Exemplo tirado de Shohamy, Elana. A Practical Handbook of Language Testing for Second Language Teachers. Tel Aviv, Israel. 1998. (edição experimental).

[15] Se a correção carecer de critérios bem estabelecidos, as notas dadas pelos corretores podem ser díspares e, portanto, a validade estará comprometida, pois não saberemos ao certo o que está sendo avaliado e como.

que não é o que queremos avaliar. Agora, se um teste for válido e tiver um bom grau de confiabilidade, mas não for prático, então esse teste também deve ser revisto. É muito difícil, se não impossível, alcançar um bom nível das três qualidades ao mesmo tempo.

O dilema pode ser visualizado da seguinte forma:

Figura 4 O dilema tridimensional de validade, confiabilidade e praticidade.



Fonte: Retorta (1996).

A figura acima nos mostra que um teste deve ter um equilíbrio entre os conceitos de validade, confiabilidade e praticidade, ainda que um deles tenha que ser sacrificado e diminuído. Mesmo assim, sempre devemos tentar um privilegiar a validade. Por exemplo: um teste de múltipla-escolha pode ter um alto grau de confiabilidade e praticidade, mas baixa validade. Por outro lado, uma entrevista de trinta minutos pode ter um alto grau de confiabilidade e validade, mas ter baixa praticidade. Em ambos os casos, um ajuste tem que ser feito em algum lugar para corrigir o desequilíbrio.

Nos anos 80, outros aspectos foram incorporados aos subconceitos de validade. Henning (1987) acrescentou a validade de resposta (response validity): até que ponto o examinando desempenhou a tarefa de maneira apropriada. Madsen (1983) falou sobre o aspecto afetivo, ou seja, o quanto um exame causa ansiedade excessiva. Hughes (1989) incluiu a validade do efeito retroativo, ou seja, o efeito que um exame causa no ensino e aprendizagem que precede ao exame.

A partir dos anos 90, um novo paradigma de validade surgiu. A validade passou a ser vista como um conceito unitário não concentrando sua análise no que um exame "mensura" *de per se,* mas na interpretação e uso que se faz dele (Bachman 1991b). Messick (1989, 1994, 1996), Bachman (1991b) e Chapelle (1990, 1993, 1994, 1999) são alguns dos pesquisadores que estão construindo um novo conceito de validação de avaliação.

2.3.8 Esquema de validação de Messick

No final dos anos 80 e começo dos 90, pesquisadores da área de avaliação em língua estrangeira lançaram um novo olhar sobre o conceito de validade. Uma noção proeminente na linguística aplicada surgiu quando Messick contestou a visão de validade tradicional na qual tipos de validade eram identificados (ver validade de conteúdo 2.3.1, de construto 2.3.2, referenciado em critério 2.3.3, de face 2.3.4 – Hughs, 1989). Para o pesquisador tal visão é inadequada (Bachman, 1990) e, em seu esquema, ele advoga a teoria unitária de validade, ou seja, o conceito enfatiza que a validade é um conceito único e não deve ser subdivido. Porém, ele distingue um número de facetas complementares dentro do conceito de validade unitária, na qual a natureza social da avaliação (valores e consequências do uso das notas) torna-se um aspecto chave. A validade de construto é essencial para cada faceta. (Yang 2006, p. 3)

As seis facetas distintas da validade que deverão fazer parte da noção de validade como um conceito unificado advogado por Messick (1996, p. 248/9; Bachman 1990, p. 236, Yong, 2006) são:

1. **O aspecto de conteúdo da validade de construto** – inclui evidências de relevância e representatividade de conteúdos para demonstrar que o teste é relevante e que cobre uma dada área de conteúdos ou habilidades.
2. **Análise de correlação ('The substantive aspect')** – inclui análises quantitativas para reunir evidências para dar sustento a notas de um exame e suas possíveis interpretações tais como confiabilidade de inter-consistência, correlação de itens, análise de fatores e de itens.
3. **O aspecto estrutural** – investiga a estrutura interna do exame avaliando até que ponto uma dimensionalidade observada das respostas coletadas é consistente com a dimensionalidade de construto que foi considerada. Nessa faceta análises qualitativas são utilizadas para

investigar os processos envolvidos em responder a um exame, empregando, para isso, abordagens tais como análise de protocolo, análise do tempo de resposta, análise de razões dadas pelo examinado por escolher uma resposta e não outra e análise de erros sistemáticos.

4. **O aspecto de generabilidade** – inclui em estudos longitudinais e transversais para examinar até que ponto as propriedades e interpretações das notas podem ser generalizadas para um dado grupo e além dele, assim como nos cenários e tarefas.
5. **Os aspectos externos (Manipulação e condições dos exames)** – inclui evidencias – oriundas de comparações 'multitrait-multimethod'[16] (MTMM: método de validação usada por Bachman & Palmer 1982, Stevenson, 1981, Swain 1990) – que sinalizam convergências entre exames bem como aspectos que os discriminam. Evidências de relevância de critérios (*criterion relevance*) e utilidades de aplicação (*applied utility*) também são analisadas.
6. **Consequências de um exame** – envolve a avaliação de valores e consequências intencionais ou não de interpretações de notas cujo foco principal está associado a questões de parcialidade de notas e interpretações, como injustiça no uso de um exame e como efeitos retroativos positivos e negativos que incidem no ensino e aprendizagem.

A validação de um exame, portanto, é uma avaliação empírica do significado e consequências da mensuração. Não devemos, também, deixar de levar em consideração fatores externos ao cenário que podem corroer ou promover a validade de interpretações e uso das notas. Para Chapelle (1999, p. 257) hoje, 'elaboradores de exames estão adotando, adaptando e contribuindo para as novas perspectivas de validação na área de educação'. Ela resume a história do conceito de validação no quadro a seguir:

[16] O MTMM é um método de validação. Vários exames com diferentes construtos são escolhidos para que cada construto seja mensurado usando vários diferentes tipos de métodos, e que evidências para a validação são confirmadas se as correlações entre os exames do mesmo construto são maiores que as correlações entre exames de diferentes construtos.

Quadro 6 Resumo de diferenças entre concepções de validade do passado e presente.

PASSADO	PRESENTE
A validade era considerada como uma característica do exame: até que ponto um exame mensura o que foi desenhado para mensurar.	A validade é considerada como um *argumento* em relação à interpretação e uso do exame: até que ponto interpretações e usos de um exame podem ser justificados.
A confiabilidade era vista como algo distinto da validade e uma condição necessária para que a validade existisse.	A confiabilidade pode ser vista como um tipo de evidência de validade.
A validade era frequentemente estabelecida através de correlações de um exame com outros.	A validade é *argumentada* embasada no número de tipos de '*rationales*' e evidências, incluindo as consequências de um exame.
A validade de construto era vista como um dos três tipos principais de validade (validade de conteúdo, de construto e referenciada em critérios).	A validade é um conceito unitário com a validade de construto central ao conceito (evidências de validade de conteúdo e referenciada em critérios podem ser usadas como evidências para dar suporte à validade de construto).
O estabelecimento de validade era considerado importante para a avaliação de exames de alta-escala e de alta-relevância.	Justificar a validade de um exame é responsabilidade de qualquer elaborador para qualquer tipo de exame.

Fonte: Chapelle (1999, p. 258).

Pesquisadores como Frederiksen and Collins (1989, p. 27) advogam a importância da validade sistêmica, ou seja, "uma avaliação, especialmente aquela autêntica e direta, tende a ser sistematicamente válida na medida em que ela induz mudanças curriculares e instrucionais em um sistema educacional que, por sua vez, fomenta o desenvolvimento de capacidades cognitivas que um exame foi elaborado para medir." Morrow (1986) também afirma que um exame deve ter validade de efeito retroativo (*washback validity*), isto porque a validade tem que estar atrelada à mensuração de quanto tal instrumento possa influenciar positivamente o ensino.

