Universidade de Lisboa
Faculdade de Letras
Departamento de Língua e Cultura Portuguesa

A dimensão diagnóstica da avaliação em PLNM e o seu uso no Ensino Secundário Português

Licenciado Tiago Alexandre Barros Teixeira de Almeida Caliço

Mestrado em Língua e Cultura Portuguesa
(Língua estrangeira / Língua segunda)
Lisboa, 2008

Universidade de Lisboa
Faculdade de Letras
Departamento de Língua e Cultura Portuguesa

A dimensão diagnóstica da avaliação em PLNM e o seu uso no Ensino Secundário Português

Dissertação de mestrado orientada pela
Prof. ª Dr.ª Maria José dos Reis Grosso.

Licenciado Tiago Alexandre Barros Teixeira de Almeida Caliço

Mestrado em Língua e Cultura Portuguesa
(Língua estrangeira / Língua segunda)
Lisboa, 2008

Aos meus pais e irmão

Agradecimentos

Em primeiro lugar, gostaria de agradecer à minha orientadora, a Professora Doutora Maria José Grosso, pela sua orientação, sempre útil e pertinente, e pelas suas (constantes) palavras de encorajamento. O que não nos mata…

Em segundo lugar, gostaria também de agradecer aos meus colegas e formadores do curso de Formação de Formadores de PLNM, por me terem permitido perceber como conjugar a realidade dos factos com a realidade dos 'papéis' exige um constante golpe de rins, muita criatividade e trabalho.

Em terceiro lugar, aos muitos alunos que na minha curta carreira como professor já tive a oportunidade de conhecer. São a lembrança constante de quão pouco sei e de quanto mais preciso de aprender para ser (realmente) útil aos que também de mim dependem para as suas aprendizagens.

Por fim, à minha família: onde tudo começa e acaba.

Resumo

As alterações na sociedade portuguesa decorrentes da imigração dos últimos trinta anos reclamam do sistema público de ensino medidas diferenciadas que permitam integrar os alunos que não têm o Português como língua materna. Neste sentido foram delineadas medidas de integração no contexto escolar, as quais visam desenvolver a competência comunicativa dos alunos recorrendo a um programa de ensino diferenciado. Para que os alunos possam receber a formação mais adequada, decidiu-se criar grupos de nível de proficiência alinhados com Quadro Europeu Comum de Referência. A colocação dos alunos nestes níveis far-se-á com recurso a testes diagnósticos, a conceber pelas escolas ou por especialistas externos.

Com este trabalho pretende-se analisar as implicações do uso de instrumentos de avaliação diagnóstica para o processo de integração linguística dos alunos, mormente no concerne ao seu desenvolvimento, aplicação e controlo de qualidade. Pretende-se contribuir para uma problematização do conceito de avaliação diagnóstica, pouco desenvolvido na área da avaliação de proficiência em língua estrangeira. Procura-se identificar quais as características internas de um teste diagnóstico e qual a natureza do constructo a medir. Investiga-se também que implicações a natureza e utilização destes testes têm para um processo de investigação da sua validade.

Conclui-se recomendando algumas linhas de acção para o desenvolvimento de testes diagnósticos em paralelo com o desenvolvimento de padrões de desempenho e itens de avaliação/ensino que possam ser usados na promoção de um ensino de qualidade e mais produtivo.

Palavras-chave: avaliação, teste, diagnóstico, constructo, validade.

Abstract

The changes that the Portuguese society has suffered as a consequence of immigration in the past thirty years demand from the public educational system distinct measures that allow the integration of students who do not have Portuguese as a mother language. Thus school integration measures have been devised, which pretend to develop the students' communicative competence by means of a differentiated syllabus. So that students can receive the most adequate tutoring, proficiency level groups aligned to the Common European Framework were created. Placement of students in such groups will be done through the use of diagnostic tests, to be conceived by schools or external experts.

This work intends to analyze the implications of the use of diagnostic assessment tools to the integration process of students, particularly in what concerns to their development, application and quality control. It is intended to contribute to the problematization of the concept of diagnostic assessment, which is barely developed in the field of foreign language assessment. A tentative identification of the internal characteristics of diagnostic tests and the nature of their construct is done. The implications of the nature and use of such tests to a validation process is also appraised.

Recommendations to procedures to be used in the development of diagnostic tests, as to the development of performance standards and assessment/teaching items that may be used in the promotion of a higher quality and more productive teaching, are present in the conclusions.

Key-words: assessment, test, diagnostic, construct, validity.

Índice

Índice de figuras e tabelas

Lista de abreviaturas

GNP: Grupo de Nível de Proficiência

QECR: Quadro Europeu Comum de Referência

L2: Língua Segunda

LM: Língua Materna

LNM: Língua Não Materna

PLE: Português Língua Estrangeira

PLNM: Português Língua Não Materna

1. Introdução

1.1 Contexto do estudo e motivação

A realidade social portuguesa mudou nos últimos trinta anos. Os fluxos migratórios inverteram-se e aquele que era um panorama linguístico (tido) por homogéneo mudou. A realidade da sociedade portuguesa é também a realidade da escola portuguesa: 90 mil dos alunos do sistema público de ensino português não têm nacionalidade portuguesa[1]. Muitos[2] têm outra língua materna que não o Português.

Não obstante, apenas recentemente[3] foram desenvolvidas medidas de integração, escolar e social, para estes alunos. Estas medidas reconhecem a necessidade do desenvolvimento de linhas de acção diferenciadas que acomodem necessidades educativas distintas das que se encontram na generalidade da população escolar portuguesa, mormente o domínio de uma língua que não é a língua materna do aluno, mas é a sua língua de escolarização, socialização e relação com as instituições públicas.

Parte essencial desta medida de integração é a organização e aplicação de um programa de ensino de Português Língua Não Materna (PLNM) que se ancila em instrumentos de avaliação próprios e numa ligação ao Quadro Europeu Comum de Referências para as Línguas – Aprendizagem, Ensino Avaliação (QECR). Neste contexto a avaliação diagnóstica tem um papel fulcral, uma vez que é o instrumento que estará na base de decisões de impacto variado na vida dos alunos e na organização do trabalho dos professores.

A avaliação não é uma área com uma forte tradição de estudo em Portugal, particularmente quando aplicada ao Português Língua Estrangeira[4]. Não obstante, a avalia-

[1] Soares *et al.* (2006:7).
[2] Os dados mais recentes (Dionízio, 2005) representam apenas uma amostra de toda a população escolar. Neste estudo foram incluídas 13,3% das escolas, estimando-se que haveria 19369 alunos nestas escolas que não tinham o Português como língua materna.
[3] Cf. Soares *et al.* (2006:7).
[4] Vd. Pascoal (1992) como um dos poucos exemplos de um estudo docimológico dedicado exclusivamente ao PLE.

ção tem recebido nos últimos anos uma crescente atenção na opinião pública portuguesa e nos *media*. O programa de PLNM que agora se desenvolve com vista à integração dos alunos tem na avaliação um dos seus elementos fundamentais, particularmente a chamada 'avaliação diagnóstica', a qual, não obstante a indefinição em muitos documentos ministeriais do seu real significado, é de importância curial no caso do PLNM.

Paralelamente, desde a sua publicação, o QECR tem assumido um papel de destaque na organização de programas de estudo e de instrumentos de avaliação em Portugal, incluindo o ensino de Português, seja como Língua Estrangeira, seja como Língua Segunda.

Contudo, o papel que o QECR tem no desenvolvimento destes planos de estudo nem sempre é claro. Por vezes reclama-se uma ligação ao mesmo com base em simples análises qualitativas de programas/testes, sem que se faça uma investigação empírica da validade dessas asserções. Por outras, não é feita sequer uma afirmação de ligação, simplesmente se invocando o 'espírito' do QECR, ou a orientação subjacente ao projecto Línguas Vivas do Conselho Europa como justificação suficiente para decisões na organização dos sistemas de ensino. Não obstante subsistirem alguns equívocos na interpretação do que as escalas do QECR[5] realmente representam (e das implicações de tal facto para a utilização do QECR como ponto de partida para o desenvolvimento de planos de estudo/instrumentos de avaliação), este documento afirma-se como a base privilegiada para a tomada de decisões que afectam a vida de milhares de alunos. O programa de PLNM agora desenvolvido é de tal facto um exemplo.

De forma análoga, muitos dos conceitos na área da avaliação de proficiência em língua estrangeira, bem como das práticas que lhes estão subjacentes, nem sempre são claros. Muitas vezes encontramos definições contraditórias de termos, ou até mesmo o

[5] Vd. os artigos de Fulcher (2004) e North (2004) no semanário *Guardian Weekly*.

recurso a instrumentos para a definição de usos da avaliação. Nem sempre os instrumentos são desenhados com recurso a uma teoria psicométrica, ou são concebidos procedimentos de investigação que permitam aquilatar da qualidade e real impacto destes instrumentos (bem como de outras práticas educativas). Existe muita informalidade no desenvolvimento de aplicação destes instrumentos, o mais das vezes recorrendo-se simplesmente à repetição de formalismos herdados, ou até mesmo ensaiando-se a emulação de técnicas de testagem sem que se considere quais os pressupostos teóricos subjacentes às mesmas.

Numa perspectiva das práticas docentes, impõe-se conhecer de que forma os professores utilizam e concebem instrumentos de avaliação, particularmente testes, e que consequências (intencionais ou fortuitas) tais práticas têm no desenvolvimento da competência comunicativa dos alunos. O desenvolvimento de instrumentos de avaliação que sejam fiáveis, válidos e com impacto sobre as aprendizagens realizadas permitirá ganhos de produtividade e, no caso do PLNM, promover o objectivo de integração. Saber como os professores manipulam os instrumentos que têm à sua disposição, particularmente testes, permite ter uma perspectiva das suas necessidades de informação sobre as características dos seus alunos, bem como de formação e actualização.

Numa outra perspectiva, a avaliação é também o momento de tomada de decisões de grande impacto sobre a vida dos alunos. Nem sempre é transparente o processo que subjaz a estas tomadas de decisão, particularmente que padrões de desempenho (este entendido como uma manifestação de uma competência que se pretende desenvolver no aluno – neste caso a competência comunicativa) são esperados dos alunos para que estes lhes vejam atribuídos direitos ou reconhecidos graus de estudo. De especial importância é a possibilidade de comparabilidade entre decisões, ou antes, entre os instrumentos e os processos que levaram a tomadas de decisão de conteúdo semelhante.

1.2 Objecto e Objectivos do estudo

Este estudo visa analisar o modelo de desenvolvimento e aplicação de testes diagnósticos em PLNM tal como delineado pelo Ministério da Educação. Após uma revisão da bibliografia nacional e internacional sobre avaliação, diagnóstica e de competência comunicativa em língua estrangeira, cotejar-se-á a orientação seguida pelo Ministério da Educação com as boas práticas recomendadas, dando particular atenção à delimitação do conceito de teste diagnóstico e ao processo de ligação de qualquer teste ao QECR. Três perguntas de pesquisa orientam este trabalho:

A. Quais são as decisões no âmbito do programa de PLNM que requerem informação obtida através de instrumentos de avaliação?
B. Quais as implicações que decorrem dos usos dados aos instrumentos de avaliação no que concerne ao seu desenvolvimento, particularmente no que se refere a garantias de validade e justiça?
C. Quais as exigências decorrentes de uma asserção de uma ligação de um programa de estudos e correspondentes instrumentos de avaliação ao QECR?

Seguir-se-á o modelo de desenvolvimento e aplicação de testes proposto por Bachman (1990), Alderson, Clapham e Wall (1995), Bachman e Palmer (1996) e Bachman (2004), particularmente no que diz respeito às características de utilidade do teste e validação. No que concerne à ligação do programa de PLNM e dos testes diagnósticos, seguir-se-á a abordagem proposta pelo Conselho da Europa (2003, 2004) no processo de ligação de testes ao QECR. Uma vez que a generalidade da bibliografia sobre avaliação em língua estrangeira se dedica sobretudo à avaliação de competências, com fins de certificação, e pouca atenção é dada ao uso diagnóstico da avaliação, seguir-se-ão as propostas de Alderson (2005) no que respeita à problematização do con-

ceito de avaliação diagnóstica, particularmente no que concerne à definição do constructo a medir e à operacionalização desse constructo

São objectivos deste estudo:

A. Contribuir para uma problematização das implicações que uma asserção de ligação ao QECR traz ao desenvolvimento de instrumentos de avaliação (bem como de programas);
B. Contribuir para uma delimitação do conceito de 'avaliação diagnóstica', particularmente em contraste com os conceitos de 'teste de colocação em nível' e 'teste de conhecimentos', no contexto de ensino de PLNM;
C. Contribuir para uma problematização das implicações que o uso de instrumentos de avaliação, diagnósticos e de conhecimentos, têm para o programa de PLNM e para o mais lato objectivo de integração dos alunos.

Concluir-se-á com uma análise das implicações do modelo seguido para o desenvolvimento de testes diagnósticos para a fiabilidade, validade, impacto e *accountability* do sistema de ensino.

2. As noções de avaliação

2.1 Avaliação

A função 'avaliar' surge, paradoxalmente, como fonte de ansiedades relativas aos papéis educativos e como instância de legitimação e validação das actuações de distintos actores: planificadores e reguladores do sistema educativo, professores, alunos e pais. De formas diversas, uns e outros olham para a avaliação ora como panaceia para ineficiências de processos, ora como 'interferência' no normal desempenho de papéis educativos.

Não obstante, a avaliação tem sido alvo de atenção crescente na opinião pública, numa época e num contexto em que a aparente falência do sistema educativo reclama melhores resultados. É assim que alguns autores, p. ex. Abrantes (2002), Alonso (2002), Pinto e Santos (2006), Roldão (2006), Santos (2007), convocam a avaliação como novo instrumento na regulação e optimização dos ensinos e aprendizagens. 'Aprendizagem auto-regulada', 'avaliação de competências', 'avaliação sumativa', quase sempre apresentada em oposição binária à 'avaliação formativa'[1], alguns dos *leitmotivs* que nos últimos anos recorrentemente surgem associados à avaliação em contextos educativos e, particularmente, no quadro de uma reorganização do sistema.

A questão 'avaliação' tem implicações que não podem ser ignoradas. Por um lado, as consequências que uma reorientação das práticas avaliativas pode trazer para o *status quo* (entendido este como a relação, essencialmente tutelar, entre poder político e instituições de ensino, professores e alunos) são um factor de incerteza que, mais ou menos justificadamente, parece criar anticorpos junto de alguns dos intervenientes em todo o processo educativo. Por outro lado, o desejo, dir-se-ia quase consensualmente partilhado por diferentes actores educativos, de ganhos de qualidade nas aprendizagens

[1] Para definições destes conceitos vd. Abrantes (2001), Despacho Normativo 1/2005, de 5 de Janeiro, Decreto-Lei 74/2004, de 26 de Março, Pinto e Santos (2006), Ribeiro e Ribeiro (1989).

e no impacto destas na vida dos alunos e, por arrasto, no funcionamento da sociedade, despertou em camadas da sociedade a consciência e esperança de que novéis e mais eficazes práticas avaliativas são condição necessária para o sucesso da escola pública. É nesta tensão dialéctica que se joga hoje o discurso em torno do papel da avaliação no sistema educativo e da relação que cada interveniente no processo com ela estabelecerá.

Curiosamente, não é incomum a circunstância de uma contribuição para tal debate não se iniciar por um passo que, porventura de tão elementar a todos pareça, é de supina relevância: definir o que é 'avaliar', no contexto vigente. De facto, como veremos, não só alguma 'abundância' terminológica e nocional rodeia o conceito, como, essencialmente, o mesmo parece definido, seja expressa, seja implicitamente, por recurso ora a *usos*, ora a *técnicas*, ora a *objectos* da avaliação .

Dir-se-á que a avaliação, independentemente do grau de atenção que, por motivos tanto técnicos, como políticos ou até ideológicos, vai merecendo ao longo do tempo, sempre, de uma forma ou de outra, esteve presente na educação, no sistema educativo português e, consequentemente, nas práticas e nas mentes de planificadores, professores e alunos. Assim sendo, considerandos sobre a sua natureza são desnecessários, porque redundantes: ninguém inicia um trabalho sobre a influência da obra de António Lobo Antunes em jovens autores portugueses definindo o conceito de livro ou de literatura, de tão correntes que os mesmos são para quem escreve e para quem lê.