Messick (1996, p. 241-255) vai mais longe e advoga que o efeito retroativo faz parte integrante da concepção de validade unitária – **consequências de um exame**. Para o pesquisador (op. cit., 245)

> A validade é um julgamento abrangente de quanto as evidências empíricas e conhecimento teórico dão subsídios às interpretações e ações adequadas e apropriadas baseadas nas notas dos exames ou outros tipos de avaliações. A validade não é uma propriedade do exame ou avaliação em si, mas sim do significado das notas. Assim, o que necessita ser validado

> não é o exame ou qualquer instrumento de observação per se, mas as inferências derivadas das notas do exame ou outros indicadores – inferências sobre as implicações das ações que a interpretação acarreta.[xix]

Para Messick (op. cit.), o elaborador deve lutar contra a baixa representatividade ou irrelevância de conteúdo em um exame. Se ele se ocupar em elaborar um exame autêntico e direto, ou seja, uma avaliação que envolve uma simulação real de tarefa, além de estar validando as habilidades e capacidades dos alunos a desenvolverem tarefas reais, ele estará facilitando que consequências positivas ocorram no ensino e aprendizado que precede ao exame e isso contribui para arguir a favor da validade de construto.

2.4 EFEITO RETROATIVO

O efeito retroativo – *washback* ou *backwash effect* – é definido pela literatura em educação e em linguística aplicada como sendo a influência que um exame, seja ele de rendimento[17] ou externo,[18] exerce sobre o ensino e a aprendizagem que o precede. Hamp-Lyons (1997, p. 295) descreve o efeito retroativo de testes como sendo um conjunto de crenças sobre as relações entre teste, ensino e aprendizagem.

Shohamy (1992, p. 513) refere-se a efeito retroativo quando ela assevera que "a utilização de testes de línguas externos afeta e direciona a aprendizagem de língua estrangeira em contexto escolar. Esse fenômeno é o resultado de uma grande autoridade que testes externos têm na vida de examinandos". Alderson e Wall (1993, p. 117) afirmam que um teste pode influenciar professores e aprendizes de línguas estrangeiras a fazer coisas que eles não necessariamente fariam caso não fossem expostos ao exame. Hughes (1989, p. 01)

---

17 Provas de rendimento são instrumentos de avaliação que têm a função de diagnosticar o que o aluno aprendeu ou deixou de aprender em sala de aula. O elaborador desse tipo de exame deveria utilizar tal instrumento para apontar e corrigir falhas no ensino/aprendizagem que precedeu a prova. Muitos professores, porém, se restringem a utilizar o exame de rendimento somente para classificar alunos em aprovados ou reprovados.

18 Exames externos são instrumentos de avaliação que ocorrem fora da instituição de origem do aluno, tais como o exame de vestibular que tem por objetivo selecionar e classificar candidatos para vagas em instituições de ensino superior; ou como os exames SAEB, ENEM e PROVÃO que têm como objetivos coletar informações para diagnosticar a qualidade dos ensinos fundamental, médio ou superior; ou ainda exames de proficiência, tais como o Exame de Proficiência em Língua Portuguesa para Estrangeiros – Celpe-Bras, que avalia a proficiência da língua portuguesa de estrangeiros que queiram estudar ou trabalhar no Brasil.

afirma que "o efeito de testes no ensino e na aprendizagem é conhecido como '*backwash*" (sinônimo de *washback*) e advoga a importância de se promover um efeito positivo (benéfico) no ensino que o precede. Bailey (1996, p. 261) também defende a ideia de promover efeito positivo quando fala sobre os quatros princípios (ver item 2. 4. 1. 1) nos quais um teste comunicativo deve ser embasado. O quarto princípio é "*Work for washback* (trabalhe para que haja efeito retroativo): testes comunicativos devem ser explicitamente desenvolvidos para trazerem efeito retroativo positivo." Para a pesquisadora, "o efeito retroativo positivo deve ser uma meta primeira para elaboradores de testes

Outros termos que têm sido associados ao conceito de efeito retroativo, revelando as inúmeras controvérsias relacionadas a sua conceituação e abrangência, como destaca Scaramucci (2004a, p. 206), são: impacto do testes (*test impact* – Bachman; Palmer, 1996); instrução guiada por medidas (*measurement driven instruction* – Popham, 1987); alinhamento curricular (*curriculum alignment* – Shepard, 1993); retorno do teste (*test feedback)*; ensinar para o teste (*teach to the test*); validade sistêmica (*systemic validity* – Frederiksen; Collins, 1989); validade consequencial (*consequential validity* – Messick, 1989); validade retroativa (*washback validity* – Morrow 1991) e ainda testes como alavancas para mudanças (*levers for change* – Pearson, 1988).

2.4.1 Efeito retroativo: positivo, negativo, ambos ou nenhum?

Vários são os fatores que podem desencadear o efeito retroativo de um exame. A literatura na área afirma que exames de alta relevância (*high-stakes* exames) – como, por exemplo, os vestibulares de instituições públicas ou o exame da OAB – tendem a provocar um efeito retroativo mais forte do que exames menos relevantes como o SAEB (ver Retorta, 2005). Messick (1996, p. 243) acrescenta uma dimensão importante ao fenômeno quando ele diz que "evidência de influência no ensino e na aprendizagem deve ser considerada como efeito retroativo somente se tal evidência pode ser relacionada com a introdução e uso de um teste". Sabe-se também que um mesmo exame pode causar efeitos positivos para algumas pessoas e não para outras. Por exemplo: um exame pode fazer com que um professor se dedique mais para preparar seus alunos para uma prova do que outros professores; ou um aluno pode estudar mais para uma prova do que outros. Um exame pode, inversamente, causar um efeito negativo em algumas pessoas, mas não em outras. Por exemplo: um professor, por falta de conhecimento das concepções nas quais os exames estão embasados, pode somente ensinar macetes ao invés de trabalhar as

competências que o exame realmente avalia, e outros, por terem um sólido conhecimento tanto da disciplina quanto das concepções de linguagem, ensino e aprendizagem que estão alinhadas com as do exame, fazem um bom trabalho com seus alunos independentemente se o exame for bom ou ruim.

Portanto, um exame pode ou não desencadear um efeito retroativo positivo no ensino/aprendizagem que o precede. Quando tal fenômeno ocorre, pode atingir, com intensidades diferentes, pessoas envolvidas no processo *stakeholders* (participantes do processo – tradução minha) como são chamadas as pessoas que podem influenciar e/ou serem influenciadas por um exame, como alunos, professores, pais, diretores/coordenadores, secretário da educação, dentre outros. Temos que ter cautela ao falarmos de efeito retroativo, pois mudanças no ensino/aprendizagem nem sempre surgem a partir de um exame. Vários outros fatores externos ao exame, como, por exemplo, diretrizes elaborados pelas secretarias municipais e estaduais, fatores econômicos e sociais da comunidade, por exemplo, podem contribuir para reformas.

2.4.1.1 Efeito retroativo positivo

Alguns pesquisadores e elaboradores de exames advogam que um exame, *a priori*, deve ser feito para produzir efeito retroativo positivo. Outros são mais cautelosos e afirmam que o efeito retroativo é um fenômeno muito complexo e, por isso, não podemos ter uma visão determinista que o exame certamente exercerá uma influência no ensino que o precede. Wall e Alderson (1993, p. 117) afirmam que testes podem ser poderosos determinadores, tanto positivos quanto negativos, do que acontece em sala de aula. Frederiksen e Collins (1989, p. 27) afirmam que exames que avaliam desempenho, especialmente aqueles que são autênticos e diretos, provavelmente terão validade sistêmica, ou seja, tais exames influenciam os currículos de um sistema educacional e induzem mudanças no ensino para que as habilidades cognitivas que o exame avalia, desenvolvam-se. Morrow (1986) diz que "a validade de um teste (washback validity) deve ser avaliada pelo grau de influência positiva que um teste provoca no ensino". No capítulo "Alcançando efeito retroativo benéfico", Hughes (1989, p. 44-47) esboça sete maneiras de promover efeito positivo:

1) Avalie as habilidades cujo desenvolvimento você quer encorajar.
2) Escolha amostras amplas e imprevisíveis.
3) Use exames diretos.
4) Faça o exame referenciado em critérios.

5) Baseie exames de rendimento em objetivos.
6) Assegure-se que o exame é conhecido e compreendido pelos alunos e professores.
7) Quando necessário, providencie assistência aos professores.

Hughes (1989, p. 47) reconhece que algumas das sugestões dadas podem ser bastante dispendiosas e irão violar o critério de avaliação da praticidade. Mas ele assevera:

> Antes de decidirmos que não podemos despender recursos para avaliar de uma maneira que promoverá um efeito retroativo benéfico, temos que nos perguntar: qual será o custo de não alcançarmos um efeito retroativo benéfico? Quando compararmos o custo de um exame com o desperdício de esforços e tempo por parte dos professores e alunos em atividades bastante inadequadas para seus verdadeiros objetivos... estamos propensos a decidir que **não** podemos no dar ao luxo de **não** introduzir um exame que possa ter efeito retroativo benéfico poderoso.

Morrow (1991) cita cinco características que uma prova deva incorporar se pretende ser considerada um bom exame e provocar, potencialmente, um efeito benéfico: o exame deve contemplar todas as habilidades (leitura, escrita, fala e compreensão auditiva); ele deve avaliar o desempenho; deve ser baseado em tarefas e referenciado em critérios e deve *refletir e encorajar boas práticas em sala de aula* (esta última um efeito positivo). Morrow (1991, p. 112) conclui que essa ligação consciente entre exame e ensino, em termos não somente de conteúdo, mas também de abordagem, é um mecanismo vital para o desenvolvimento da educação.

Canale e Swain (1980), Swain (1984, p. 185-201), dentre outros pesquisadores, afirmam que o desenvolvimento de um exame de língua deve ser elaborado para provocar um efeito retroativo positivo. Bailey (1996, p. 261) diz que efeito retroativo positivo deve ser um objetivo primário para um elaborador.

Então, quais são os possíveis efeitos positivos ou benéficos? Bailey (1996, p. 264), ao sugerir um modelo básico do efeito retroativo, afirma que um exame pode influenciar quatro tipos de 'participantes' do processo de avaliação: os alunos, os professores, os elaboradores de materiais junto com os técnicos que desenvolvem currículos e os pesquisadores. Um exame pode fazer os alunos estudarem mais e desenvolverem competências da língua que serviriam para o uso do dia-a-dia, por exemplo. Nos professores, o exame pode induzi-los a buscar embasamento teórico para compreender as concepções que estão por trás

do exame e de boas técnicas de ensino. Para os elaboradores de materiais e técnicos de educação, o exame pode nortear seu trabalho, assim como dar embasamento suficiente para que eles possam desenvolver materiais e documentos coerentes e eficientes, além de fazê-los buscar compreensão teórica de exame, de ensino e de aprendizagem. Para os pesquisadores, um exame pode fornecer dados suficientes para que eles possam retroalimentar todos os participantes com resultados importantes para o melhoramento do ensino/aprendiagem/avaliação.

Bailey (1996, p. 268-272), além de falar sobre os participantes do processo de avaliação, também advoga quatro fatores que podem promover um efeito retroativo positivo. O primeiro é 'objetivo de aprendizagem de uma língua', ou seja, um exame pode promover (ou impedir!) um cumprimento de objetivos educacionais. O segundo é 'aumento de autenticidade': um exame deve reproduzir os tipos de situações que os alunos participam no dia-a-dia. Haverá uma congruência total entre avaliação e a vida real quando uma tarefa pedida em um exame corresponder a uma situação idêntica que o aluno desenvolverá no futuro. O terceiro é 'autonomia do aprendiz e auto-avaliação': um exame pode desencadear nos alunos um mecanismo de auto-avaliação. Isso poderá permitir a eles assumir mais responsabilidade na avaliação de sua proficiência, diagnosticar suas áreas mais fracas e obter uma visão realista de suas habilidades e perfis. Permitirá também que eles conheçam sua proficiência atual em relação ao nível que desejam chegar, e um exame pode ajudá-los a motivar-se e orientar-se por objetivos. O quarto e último efeito são 'relatórios de desempenho' – os resultados de um exame tem mais chances de promover um efeito retroativo positivo se as informações dos documentos forem detalhadas, inovadores, relevantes e diagnósticos. A partir de uma visão minuciosa que um bom relatório possa oferecer muitas decisões de mudanças, para melhor, podem ser tomadas.

Para melhorar um sistema de avaliação, Kellaghan e Greaney (1992, p. 7) apresentam algumas recomendações que, acreditam eles, diminuem os efeitos negativos de um exame em sala de aula:

1) Exames devem refletir todo o currículo, não meramente um aspecto limitado dele.
2) Habilidades cognitivas de alto nível devem ser avaliadas para assegurar que foram trabalhadas em sala de aula.
3) Habilidades que serão avaliadas não deverão ser limitadas às áreas acadêmicas, mas também devem ser relevantes em relação às tarefas que acontecem fora da escola.

4) Uma variedade de avaliações diferentes deve ser usada, incluindo exame escrito e oral.
5) Ao avaliar resultados de exames publicados e classificação nacional, outros fatores além do esforço do professor devem ser levados em consideração.
6) Dados detalhados devem oferecer às escolas informações sobre o nível de desempenho dos alunos e áreas de dificuldades.
7) Estudos sobre validade prognóstica de exames públicos devem ser conduzidos. (Isso é para ver se os exames estão satisfazendo seus propósitos.
8) A competência profissional de autoridades que trabalham com exames necessita ser desenvolvida, especialmente na área do construto do exame.
9) Cada grupo de examinadores deve ter capacidade para fazer pesquisas, isto é, para investigar, dentre outras coisas, o impacto que o exame tem no ensino.
10) Autoridades da área de exames devem trabalhar juntamente com organização de currículos e administradores educacionais.
11) Uma rede profissional regional deve ser criada para iniciar programas de trocas e compartilhamento de interesses comuns.

Wall (1996, p. 346), ao falar sobre o projeto Sri Lanka (Alderson e Wall, 1990, 1991), chama a atenção sobre os seguintes efeitos positivos que um exame pode causar:

1) Fornecer uma avaliação válida e confiável do rendimento do aluno.
2) Fornecer informações para nivelamento, ou seja, estipular um patamar de proficiência que o aluno deva atingir.
3) Avaliar o rendimento de alunos mais fracos assim como dos mais fortes.
4) Encorajar atitudes positivas e atividades de sala de aulas bem sucedidas.
5) Avaliar a eficiência e eficácia do curso e dos professores.
6) Restaurar confiança pública do exame.