No entanto, nada garante que o conceito e, especialmente, a *representação do mesmo* que cada um destes intervenientes partilha sejam idênticos, o que pressupõe, no desenvolvimento e aplicação de qualquer sistema ou procedimento de avaliação, um problema aparentado ao da fiabilidade. A esta observação talvez deva acrescer outra: uma prática regida pelo costume, pela imitação e por um empirismo limitado à experiência pessoal ou de uma comunidade restrita em que um se encontra carece de capaci-

dade de generalização; problema, em última análise, da validade das práticas. Se assim é com o conceito de avaliação *latu sensu*, também será com as distintas variações do mesmo.

Uma definição possível de avaliação é a que decorre do Despacho-Normativo 30/2001[2]. Esta definição é importante, pois vincula, por mecanismo administrativo, os milhares de docentes do Ensino Básico a uma concepção de avaliação. Embora não defina explicitamente o que é avaliação, delimita as finalidades, objecto(s) e princípios da avaliação. A finalidade primeira é regular a "prática educativa, permitindo uma recolha sistemática de informações que, uma vez analisadas, apoiam a tomada de decisões adequadas à promoção da qualidade das aprendizagens". Como veremos, para alguns autores esta 'finalidade' da avaliação é a definição do próprio acto de avaliar. Entre as decisões que a avaliação visa informar encontram-se "apoiar o processo educativo", "certificar as diversas competências adquiridas pelo aluno" e "contribuir para melhorar a qualidade do sistema educativo".

Os objectos da avaliação educativa no Ensino Básico são "as aprendizagens e as competências definidas no currículo nacional". Entre os princípios em que a avaliação se ancila estão a "consistência entre processos de avaliação e as aprendizagens e competências pretendidas", a "primazia da avaliação formativa", a "valorização da evolução do aluno", a "transparência do processo de avaliação" e a "diversificação dos intervenientes".

Entre o fim, o objecto e os modos (ou meios?) de avaliação, a primeira parte deste Despacho proporciona uma definição do que é a avaliação, mas, simultaneamente,

[2] Embora revogado pelo despacho normativo 1/2005 de 3 de Janeiro, opta-se por nesta fase se analisar esta versão dos princípios orientadores da avaliação no Ensino Básico por três razões: 1) com a excepção da introdução da avaliação sumativa externa e da reapreciação da noção de avaliação diagnóstica, os despachos são essencialmente iguais, 2) esta versão foi acompanhada por um documento divulgador com artigos de vários autores ligados à avaliação educacional, 3) não obstante uma história da legislação sobre avaliação transcender o escopo deste trabalho, é maior a mudança entre a orientação que este despacho revoga do que entre este e o que lhe sucede.

um programa do que deve ser o *uso* da avaliação, as suas modalidades privilegiadas e qual o impacto esperado desta sobre o sistema educativo.

No entanto, resta a questão dos utilizadores da avaliação. Tomando-se a noção de avaliação como um processo de *manuseio* de informação, há que considerar a perspectiva do utilizador sobre a informação, ou antes, prever as diferentes necessidades e capacidades dos utilizadores e qual o conteúdo e forma da informação que requerem. A noção de 'intervenientes' não é suficiente neste caso. Não decorre do despacho uma definição clara de intervenientes, antes uma listagem destes e da sua forma de relacionamento com a avaliação, que parece aproximar-se da noção de 'stakeholders', ou interessados, como Bachman e Palmer (1996:31) a definem: aqueles que, directa ou indirectamente, têm um interesse em ou serão previsivelmente afectados pelos resultados da avaliação; mormente professores e alunos, mas também encarregados de educação, técnicos educativos e a tutela.

Uma outra definição possível de avaliação é a que encontramos em Peralta (2002:27): Avaliar é "a recolha sistemática de informação sobre a qual se possa formular um juízo de valor que facilite a tomada de decisões". Curiosamente, a definição oferecida pela autora, que a apresenta como "muito geral e, aparentemente, consensual", correlaciona-se fortemente com a 'finalidade' da avaliação apresentada no Despacho Normativo. De facto, com a pequena *nuance* do juízo de valor, são quase idênticas. No caso presente, a autora problematiza a noção da avaliação em função do **objecto** a avaliar, a competência, tal como é modernamente compreendida nas orientações curriculares. Alertando para a dificuldade de avaliar competências complexas, conclui que "o que podemos avaliar, pela observação do modo como um aluno age, em situações simuladas e intencionalmente construídas, é um conjunto de desempenhos ou o desempenho

global do aluno, a partir do qual podemos fazer generalizações sobre a sua competência".

Ora desta advertência decorre que a avaliação educativa, visando aceder a um determinado objecto intangível (a competência), tem de se reduzir a avaliar determinados comportamentos, inferindo-se depois um determinado estágio de desenvolvimento dessa competência, em função da sua manifestação (o comportamento). Partindo dessa constatação tomar-se-ão depois decisões, com base em determinados critérios de decisão. Deste modo, parece ser possível reconhecer dois momentos, ou duas subtis modalidades de avaliação: reconhecer o objecto[3] em função de um critério de semelhança ou inferência, decidir o que fazer, em consequência do cotejo dessa realidade com um determinado conjunto de regras pré-estabelecidas.

Em algumas tradições avaliativas, mais próximas da psicometria[4], esta destrinça parece resultar mais clara. Note-se a definição que o Quadro Europeu Comum de Referência (doravante 'QECR'), na sua versão inglesa, apresenta daquilo que na versão portuguesa é apenas denominado como 'avaliação':

> *"Assessment is used in this chapter in the sense of the assessment of the proficiency of the language user. All language tests are a form of assessment, but there are also many forms of assessment (e.g. checklists used in continuous assessment; informal teacher observation) which would not be described as tests. Evaluation is a term which is again broader than assessment. All assessment is a form of evaluation, but in a language programme a number of things are evaluated other than learner proficiency. These may include the effectiveness of particular methods*

[3] Termo também utilizado é 'constructo'. A este tópico dedica-se o Capítulo 4.
[4] Alderson, Clapham e Wall (1995), ALTE members (1998), Bachman (2004), Bachman e Palmer (1996).

or materials, the kind and quality of discourse actually produced in the programme, learner/teacher satisfaction, teaching effectiveness, etc."[5]

Resultando algo circular, a definição de 'assessment', aproxima-se do primeiro momento do que é avaliação: conhecer. No caso, conhecer a proficiência que alguém demonstra ter numa determinada língua estrangeira, ou o seu nível de desenvolvimento da competência comunicativa. 'Evaluation', por seu turno, já se aproxima do segundo momento do acto de avaliar: atribuir valor, significado externo à simples descrição do objecto de estudo, retirar daí consequências e, porventura, tomar decisões informadas. Por outro lado, aparentemente, 'assessment' e 'evaluation' parecem distinguir-se pelos objectos avaliados: 'assessment' recai essencialmente sobre a competência do aprendente/utilizador da língua, 'evaluation' sobre os elementos do contexto em que a aprendizagem tem lugar.

Na tradução portuguesa, esta subtileza terminológica perde-se: 'assessment' e 'evaluation' são traduzidas como 'avaliação', perdendo-se uma tenuidade implícita na definição proposta pelo Conselho da Europa: que só há 'juízo de valor' no momento de 'evaluation', não no momento anterior (e que o informa), de 'assessment'[6]. Retornaremos à noção de teste mais tarde, aquando de uma discussão mais aturada sobre técnicas e procedimentos.

Notemos uma outra definição destes dois conceitos, desta feita proposta pela *Association of Language Testers in Europe* (ALTE)[7]:

---

[5] Conselho da Europa (2001:177).

[6] Na tradução portuguesa, encontramos: "Neste capítulo, o termo 'avaliação' é usado no sentido de avaliação da proficiência do utilizador da língua. Todos os testes de língua são uma forma de avaliação, mas há também muitas outras formas de avaliação (p. ex.: as listas de verificação usadas na avaliação contínua, a observação informal do professor), que não são descritas como testes. A avaliação é um termo mais vasto que a testagem. Qualquer testagem é uma forma de avaliação, mas num programa de língua existem muitos outros aspectos, para além da proficiência do aprendente, que também são avaliados – p. ex.: a rentabilidade de determinados métodos ou materiais, o tipo, a qualidade do discurso efectivamente produzido, a satisfação do professor e do aprendente, a eficácia do ensino, etc."

[7] ALTE members (1998:135, 144).

Assessment: In language testing, the measurement of one or more aspects of language proficiency, by means of some form of test or procedure.

Evaluation: Gathering information with the intention of using it as a basis for decision-making. In language testing, evaluation may focus on the effectiveness or impact of a programme of instruction, examination, or project[8].

Mais uma vez, uma distinção é feita entre o momento de medição de aspectos da proficiência, e o momento de tomada de decisão. Poder-se-á objectar que tal distinção é espúria: mesmo que se considere que 'assessment' é apenas um sinónimo para 'gathering information', naturalmente a recolha de informação visa auxiliar ou fundamentar a tomada de informação; por outras palavras, ninguém faz 'assessment' se não pretender usar os seus resultados em determinado acto de 'evaluation'. No entanto, o que esta distinção permite (ou antes almeja) afirmar é que é possível descrever determinadas características de um ser humano, não físicas e não directamente mensuráveis, através de procedimentos que, não sendo objectivos como as ciências ditas exactas concebem o termo, não têm de ser, necessariamente, juízos de valor, como são entendidos pela Sociologia, a Moral ou até mesmo o Direito. Isto porque essa descrição de uma característica do indivíduo se baseia em instrumentos tidos por fiáveis e 'imparciais', ou seja, em que a inevitável subjectividade que qualquer classificação de um comportamento humano (e da suposta competência de que este é uma manifestação) por outro ser humano implica é conhecida e controlada. Para que a subjectividade possa ser conhecida e controlada, é necessário que se estabeleçam procedimentos sistemáticos e funda-

[8] Também na versão portuguesa destas entradas, 'avaliação' é o termo utilizado para ambas as definições, ainda que apresentadas em alíneas distintas.

mentados que servem de base à recolha de informação: ou seja, um instrumento de 'assessment'.

Por outro lado, muitas vezes a avaliação pode ser feita sem recurso a informação sistemática e fundamentadamente recolhida. É o caso, por exemplo, das decisões que um professor toma no âmbito da sala de aula em função da sua observação impressionista da reacção dos alunos à introdução de um novo tópico introduzido. Há recolha de informação e uma tomada de decisão (retomar a apresentação do tópico, passar a um tópico seguinte no programa, p.ex.). Mas se esta tomada de decisão não se ancila em informação recolhida de forma sistemática e com um fundamento teórico sólido não podemos propriamente falar em 'assessment'.

Noutra perspectiva, a informação recolhida através de um processo sistemático e teoricamente fundamentado pode não ser usada para auxílio à tomada de decisão. É o caso de testes desenhados com o fim de investigar uma determinada hipótese sobre a natureza do conhecimento ou uso da língua.

Bachman (2004:7), autor consistentemente citado na área da avaliação em língua, define hoje os conceitos de 'assessment' e 'evaluation' através de uma relação instrumento/uso. Assim, 'assessment' será "the process of collecting information about a given object of interest according to procedures that are systematic and substantively grounded". Por outro lado, considera que "evaluation, which involves making value judgment and decisions, can be best understood as a *use* of assessment" (*idem,* 9) (itálico no original). Esta relação instrumental não é exclusiva: nem todas as formas de 'assessment' são usadas em 'evaluation', nem sempre se faz 'evaluation' com base em informação conseguida através de 'assessment'. Por outro lado, ao contrário de posições anteriores[9], Bachman não considera que todas as formas de 'assessment' impliquem

[9] Cf. Bachman (1990, cap. 2)

uma medição. De facto, a medição ou quantificação, implicando a atribuição de valores numéricos a características de pessoas, não é uma forma exclusiva de 'assessment'. A forma como a informação recolhida é apresentada aos seus utilizadores pode ser numérica, mas também verbal e qualitativa, ou até mesmo pictográfica.

O esquema seguinte, adaptado de Bachman (2004), permite recortar mais claramente as fronteiras entre a dimensão 'instrumento' e a dimensão 'uso' daquilo que em português sói nomear-se como avaliação.

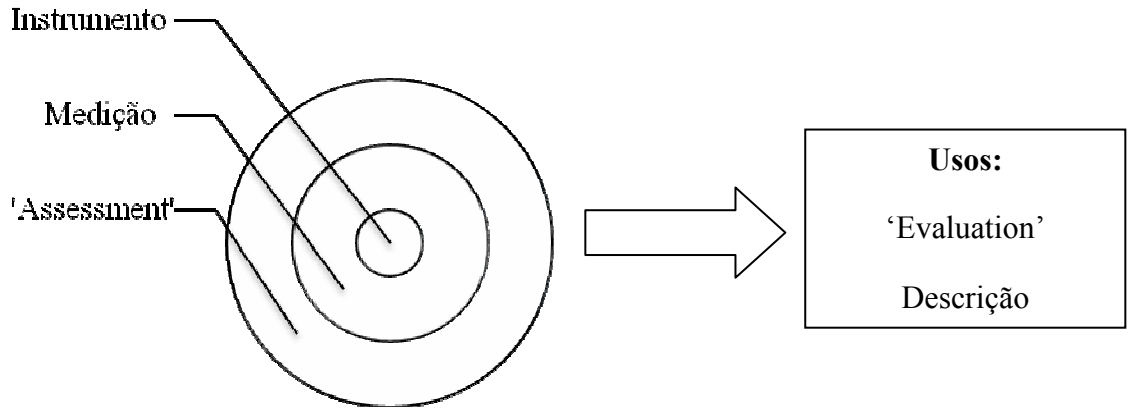

Fig. 1: relação entre 'assessment' e 'evaluation'.

Temos assim que, não só encontramos duas dimensões, independentes ainda que complementares, para o que entendemos como avaliar, mas que os critérios de decisão que orientam os processos internos a cada uma dessas dimensões também são independentes e porventura de natureza distinta. Se entendermos 'assessment' como um sinónimo de medição, então a base para a nossa caracterização do objecto é uma escala numérica. Se estendermos o nosso entendimento de 'assessment' a outras caracterizações que não necessariamente numéricas, então poderemos ter, por exemplo, um racional qualitativo e verbal. Por outro lado, a avaliação, entendida aqui como o processo de tomada de decisão, terá de recorrer a um conjunto de critérios de decisão que são independentes daqueles que orientaram a caracterização do objecto. Em traços

gerais, ‘assessment’ responde à questão ‘como é?’, ao passo que ‘evaluation’ responde à questão ‘o que fazer então?’.

A distinção entre estes dois momentos, note-se, não é de somenos importância. O juízo de valor não se confunde, não se pode confundir, com o juízo de facto, ainda que este tenha como objecto um conceito abstracto fisicamente manifestado e não uma característica ‘directamente observável’. Confundi-los é, no limite, fazer os dados adequarem-se às conclusões, ou, pelo menos, recusar qualquer hipótese de conhecimento, ainda que parcial e grosseiro, sem que se recorra a um juízo de valor.

Usando uma metáfora externa à área educacional, não se admite de forma alguma que se tomem decisões sobre a localização, forma e modo de construção de uma ponte sem que se tenham por adquiridos e fiáveis conceitos como distância ou massa, nem sem que se esteja na posse de instrumentos de agrimensura considerados fiáveis e úteis. Tão-pouco se tomará essa decisão sem que estabeleçam critérios de valor, que permitam escolher entre localizações alternativas em função do impacto e custo comparados de uma e outra localização. Não obstante este segundo momento de juízo de valor, há sempre que determinar *a priori* o que se entende por impactos e por custos, quais os que se consideram relevantes para o contexto em questão e como os caracterizar: medir, descrever, ‘avaliar’ no sentido descritivo. Numa fase subsequente ter-se-á de escolher um critério de selecção entre as diferentes hipóteses. Este critério não se confunde com o critério de caracterização de cada uma das hipóteses. Uma e outra localização custarão tanto e tanto a construir e suportarão tal e tal volume de tráfego, independentemente do critério de selecção de entre cada localização, o qual se baseia, por exemplo, no valor relativo que se dá entre o ganho marginal de construção numa determinada localização e o impacto da mesma.