Shohamy (1992, p. 515), ao propor um modelo de retro-alimentação de informação para a avaliação e diagnóstico de aprendizes de língua estrangeira, enfatiza a necessidade de haver uma conexão entre exame e currículo. Seu

modelo é baseado em seis princípios nos quais um exame deve embasar-se para provocar um efeito positivo. Caso um dos princípios seja negligenciado, um efeito negativo muito provavelmente surgirá. Os princípios que devem ser respeitados são:

1) *Rendimento escolar e proficiência*: ela afirma que nem sempre o que se aprende na escola (rendimento escolar) é o que é necessário para se usar na vida real (proficiência). Nem sempre o que o exame exige de um aluno é o que ele realmente vai precisar para vida. Exame, ensino, aprendizagem devem convergir com necessidades reais da vida.
2) *Informação diagnóstica*: muitas vezes os resultados de um exame não são utilizados para diagnosticar problemas (em várias dimensões – aluno, professor, escola, currículo etc.), para que sejam solucionados. A função diagnóstica é primordial para que qualquer sistema seja avaliado e retro-alimentado com informações para que ações corretivas sejam tomadas.
3) *Conexão do ensino com aprendizagem*: mudanças na instituição acontecerão de acordo com informações obtidas dos resultados do exame. O ensino deve estar em sintonia com as expectativas dos alunos e vice-versa.
4) *Envolvimento dos agentes que podem trazer mudanças*: um teste somente terá um impacto instrucional positivo, se os agentes (professores, administradores da escola, técnicos e secretários do departamento de educação) envolverem-se no processo, pois são eles que deverão conduzir as mudanças.
5) *Necessidade de informações comparativas*: um exame ideal será aquele referenciado em norma e em critério para que seus programas possam usar os resultados para avaliar seu sucesso em relação ao seu próprio programa e outros programas.
6) *Necessidade de exames comunicativos*: um exame ideal refletirá teorias de linguagem atuais e se concentrará em tarefas e situações de linguagem autênticas.

Todos os princípios, recomendações e fatores supracitados que promovem o efeito retroativo podem desencadear um efeito contrário, negativo ou maléfico, se mal interpretados ou negligenciados.

2.4.1.2 Efeito retroativo negativo

Um exame de alta relevância pode exercer vários tipos de efeitos negativos como, ansiedade nos alunos e professores, ensino de macetes, o uso de somente um método de avaliação de leitura (a múltipla-escolha, por exemplo) em sala de aula, estreitamento de currículo, dentre outros. Muitas vezes, efeitos retroativos negativos surgem pelo fato de os participantes (principalmente professores, elaboradores de material, diretores/coordenadores de escola, autoridades das secretarias de ensino, dentre outros) desconhecerem as concepções correntes que subjazem a um exame. Muitos profissionais, por desconhecerem as novas tendências em avaliação ficam presos e reféns de abordagens e técnicas antigas e ultrapassadas.

2.4.1.3 Dimensões do efeito retroativo

O efeito retroativo apresenta várias dimensões. Watanabe (2004, p. 20-21) descreve cinco dimensões diferentes que o efeito pode ter. Cada dimensão representa um dos vários aspectos de sua natureza como, por exemplo:

> *Especificidade* – pode ser geral ou específico. Geral é quando um exame produz mudanças independentemente dos conteúdos e habilidades que ele avalia. Específico, por outro lado, é quando um exame leva em conta o conteúdo do teste ou um aspecto específico desse conteúdo.
>
> *Intensidade* – pode ser forte ou fraca. O efeito pode ser forte ou fraco dependendo da intensidade com que se manifesta na sala de aula. É forte quando determina praticamente todas as atividades de sala de aula ou todos os participantes; e fraco quando afeta parte da aula e/ou somente alguns professores e alunos.
>
> *Extensão* – pode ser longa ou curta. O efeito pode durar um período longo ou curto de tempo. No caso do vestibular, se, por exemplo, o efeito durar todo ensino médio como é o caso de algumas escolas que têm o *primeirão*, o *segundão* e o *terceirão* – os três anos voltados para o preparo para o vestibular, então podemos dizer que tem um efeito longo. Por outro lado, se a escola resolve somente preparar seus alunos para o vestibular no último semestre do terceiro ano, podemos dizer que o efeito é curto. Uma outra perspectiva dessa dimensão pode ser exemplificada quando um efeito que se observa logo após a introdução de um exame, diminui com o tempo, e, portanto, é curto; e longo, quando esse efeito permanece por um longo tempo depois de o exame ter sido implementado.

> *Intencionalidade* – pode ser intencional ou não intencional. Alguns exames provocam efeito retroativo mesmo não tendo a intenção como advogam muitos elaboradores de vestibulares. Por outro lado, alguns exames têm a intenção de provocar mudanças no ensino. (ver neste capítulo o item 2.4.1.1).
>
> *Valor* – pode ser positivo ou negativo. Um exame pode provocar mudanças para o bem – efeito benéfico ou positivo, ou mudanças que prejudicam alunos, professores e cursos – efeito maléfico ou negativo. O efeito retroativo tem sido associado à perspectiva intencional, ou seja, positivo se o exame for desenhado para causar mudanças, e negativo com a dimensão não intencional (ver neste capítulo os itens 2.4.1.1 e 2.4.1.2).

Como afirma Scaramucci (2004, p. 206), a dimensão positiva/negativa é complexa, pois envolve julgamento de valor. Um exame pode ser visto como positivo por alguns professores e negativo por outros; ou ser positivo para os professores e negativo para os diretores de escola, por exemplo.

Foucault (1979) propõe uma dimensão de *dominação e poder* do exame quando o vê como instrumento mais eficiente pelo qual a sociedade impõe disciplina e que contém todos os aspectos necessários para poder e controle. Também Shohamy (1994) apresenta evidências do discurso de avaliação para mostrar que "a autoridade (*decision-maker*) usa testes para poder e controle, especificamente para observar, vigiar, classificar, normatizar, julgar e punir". Shohamy (1993b) acrescenta que "autoridades da educação (*policy-makers*) de agências centrais, conscientes do poder de um teste e de seu efeito retroativo, usam-no para manipular sistemas educacionais, para controlar currículo e para impor novos livros-textos e novos métodos de ensino". No Brasil, a política de exames, como Enem,[19] SAEB,[20] Enade,[21] e vestibulares de várias instituições renomadas, ainda não contempla o efeito retroativo como um instrumento direcionador de currículos, ementas ou livros didáticos.

2.4.2 Estudos sobre efeito retroativo: breve panorama histórico

O efeito retroativo, no passado, era visto por um grupo de especialistas em avaliação como um fenômeno determinista, ou seja, acreditava-se que todos

[19] Enem – Exame Nacional do Ensino Médio

[20] SAEB – Sistema Nacional de Avaliação da Educação Básica

[21] Enade – Exame Nacional de Avaliação de Desempenho de Estudantes – faz parte do Sistema Nacional de Avaliação da Educação Superior – Sinaes.

os testes afetavam professores e alunos da mesma forma, e, consequentemente, o ensino e a aprendizagem. Tal efeito era visto por alguns como algo que direcionava a atenção dos professores para o conteúdo do teste, e, assim, o exame ditava as atividades nas escolas, ou seja, o que era avaliado era o que era ensinado. Pensava-se também (Ebel, 1979; Khaniya, 1990b; Pearson, 1988) que todo e qualquer exame influenciava atitudes, comportamento, e motivava professores, aprendizes e pais. Achava-se que os todos alunos estudariam mais quando soubessem que enfrentariam exames. Além disso, julgava-se (Davies, 1985; Morris, 1972; Wong, 1969) que os exames tinham o poder de sempre causar inovações curriculares ou que tais instrumentos eram necessários para assegurar que o currículo fosse colocado em prática. E ainda, alguns pesquisadores (Khaniya, 1990b; Pearson, 1988; Swain, 1985) eram adeptos de uma visão que afirmava que o efeito retroativo era um atributo inerente a qualquer exame.

Alderson e Wall (1993) traçam um panorama histórico sobre o efeito retroativo em língua estrangeira. Eles afirmam que um dos primeiros estudos feitos sobre o assunto na área de língua estrangeira foi conduzido por Airasian, Kellaghan e Madaus, intitulado "*The Effects of Standardized Testing*", publicado em 1982, porém planejado e executado nos anos 70. Tal investigação tinha como objetivo estudar o impacto que 'exames padrões' (*standardized tests*) tinham sobre as escolas irlandesas. Foi um estudo longitudinal, executado de 1974 a 1977, e com um desenho quantitativo, uma vez que havia um grupo de experimento e outro de controle. Alderson e Wall (1992) citam duas críticas que foram feitas em relação ao trabalho: primeiramente, como havia um grupo de experimento, a situação ficou bem artificial, ou seja, não havia a pressão e medo da aprovação ou reprovação, não havia a pressão para a aceitação ou recusa na entrada em nível secundário ou terceiro grau, ou qualquer outro tipo de consequências que naturalmente acontecem em ambientes comuns. Uma segunda crítica foi feita ao problema de 'controlar' as variáveis dependentes que eram as notas dos professores, as notas dos exames ou respostas dos questionários. Pouco se sabe sobre o que aconteceu nos grupos de experimento e de controle em termos comportamentais. O uso mínimo de observação de sala de aula não possibilitou ao estudo identificar outros aspectos que pudessem contribuir para o impacto como dizem Alderson e Wall (op. cit.) e Bailey (1996).

O uso de um experimento verdadeiro dificulta o estudo do efeito retroativo, pois o efeito não pode ser inteiramente separado de outras variáveis que influenciam o ensino e aprendizagem, algumas variáveis com relações ainda desconhecidas ao fenômeno.

Outras investigações da época foram a de Wesdorp (1982), nos Países Baixos, que averiguou a introdução de testes de múltipla-escolha para avaliar a língua materna e estrangeira; a de Hughes (1988), na Turquia, para estudar se inovações feitas em testes trariam mudanças nos currículos escolares; Khaniya (1990), no Nepal, que investigou o efeito retroativo da 'School Leaving Certificate' ao final do ensino secundário para entrar no nível superior. Herman and Golan (1993), também no Nepal, relataram que exames públicos afetaram as percepções de currículos da maioria das professoras. Todos esses estudos eram quantitativos e não incluíram observação direta do pesquisado e de eventos de sala de aula.

A partir de um trabalho publicado por Alderson e Wall (1992), intitulado *Does washback exist?*, o conceito de efeito retroativo começou a ser questionado. Baseados nas pesquisas pioneiras sobre o assunto – Alderson e Wall (1990, 1991); Kellaghan, Madaus e Airasian (1974-77); Hughes (1988); Khaniya (1990b); Smith, M.L. (1991); Wesdorp (1982) –, os pesquisadores começam a colocar as crenças do passado em xeque, pois alguns dos pressupostos que anteriormente acreditavam ser inerentes aos exames foram desmontados nos estudos posteriores a esse trabalho de Alderson e Wall. As investigações sobre o efeito (Alderson, 1992; Cheng, 1998; Gimenez, 1988, 1997, 1998, 1999; Hamp-Lyons, 1996; Scaramucci 1992, 1996, 1997, 1998 a, b, 1999 b; c. 2002; Wall, 2000; Watanabe, 1996) nos permitiram ver que o efeito retroativo não é um fenômeno completamente entendido e consolidado. Algumas hipóteses levantadas sobre o que o efeito pode causar no ambiente escolar não foram inteiramente confirmadas nem refutadas. Os pesquisadores começaram a perceber que outros fatores, sociais e psicológicos, também podem interferir no ensino/aprendizagem, fatores esses que não eram originários dos exames e, portanto, não podem ser atribuídos ao efeito retroativo do exame em si. As poucas investigações empíricas sobre o efeito, bem como o pouco conhecimento científico sobre o que realmente ele é e como se realiza na educação, faz surgir outros estudos sobre tal fenômeno. Alderson (1993, p. 06) afirma que "o que nós sabemos sobre o efeito retroativo é que tal fenômeno não é tão simples assim: o que influencia e como, quando etc., professores e alunos mudam seu comportamento e crenças é certamente complexo."

Alderson e Wall (1993, p. 7-9) fazem um convite aos pesquisadores na área de avaliação: desenvolver estudos sobre o efeito retroativo para confirmar ou refutar as quinze hipóteses por eles levantadas a respeito do efeito a partir de pesquisas feitas em Sri Lanka, no início dos anos 90. Suas hipóteses são:

1. Um teste influenciará o ensino;
2. Um teste influenciará a aprendizagem;
3. Um teste influenciará **o que** os professores ensinam;
4. Um teste influenciará **como** os professores ensinam;
5. Um teste influenciará **o que** os alunos aprendem;
6. Um teste influenciará **como** os alunos aprendem;
7. Um teste influenciará o **ritmo** e a **sequência** do ensino;
8. Um teste influenciará o **grau** e a **profundidade** do ensino;
9. Um teste influenciará o **ritmo** e **sequência** da aprendizagem;
10. Um teste influenciará o **grau** e a **profundidade** da aprendizagem;
11. Um teste influenciará **atitudes** em relação ao conteúdo, método, etc. do ensino e da aprendizagem;
12. Testes que têm consequências importantes causarão efeito retroativo;
13. Testes que não têm consequências importantes não causarão efeito retroativo.
14. Testes causarão efeito retroativo em todos os aprendizes e professores;
15. Testes causarão efeito retroativo em alguns dos aprendizes e professores.

Alderson e Hamp-Lyons (1996) chegam à conclusão de que "a existência de um teste, por si só, não garante o efeito retroativo, nem positivo e nem negativo", e sugere a expansão das hipóteses de efeito retroativo para:

16. Testes provocarão tipos e intensidade diferentes de efeito retroativo em alguns professores e aprendizes do que em outros;
17. A intensidade e tipo de efeito retroativo irão variar de acordo com:
    a) o *status* do teste;
    b) a quantidade de informação disponível sobre o teste;
    c) até que ponto o teste vai contra a prática de ensino corrente;
    d) até que ponto os professores estão dispostos e capazes para inovar.

Scaramucci (1998a, b, 1998/1999, 1999c, 2001/02, 2002a, 2004a, b) confirmou algumas hipóteses quando chegou às seguintes conclusões:

1. Mudanças introduzidas pelos exames não são suficientes para garantir inovações no ensino;

2. Um mesmo exame pode ter efeitos de intensidades diferentes em contextos diversos, pois há forças diferentes agindo, como, por exemplo, as diferentes formações de professores.

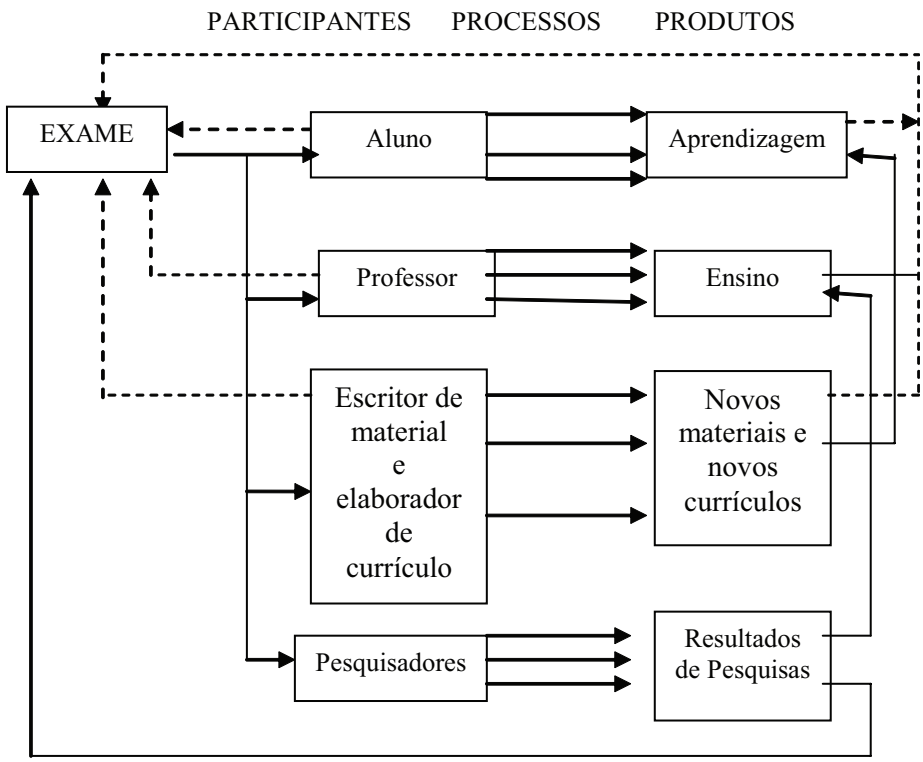
Gimenez (1999, p. 36) confirmou algumas hipóteses quando ela chegou às seguintes conclusões:

1. Partes de um teste terão efeito retroativo;
2. O efeito retroativo é dependente das crenças do professor a respeito do ensino;
3. O efeito retroativo é dependente das crenças do professor a respeito das chances de aprovação de seus alunos.