Analogamente, como se pode ambicionar seleccionar candidatos a programas de ensino, regular práticas de ensino, monitorar a qualidade e a natureza das aprendizagens, potenciar a auto-regulação das aprendizagens, atribuir certificados com implicações na vida académica e profissional, reorganizar sistemas de ensino e atribuir responsabilidades aos seus intervenientes sem que se defina, com aceitável grau de clareza, rigor e segurança, *o que* é necessário que alguém seja capaz de fazer em determinadas circunstâncias e *como* se chegou a conclusões que permitam afirmar com justiça que consequências um indivíduo deve ou não sofrer em função da descrição que fazemos de características do mesmo? Os meios servindo os fins, dir-se-á que, num primeiro momento, impõe-se identificar as tarefas de tomada de decisão que se crê beneficiarem de informação. Depois, caracterizar esta informação em natureza, âmbito e qualidade. De seguida definir o objecto que servirá de fonte a esta informação e, por fim, aplicar aqueles instrumentos que demonstrem ser de maior valia para as tarefas em causa. Para que se avalia? Se é necessário avaliar, para o que se deve olhar e de que forma se pode recolher a melhor informação? Numa fase subsequente (de meta-avaliação?): os instrumentos usados e o *uso* que deles se fez podem ser considerados fiáveis, válidos, económicos e com impacto, é dizer, úteis?

A montante desta questão está o problema da definição conceptual dos termos e, porventura tão importante, da conotação que os mesmos têm em diferentes intervenientes no processo educativo. Sendo polissémicas, e por vezes contraditórias, as definições de avaliação que encontramos, não é razoável esperar que uma comunidade tão abrangente como é a comunidade escolar partilhe as mesmas denotações e conotações do termo. Tão-pouco, que uma simples definição dos termos, ainda que por processo normativo-administrativo seja suficiente para a) o seu cabal recorte, b) conhecimento partilhado e c) utilização sistemática ao longo de todo o sistema educativo.

Deste modo, uma escolha impõe-se e, no decurso deste trabalho, utilizarei a dicotomia ‘assessment’/’evaluation’, tal como é apresentada por Bachman. Há falta de melhores termos, utilizarei ‘aferição’ como sinónimo para ‘assessment’ (avaliação em sentido estrito) e ‘avaliação’ como sinónimo para ‘evaluation’ (avaliação em sentido lato). O esquema seguinte é uma representação do modelo dicotómico de avaliação que seguirei, nas suas diferentes componente e relações.

Neste modelo, o uso dado à informação que é resultante da aferição é externo a esta, o que implica que é necessário encontrar uma ligação entre o constructo aferido, a técnica de aferição utilizada, os resultados da aferição e o uso dado a estes.

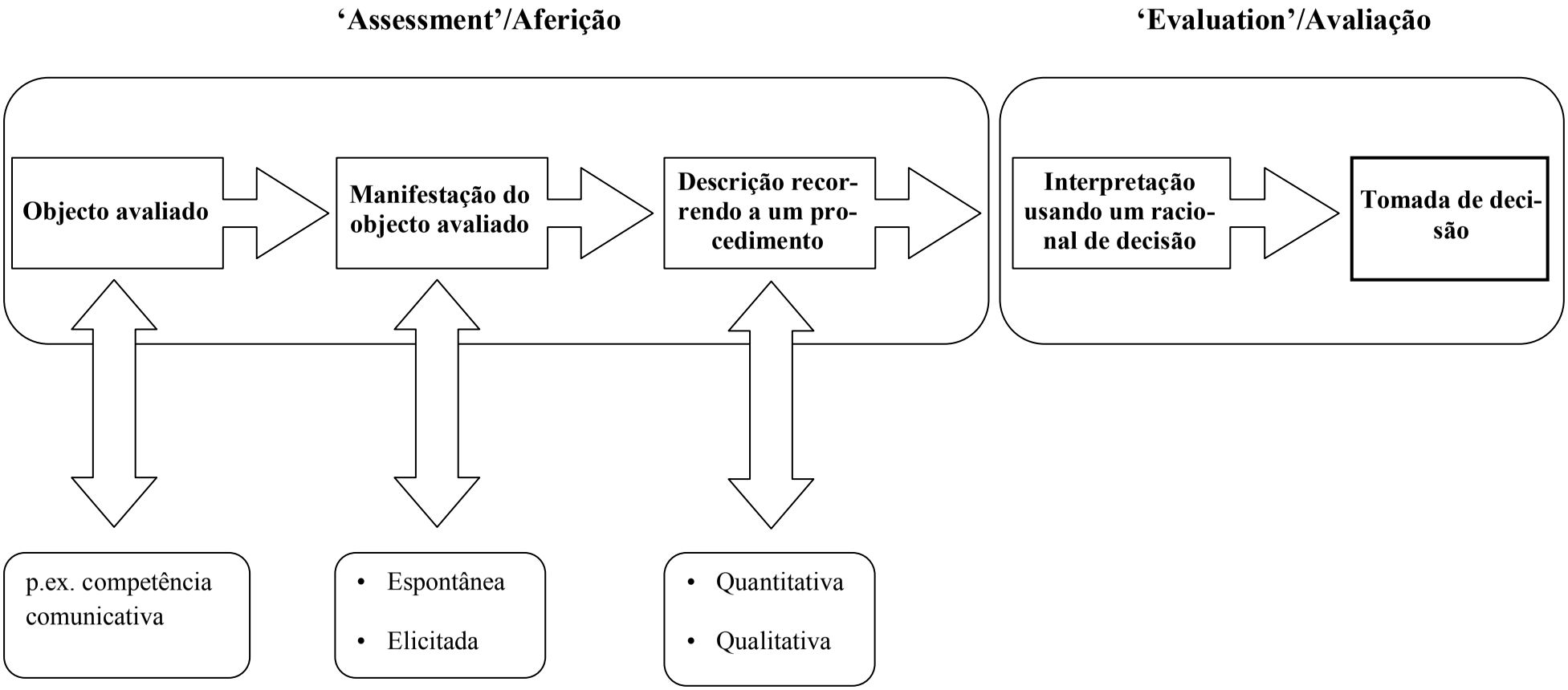

Fig. 2 – Esquema conceptual de ‘avaliação’

2.2 Usos da aferição

A informação que a aferição devolve, quando utilizada num acto avaliativo, servirá para orientar diferentes processos de decisão. Na avaliação educativa, três dos usos mais comuns são a certificação (de competências e/ou de aprendizagens), a selecção e a regulação das aprendizagens.

Uma hipótese operacional é que diferentes tipos de decisão requerem diferentes tipos de informação e, consequentemente, distintas técnicas de aferição devem ser empregues. Talvez esta linha de raciocínio seja uma explicação possível para a associação que comummente se faz entre determinadas técnicas de aferição e correspondentes usos, como é o caso do chamado 'teste', às decisões de certificação, e das 'fichas de trabalho' ou 'portefólios' à chamada 'avaliação formativa', mesmo que nem sempre se verifique uma cabal diferença entre o objecto que se pretende aferir e a técnica empregue num e noutro caso[10].

Outra abordagem possível, que acentua a atenção dada à tipologia e uso da informação, é que diferentes fontes de informação (entenda-se, diferentes instrumentos de aferição) podem ser utilizadas para alimentar um processo de decisão e que a mesma fonte de informação pode ser utilizada em distintos processos de decisão. Como veremos no capítulo 5, independentemente do declarado fim do instrumento de aferição utilizado, o que é necessário para que uma decisão seja justa e válida é demonstrar a adequação e apropriação das decisões tomadas com base nos critérios de decisão e na informação utilizados.

De qualquer forma, há que tipificar o acto avaliativo em si, para que depois se escolham a informação e a forma de a obter mais adequadas e económicas. Uma maneira possível

[10] Cf. Roldão (2006:43): "Julgo até que foi desta lufada de conhecimento teórico sobre avaliação […] que entrou nas escolas o hábito de usar as hoje famosas «fichas», inicialmente utilizadas, por oposição aos «testes» […] para pequenas avaliações ditas formativas ao longo das unidades de aprendizagem das diferentes áreas ou disciplinas. […] E aí estão as «fichas», numerosas, que se vão fazendo como exercícios de treino e de verificação mais segmentada (o que pode ter um pouco de utilidade), mas que, em última instância, não se usam como formativas, porque não são utilizadas para identificar, explicar e regular as dificuldades surgidas e remediá-las […]."

de caracterizar os usos avaliativos da aferição é distingui-los através de dois critérios básicos: a importância relativa e a perspectiva do utilizador.

Bachman (2004) comenta um modelo comum de importância relativa de decisões: decisões de baixo e elevado impacto. *Grosso modo*, distinguem-se pela influência, facilidade de rectificação e custos associados. As decisões de elevado impacto influem de forma significativa sobre aspectos importantes da vida de um indivíduo (atribuição de direitos e deveres políticos, aceitação num programa de estudo, certificação de aprendizagens com influência na vida profissional, p. ex.), dificilmente são corrigíveis e têm um elevado custo associado à uma tomada de decisão errada. As decisões de baixo impacto terão características opostas e normalmente estão associadas a contextos de formação (inclusão num determinado grupo de nível, programação das tarefas a executar no contexto de um programa educativo).

Elevado Impacto	**Baixo Impacto**
Grande influência sobre a vida	Pouca influência sobre a vida
Erros de decisão dificilmente corrigíveis	Erros de decisão facilmente corrigíveis
Elevado custo associado a decisões erradas	Baixo custo associado a decisões erradas

Tabela 1 – Importância relativa de decisões, Bachman (1990) (adaptado)

Decorre que decisões com elevado impacto reclamam da sociedade a garantia (ou talvez apenas a convicção) de que são tomadas com base em dados fiáveis, seguros e seguindo procedimentos de aferição e tomada de decisão claros, transparentes e robustos. Assim, as decisões de elevado impacto correlacionam-se fortemente com o uso de instrumentos de aferição que se têm ora por 'objectivos', ora por 'rigorosos', usualmente chamados 'testes' ou 'exames'.

Mais uma vez a abundância nocional impede descortinar com rigor o que se entende por teste. Bachman (1990:20), citando Carroll (1968), considera que um teste é um procedimento desenhado de tal forma que elicita num dado sujeito um comportamento desejado que

permite fazer inferências sobre suas certas características. Nesta acepção lata, um teste pode ter imensas formas: pergunta/resposta, textos com espaços elididos, entrevista estruturada acompanhada por critérios de classificação da linguagem produzida, estímulos à produção escrita, etc. Exame, por seu turno, será um conceito de ordem social: um teste utilizado como fonte de informação para decisões de elevado impacto, provavelmente com uma dimensão pública na sua utilização (p. ex., os Exames Nacionais do Ensino Secundário).

No entanto, decisões de elevado impacto podem ser tomadas com base noutro tipo de aferições, como é o caso do uso de portefólios, narrativas de vida ou recomendações[11]. Não obstante, no contexto educacional as decisões de elevado impacto tendem a ser tomadas com base em 'exames', que usualmente têm a forma de testes escritos[12], os quais incluem uma variedade de técnicas. Decorre do raciocínio económico que aquelas actividades que previsivelmente têm maior impacto sobre a sociedade reclamam maiores recursos e por esta razão muita da investigação feita em torno da avaliação educacional versa sobretudo sobre a testagem, particularmente enquanto fonte de informação para a tomada de decisões de elevado impacto. No entanto, as preocupações que se aplicam à testagem aplicam-se a qualquer instrumento de aferição, embora naturalmente o âmbito e minúcia com que são tratadas variem em função do valor relativo que se atribui a cada acto avaliativo e, consequentemente, aos instrumentos que o informam. Refiro-me às questões relacionadas com aquilo que Bachman e Palmer (1996) entendem por 'utilidade do teste': a fiabilidade, a validade, a autenticidade, a interactividade, o impacto e a exequibilidade.

Outra forma de caracterizar o uso da aferição é a perspectiva do utilizador. Um mesmo instrumento de aferição devolve informação que é utilizada por diferentes sujeitos como instrumento que visa servir fins distintos. Em tese, um instrumento de aferição é construído de

[11] É o caso do projecto *Novas Oportunidades*, que recorre a um referencial comum para validar aspectos da experiência profissional e pessoal de cada indivíduo como evidências do desenvolvimento de competências.
[12] O projecto *Provas de Expressão Oral* para o Ensino Secundário, o qual recorre a à classificação de entrevistas estruturadas como fonte de evidências para decisão de progressão/retenção.

tal forma que a informação que devolve seja modelada da forma mais produtiva possível para quem usar a informação disponibilizada. Usualmente, o utilizador da informação é um 'avaliador', no sentido que é alguém que tomará uma decisão, e este não é o sujeito que se submeteu à aferição. Tal é o caso, por exemplo, do uso dos testes em sala de aula, para propósitos ditos 'sumativos' ou até mesmo 'formativos': o utilizador do teste é usualmente o professor, que utilizará a classificação obtida no teste (uma forma de modelação da informação) para decidir, por exemplo, aprovar ou reprovar um aluno (acto avaliativo, com base num racional externo ao acto aferitivo: a classificação considerada mínima para propósitos de aprovação).

No entanto, nada obsta que o mesmo instrumento de aferição seja usado por utilizadores distintos que podem ou não ser avaliadores. Suponha-se o uso da classificação num dado teste como base para acção futura do professor (retomar o ensino de conteúdos programáticos não dominados pelos alunos, avançar para a unidade seguinte, p. ex.), do aluno (escolher quais os pontos do programa a que deve dedicar mais tempo e trabalho, escolher entre técnicas de aprendizagem que considere mais produtivas) ou de investigadores (saber, numa dada população, que aspectos do programa consistentemente demonstram ser mais problemáticos). O que acontece é que tradicionalmente o teste é, dir-se-ia holisticamente, visto como um acto avaliativo completo e que o avaliador é o professor. Logo, a informação deve ser modelada apenas tendo em conta a perspectiva do professor e um processo de tomada de decisão económico e inequívoco ($\geq 50\%$ = aprovação). Nada confirma (ou infirma, também) que essa modelação da informação seja igualmente útil e produtiva para outros utilizadores, e. g., um aluno em busca de *feedback* sobre o seu desempenho como forma de regular as suas aprendizagens.

De todo o modo, a perspectiva do utilizador influi necessariamente sobre as decisões de *design* do instrumento de aferição, mormente na forma como o resultado é apresentado. Um resultado quantitativo, por hipótese, é uma forma económica de modelar informação para

uma decisão do tipo aprovação/reprovação, por exemplo. Não obstante, nada impede o aferidor de modelar a informação conseguida a partir da prestação do avaliado de formas distintas, por hipótese, através de uma descrição verbal das competências que se demonstrou melhor serem aplicadas e das que podem beneficiar de intervenção futura. Nada impede a não ser, talvez, considerações de ordem económica: a convicção de que o custo marginal de modelar a mesma informação de formas distintas não é superado pelos ganhos marginais que daí possam advir (melhores e mais produtivas aprendizagens). E o simples facto de o aferidor/professor desconhecer ou dominar qualquer outra técnica que não uma classificação numérica.

Referi-me aos termos 'avaliação sumativa' e 'avaliação formativa', cujo alcance no âmbito deste trabalho urge recortar. O mencionado Despacho-Normativo 30/2001 estabelece a primazia desta sobre aquela (cf. art.º 6º e 13º). *Grosso modo*, entende-se por 'avaliação formativa' a regulação das aprendizagens e do ensino. Não se especificam técnicas ou procedimentos. O objecto da avaliação tão-pouco é definido, embora pareça incidir sobretudo sobre as aprendizagens dos alunos (cf. art.º 16º). Os utilizadores da informação são professores, alunos e encarregados de educação "de modo a permitir rever e melhorar os processos de trabalho" (Art.º 18º).

Já a 'avaliação sumativa' visa, com periodicidade pré-estabelecida, sintetizar "as informações recolhidas sobre o desenvolvimento das aprendizagens e competências definidas para cada área curricular" (Art.º 22º). Não se impõe a utilização de qualquer instrumento de aferição específico, ou se estabelecem padrões de desempenho que possam orientar as escolas no processo de avaliação, o que parece contrariar a ideia que avaliações de elevado impacto *têm de* socorrer-se de determinadas técnicas de aferição (e.g., teste escrito). Tão-pouco, que a informação tem de ser modelada exclusivamente em termos numéricos, ideia reforçada pelo art.º 30º que postula que "No primeiro período do 5º e 7º anos de escolaridade, a avaliação sumativa poderá […] não conduzir à atribuição de classificações ou menções, assumindo a

sua expressão apenas carácter descritivo". Questão que fica por responder é em que situações, com base em que critérios de decisão e com vista a que objectivos. A 'avaliação sumativa' (ou antes a aferição sumativa), neste contexto, é também sinónima de avaliação de elevado impacto, uma vez que os resultados obtidos são utilizados na tomada de decisão sobre a progressão e retenção.