Bailey (1996, p. 264) cita Hughes (1989) e propõe um modelo básico de efeito retroativo no qual ela mostra um mecanismo no qual o efeito opera. Primeiramente, é necessário distinguir três elementos: os participantes, os processos e os produtos no ensino e aprendizagem e reconhecer que as três podem ser afetadas pela natureza do teste. De acordo com Hughes, os participantes incluem alunos, professores, administradores, autores de materiais e editoras, cujas percepções e atitudes em relação ao seu trabalho podem ser afetadas pelo teste. Dentro do processo, Hughes inclui qualquer ação tomada pelo participante que possa contribuir para o processo de aprendizagem. Tais processos incluem desenvolvimento de materiais, desenvolvimento de métodos, mudanças na metodologia do ensino, o uso de estratégias de aprendizagem e/ou de fazer testes (*test-taking strategies*). E, finalmente, o produto que se refere a 'o que foi aprendido' (fatos, habilidades etc) e a qualidade do que foi aprendido (fluência). Hughes (1989) afirma:

> A tricotomia em participantes, processo e produto nos permite construir um modelo básico do efeito retroativo. A natureza do teste pode primeiramente afetar as percepções e atitudes dos participantes em relação as suas tarefas de ensino e aprendizagem. Essas percepções e atitudes, por sua vez, afetam o que os participantes fazem quando fazem seu trabalho (processo) incluindo praticar os tipos de itens que são encontrados no teste que, por sua vez, afetará o resultado da aprendizagem, o produto do trabalho. (tradução minha)[xx]

Figura 5 Um modelo básico de efeito retroativo.



Fonte: BAILEY, 1996, p. 264.

Observando o modelo de Bailey, pode-se notar que as linhas pretas mostram quem influencia quem. Por exemplo: o exame influencia alunos, professores, escritores de materiais e elaboradores de currículos além dos pesquisadores. Esses por sua vez influenciam a aprendizagem, o ensino, novos materiais e currículos e novas pesquisas respectivamente. Os resultados de pesquisa influenciam tanto o ensino quanto o exame e novos materiais e currículos exame influencia o ensino. As linhas pontilhadas referem-se às influências que podem ocorrer, mas não que necessariamente acontecem. Então o aluno e sua aprendizagem, novos materiais e currículos podem provocar mudanças no exame.

Burrows (2004, p. 125-7), ao propor um novo modelo do efeito retroativo, compara-o com o modelo tradicional e os modelos dos anos 90. Para a pesquisadora, o novo modelo de efeito retroativo tem que levar em conta os relatos das crenças dos professores bem como as consequentes respostas às mudanças.

Figura 6 Teoria tradicional do efeito retroativo: modelo de estímulo-resposta.

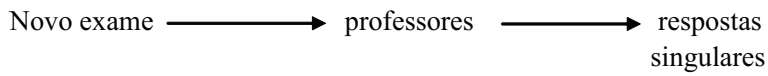


Figura 7 Efeito retroativo dos anos 90: um modelo da 'caixa preta'.

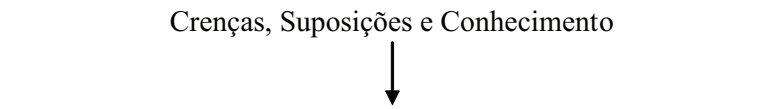
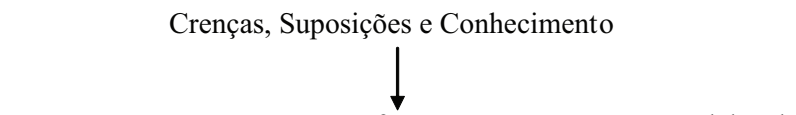


Figura 8 Uma proposta de efeito retroativo: um modelo de inovação de currículo.



O modelo tradicional tinha o efeito como um fenômeno determinista, ou seja, dado um exame de alta relevância bem desenvolvido, haveria consequentemente efeito retroativo positivo. O efeito positivo ou negativo dependia da qualidade do exame e não dos participantes do processo. Dessa concepção surgiu o conceito de *validade retroativa* de Morrow (1986) *e trabalhando para efeito retroativo* de Hughes (1989) e Swain (1985). O modelo da caixa preta foi alimentado com dados objetivos, observados e colhidos a partir de evidências empíricas. O artigo de Alderson e Wall (1992), levou a muitas pesquisas. Os resultados indicavam a existência de respostas individuais à implementação de um exame. A descoberta de que nem todos os professores envolvidos nos estudos respondiam de uma forma uniforme contradisse o modelo tradicional, já que os resultados indicaram que respostas singulares ao efeito não eram padrões.

O modelo de inovação de currículo relaciona o efeito retroativo à inovação de currículo e crenças, suposições e conhecimentos dos professores. Ele incorpora a visão de que a análise qualitativa das respostas dos professores à introdução de um novo exame ou sistema de avaliação pode revelar padrões

nas respostas. Tal modelo derivou de análises de respostas individuais de professores de padrões de comportamento. Ele é fundamentado na noção de que o efeito é uma forma de provocar mudanças na educação (Wall, 1996) e prevê que modelos de comportamento, propostos por outras áreas da educação (Markee, 1997; McCallum et al., 1995), podem ser aplicados ao efeito retroativo, uma vez que estudos do efeito e inovação de currículo têm em comum o exame de relevância que provoca mudanças educacionais no ensino. Um problema que pode surgir nesse modelo pode estar atrelado ao fato de que um único modelo de comportamento seria conceitualmente incompleto, pois pode não dar conta de variações no comportamento de professores que surgem das diferenças de objetivos, julgamentos e decisões.

2.4.3 Metodologia das pesquisas sobre efeito retroativo

Atualmente, os pesquisadores propõem a inclusão de alguns procedimentos quando são realizadas pesquisas sobre efeito retroativo. Tais procedimentos metodológicos sugeridos, e que faltaram em pesquisas anteriores aos anos 90, são a observação de aula e a triangulação dos dados dessa observação, acompanhadas pelas percepções dos professores e dos alunos. Para Bailey (1996, p. 273),

> os mais completos desenhos de pesquisa para o efeito retroativo incluem tanto as observações em sala de aula como perguntas aos participantes sobre suas visões e experiências (através de entrevistas ou questionários) para determinar se o ensino e/ou aprendizagem estão evidentemente ligados à introdução e uso de um determinado teste/exame.

As observações são registradas em fita cassete e são feitas anotações em diários. Alguns pesquisadores optam por utilizar **esquemas** de observação em sala de aula para estruturar tais observações (*class observation scheme*), como o COLT – *Communicative Orientation od Language Teaching Observation Scheme* (Spada; Frohlich, 1995) e o Instrumento de Observação – *Observation Instrument* – desenvolvido pela University of Lancaster e utilizado pela University of Cambridge Local Examinations Syndicate (UCLES) (Alderson e Banerjee, 2001; Savilli, 2000). As percepções dos professores e alunos são levantadas através de questionários e entrevistas – antes e depois da observação de aulas. Bailey (1996, p. 275) também afirma que "pesquisas sobre o efeito retroativo, necessariamente e por definição, devem ser longitudinais", e a observação em sala de aula deve ser um instrumento de coleta primária, ou seja, essencial para a investigação na área.

A partir das considerações metodológicas feitas por Bailey (1996) e Alderson e Wall (1993), surgiram, nos anos 90, várias pesquisas sobre o efeito retroativo de exames de EFL/ESL, com desenhos mais qualitativos, como os estudos de Shohamy (1996, p. 298-317); Alderson e Hamp-Lyons (1996, p. 280-297); Wall e Alderson (1993); Watanabe (2004, p. 129-146); Cheng (2004); Ferman (2004); Qi (2004, p. 171-190); Burrows (2004, p. 113-128); Hayes e Read (2004, p. 97-112); Saville e Hawkey (2004, p. 73-96); Stecher, Chun e Barron (2004, p. 53-72). No Brasil, as pesquisas sobre o efeito retroativo foram as de Scaramucci (1998 a ,b; 1999 b, c; 2002 a, b; 2004) e Gimenez (1997, 1998, 1999).

Porém, apesar de advogar as observações em sala de aula e estudos longitudinais como procedimentos adequados para a investigação do efeito retroativo, Wall e Alderson (1993) são cautelosos ao afirmar que as mudanças observadas em sala de aula não podem ser somente atribuídas a um exame em específico, uma vez que há falta de controle sobre outras variáveis. Tal método de coleta não daria conta de variáveis como conhecimento prévio dos alunos da língua estrangeira, conhecimento linguístico e teórico dos professores, infra-estrutura da escola, políticas de educação federais e estaduais, percepção dos pais e alunos do exame, percepção das autoridades locais do exame, percepção dos elaboradores de materiais do exame. Tais aspectos influenciam o ambiente escolar, mas não necessariamente estão preconizando o ensino balizado pelo exame de vestibular.

2.4.4 Estudos sobre o efeito retroativo no Brasil

No Brasil, a investigação sobre o efeito retroativo de exames de alta relevância, como o vestibular, é recente. As pioneiras no desenvolvimento desses estudos são Scaramucci (1992, 1998 a, b; 1999 b, c; 2002 a, b; 2004) e Gimenez (1988, 1997, 1998, 1999). Scaramucci investigou o efeito retroativo do exame de inglês do vestibular da Unicamp no ensino médio em uma escola pública, em uma privada e em um curso pré-vestibular de Campinas, São Paulo. Gimenez investigou o efeito retroativo do vestibular da Universidade Estadual de Londrina – UEL –, em cursinho pré-vestibular, bem como no ensino médio, na cidade de Londrina, no Estado do Paraná. Scaramucci (2002) chegou às seguintes conclusões sobre efeito retroativo do exame de língua estrangeira[22] do vestibular da Unicamp:

[22] O exame de língua estrangeira da Unicamp avalia a competência em leitura. No *site* da Comvest, a leitura é definida como o resultado de uma operação de atribuição de sentido que

1) A hipótese 'um teste influenciará **o que** os professores ensinam' foi parcialmente confirmada. Scaramucci (2002) notou que o impacto do exame de inglês do vestibular da UNICAMP no ensino médio foi parcial. Apenas alguns aspectos do exame foram percebidos e seu efeito identificado, tais como o enfoque do ensino voltado para o conteúdo do exame (leitura), a utilização em sala de aula dos mesmos tipos de textos empregados no exame, o treinamento em aula de questões e respostas escritas em língua portuguesa como no exame, e a inclusão, nos testes de rendimento, do método de avaliação do exame vestibular, ou seja, perguntas abertas também foram adaptadas para atividades de sala de aula.
2) A hipótese 'testes provocarão tipos e intensidades diferentes de efeito retroativo em alguns professores e aprendizes do que em outros' também foi averiguada. Percebeu-se que o efeito retroativo da prova de língua inglesa do vestibular da Unicamp foi maior no curso pré-vestibular. E menos intenso na escola pública. A escola particular ficou no meio-termo. Pode-se atribuir o fraco efeito na escola pública a várias razões: a infra-estrutura da escola pública é precária, o professor possui pouco entendimento das concepções de leitura implícitas no exame da Unicamp, falta de material adequado, visão mais tradicional do ensino de EL do professor da escola pública, dentre outros fatores.
3) A hipótese 'um teste influenciará **como** os professores ensinam' não foi confirmada, pois o exame não teve efeito retroativo na metodologia que o professor utilizou nem na abordagem. Apesar do exame de vestibular da Unicamp ser embasado em uma concepção na qual a leitura é vista como construção de sentidos, os professores continuaram suas aulas em um paradigma estruturalista – leitura como decodificação, com leitura dos textos em voz alta, seguida de suas traduções. Pôde-se perceber que nenhum dos professores parece ter entendido a visão de leitura implícita no exame. Portanto, a influência do exame na metodologia não existiu.

atua sobre o texto em sua globalidade, recuperando seu funcionamento. Ela não é uma tarefa passiva de simples decodificação de sentido. Trata-se de uma compreensão ativa que resulta na produção de um texto novo pelo leitor (assim é que diferentes leitores podem produzir leituras diferentes do mesmo texto, o que não significa, em outro extremo, que qualquer leitura possa ser feita).

Gimenez (1999), em sua investigação do efeito retroativo do vestibular da UEL no ensino médio chega às seguintes conclusões:

1) Foi novamente confirmada a hipótese 'um teste influenciará **o que** os professores ensinam', pois se percebeu que o conteúdo cobrado no exame foi cobrado em sala de aula. Além disso, o exame teve efeito retroativo sobre a avaliação, pois o simulado do exame foi usado como um instrumento de avaliação.
2) A hipótese 'um teste influenciará **como** os professores ensinam' foi confirmada somente no curso pré-vestibular, não na escola pública, isto é, o exame não teve efeito retroativo sobre a metodologia de ensino no nível médio da escola pública, mas teve algum efeito sobre a metodologia de ensino de cursinho.

Existem convergências nas conclusões de ambas pesquisadoras brasileiras como:

a) os exames estudados exerceram uma influência, mesmo que parcial, no conteúdo ensinado, confirmando assim que 'um teste influenciará **o que** os professores ensinam', mas não necessariamente '**como** os professores ensinam', pois somente um cenário foi afetado em relação à metodologia e mesmo assim parcialmente.
b) As percepções dos professores, alunos, instituições de ensino, diretores de escolas, escritores de livros didáticos, pais e qualquer pessoa envolvida direta ou indiretamente com os exames variam em relação ao efeito retroativo, e por isso causarão tipos e intensidades de efeitos diferentes.
c) o efeito retroativo potencial dos vestibulares estudados é um fenômeno complexo, não determinista, que merece mais investigações.

O efeito retroativo de um exame é consensualmente reconhecido como um fenômeno complexo. Além disso, é bastante difícil prever todas as variáveis que podem influenciar o ensino que precede um exame. Quando escolhemos investigar o efeito retroativo de um exame de vestibular, que muitas vezes é de alta-relevância (*high-stake exam*), devemos adotar os procedimentos defendidos pelos pesquisadores da área de língua estrangeira, como observação de sala de aula, para triangular os dados com as percepções de alunos e professores levantados através de questionários/entrevistas. Entretanto, além disso, como pode haver variáveis, muitas vezes ainda desconhecidas pelo pesquisador,

parece ser prudente investigar outros agentes envolvidos no processo, como as percepções dos diretores e coordenadores de escola, dos pais dos alunos, dos escritores de materiais didáticos, das autoridades estaduais e federais responsáveis pelo ensino de língua estrangeira nas escolas públicas, além de fazermos uma análise detalhada de todos os documentos oficiais: ementas e planejamentos das escolas e parâmetros e diretrizes federais e estaduais. Quanto maior for a abrangência de sujeitos estudados, melhor o fenômeno será compreendido e, assim, poderemos visualizar e entender outras variáveis envolvidas no processo ensino/avaliação que não seria possível de outra forma.