Há que usar de alguma cautela na aceitação destas distinções. Por um lado, podem ser artificiais e apenas desiderativas. Nada garante que a 'avaliação sumativa' (entendida aqui como os instrumentos de aferição usados para os actos de avaliação sumativa) não possa ser também usada como fonte de informação para a 'avaliação formativa'. Não é incomum escolas e professores reajustarem os seus procedimentos em função de resultados dos alunos em momentos de 'avaliação sumativa'. Aliás, é até questionável se não será esta a principal fonte de informação para avaliações de procedimentos didácticos, escolha de manuais, abordagens dos programas, planificação do conteúdo e método a utilizar nas unidades didácticas que se seguem no currículo, por exemplo. Acresce ainda o facto de, como Roldão (2006) alerta, muitas vezes aquilo que se considera ser 'avaliação formativa' (recolha de informação que depois de processada é devolvida a professores e alunos por forma a que possam usá-la na optimização das aprendizagens) não passar de 'avaliação sumativa' aplicada com uma periodicidade mais imediata. Mais uma vez: nada obriga a que um mesmo instrumento não possa ser usado com fins distintos (aprovar/reprovar *vs* planear o trabalho futuro), mas importa demonstrar que *de facto* pode ser usado com ganho em fins distintos, ou antes, que as interpretações que se fazem são adequadas e apropriadas ao contexto em causa, enfim, que o uso do instrumento de avaliação é válido. Além disso, no limite, toda a avaliação regula alguma coisa (logo, é formativa), sejam os processos internos de aprendizagem do aluno, sejam as escolhas que escolas e professores fazem na planificação e execução do seu trabalho.

Por outro lado, dado o acima exposto, é questionável o porquê da primazia dada à 'avaliação formativa'. Por que razão se considera que a aferição de aprendizagens e/ou competências dentro do contexto imediato do processo de ensino é mais importante do que uma sua 'certificação'? Roldão (2006:41) defende que "*é na avaliação reguladora que se confirma a avaliação certificativa e vice-versa*" (itálico no original), porque de facto se ambas as dimensões não estiverem ligadas muito dificilmente se poderá demonstrar a sua utilidade. Outras evidências serão necessárias (por exemplo, a aplicação de competências desenvolvidas na escola em situações de 'vida real'), mas sem uma ligação entre aquilo que o aluno demonstra ser capaz de fazer no contexto da aprendizagem e em momentos de aferição externos a esta, reduzem-se ambas as modalidades de avaliação a uma espécie de tautologia desligada da realidade educativa: o teste que se usa para 'regular o aluno' e o teste que se usa para 'certificar o aluno' são úteis e válidos porque são os que se usam nessas situações.

Uma modalidade ou função da chamada avaliação formativa é a 'avaliação diagnóstica'. O Decreto-Lei 74/2004 de 26 de Março, que estabelece os princípios orientadores do currículo e da avaliação no Ensino Secundário, refere-se explicitamente a esta "função" da avaliação formativa no seu artigo 11º. Não obstante, o Despacho Normativo 1/2005 de 5 de Janeiro contrapõe a avaliação diagnóstica às modalidades sumativa e formativa. O capítulo seguinte dedicar-se-á a uma tentativa de delimitação do significado do termo e à sua aplicação no contexto do ensino de PLNM no Ensino Secundário.

3. A avaliação diagnóstica em PLNM no Ensino Secundário

3.1 Noções de avaliação diagnóstica

3.1.1 Na bibliografia nacional e internacional

A metáfora da diagnose, quando aplicada à avaliação educacional, não é de fácil recorte. Na taxonomia, a diagnose é o acto de distinguir espécies pelas suas características. Na Medicina, procurar a natureza e a causa de uma afecção. Assim, parece que no âmago do significado estão os conceitos 'identificação' e 'distinção'. No entanto, é questionável se esta não será uma especificidade de toda a aferição: não identificar espécies, mas identificar características no constructo que se pretende medir e, porventura, distingui-las de outras características, em função do seu grau de desenvolvimento; ou distinguir alunos entre si pelas características/grau de desenvolvimento de características que demonstram. Talvez fosse este o entendimento de Bachman (1990:60) quando defendeu:

> *"[V]irtually any language test has some potential for providing diagnostic information. A placement test can be regarded as a broad-band diagnostic test in that it distinguishes relatively weak students from strong students so that they can be provided learning activities at the appropriate level. […] A detailed analysis of student responses to the questions on placement and readiness tests can also provide more specific information about particular areas of weakness. When we speak of a diagnostic test, however, we are generally referring to a test that has been designed and developed specifically to provide detailed information about the specific content domains that are covered in a given program or that are part of a general theory of language proficiency. Thus, diagnostic tests may be either theory or syllabus-based."*

O Glossário ALTE (1998:392) define teste diagnóstico como "Usado para determinar os pontos fortes e fracos de um aprendente. Os resultados podem ser úteis na tomada de decisões relativamente à formação, aprendizagem ou ensino futuros". Temos então que um teste diagnóstico permite distinguir entre pontos fracos e fortes relativos e/ou absolutos, implica uma análise detalhada das respostas dos alunos, é orientado ou por conteúdos curriculares, ou por uma teoria da competência em língua, e visa orientar práticas educativas futuras.

Alderson (2005) pretende problematizar o conceito e na sua análise destas e de outras definições de testes diagnósticos conclui que, não só as definições do termo parecem ser contraditórias e pouco desenvolvidas, como não parece ter havido particular interesse em desenvolver (e portanto estudar) testes diagnósticos. Por outro lado, como podemos verificar pelas palavras de Bachman (1990), há uma forte sobreposição entre os conceitos de 'teste diagnóstico' e 'teste de colocação'[1], facto que requer mais investigação (cf. Bachman 2004:11 e McKay 2006:22-23). Possivelmente o conceito de custo de oportunidade justifica o facto: se testes diagnósticos fazem parte da avaliação de baixo impacto, porque são avaliação formativa, a sua importância não justificará o dispêndio de recursos necessários ao desenvolvimento de outros instrumentos de avaliação (cf. McKay 2006:294).

Alderson (2005) prossegue com uma análise do possível conteúdo e tipos de tarefas a utilizar num teste diagnóstico. Não chegando a uma conclusão clara, sugere que um teste diagnóstico poderá medir o mesmo que um teste de conhecimentos[2], mas apresentar a informação na forma de *feedback* a alunos, planificadores de cursos e professores, para que possam optimizar as aprendizagens. Neste sentido, não se pode pro-

[1] ALTE (1998:391) "**teste de colocação** sin: teste de nível. Teste aplicado com o objectivo de colocar os estudantes num grupo com o nível que corresponda ao seu grau de conhecimentos e capacidades."
[2] ALTE (1998:391) "**teste de conhecimentos** Teste destinado a medir os conhecimentos adquiridos por um candidato num curso, num manual, etc., ligado, por isso, a um curriculum."

priamente falar de um teste diagnóstico, mas de 'avaliação diagnóstica', uma vez que o que está em causa é um determinado *uso* de um instrumento de aferição (o teste) e não uma característica intrínseca do teste *per se*.

Por outro lado, Alderson (2005) sugere que existe uma correlação entre testes ditos diagnósticos e conteúdos 'discretos', é dizer, tarefas de aferição que se concentram em conteúdos detalhados do programa e/ou da competência linguística (uso de preposições, verbos auxiliares modais e de tempo, conjunções coordenativas e subordinativas, articulação de sons, p. ex.). Neste sentido, não só há avaliação diagnóstica, mas também há de facto uma forma de aferição diagnóstica: testes que, no lugar de se concentrarem em capacidades de alto nível (ler para identificar o argumento principal de uma tese, usar a estrutura e organização de um texto para influenciar o leitor, p. ex.), se concentram em elementos mais discretos que por sua vez formam uma competência mais lata. Este é um problema premente: saber qual o constructo da aferição diagnóstica e de que forma se pode distinguir (ou não) do constructo da aferição usada para fins sumativos. Em contraposição, Buck (2001:256) aventa a hipótese de as sub-competências nem sequer existirem – seriam apenas metáforas para a descrição do que implica a competência da compreensão do oral – e, portanto, não serem passíveis de aferição, o que traria implicações para a definição do constructo e, a jusante, para a validação do uso dos instrumentos de avaliação, *maxime* inferências sobre a natureza da competência aferida.

Esta análise retoma trabalhos anteriores do autor, nomeadamente Alderson (2000), em que defende que o uso de taxonomias de competências e micro-competências tem potencial diagnóstico. O problema está em saber se tais taxonomias são 'naturais', se realmente têm fundamentação empírica e se baseiam na demonstração de uma teoria da competência e do seu desenvolvimento, ou se não serão meramente desiderativas, limitando-se a reflectir uma *praxis* de organização do ensino

(Alderson 2000:11). Mas mais importante é associação que o autor faz entre micro-competências, aferição recorrendo a itens discretos e avaliação diagnóstica (Alderson 2000:148), ponto de vista em que é seguido por autores como Buck (2001:97, 256), Weigle (2002:114-121) Luoma (2004:76-78) e McKay (2006:194, 223).

Por outro lado, uma vez que o fim principal da avaliação diagnóstica é identificar causas distintas para falhas ou incapacidades semelhantes, propondo de seguida um caminho de "remediação"[3] adequado, impõe-se sempre alguma forma de teoria da proficiência e do seu desenvolvimento (Alderson 2000:20), para além de um mero percurso de ensino pré-estabelecido. Caso contrário, a avaliação diagnóstica seria uma espécie de profecia auto-realizada, um ciclo tautológico que se retro-alimentaria das hipóteses que os professores colocam e dos resultados dos instrumentos de avaliação que usam para confirmar (em lugar de testar) essas hipóteses.

Outra característica da avaliação diagnóstica que parece ser partilhada pelos autores é que a mesma leva sempre a algum tipo de diferenciação pedagógica, em menor ou maior grau. Buck (2001:97) refere-se à maior eficácia nos processos de ensino, Luoma (2004:76) a *feedback* detalhado que permita aos alunos perceber quais as áreas (da tarefa de comunicação e da competência linguística) que requerem mais atenção, Purpura (2004:156) considera que o diagnóstico leva sempre a alguma forma de *tutoring* (não se confundindo com as acções de seleccionar, colocar em grupos de nível ou controlar a progressão no desenvolvimento da proficiência), McKay (2006:145), embora defenda que também possa existir diagnóstico durante o processo de ensino, aproxima-se de Purpura (2004) ao defender que o essencial do diagnóstico é a planificação do ensino futuro, optimizando-o e adequando-o às necessidades e perfis dos alunos.

[3] Os termos 'remediation' e 'tutoring' são utilizados como sinónimos da acção que decorre do processo de diagnóstico. Por exemplo, Alderson (2000:11): "Such lists or taxonomies […] suggest the possibility of diagnosing a reader's problems, with a view to identifying remediation"; Purpura (2004:156): "[…] diagnosis of a student's grammatical strengths and weaknesses so tutoring can be recommended".

Na bibliografia de língua portuguesa, Ribeiro e Ribeiro (1989) dedicam toda uma secção do capítulo sobre avaliação à noção de avaliação diagnóstica. Das palavras dos autores é possível extrair duas conclusões: existe diagnóstico no sentido de avaliação (um uso específico de informação) e de aferição (são criados instrumentos distintos que visam constructos também eles distintos da avaliação formativa e sumativa).

Neste sentido, para Ribeiro e Ribeiro (1989:342) a avaliação diagnóstica, formativa e sumativa são tipos complementares e distintos, definindo aquela da seguinte forma:

> *"A avaliação diagnóstica tem como objectivo fundamental proceder a uma análise de conhecimentos e aptidões que o aluno deve possuir num dado momento para poder iniciar as novas aprendizagens."*

Esta definição parece aproximar-se das perspectivas já analisadas, no sentido em que se considera que deve ocorrer antes das aprendizagens programadas e que o ensino a encetar deve ser planeado e executado tendo em conta os resultados devolvidos pela aferição diagnóstica. McKay (2006:145) aproxima-se desta visão: a avaliação diagnóstica deve ter lugar sempre que seja útil planear o ensino (no início do ano lectivo, bem como em qualquer outro momento).

De igual forma, a aferição diagnóstica poderá também tentar saber do conhecimento que os alunos já têm das aprendizagens futuras, podendo desta forma optimizar-se o trabalho a desenvolver (Ribeiro e Ribeiro, 1989:344). Assim, o teste diagnóstico deve versar

> "*sobre um conteúdo restrito de objectivos em volta dos quais se organizam grupos de perguntas, muitas vezes várias perguntas sobre o mesmo objectivo. Tem, assim, uma estrutura de malha apertada, que pretende analisar pouco mas em profundidade*" (Ribeiro e Ribeiro, 1989:345).

No entanto, uma vez que Ribeiro e Ribeiro (1989) defendem que os testes diagnósticos não devem ser classificados, não é claro de que forma deve ser codificada (e posteriormente usada) a informação que potencialmente carregam. De notar que Weigle (2002:124) chama a atenção para a importância de ter em mente o utilizador previsto da informação que o teste diagnóstico produz. De igual forma, Luoma (2004:76-78), ao referir-se ao *feedback* detalhado que deve ser dado aos alunos, implica que alguma forma de codificação deve ser utilizada, no caso, o uso de listas analíticas, de classificação numérica e com elevado grau de discriminação (complementadas com descrições verbais do que significa cada classificação).

Da bibliografia analisada pode então inferir-se que a noção de teste/aferição/avaliação diagnóstica parece correlacionar-se com a ideia de aferição detalhada de conteúdos/sub-competências *antes do facto*, a informação conseguida servindo para a tomada de decisões que influenciam o conteúdo, ritmo e metodologia do ensino – o facto. Estas informações muito provavelmente são conseguidas recorrendo a instrumentos de aferição diferenciados, que ora recorrem a itens discretos, ora visam conteúdos/competências que precedem, e são condição necessária à aprendizagem de, conteúdos/competências futuros. Os utilizadores tanto poderão ser professores como alunos, embora pareça ser dada a primazia àqueles. Não é claro que formas, se algumas, de codificação da informação caracterizam a informação diagnóstica, embora alguns autores pareçam apontar na direcção de escalas analíticas e detalhadas.

3.1.2 Nos documentos ministeriais

Em linha com as definições apresentadas, analisando os documentos ministeriais, parece existir uma forte correlação entre aquilo que se considera teste/avaliação diagnóstica e os fins formativos da aferição discutidos na secção 1.2. De facto, uma análise das definições de avaliação diagnóstica, formativa e sumativa permite-nos reco-

nhecer que há uma forte sobreposição entre aqueles dois conceitos. O Anexo 1 apresenta uma comparação das definições propostas para os conceitos, tanto nas versões de 2001 e 2005 das orientações para o Ensino Básico, como no Decreto-Lei 74/2004, que se aplica ao Ensino Secundário.

Mais uma vez, não parece possível uma delimitação inequívoca dos conceitos. Enquanto no Ensino Secundário 'diagnóstico' é uma *função* da 'avaliação formativa', no Ensino Básico, desde 2005, 'avaliação formativa' e 'diagnóstica' são apresentadas como *modalidades* da avaliação, juntamente com a avaliação sumativa. Curiosamente, na versão de 2001 das orientações para a avaliação no Ensino Básico, a avaliação diagnóstica era uma "vertente" da avaliação formativa. Simplesmente, entre a versão de 2001 e de 2005 das orientações para o Ensino Básico, entendeu-se isolar a avaliação diagnóstica como mais uma categoria taxonómica, sem que surjam diferenças substanciais de linguagem na definição do que é a 'avaliação diagnóstica', ou se apresente qualquer justificação.

Todavia, a definição de avaliação diagnóstica apresentada parece aproximar-se muito da proposta de definição de *uso da aferição* feita por Bachman (1990): adoptar estratégias que visem reorientar as práticas pedagógicas, optimizando-as. Essa é a definição de avaliação formativa nos princípios orientadores para a avaliação no Ensino Secundário. Como se distinguem uma da outra não parece possível dizer, tanto mais que as orientações para o Ensino Básico pressupõem uma articulação da avaliação diagnóstica com a avaliação formativa, articulação essa que também não é caracterizada. Embora pareça claro que a avaliação diagnóstica é um processo de avaliação, no sentido em que se tem usado o termo 'avaliação' no decurso deste trabalho, não é possível dizer se tal processo de decisão se baseia em instrumentos de aferição diferenciados, como Alderson (2000 e 2005), e Ribeiro e Ribeiro (1989) sugerem. Sobretudo, e porventura

mais importante, não é claro que critérios devem orientar as tomadas de decisão que levarão a uma reorientação das práticas educativas.