Neste capítulo, apresentei uma retrospectiva histórica da avaliação em língua estrangeira e sua relação com as concepções de linguagem e abordagens de ensino/aprendizagem. Também fiz um breve apanhado de modelos e visões de leitura e optei por um arcabouço teórico para este estudo. Além disso, discorri sobre a Matriz de Questões de Respostas Abertas de Cherem e Nery (1993) e Nery (2003) para que pudéssemos analisar as provas de inglês do vestibular da UFPR. Depois disso, defini os conceitos de validade, confiabilidade e praticidade, que são fatores importantes a serem considerados em uma avaliação. Por último, fiz um apanhado histórico do conceito efeito retroativo e apresentei as pesquisas feitas sobre o fenômeno.

NOTAS

[i] 'These shifts in emphasis in language teaching have inevitably had consequences for language testing. Testing techniques and theories, however, have been rather more resistant to change than theories about methodology and course design. This is principally because modern language testing is based on principles which, like the old 'structural' syllabuses, take as their starting point a description of the language independent of any particular use of it.'

[ii] 'There is an intrinsic reciprocal relationship between research in language acquisition and developments in language teaching on the one hand, and language testing on the other. That is, language testing both serve and is served by research in language acquisition and language teaching.'

[iii] 'A number of factors contributed to the development of interest in systematic 'scientific' language testing after the war. Wartime language programs in the United States and elsewhere and the growth of international agencies gave new importance (and funds) to language teaching projects. Methods of evaluating the effectiveness of these projects were required and the work done in the United States during this period quickly became the prevailing orthodoxy in the field of language testing. It would be difficult to exaggerate the extent to which current ideas about language testing have been influenced by this approach.'

[iv] "The 1920s and 30s saw a great vogue for psychological testing. Large numbers of tests investigating every aspect of the psyche form intelligence to job aptitudes were produced and

millennial predictions were sometimes made about the social benefits that large-scale testing of this kind would bring. Few of these tests actually delivered the miraculous solutions which had been promised but they survive today in the form of intelligence tests and, in a less serious form, as magazine quizzes of the kind 'Are you a good husband?'

[v] The psychometric tradition in psychology provided the tools for producing and developing tests. What was required was a basis for the content of the tests, which were produced. What kind of thing should be tested in a language test? Here, naturally enough, use was made of the same framework that was being on the work of the teaching programmes: a language description broadly based on the work of American 'structuralist' linguists. Crudely expressed, the analysis used involved breaking the language system down into small bits, and then describing the ways in which these bits could be put back together again to make stretches of speech The description was hierarchical in shape, the basic 'bits' being phonemes at the bottom of the pyramid which combines to produce morphemes, which….combine…etc.

[vi] Discrete point analysis necessarily breaks the elements of language apart and tries to teach them (or test them) separately with little or no attention to the way those elements interact in a larger context of communication. What makes it ineffective as a basis for teaching or testing languages is that crucial properties of language are lost when its elements are separated. The fact is that in any system where the parts interact to produce properties and qualities organizational constraints themselves become crucial properties of the system, which simply cannot be found in the part separately.

[vii] 'If a communicative approach to second language teaching is adopted, then principals of syllabus design must integrate aspects of both grammatical competence and sociolinguistic competence. Furthermore, teaching methodology and assessment instruments must be designed so as to address not only communicative competence but also communicative performance, i.e., the actual demonstration of this knowledge in real second language situations and for authentic communication purposes. It is also important to keep in mind that one cannot directly measure competence: only performance is observable.'

[viii] The concept of an integrative test was born in contrast to the definition of a discrete point test. If discrete items take language skill apart, integrative tests put it back together. Whereas discrete items attempt to test knowledge of language one bit at a time, integrative tests attempt to assess a learner's capacity to use many bits all at the same time, and possibly while exercising several presumed components of a grammatical system, and perhaps more than one of the traditionally recognized skills or aspects of skills.

[ix] 'What Oller said, in brief, was that language proficiency is indivisible, that tests only differ in their effectiveness at measuring this one factor, and that the elaborate apparatus of dimensions and tests used by the pychometrists could be replaced by one test which would directly tap the single indivisible faculty of language proficiency. Tests which were capable of doing this, Oller called 'integrative'; they included 'cloze' tests in which the candidate had to restore words blanked out at regular intervals in a text, and dictation, in which the candidate had to write down the words of a text read aloud.'

[x] Although Oller has claimed that his integrative tests represent total language proficiency better than any other single test or combination of tests, this is not in itself an argument in favor of the unitary competence hypothesis, as measures such as cloze and dictation are so

integrative that they contain most or all language abilities anyway. High correlations between cloze and other measures may only reflect that they are measuring different skills which are highly correlated among individuals; however, this does not mean that there will be no individuals whose performances in the various skills differ considerably.

[xi] 'One of the characteristic features of the communicative approach to language teaching is that it obliges us (or enables us) to make assumptions about the types of communication we will equip learners to handle. This applies equally to communicative testing.'

Um dos aspectos característicos da abordagem comunicativa para o ensino de língua é que tal abordagem nos força ou nos capacita a fazer suposições sobre os tipos de comunicação que nossos aprendizes necessitarão. Isso também se aplica a testes comunicativos.

[xii] Although language testing specialists have probably always recognizes the need to base the development and use of language tests on a theory of language proficiency, recently they have called for the incorporation of a theoretical framework of what language proficiency is with the methods and technology involved in measuring it.

[xiii] Communicative testing, as well as being concerned with what the learner knows about the form of the language and about how to use it appropriately in contexts of use (competence), must also deal with the extent to which the learner is actually able to demonstrate this knowledge in a meaningful communicative situation (performance), i.e. what he can do with the language – his ability to communicate with ease and effect in specified sociolinguistic settings.

Testes comunicativos têm a preocupação em descobrir o que o aprendiz sabe sobre a forma da língua e sobre como usá-la apropriadamente em contextos de uso (competência), e também têm que lidar como o aprendiz pode, de fato, demonstrar esse conhecimento numa situação comunicativa significativa (desempenho), isto é, o que ele possa fazer com a língua – sua habilidade para comunicar com facilidade e eficiência em cenários sociolinguísticos específicos.

[xiv] Other than serious marker reliability problems, associated with the assessment of performance, the major issue affecting an adoption of a `communicative' approach to language testing is the generalisability of the results produced by a test.

[xv] 1) 1) The behavior domain to be tested must be systematically analysed to make certain that all major aspects are covered by the test items, and in the correct proportions;

2) The domain under consideration should be fully described in advance, rather than being defined after the test has been prepared;

3) Content validity depends on the relevance of the individual's test responses to the behavior area under consideration, rather than on the apparent relevance of item content.

[xvi] A test, part of a test, or a testing technique is said to have construct validity if it can be demonstrated that it measures just the ability (not the content) which it is supposed to measure. The word `construct' refers to any underlying ability (or trait) which is hypothesised in a theory of language ability.

[xvii] The term construct refers to a psychological construct, a theoretical conceptualisation about an aspect of human behavior 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 eviden-

ce to support the contention that a given test indeed measures the psychological construct the makers intend it to measure. The goal is to assure that the scores mean what we expect them to mean.

[xviii] is not validity in the technical sense; it refers, not to what the test actually measures, but to what it appears superficially to measure. Face validity pertains to whether the test `looks valid' to the examinees who take it, the administrative personnel who decide on its use, and other technically untrained observers. Fundamentally, the question of face validity concerns rapport and public relations.

[xix] Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions base on test scores or other modes of assessment. Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. Hence, what is to be validated is not the test or observation device per se but rather the inferences derived from test scores or other indicators – inferences about score meaning or interpretation and about the implications for action that the interpretation entails.

[xx] The trichotomy into participants, process and product allows us to construct a basic model of backwash. The nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks. These perceptions and attitudes in turn may affect what the participants do in carrying out their work (process), including practicing the kind of items that are to be found in the test, which will affect the learning outcomes, the product of that work.