Assumindo que a avaliação diagnóstica tem à sua disposição meios de aferição considerados úteis, à luz de que critérios de decisão se os usa? O Despacho Normativo 1/2005 refere que compete às escolas, no início de cada ano lectivo, estabelecer 'critérios de avaliação' (Art.º 15º). Não é claro se estes critérios de avaliação são especificações técnicas a utilizar na criação de instrumentos de aferição (definições do constructo a medir, técnicas e procedimentos a utilizar) ou critérios de decisão (por exemplo, níveis de desempenho considerados mínimos para a progressão de ano e sua definição operacional em instrumentos de aferição). Se forem critérios de decisão e se estes forem, por hipótese, padrões ou níveis de desempenho, também não é claro como é que a escola deve identificar e caracterizar esses padrões ligando-os com os resultados das aferições. Ou seja, se parece ser dada relativa latitude de decisão às escolas na elaboração e aplicação de instrumentos de aferição (incluindo hipotéticos instrumentos 'diagnósticos'), não resulta que estes sejam utilizados na avaliação em cotejo com os mesmos critérios de decisão, seguindo os mesmos princípios e buscando atingir os mesmos objectivos de desempenho, em todo o sistema de ensino.

É defensável que, como Alderson (2005:6) e McKay (2006:294) sugerem, haja uma decisão de 'custo comparado' a operar. É dizer: a menor importância relativa da avaliação formativa permite um relaxamento dos procedimentos de fiabilidade e validade usualmente aplicados na avaliação sumativa, mormente aquela que visa certificar competências, logo de elevado impacto. No entanto, duas objecções se colocam. Uma é que, se uma prática é suficientemente importante para merecer aparato normativo, até com a dignidade de Decreto-Lei, então também é suficientemente importante para ser informada por investigação científica e práticas pedagógicas validadas e úteis, a gestão

da coisa pública não se podendo fazer com base em crenças não demonstradas e empiricamente sustentadas. Outra, relembrando as palavras de Roldão (2006), que a avaliação formativa e sumativa estão indelevelmente ligadas, uma e outra servindo como argumentos de validação recíprocos, logo implicando que ambas merecem igual dignidade e, acima de tudo, reclamam igual atenção pelo impacto agregado que têm sobre as aprendizagens e todo o sistema de ensino.

Na secção seguinte proceder-se-á a uma análise da avaliação diagnóstica no ensino de PLNM na escola pública portuguesa e de como as questões ora levantadas têm implicações próprias nesta área de ensino/aprendizagem.

3.2 Avaliação diagnóstica e PLNM: Orientações e aparato legislativo

Recentemente, o PLNM foi reconhecido como área curricular no Ensino Secundário, com equivalência à disciplina de Português, através do Despacho Normativo 30/2007, de 10 de Agosto. Este facto vem na senda de trabalhos que remontam a 2003, nomeadamente Soares *et al*. (2005), Leiria *et al.* (2006) e Pascoal e Oliveira (2006). Já em fins de Abril de 2008 surgem as "*Orientações Programáticas de Português Língua Não materna (PLNM) – Ensino Secundário"*. Será com base nestes textos que procederei a uma análise dos usos e instrumentos de aferição do PLNM no Ensino Secundário português, especialmente a avaliação diagnóstica.

Antes de se identificarem os potenciais usos da aferição e modalidades de avaliação que os documentos prevêem, há que dar conta de quais são as intenções declaradas da criação de uma área curricular como o PLNM. Soares *et al.* (2005) no seu 'Documento orientador' procedem a um "Diagnóstico[4] da situação actual" (p. 7) a partir do qual extraem princípios e objectivos que visam "Oferecer condições equitativas para assegurar a integração efectiva dos alunos, cultural, social e académica, independente-

[4] Termo que nesta utilização particular parece querer significar 'descrição' ou ' caracterização'.

mente da sua língua, cultura, condição social, origem e idade" (p. 10). Deste objectivo geral são extraídos corolários, seguindo princípios de integração, igualdade, interculturalidade e qualidade, que orientarão as medidas a tomar.

De notar que, embora o documento seja apresentado como uma análise e orientação para o PLNM no currículo nacional, o principal objectivo é a integração, aparecendo o domínio da língua como um meio (porventura o privilegiado ou decisivo?) para este fim. Este facto não é de somenos importância, pois como veremos a validade do uso de um instrumento de aferição (e concomitantemente dos referenciais em que se baseia) implica uma ligação entre os resultados que devolve e as inferências, usos e consequências que neles têm base. Deste modo, um instrumento de aferição e, em mais larga medida, todo um processo de avaliação e de organização curricular necessitam de validar as suas práticas em função dos objectivos declarados e das consequências efectivamente observadas. No caso presente, a noção de 'integração' (a qual nunca é definida explicitamente, seja através de conteúdos ou de níveis de desempenho) é essencial a qualquer validação que se queira fazer da prática educativa em torno do PLNM, 'avaliação' (em sentido estrito ou lato) incluída.

Sendo que as necessidades dos alunos que não têm o Português como língua materna são de natureza linguística, curricular e de integração, impõem-se medidas de diferenciação, que extravasam o domínio do conhecimento da língua. Estas setes medidas são divididas nas categorias 'acolhimento' e 'escolarização', sendo que a criação e aplicação de um Teste Diagnóstico se inclui na primeira categoria. Por outro lado, prevê-se a elaboração de orientações nacionais e da caracterização de grupos de nível de proficiência (necessidade a que o documento de Leiria *et al*. (2008) vem responder, no caso do Ensino Secundário), medidas de escolarização que se articulam com a necessária avaliação das aprendizagens. A ideia subjacente é que características distintas de um

determinado aluno (ou conjunto de alunos) implicam medidas de acomodação que lhe permitam atingir padrões de desempenho comparáveis aos dos seus pares (neste caso alunos que têm o Português como língua materna). O Despacho Normativo 30/2007 vem confirmar esta hipótese, ao estabelecer 3 Grupos de Nível de Proficiência (GNP), os quais implicam medidas de diferenciação com alcance e âmbito distinto. Assim, os alunos inseridos nos grupos de Iniciação (equivalente a A1/A2, no esquema proposto pelo QECR) e Intermédio (B1) vêem as aulas de Língua Portuguesa substituídas por Português Língua Não Materna (PLNM), ao passo que os alunos integrados no grupo Avançado (B2/C1) devem seguir o currículo da disciplina de Língua Portuguesa, embora beneficiando de uma unidade lectiva semanal extra "para o desenvolvimento de competências de leitura literária e conhecimento do cânone literário" (Leiria *et al.,* 2008:4).

A integração dos alunos num destes três grupos de nível, ao serem pela primeira vez integrados no sistema de ensino português, faz-se em função da sua prestação num teste diagnóstico a realizar na escola. Os testes não são concebidos pelos professores, mas são "aplicados e avaliados (*sic*) por estes" (Soares *et al*, 2005). A principal consequência que parecem ter é a integração do aluno num determinado grupo de nível. Neste sentido, a noção de teste diagnóstico parece estar mais próxima da de teste de colocação em nível, uma vez que não há qualquer referência ao recurso a outras características dos alunos na criação de grupos (como por exemplo grupos baseados nas línguas maternas dos alunos, ou áreas de estudo do ensino secundário). Pelo contrário, a Medida 2 de Escolarização prevê que os alunos sejam agrupados em níveis de proficiência, indiciando que esta será a melhor forma de potenciar as aprendizagens dos alunos e a sua integração.

No entanto, a Medida 2 prevê também que "a planificação do trabalho para cada GNP deverá ser feita tendo em contas as características individuais dos alunos e do gru-

po que os integram, bem como as orientações para o Português Língua não Materna" (Soares *et al.*, 2005:16). Tal orientação parece pender já mais no sentido de diagnóstico que Alderson (2005) perfilha. De todo o modo, resta saber se 1) é possível utilizar de forma simultaneamente económica e válida um mesmo instrumento de aferição para dois fins distintos, ainda que complementares, e se 2) tal procedimento não é simplesmente o que já acontece em todas as disciplinas do Ensino Secundário, haja ou não instrumentos de avaliação ditos diagnósticos formalmente introduzidos no plano de trabalho.

O texto de Pascoal e Oliveira (2006) visa completar e aprofundar as considerações sobre o teste diagnóstico de Soares et al. (2005), embora não resulte claro se se destina aos utilizadores finais do teste diagnóstico (os professores) ou à tutela e futuros elaboradores de testes diagnósticos.

De todo o modo, não é apresentado um modelo de teste diagnóstico ou sequer se define (conceptual e operacionalmente) o constructo a medir, se bem que se façam considerações sobre as técnicas a utilizar preferencialmente (itens de completação e de correspondência, p. ex.), as habilidades a serem alvo da aferição (expressão escrita, compreensão da leitura, funcionamento da língua, compreensão do oral e expressão oral) e se estabeleçam a escola e os amigos como os domínios de uso da língua a privilegiar, não se excluindo a "sociedade em geral". Sublinha-se a importância da adequação da forma e conteúdo à faixa etária e perfil cultural dos alunos na concepção dos testes.

Admite-se que, caso os alunos já estejam integrados numa turma, o diagnóstico se faça com base em quaisquer elementos que o professor de PLNM considere adequados, até mesmo um teste diagnóstico estruturado. Caso o aluno ainda não esteja integrado, deverá ser sujeito a avaliação diagnóstica com base num teste e entrevista estrutura-

dos, como o Documento Orientador prevê. Mais uma vez, não é apresentada qualquer tipo de estrutura definitiva, tão-pouco se indica a quem caberá tal ónus.

Em paralelo com Alderson (2000 e 2005) os autores parecem perfilhar a ideia que o teste diagnóstico versa analisar várias competências em detalhe (cf. p. 33). Contudo, propõem que o mesmo teste sirva também como verificação "da competência alcançada após o programa de apoio em língua portuguesa e a frequência do sistema educativo. Esta dupla aplicação permitirá medir o progresso dos alunos e comparar o tipo de textos escritos e orais dos dois momentos de medição da competência" (p. 34).

Esta questão, como veremos, é premente, uma vez que a validação dos usos dos instrumentos de aferição diagnóstica (a entrevista e o teste) implicará uma análise das consequências: inclusão num determinado GNP, sujeição a um dado programa de ensino e concomitante variação na proficiência. Ora só será possível dar conta de tal variação se se estiver na posse de instrumentos de medição da proficiência que se refiram a uma escala comum (baseada no QECR, p. ex.) e cujas fiabilidade e validade sejam demonstradas. Essas são condições necessárias à comparabilidade dos resultados obtidos em cada um dos instrumentos. Os autores abordam este obstáculo (de manuseio tecnicamente difícil) recorrendo ao mesmo instrumento de aferição, facilitando assim quaisquer comparações entre os resultados (numéricos, descritivos) que devolva, quer num momento pré como pós instrução. Não é no entanto claro se é possível garantir a validade do constructo para ambos os usos do teste, mormente na cobertura e relevância do conteúdo (*vide* cap. 5 deste texto).

De qualquer modo, não sendo proposto um formato definitivo de teste ou sequer uma definição do constructo, é possível que futuras formas do teste possam servir ambos os fins, comprometendo-se diferentes dimensões da sua utilidade (no sentido que Bachman e Palmer, 1996 dão ao termo) em cada uso. Ainda assim, evidências de vali-

dade terão de ser investigadas em ambos os usos e, como veremos, muitas vezes a correlação de resultados entre distintos instrumentos de aferição que visam o mesmo constructo é usada como argumento de validade (bem como de fiabilidade).

De notar que Pascoal e Oliveira (2006:35-37) incluem no conceito de diagnóstico mais do que a simples caracterização de aspectos do desenvolvimento de uma competência comunicativa ou simplesmente linguística. O conceito é alargado à caracterização de outras competências dos alunos e de aspectos do seu percurso pessoal e escolar. Neste sentido, diagnóstico afasta-se da concepção mais psicométrica de Alderson (2005), mantendo no entanto o essencial da metáfora: saber com suficiente grau de detalhe o que caracteriza e distingue cada indivíduo, como forma de melhor adequar as práticas educativas futuras à sua necessidade. Contudo, não é apresentado um modelo de utilização de toda esta informação, ficando ao critério dos professores que dimensões privilegiar e de que forma. Aspecto a que os autores dão alguma relevância é o uso do instrumento de aferição e a necessidade de garantir a sua utilidade aos seus utilizadores preferenciais: os professores. O treino (cf. pp. 37-38) parece ser, então, uma necessidade inerente a todo o processo. Resta saber se ficará a cargo destes decidir quais as linhas de acção a seguir dadas as características de cada aluno, ou se, pelo contrário, a acção de diagnóstico será completada com orientações claras dos corolários a retirar de cada realidade diagnosticada.

De facto, não é razoável crer-se que, mesmo que haja um conjunto único de instrumentos de diagnóstico disponibilizados aos professores e que estes sejam treinados no seu uso de tal forma que se garanta um elevado grau de fiabilidade, o *uso* dos resultados do diagnóstico seja o mesmo sem critérios de decisão inequívocos do que implica este ou aquele resultado no diagnóstico. É isto que está no âmago do conceito de validade: não apenas o que está no teste, mas o que decorre dos resultados desse teste, neste

caso, que escolhas na planificação e execução do trabalho que cada professor faz em função dos resultados do diagnóstico.

Por fim, o texto de Leiria *et al.* (2008) retoma a questão da avaliação no ensino de PLNM, dedicando particular atenção à avaliação diagnóstica. Distingue-a da avaliação para transição de nível e ao registo do perfil sociolinguístico do aluno, indiciando que a progressão de nível se fará com base em instrumentos de aferição distintos dos utilizados no acto de diagnóstico. Embora nunca citando Pascoal e Oliveira (2006), Leiria *et al.* (2008) aproximam-se da sua abordagem no que toca aos *usos* da avaliação de diagnóstico: estimar o nível de proficiência do aluno, identificar os pontos fortes e fracos do aluno, identificar a língua materna do aluno e identificar e caracterizar outros aspectos do seu percurso escolar e pessoal que possam ter influência nas aprendizagens. Decorre também da abordagem de Leiria *et al.* (2008) que a inclusão num GNP adequado é o fim principal (e garante) ao desenvolvimento linguístico.

Não apresentando também uma estrutura de teste(s) a utilizar, Leiria *et al.* (2008) seguem de perto a abordagem de Soares et al. (2005) e Pascoal e Oliveira (2006): entrevista, a qual servirá para "identificar o tipo de teste diagnóstico adequado ao aluno" (p. 30), e teste diagnóstico, este dividido em teste oral (competência de compreensão do oral) e teste escrito (competências de compreensão e produção oral e escrita). Não são indicados quaisquer padrões, sejam de conteúdo ou de desempenho, que possam permitir decidir pela inclusão do aluno em cada grupo de nível. Assim como Pascoal e Oliveira (2006), Leiria *et al.* (2008) recomendam tipologias de 'exercícios' que podem ser usados nos testes, nunca correlacionando cada tipologia com competências ou sub-competências a avaliar. Sugere-se que sejam criadas "várias versões com graus de complexidade crescente" (p. 31) para cada competência a avaliar, mas não é

claro se esta complexidade é sinónimo de diferente níveis de referência ou diferentes padrões de desempenho dentro de cada nível de referência.

Não são referidos domínios de uso de língua ou tipologias de texto mas, uma vez que o documento de Leiria *et al.* (2008) pretende orientar os professores na organização do currículo em cada GNP, pode inferir-se que os conteúdos apresentados para cada GNP, assim como as correspondentes tipologias de texto, deverão servir de guia para os conteúdos dos testes.

De referir, também, que se considera que um aluno deve ser integrado num dado GNP quando demonstre "não possuir as competências linguísticas nos domínios de compreensão, expressão e interacções orais e da compreensão e expressão escritas que caracterizam" o nível subsequente (p. 31), o que implica que se pressupõe um *continuum* de desenvolvimento da proficiência implícito na organização dos níveis QECR. Admite-se o desenvolvimento assimétrico das diferentes capacidades (compreensão, expressão e interacção oral, compreensão e expressão escrita) e recomenda-se que no trabalho nos níveis de Iniciação seja dada primazia ao desenvolvimento das capacidades ligadas à oralidade. No entanto, não é claro com base em que critérios de decisão se deve optar pela mudança do aluno para o GNP subsequente, haja ou não um desenvolvimento assimétrico das competências: o aluno já é B1 em compreensão e expressão oral, mas ainda A2 em expressão escrita e domínio do vocabulário e gramática, por hipótese.

Este é um aspecto que não pode ser descurado, dado o impacto que poderá ter na vida académica e/ou profissional dos alunos. Ao contrário de Pascoal e Oliveira (2006) que sugerem que o mesmo instrumento de aferição seja usado nos momentos de diagnóstico e de verificação das aprendizagens, Leiria *et al.* (2008) admitem que os professores criem instrumentos distintos. Resta saber que garante se terá então de comparabi-

lidade entre os resultados. Caso seja deixado exclusivamente ao critério do professor a escolha do conteúdo e a estimativa do nível de dificuldade dos instrumentos de avaliação que usar para a decisão de progressão entre GNPs, corre-se o risco de alunos que tenham o mesmo nível de desenvolvimento de proficiência em PLNM sejam sujeitos a decisões distintas dada a natureza dos instrumentos de aferição a que são sujeitos. Esta consequência é tanto mais preocupante quanto as orientações ministeriais para o Ensino Secundário prevêem que os alunos integrados no nível Avançado seguem normalmente o programa de Português (embora beneficiando de uma unidade lectiva extra por semana) e são sujeitos aos Exame Nacional de Língua Portuguesa, tal como os alunos que têm o Português como língua materna. Ora, independentemente de se considerar que um aluno com grau de proficiência comparável ao nível B2 tem tanta capacidade linguística/comunicativa para se sujeitar ao Exame Nacional como um aluno de nível C1 ou nativo, se não se conseguir demonstrar que todos os alunos que efectivamente estão nesse patamar da competência comunicativa estão incluídos no GNP correspondente, criam-se injustiças entre os alunos que têm PLNM (os alunos que não deveriam ainda ser integrados neste nível mas são) e entre estes e os alunos nativos do Português (os que deveriam ser integrados no nível Avançado e se mantêm num nível mais 'fácil', beneficiando de um regime de certificação de competências mais favorável). Apenas será possível garantir a justiça das decisões tomadas, assim como validar o pressuposto de que um aluno de nível B2 tem um desenvolvimento da sua competência comunicativa que lhe permite em condições de igualdade resolver o Exame Nacional, se se estiver na posse de instrumentos de aferição fiáveis e comparáveis. Não é claro como é possível atingir esse objectivo deixando ao critério de cada professor em cada escola decidir qual a forma de aferir o nível do aluno.

Em todo o caso, a esta acresce outra questão premente, que é de saber da inclusão ou exclusão dos alunos no programa de apoio em PLNM. Embora Leiria *et al.* (2006:26) chamem a atenção para o facto de este ser o objectivo principal da avaliação dita diagnóstica, nem este texto, nem Pascoal e Oliveira (2006), nem Leiria *et al.* (2008) estabelecem critérios claros de decisão para a inclusão/exclusão no programa de PLNM ou para a progressão entre GNPs. Neste último caso, impõe-se a rápida criação de padrões de desempenho que permitam com clareza afirmar em que GNP o aluno deve ser integrado. Mas mais importante, dada a complexidade da realidade da imigração em Portugal e do consequente panorama linguístico (cf. Leiria *et al.* 2006, especialmente as páginas 13 a 24), a ausência de critérios claros de decisão de inclusão/exclusão poderá levar a injustiças de difícil resolução.

3.3 Discussão

A questão da relação entre testes diagnósticos e o ensino em PLNM redunda na concepção do que é aprender uma língua que não a materna (mormente num contexto de imersão e desejável 'integração') e do que se pode fazer para potenciar essa aprendizagem.

Em primeiro lugar, a concepção de aferição que tenho seguido ao longo este texto é devedora, sobretudo, da testagem, especialmente aquela desenvolvida com vista a ser usada em contextos de elevado impacto (certificação de competências, selecção de candidatos a integrar em planos de estudos/emprego). A 'revolução comunicativa' sublinhou a importância da língua como evento social, uma 'abordagem orientada para acção', para usar a terminologia do QECR, vindo a testagem dar conta desta preocupação: testes (usado aqui o termo no seu sentido mais lato) que visam antes de mais possibilitar fazer inferências sobre o uso da língua que um utilizador será capaz de fazer em situações de vida real. Testes concebidos em paralelo com a noção de que o ensino deve

reflectir, potenciar e privilegiar o uso da língua, em lugar da análise e treino de aspectos da componente linguística subjacente à comunicação verbal.

A noção de validade, aplicada à avaliação da proficiência em língua não materna, visa então, em boa medida, dar conta desta 'capacidade' do teste: fazer previsões informadas sobre o uso futuro da língua em contextos externos ao ensino/avaliação. Tal abordagem 'holística', que privilegia o que se consegue fazer em determinados domínios de comunicação, levou, talvez, a uma desvalorização do conhecimento da língua (a competência linguística de Bachman, por exemplo) como objecto das aferições, em benefício de um constructo mais lato que abarca esta e outras competências (embora não necessariamente dando conta de como interagem entre si): a competência comunicativa. O advento do QECR é o corolário desta abordagem. Não sendo uma teoria da comunicação (em língua materna ou não materna) ou da aprendizagem da língua não materna, o QECR antes almeja que diferentes abordagens teóricas possam interagir através do mecanismo dos descritores de proficiência comunicativa. O comportamento observável funciona então como o máximo denominador comum entre distintas formas de perceber e manipular o fenómeno da aprendizagem de uma língua estrangeira e de aquilatar a utilidade dos instrumentos usados.

Sendo certo que não se pode atribuir aos autores do QECR a responsabilidade por se ver nele capacidades que não tem, o facto é que o QECR não isenta os seus utilizadores de fazerem o 'trabalho de casa', de afirmarem com clareza e de forma sindicável pelos seus pares que corolários extraem, por hipótese, dos descritores apresentados para uma dada capacidade e de quais as implicações que terão num determinado contexto de ensino/aprendizagem e domínio de uso da língua. Tão-pouco isenta os utilizadores do QECR, *maxime* organizadores de cursos e avaliadores, de declararem qual o modelo teórico de desenvolvimento de competências que seguem (se algum) e que evidências

têm que permitem confirmar a adequação do mesmo ao uso em causa e infirmar a utilização de modelos distintos, pelo menos com base num argumento de produtividade.

Em segundo lugar, a ausência de uma teoria sólida do que é a aprendizagem de uma língua não materna parece desencorajar à partida a criação de instrumentos de diagnose aparentados àqueles que se conhecem para as afasias em língua materna, por exemplo. Informados por uma teoria do desenvolvimento da fala e do domínio ortográfico, estes visam elencar aqueles componentes da competência cujo desenvolvimento atípico impede uma proficiência plena ou 'normal'. Levando-se ao limite a ideia, comum na abordagem comunicativa, que uma proficiência não 'normal' não é sinal de incapacidade, mas antes de um estágio intermédio de desenvolvimento abaixo do objectivo estipulado, a necessidade de criação de instrumentos puramente diagnósticos tornar-se-ia quase espúria: independentemente da causa ou sintoma, com maior ou menor dispêndio de tempo e outros recursos o nível desejado será sempre alcançado.

Se adicionarmos a este facto a constatação da heterogeneidade que, no caso presente, a população escolar denota, então duas das principais características do teste diagnóstico tornam-se quase impraticáveis: a aferição "um para um" e a concepção de planos de estudo-ensino diferenciados. A solução passará, o mais das vezes, por um compromisso: estabelecem-se grandes grupos de aprendentes que (pelo menos teoricamente) denotam as mesmas 'dificuldades', optando-se em seguida por planos de estudo que, *grosso modo*, se adaptem a estes perfis. Se substituirmos 'dificuldade' por 'pertença a um nível de proficiência' (este entendido como a incapacidade de executar determinadas tarefas de comunicação, ou o sub-desenvolvimento de uma qualquer competência, como Leiria *et al.,* 2008:31), juntamente com a concepção, comum em alguns embora infundada, que os níveis QECR representam um *continuum* 'natural' de desenvolvimento de proficiência em LNM, está encontrada a explicação para a sobreposição

dos conceitos de teste 'diagnóstico' e de 'colocação em nível'[5]. Pode ser que tal abordagem se mostre produtiva, no mínimo económica, no entanto o uso do rótulo 'diagnóstico' promete bem mais do que aquilo que, no contexto presente, efectivamente pode oferecer.

Em terceiro lugar, o ensino de uma L2 não pode ser, no contexto de um sistema de ensino, comparável ao de uma qualquer outra disciplina. Ainda que se admita a exequibilidade de se estabelecerem programas únicos que se apliquem a toda a população escolar por igual e com igual grau de sucesso numa disciplina como, por exemplo, História, tal nunca seria possível no caso presente, dada a heterogeneidade da população escolar em causa. Alunos com diferentes línguas maternas e tradições educativas requererão distintos períodos de tempo e tarefas de ensino de forma a alcançarem o mesmo nível de proficiência. É aqui que documentos como o QECR, complementado com orientações programáticas exaustivas e, sobretudo, exemplos claros e contextualizados de produções linguísticas que ilustrem o âmbito e alcance dos descritores no contexto de aplicação em causa, são de particular utilidade. Especialmente, a se estabelecerem padrões de desempenho que se possam usar como base para decisões de médio/elevado impacto (como são a progressão para um grupo de nível e/ou a certificação de competências), exemplos de produções linguísticas são essenciais de forma a proporcionar uma utilização fiável e justa dos instrumentos de aferição. Tanto mais quando há uma grande instabilidade, natural, na concepção e aplicação *ex-nulo* de orientações programáticas e instrumentos de aferição, dificuldade essa que é função da disparidade de perfis e competências dos recursos humanos que são chamados a executar, simultaneamen-

[5] Alderson, Clapham e Wall (1995:12) oferecem uma análise semelhante: "These more specific diagnostic tests are not easy to design since it is difficult to diagnose precisely strengths and weaknesses in the complexities of language ability. For this reason there are very few purely diagnostic tests. However, achievement and proficiency tests are themselves frequently used, albeit unsystematically, for diagnostic purposes."

te, as funções de elaboração, aplicação e avaliação de programas, materiais de ensino e instrumentos de avaliação.

Aquilo em que o QECR não é, ou apenas marginalmente será, útil é na concepção, aplicação e monitorização da planificação individual de cada professor, a qual pretende levar, de forma diferenciada ou não, todos os alunos ao mesmo patamar de proficiência. Seria aqui que a avaliação diagnóstica, e não a aferição na simples forma de testagem ou entrevista estruturada com fins de colocação em grupos de nível, seria da maior utilidade: orientar professores num processo de tomadas de decisão que culminará, algures no futuro, com uma certificação das competências dos alunos em momentos de avaliação sumativa externa. Tal processo implicaria, sucintamente, as seguintes fases: diagnóstico, selecção e organização dos conteúdos e métodos de ensino a usar com os alunos (com componentes de grupo e individuais), aplicação, monitorização (incluída nesta alterações à planificação do trabalho que se considerem necessárias) e certificação. Resta saber se, dada a realidade dos recursos humanos e materiais disponíveis, tal abordagem é exequível e sequer geradora de aprendizagens mais produtivas e, em última análise, de 'integração'. Tanto mais quando, dados os documentos publicados, não é ainda claro se caberá ao não aos professores a concepção dos instrumentos de aferição diagnóstica, formativa e sumativa a utilizar, ou se, pelo contrário, se limitarão a usar os instrumentos que lhes forem disponibilizados[6].

De qualquer forma, transcenderá sempre a capacidade de qualquer professor, ou grupo de professores (mesmo que numa escola/agrupamento de escolas) a validação do uso dos instrumentos de aferição, dada a natureza e quantidade da informação a recolher. A esse particular dedicar-se-á o Capítulo 5. No entanto, antes que se pretenda vali-

[6] Numa fase adiantada da redacção deste trabalho (Julho de 2008), foi publicada no sítio da Direcção-Geral de Inovação e Desenvolvimento Curricular uma informação dando conta da elaboração de testes por uma equipa de especialistas, os quais serão disponibilizados às escolas após uma fase de experimentação e subsequente aperfeiçoamento. Não são, todavia, disponibilizadas quaisquer informações sobre o formato e conteúdo dos testes.

dar o uso de instrumentos de avaliação é preciso construí-los e para isso é necessário, antes de mais, estabelecer qual o constructo a aferir e tipificá-lo. O capítulo seguinte dedica-se precisamente a esta questão.

4. As especificações dos testes e o constructo a medir

4.1 Especificações e estrutura dos testes

Um instrumento de aferição visa recolher de forma sistemática e fundamentada informação que possa ser usada na tomada de decisões. No caso em estudo, de âmbito educacional, são três as decisões mais importantes, do ponto de vista do impacto sobre o aluno: inclusão num determinado GNP, progressão para o GNP seguinte, exclusão do programa de PLNM. Correlacionadas com estas estão duas outras utilizações, ancilares: a monitorização da progressão da aprendizagem dos alunos, o apoio à escolha de metas de aprendizagem e planificação do ensino. Em quase todos os casos o utilizador da informação é o professor (incluir/excluir o aluno no programa de PLNM/GNP, decidir sobre a progressão do aluno e planificação do ensino), ao passo que o aluno necessita de informação para monitorizar a sua progressão, com fins de motivação, organização das suas aprendizagens, readequação de estratégias de aprendizagem/de comunicação, etc.

Desta forma, é necessário tipificar o tipo de informação necessária para em seguida se tomar a respectiva decisão. Os documentos de Soares et al. (2005) e Leiria *et al.* (2008) estabelecem que a inclusão num GNP se faz em função do grau de proficiência estimado do aluno, em termos comparáveis com o QECR (Cf. secção 3.2 deste texto). Adicionalmente, Leiria *et al.* (2008:31) estabelecem que os alunos se consideram posicionados num nível quando "não possuem as competências linguísticas nos domínios da compreensão, expressão e interacção orais e da compreensão, expressão e interacção orais que caracterizam" o nível subsequente. Ora tal definição é útil na medida em que permite estabelecer critérios de decisão muito claros: proficiência abaixo do patamar mínimo de proficiência do nível X implica colocação no nível X-1. Mais ainda, uma vez que o Despacho normativo 30/2007 prevê que a progressão pode ser feita em qualquer momento do ano lectivo, desde que se demonstre que o aluno já ultrapassou no

patamar mínimo do novo nível, aparentemente qualquer imprecisão na colocação pode ser economicamente corrigida.

De todo o modo, é necessário estabelecer instrumentos que sejam úteis a estas tomadas de decisão. Segue-se aqui a definição de Bachman e Palmer (1996) para utilidade: fiabilidade, validade do constructo, autenticidade, interactividade, impacto e exequibilidade.

As noções de fiabilidade e validade serão analisadas no capítulo seguinte. A autenticidade liga-se com a relação que se estabelece entre as tarefas que são apresentadas num momento de aferição e as tarefas de uso da língua sobre cuja capacidade de desempenho se pretende fazer inferências. Este conceito de autenticidade aproxima-se muito da dimensão substantiva da validade do constructo, como apresentada por Messick (1989). No entanto, merece também ser considerada individualmente se tivermos em conta que a ligação entre os processos cognitivos necessários à resolução de um item de teste e à tarefa de uso de língua com que aquele se relaciona pode ter graus ou justificações distintos. Por exemplo, no caso de um teste de colocação em GNP a autenticidade pode limitar-se a uma amostragem de aspectos de baixa-ordem que compõem a capacidade de expressão escrita, mas que têm elevado grau de previsão do desempenho futuro do aluno ao longo do programa de estudos. No caso da aferição de desempenho[1] a autenticidade da tarefa (entendida como semelhança na forma e nos processos cognitivos activados na sua resolução com os de uma tarefa de uso da língua) tem uma importância maior na criação de um argumento de validação: exemplos de produções orais ou escritas são tidas como exemplos mais representativos da real capacidade do aluno de realizar determinadas tarefas, se não mesmo do seu grau de desenvolvimento da competência.

[1] Cf. ALTE members (1998:392): teste de *performance*

A noção de autenticidade prende-se com a qualidade de interactividade: a medida em que as características individuais do aluno são de facto agenciadas no momento de resolução de um item de teste. Contudo, a interactividade é uma função única de cada tarefa[2]: diferentes tarefas (sejam de aferição sejam de uso 'real' da língua) reclamam distintos graus de interactividade e, por isso, deve ser tida em conta no momento da selecção de itens de teste/tarefas. Qualquer tarefa varia em autenticidade e interactividade em relação a outras tarefas com que partilhe dimensões do constructo a medir. Desta forma, é necessário ter em conta no momento da selecção de itens de aferição que dimensões se pretende favorecer, em função dos usos previstos do teste, a expensas de outras. Essa decisão é relativa e prende-se com as necessidades da criação de um argumento de validação: será difícil justificar a validade do constructo de um teste que visa certificar a capacidade de interacção oral com recurso a tarefas de baixa interactividade – leitura de um texto em voz alta, resposta a perguntas directas que não têm em conta respostas anteriores do aluno, recitação de um texto que o aluno poderá ter memorizado antes do momento de aferição.

A qualidade de impacto é de particular importância para um teste que se pretende diagnóstico. Bachman e Palmer (1996:29-35) apresentam um modelo descritivo dos efeitos que um procedimento de aferição tem sobre os envolvidos no processo de educação/avaliação. Os impactos podem verificar-se tanto a nível macro como micro. Aquele entendido como a organização do sistema de ensino, os objectivos gerais estabe-

[2] A noção de tarefa pode ser operacionalizada de duas formas. Na terminologia do QECR, uma tarefa é uma acção executada por um ou mais indivíduos, utilizando estrategicamente as suas próprias competências para atingir um determinado resultado (Cf. Conselho da Europa, 2001:29). Bachman e Palmer (1996:43-44) oferecem uma definição semelhante: as tarefas de comunicação estão intimamente ligas com situações específicas, são orientadas por objectivos e implicam a participação activa dos utilizadores da língua.
Outra operacionalização da noção de tarefas é a sua aplicação à construção de testes. Neste sentido uma tarefa é a combinação de instruções, material-estímulo e resposta. A tarefa visa elicitar um comportamento que permitirá fazer inferências sobre o grau de desenvolvimento de uma competência do indivíduo. (Vd. ALTE members, 1998:389, Bachman e Palmer, 1996:43-60). Uma tarefa de teste será tão mais autêntica quanto modular as características da tarefa de uso da língua que está na sua base.

lecidos por órgãos tutelares, ou a criação de conteúdos/padrões de desempenho curriculares. Este como a influência directa sobre alunos e professores que são os utilizadores imediatos de um qualquer instrumento de aferição. No caso de um teste diagnóstico podemos considerar que a dimensão impacto tem de ser elevada ao nível micro para se considerar o teste útil. Relembrando a definição de Alderson (2005) para diagnóstico – a investigação de causas distintas para comportamentos/grau de desenvolvimento de competências que se pretende desenvolver –, podemos inferir que um teste diagnóstico será tão mais útil quanto se demonstrar que há alterações efectivas no comportamento de professores (planificação do trabalho, selecção de conteúdos/materiais/tarefas, alocação de tempo, escolha de metodologias) e alunos (desenvolvimento mais eficiente e eficaz da competência, selecção de estratégias mais produtivas, por exemplo).

Por fim a exequibilidade, que podemos considerar como a noção económica de custo de oportunidade aplicada à organização do sistema de ensino. Os recursos são limitados e como tal devem ser aplicados onde demonstrem ter um maior ganho marginal. Por analogia, o tempo dispendido na concepção e, sobretudo, na aplicação de um qualquer procedimento de aferição tem de ser compensado por informação significativa e produtiva. No caso da instituição de procedimentos diagnósticos que vão além do simples agrupamento de alunos em GNPs, será necessário conceber procedimentos cuja maior morosidade na aplicação não seja um óbice ao tempo e esforço que poderia ser empregue com o mesmo ganho seguindo procedimentos anteriores de seriação e agrupamento.

Nenhuma destas qualidades é condição suficiente para justificar a implementação de um qualquer instrumento de aferição, antes todas as qualidades têm de ser consideradas e por vezes a optimização de uma pode levar a uma diminuição do grau de outra. O caminho a seguir é a descrição e caracterização de cada um dos usos dos ins-

trumentos de aferição, inferindo-se em seguida quais as qualidades que são factores determinantes para o sucesso da decisão.

4.1.1 Especificações para um uso 'diagnóstico'

Como vimos no capítulo anterior, aquilo que em Soares et al. (2005) e no Despacho Normativo 30/2007 é considerado avaliação diagnóstica correlaciona-se sobretudo com o acto de colocação em nível. Este será talvez o uso mais comum dos testes chamados diagnósticos e assemelha-se àquilo a que Alderson (2005:77) chamou "macro-level diagnosis": relacionar o desempenho de um aluno com um determinado padrão de competência ou desempenho. Neste sentido, a escolha e programação do trabalho a desenvolver com os alunos é um uso secundário do instrumento de avaliação e não é claro de que forma os professores utilizarão a informação recolhida, ou que forma esta terá.

Seguindo-se a hipótese aventada por Alderson (2005), que um 'verdadeiro' diagnóstico distingue entre causas para um mesmo comportamento, porventura explicando-as, e orientando o professor para os caminhos a trilhar, então um procedimento diagnóstico deveria incluir itens de aferição para cada uma das sub-componentes das competências do constructo. Contudo, no caso presente tal procedimento seria de muito difícil execução. Em primeiro lugar, porque não se está na posse de nenhuma teoria comummente aceite sobre a natureza da competência de compreensão da leitura, do discurso oral, etc., ou de qualquer taxonomia para estas ou outras competências; em segundo lugar, porque mesmo que tais instrumentos fossem desenvolvidos (porventura com custos pouco justificáveis) ainda seria necessário estabelecer critérios de decisão que permitissem aos professores utilizar de forma significativa e produtiva essa informação. Não obstante a referência feita à adequação do ensino ao conjunto dos alunos, a realidade é que a função de diagnóstico, ou seja a identificação de causas distintas para

comportamentos semelhantes e a consequente remediação com procedimentos adequados, não pode ser integrada em nenhuma teoria explicativa ou procedimentos partilhados. Avaliação diagnóstica, no contexto actual, é um sinónimo para colocação em nível e é com economia de meios humanos e materiais e certeza nas escolhas feitas que essa mesma acção deve ser executada. O teste de colocação em GNP deve ser de manuseio (aplicação e classificação) tão fácil que permita a sua utilização tanto em grupos numerosos no início do ano, como com indivíduos isolados no momento de chegada ao sistema de ensino. Independentemente do momento de aplicação, os resultados que devolver devem permitir de forma clara e eficiente decidir pela colocação do aluno em qualquer GNP.

Assim, a abordagem mais produtiva, do ponto de vista dos elaboradores e utilizadores de instrumentos de aferição, seria a criação de um único instrumento que permitisse decidir sobre a colocação a progressão, como Pascoal e Oliveira (2006) recomendam. Contudo, a progressão, especialmente a progressão entre o nível B1 e B2 implica uma inferência sobre o desempenho futuro do aluno em situações de uso da língua de 'vida real': a sala de aula que não a de PLNM, antes de mais, o contexto escolar fora da sala de aula, a sociedade portuguesa. Ora tal inferência reclama um grau de certeza que implicará, muito provavelmente, o uso de itens de aferição que visem capacidades de elevada ordem e um bom grau de fiabilidade das aferições feitas (cf. secção seguinte).

Com efeito, o momento e modo como a aferição para a colocação e progressão de nível se realizarão são bastante distintos. A colocação será feita, no mais das vezes, no início do ano lectivo, por ventura por um conjunto reduzido de professores, num período de tempo curto e decidindo sobre um número elevado de alunos. Neste caso a exequibilidade é uma característica essencial do procedimento a utilizar, tanto mais se tivermos em conta que uma entrevista inicial é, por disposição normativa, parte necessá-

ria deste procedimento. Deste modo, a orientação mais económica será a definição de um modelo de teste relativamente curto, de fácil aplicação e classificação e que permita com elevada fiabilidade calcular o nível de proficiência do aluno em relação com a escala geral do QECR, ou qualquer outra que se entenda utilizar, equacionada ou não com o QECR.

A experiência do projecto DIALANG[3], particularmente os dados recolhidos com a calibragem dos itens de língua inglesa, sugere uma forte correlação entre a proficiência em itens de vocabulário e gramática e a proficiência nas capacidades tradicionais. Alderson (2005:261) vai mais longe e aventa a hipótese de o diagnóstico da competência comunicativa dever versar mais sobre aspectos linguísticos e de baixa ordem do que sobre capacidades de elevada ordem, como as capacidades 'tradicionais' de leitura, escrita, compreensão e expressão/interacção orais.

Visto que não há nas Orientações Programáticas para o Ensino Secundário uma definição clara das tarefas de uso da língua, dos domínios de comunicação e de conteúdos programáticos, é à partida difícil saber o que se correlaciona com o quê. Contudo, será necessário em algum ponto definir tarefas de uso da língua, ou pelo menos tarefas de testagem, quando se estabelecerem padrões mínimos de desempenho (Cf. secção 5.1). Ao se definirem estas tarefas é possível também a criação de itens de vocabulário e gramática que versem sobre capacidades tidas por essenciais para as actividades de compreensão da leitura e produção escrita (reconhecimento de vocabulário, marcadores de coesão, p. ex.). Se todos os itens forem calibrados numa escala comum e se se investigarem as correlações entre o desempenho em itens de capacidades de elevada ordem e

[3] O projecto DIALANG consiste num sistema de testes diagnósticos aplicados pela Internet. Os testes estão disponíveis em 14 línguas da União (incluindo o Português) e aferem a competência dos utilizadores em Compreensão do Oral, Leitura, Gramática, Vocabulário e Expressão Escrita. Inclui também uma componente de auto-avaliação. Os resultados são expressos com relação ao QECR, utilizando as suas escalas de descritores, e não são utilizados quaisquer valores numéricos. Para mais informações vd. www.dialang.org e, sobretudo, Alderson (2005).

itens 'linguísticos', é possível criar testes de colocação com elevado grau de fiabilidade na capacidade de previsão do comportamento futuro do aluno durante o programa de PLNM, contando que este último seja estável e do conhecimento do professor.

Neste sentido, a fiabilidade, o impacto e a exequibilidade seriam as qualidades do teste de colocação em nível a maximizar, com sacrifício de alguma interactividade, autenticidade e até de partes da validade do constructo. De toda a forma, o importante neste caso é obter informação que permita agrupar os alunos em grupos coesos, com pouca variação entre a proficiência média dos alunos, ter alguma ideia sobre quais as capacidades que requerem maior atenção (embora no caso dos níveis A1 e A2, dadas as orientações programáticas, a expressão oral recolher sempre a primazia) e estabelecer algumas hipóteses, ainda gerais e imprecisas, sobre a natureza das dificuldades específicas de cada um dos alunos e de como as abordar.

4.1.2 Especificações para um uso de progressão

Como decorre da secção anterior, a decisão de progressão poderá beneficiar do recurso a instrumento de aferição distintos dos utilizados para a decisão de colocação em GNP. A decisão de progressão não é igual no impacto que terá sobre os alunos. A progressão até ao nível B1 é de impacto relativo em comparação com progressão para o nível B2, dada a alteração no regime de avaliação sumativa a que os alunos se sujeitam. Deste modo um particular cuidado se impõe, especialmente ao demonstrar a validade do uso da informação recolhida, mormente que todas as evidências relevantes foram recolhidas.

Uma vez que a progressão entre níveis se pode dar em qualquer momento do ano lectivo, a mesma se reportando a um nível de proficiência percebido e não a uma prestação num instrumento de aferição que versa apenas sobre o conteúdo do programa lec-

cionado, é possível maximizar as qualidade de validade do constructo, autenticidade e interactividade, com sacrifício de alguma economia de meios. É precisamente no aspecto da autenticidade que se pode considerar alargar a noção de aferição a mais do que o tradicional teste de papel e caneta.

De facto, a decisão que se toma é a de progressão, com base no pressuposto que o aluno atingiu o patamar máximo do nível em que se encontra e que, então, está na posse dos pré-requisitos necessários ao início do estudo num novo GNP. No caso de capacidades de produção é possível seguir duas orientações: uma, criar um teste que elicite comportamento linguístico que se considere uma evidência fiável do grau de desenvolvimento da competência do aluno, outra, recorrer a uma recolha sistemática de produções dos alunos que, colectivamente classificadas e consideradas, se considerem evidência da progressão do aluno e do seu actual grau de desenvolvimento da competência. A segunda abordagem, ainda que menos económica, tem a vantagem de basear as decisões numa amostragem mais representativa das produções do aluno e de permitir equacionar o peso relativo que cada tarefa deverá ter na construção das inferências sobre 1) o desenvolvimento da competência ao nível do padrão de desempenho estabelecido, 2) o desempenho futuro fora do contexto das aulas de PLNM.

No caso da expressão escrita, um portefólio de escrita é uma abordagem produtiva e económica: ao mesmo tempo que se desenvolvem actividades que visam a aprendizagem, produz-se e analisa-se informação que é significativa para uma tomada de decisão – faz-se aferição. Esta abordagem é muita próxima da ideia de Alderson (2005), que o diagnóstico de competências é, na verdade, uma redução da distância entre ensino e avaliação. Não obstante, critérios claros de classificação e amostragem das produções dos alunos terão que ser desenvolvidos (Cf. Cap. 4 deste texto), assim como critérios de decisão: todas as produções devem ser consideradas, ou só aquelas em que o aluno

demonstrou maior proficiência? Deve o aluno demonstrar igual grau de desenvolvimento em todos os formatos textuais importantes, ou apenas uma proficiência média com base em amostras de cada formato? Que formatos textuais devem ser incluídos[4]? São questões como estas que devem guiar o processo de análise de necessidades anterior à definição do programa e à criação dos respectivos instrumentos de aferição.

No caso da produção oral, um procedimento semelhante pode ser aplicado. Regularmente, no decurso das suas actividades de leccionação, o professor procede a uma aferição do desenvolvimento da proficiência do aluno com recurso a escalas calibradas de acordo com o QECR. É até possível usar as mesmas escalas utilizadas na entrevista diagnóstica e até desenvolver, para cada nível, escalas analíticas específicas de cada tarefa do constructo[5]. Não obstante, os mesmos cuidados na definição da amostra e na tomada de decisão que se aplicam à produção escrita também se aplicarão neste caso.

Por fim, nas capacidades de recepção (oral e escrita) será de todo o interesse proceder a uma aferição com recurso a itens de teste/testes escritos, cuja fiabilidade e grau de descriminação sejam bem conhecidos, porventura complementados por itens de vocabulário e gramática que se considerem importantes e que podem nem sempre ser aferidos nas tarefas de leitura e compreensão do oral. De todo o modo, antes que a estrutura de tais testes seja definida é necessário definir o constructo, o objecto das aferições a realizar, questão que a próxima secção trata.

4.2 Definição do constructo

Seguindo Bachman (1990), que defende que um teste diagnóstico pode ser 'syllabus-oriented' ou 'theory-oriented', e após análise dos documentos que conformam o

[4] Cf. Leiria *et al.* (2008:26, 29) para listagem de formatos textuais incluídos no programa de PLNM.
[5] Cf. Luoma (2004, Caps. 3 e 4)

ensino de PLNM na escola pública, há que recortar a natureza do constructo que estará na base da avaliação, diagnóstica e de proficiência/sumativa, em PLNM. Neste sentido, três questões fundamentais se impõem: que fonte de informação usar na definição do constructo, como definir conceptual e operacionalmente o constructo e saber se uma única operacionalização é necessária ou suficiente para cada um dos usos de informação previsto. Nesta secção, procurar-se-á contribuir para uma resposta a estas questões através de uma tentativa de definição do constructo.

A definição do constructo a aferir tem implicações que transcendem a simples validade do uso dos instrumentos de aferição a desenvolver. De facto, a definição do constructo é um passo essencial para a posterior validação de todos os elementos que constituem o ensino/aprendizagem de PLNM: o programa (quando existente), os padrões de desempenho tidos como mínimos aceitáveis (quando estabelecidos), as decisões tomadas por professores e órgãos tutelares. Sem se estabelecer o que a aferição tentará medir e que papel esse objecto tem nas interacções sociais que se pretende potenciar com o programa de ensino, apenas ficaremos na posse de elementos esparsos, que não se sujeitam a nenhuma lógica unificadora. O resultado num teste, um resultado noutro teste, certos desempenhos em situações de 'vida real', consequências que são função de uma qualquer avaliação sumativa serão apenas eventos desconexos se não houver um elemento unificador, o argumento de validação, que una os pressupostos teóricos, as evidências empíricas, decisões e consequências de forma inteligível e sindicável por todos os intervenientes. A definição do constructo é, então, a primeira condição, necessária mas longe de suficiente, para que se possa montar um sistema de aprendizagem/ensino/avaliação válido, socialmente útil e justo.

Tal necessidade decorre do simples facto de a aferição/avaliação (assim como o ensino) não ser indiferente à natureza do objecto sobre o qual se crê actuar. Concepções

sobre o que é ser capaz de usar o PLNM, em situação escolar ou com vista ao mais lato objectivo de 'integração', de como essa capacidade ou competência se desenvolve e se manifesta implicam meios coesos de interacção com ela, assim como critérios de decisão racionais, justos e adequados.

A definição do constructo conforma a interpretação que se faz do significado das classificações num teste. Suponha-se que se considera que o teste de colocação em nível deve ser constituído por 20 itens de funcionamento da língua, 10 sobre vocabulário e 10 sobre gramática. Suponha-se também que após aplicação no teste se verifica que há uma correlação elevada (.85) entre os resultados numa e noutra parte. Que dizer da validade do teste? Se se tiver definido ao nível do constructo que o domínio do vocabulário e da sintaxe, por hipótese, são competências distintas, então talvez se deva investigar se a correlação entre os resultados no teste se deve à técnica utilizada, a uma definição pouco rigorosa do constructo, ou a uma especificação do teste tão lata que qualquer item pode acabar por aferir qualquer coisa. Sem uma definição de constructo não é possível um argumento de validação. Quando muito será possível dizer que os alunos que tiverem o resultado *x* no teste A também tiveram o resultado *y* no teste B, mas isso não diz nada sobre a validade do constructo do teste, especialmente na sua dimensão substantiva (Cf. Cap. 5 deste texto).

O primeiro passo para a definição do constructo a medir é identificar e descrever os domínios de uso da língua. Estes subordinam-se aos objectivos do programa de PLNM, o primeiro dos quais é, recorde-se, a integração. Assim, uma análise dos documentos ministeriais oferecerá uma primeira orientação na definição do constructo.

Todavia, tal análise não é particularmente produtiva. São escassas a referências a domínios de uso de língua, com a óbvia excepção do contexto escolar. De facto, Leiria *et al.* (2008:4) relembram que por disposição administrativa (Despacho Normativo

30/2007) um terço da carga horária do programa de PLNM é dedicada a "trabalho da língua portuguesa enquanto língua veicular de conhecimento para as outras disciplinas do currículo". Esta definição, ainda que lata, de conteúdos de ensino permite recortar um dos domínios de uso da língua para o qual será necessário fazer inferências: o educativo. Este domínio poderá ser caracterizado em função de situações, tipos de texto e papéis em que o aluno se verá envolvido. Leiria *et al.* (2008) dão conta desse facto; na secção 3.1.2 descrevem o uso da língua enquanto língua de escolarização, denotando diferenças nas tipologias de textos e tarefas comunicativas em função de áreas do currículo. Decorrem prioridades metodológicas (privilégio do modo oral) mas, mais importante para o tópico em análise, decorrem também considerações na definição de sub-componentes das competências e da sua articulação com conteúdos do programa.

Não obstante, nenhum outro domínio de uso é definido, tendo-se de inferir quais os domínios de uso dos conteúdos programáticos apresentados em Leiria *et al.* (2008). É possível que a inclusão de tais domínios – privado e público – decorra do próprio conceito de integração, ou até mesmo que sejam de tal forma essenciais à aprendizagem de uma LNM que têm sempre de ser incluídos, de alguma forma, num qualquer programa. De qualquer modo, não são indicadas tarefas específicas em nenhum domínio, antes competências/conteúdos que se correlacionam com um ou outro nível, sem indicação das tarefas de uso da língua a que se referem (Cf. Leiria *et al.*, 2008:17-22).

Considerando-se que o programa de PLNM apenas tem uma estrutura, ainda que muito lata, até ao nível B1 (não são apresentadas quaisquer orientações para os níveis B2 e C1, apesar de o Despacho Normativo 30/2007 prever uma hora extra semanal de apoio para os alunos de PLNM já integrados nas turmas de Língua Portuguesa) e que só é possível encontrar alguma distinção nos conteúdos a partir do nível B1 (domínio metalinguístico e metadiscursivo), uma hipótese a explorar nesta fase é o recurso ao

Nível Limiar como fonte de um constructo. É um documento que pelo menos alguns dos professores de PLNM já conhecerão (a par do *Português Fundamental*) e que, ainda que não esteja equacionado com o QECR, fornece uma base de fácil manuseio e que com economia pode ser aplicada a cada um dos níveis. Outra vantagem do recurso ao *Nível Limiar* é que este, ao contrário de qualquer outro documento de momento disponível, inclui um guia para os utilizadores que permite o cruzamento de tipologias de texto com actos de fala, noções, gramática, etc. Tendo em conta que nesta fase os professores terão de desempenhar os papéis de organizadores de cursos, elaboradores de materiais didácticos/instrumentos de aferição e docentes, o *Nível Limiar* é um atalho produtivo a explorar. Numa fase posterior de definição de padrões de desempenho e calibração de itens (Cf. Cap.5 e Conclusões), os materiais desenvolvidos com base no *Nível Limiar* poderão ser calibrados numa escala comum, assim como os testes e programas poderão ser ligados com o QECR.

Contudo, o *Nível Limiar* não resolve o problema mais premente nesta fase: definir quais as tarefas de uso da língua que são prioritárias e operacionalizá-las em itens que possam ser usados com proveito tanto como exercícios em sala de aula como em momentos de aferição. De particular interesse são as tarefas de produção e interacção. Seria de todo o proveito cruzar as tipologias de texto apresentadas em Leiria *et al.* (2008) para cada nível com tarefas reais de escrita com que os alunos se depararão no contexto académico[6]. De igual forma, caracterizar desempenhos prototípicos de alunos nativos, no que concerne ao uso da língua, usando-os como referência para os padrões de desempenho a esperar dos alunos de PLNM. *Mutatis mutandis*, a mesma consideração pode ser feita para a expressão e interacção oral, particularmente no que se refere à adequação do registo.

[6] Cf. Fernandez (2003), em que é avançada uma proposta de desenvolvimento do QECR por tarefas comunicativas.

5. Investigar a Validade e estabelecer padrões de desempenho

"It is responsibility of the test developers to go beyond mere assertions of reliability and construct validity, and to provide evidence to test users that demonstrates *that their tests have the qualities the developers claim."*

Bachman (2004:5)

Samuel Messick (1988:13) definiu validade como "an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment". Esta avaliação da validade de um teste pode ser comparada a um processo de investigação científica, em que hipóteses (as asserções feitas sobre a adequação do uso de um instrumento de aferição) são confrontadas com dados empíricos e análises qualitativas. Este é um processo iminentemente social: um teste não é válido só porque é psicometricamente bem construído, um teste é válido se for reconhecido pela comunidade (política, escolar, científica) em que é aplicado como uma boa medida de competências e uma boa fonte de informação para as decisões que com base nele têm de ser tomadas. Desta forma, a validação do uso de um teste é um processo simultaneamente retórico e empírico.

Messick (1992:7) relembra também a importância de se considerarem qualidades dos testes como a fiabilidade, validade, comparabilidade e justeza não porque sejam requisitos psicométricos, mas porque, numa mais larga medida, são valores sociais que conformam o próprio contexto (político e educacional) em que os testes são usados. Embora defenda também que em contextos educacionais de baixo impacto[1] os padrões de exigência dessas qualidades possam ser algo relaxados, o facto é que é sempre necessário demonstrar que estão presentes em qualquer instrumento de avaliação e em grau

[1] Cf. Secção 1.2

mínimo suficiente para o contexto de uso. Estabelecer que grau mínimo é esse é toda uma outra questão…

Ainda assim, há que nunca perder de vista que a noção de baixo e elevado impacto não é dicotómica e porventura a interacção entre usos de distintos instrumentos de avaliação é, de forma agregada, de muito elevado impacto. Considerando que a avaliação de baixo impacto é aquela que tem consequências de pouca monta e facilmente alteráveis para o sujeito, poder-se-ia incluir nesta categoria, à partida, a avaliação diagnóstica. Contudo, como vimos, no caso em análise os usos de instrumentos de aferição diagnóstica (e de conhecimentos) têm implicações distintas conforme o grau de desenvolvimento da competência.

Recorde-se que o teste diagnóstico pode implicar a inclusão ou exclusão do aluno no programa de PLNM, com consequências no regime de avaliação sumativa a que se sujeita. Não são despiciendas as consequências de se incluir no nível B2 (sujeitando assim o aluno ao Exame Nacional de Língua Portuguesa) um aluno de nível B1, só porque não se cuidou de saber do grau de fiabilidade, descriminação e imparcialidade do teste usado para tomar essa decisão. E, ainda que se afirme que a colocação é facilmente alterável no contexto do sistema de ensino português, é preciso documentar as práticas que efectivamente se verificam nas escolas e verificar se tal asserção é verdadeira. Por outro lado, note-se que há uma legítima expectativa, por parte do público-alvo deste programa e num mais lato sentido da sociedade, de uma coerência de procedimentos. Um aluno deve ser incluído num qualquer nível com base numa estimativa semelhante da sua proficiência e da aplicação sistemática dos mesmos critérios de decisão, independentemente do momento ou lugar em que tal decisão ocorra.

De notar também que para além de um falso positivo (colocar em B2 quem tem um nível de proficiência inferior) também um falso negativo (manter em B1 um aluno

que já ultrapassou esse limiar de proficiência) acarreta questões de justiça: na prática, sujeita-se a um regime de certificação de competências mais favorável um aluno que deveria realizar o exame nacional de Língua Portuguesa como todos os outros alunos. Questões como esta não se restringem a uma análise empírica, mormente estatística, embora tenham implicações técnicas: como todos os instrumentos de aferição têm uma margem de erro inerente, é necessário decidir se este erro beneficia falsos positivos ou negativos. Esta decisão prende-se com a noção de validade como de justeza: que uso é mais adequado neste contexto? A resposta a esta questão tem de ser atalhada de forma pública e sindicável.

É por este motivo que defender um menor grau de exigência dos padrões de qualidade dos instrumentos de aferição em função do seu uso imediato é de difícil sustentação.

As imprecisões resultantes de aferições feitas com base em instrumentos de rigor desconhecido, ou o impacto resultante das decisões baseadas nestes instrumentos, agregam-se de uma forma que pode ser imprevisível e, só porque ninguém as estimou até hoje, não significa que não sejam geradoras de grandes ineficiências no ensino e, sobretudo, de injustiças. É de relembrar, mais uma vez, as palavras de Roldão (2006:41): "*é na avaliação reguladora que se confirma a avaliação certificativa e vice-versa*" (ênfase no original).

Deste modo, talvez fosse de considerar que a noção de validação, nas suas dimensões evidencial e consequencial, poderia ser usada com proveito não apenas na sindicância do uso de instrumentos de aferição, mas em todo o processo de ensino, aprendizagem e avaliação, esta última particularmente na sua componente de certificação. De todo o modo, no caso presente a questão é saber como validar o uso de um hipotético teste 'diagnóstico' quando este começar a ser aplicado nas escolas portugue-

sas. Este capítulo tentará apontar alguns caminhos para essa investigação, relacionando sempre a noção de validação com a criação e definição de padrões mínimos de desempenho.

De notar também que a validade, para além de não ser uma característica do instrumento de aferição (antes dos seus usos), não é uma qualidade que se possa confirmar ou infirmar, dicotomicamente, num qualquer ponto no tempo. Antes, a validação é um processo contínuo, que informa e é informado por aspecto sociais como psicométricos, e que pode, dadas as alterações de circunstâncias e/ou a recolha de novas evidências, fazer variar em grau a nossa perspectiva sobre a validade. Assim, a validação pode ser vista como um processo inerente à programação, execução e controlo de todo o processo de ensino e avaliação, constantemente colocando-se a diferentes actores asserções e contra-asserções que pretendem confirmar ou infirmar aspectos da sua actuação. No caso presente, preocupa-nos o teste 'diagnóstico', mas, como veremos, dada a natureza dos dados quantitativos e qualitativos a recolher, a validação do seu uso é também em boa medida a validação dos testes de avaliação de conhecimentos, dos exames nacionais de certificação de competências e das decisões que os professores tomarem na planificação e execução do seu trabalho. Deste modo, as conclusões tiradas de diferentes investigações sobre a validade dos instrumentos de aferição desenvolvidos devem ser formadas com cautela, sendo certo, porém, que um processo de validação transparente, bem documentado e partilhado entre tutela e escolas é um primeiro e importante passo para ganhos de qualidade nos processos de ensino e, em última análise, nas aprendizagens feitas e na integração dos alunos.

Por fim, de notar que a fiabilidade dos instrumentos de aferição usados é condição *sine qua non* para a sua validade. Se não se demonstrar que os resultados num determinado teste são função da interacção entre o mesmo e a competência do aluno (e

não resultado de erros de medição), não há maneira de se aquilatar da validade do mesmo, uma vez que as variações dos resultados e as inferências neles baseadas poderiam ser consideradas função de factores externos e irrelevantes para o constructo a medir. Visto que controlar a fiabilidade das aferições depende, em grande medida, dos tipos de itens utilizados e ainda não estamos na posse de um modelo definitivo de teste diagnóstico, opta-se por não se dedicar uma secção a este problema. De qualquer modo, independentemente do formato final a ser utilizado, dois cuidados se impõem. O primeiro é garantir que factores tecnológicos não são fonte de infiabilidade nas medições. O recurso a itens de classificação 'objectiva' (Verdadeiro/Falso, escolha múltipla, por exemplo) e a formatos únicos de teste é uma forma de obviar este problema, ainda que, possivelmente, a troco de alguma autenticidade das tarefas. O segundo, e porventura mais importante, é o factor humano. Eventualmente, no momento do diagnóstico ou mais tarde no momento de aferição dos conhecimentos, serão usados itens de classificação aberta, particularmente no domínio da produção escrita e oral. Neste caso, apenas será possível garantir patamares mínimos de fiabilidade através do uso de tarefas comparáveis e de procedimentos de classificação igualmente replicados por todos os aferidores (neste caso, os professores). O treino tem um papel fundamental neste particular: não é razoável esperar-se que centenas de professores apliquem de forma comparável os mesmos critérios de classificação se não forem sujeitos a algum procedimento de familiarização e treino. Como veremos, a instituição de padrões de desempenho e o uso de exemplos são também de importância curial neste caso.

5.1 Padrões de desempenho

Uma condição essencial para a validação dos testes diagnósticos, bem como das Orientações Programáticas para o Ensino Secundário e da asserção feita no Despacho Normativo 7/2006 que um aluno com nível B2 ou C1 tem um domínio do Português que

lhe permite seguir o programa nacional de Língua Portuguesa, é estabelecer uma ligação entre todos estes documentos e o QECR, uma vez que se entendeu usar o mesmo como referência para a organização de todo o programa de PLNM. O primeiro passo para esta ligação é a instituição de padrões de desempenho claros e inequívocos que se apliquem ao uso da língua nos contextos pertinentes. Sem estes padrões de desempenho, estaremos apenas na posse de elementos esparsos e praticamente desgarrados de qualquer modelo teórico.

Estes padrões de desempenho podem ser considerados versões operacionais dos descritores do QECR. Seja no momento de classificar produções linguísticas dos alunos (em formato escrito ou oral), seja no momento de classificar testes ('diagnósticos', de competências), é necessário haver um indicador claro que funcione como aplicação do descritor QECR. Esse indicador é o padrão de desempenho e a sua delimitação e descrição é o primeiro passo para a fiabilidade dos instrumentos de aferição. Este indicador de desempenho pode ter a forma de uma classificação mínima ou máxima num teste de formato e conteúdo padronizado, ou uma classificação de uma produção, escrita ou oral, elaborada com recurso a escalas desenvolvidas e calibradas com base no QECR. Em qualquer caso, a criação de padrões de desempenho requererá:

1. níveis de desempenho (neste caso os níveis QECR serão os usados);
2. descritores de desempenho (que poderão ser os dos níveis QECR, muito embora fosse de explorar a hipótese de criar descritores intermédios, particularmente para tarefas de produção/interacção oral e escrita do domínio educativo);
3. exemplos de produções;
4. classificações em testes que podem ser usadas na decisão de inclusão, progressão, exclusão em GNPs.

Se tais indicadores não existirem, não é possível comparar as decisões tomadas em diferentes momentos e por diferentes decisores, uma vez que não há garantia que todos estejam a aplicar da mesma forma o descritor QECR. O segundo passo será demonstrar com aceitável grau de certeza que o nível de proficiência dos alunos é estimado com igual rigor pelos aferidores, ou seja, que todos os aferidores compreendem o âmbito e o alcance do padrão de desempenho e o aplicam de forma sistemática e coerente. Demonstrando-se a fiabilidade, está o caminho aberto para se investigar a validade dos usos dos testes.

O processo de definição e delimitação dos padrões de desempenho é complexo e moroso. O Conselho da Europa elaborou um manual e um suplemento de referência (Cf. Conselho da Europa 2003, 2004) que visam auxiliar a ligação entre instrumentos de avaliação e o QECR. Vários procedimentos são recomendados, sendo que as secções dedicadas à definição de padrões de desempenho são particularmente detalhadas (Conselho da Europa 2003:Cap. 5, 2004:Secção B). Essencialmente, três caminhos são possíveis: ligar o instrumento de aferição ao QECR através de outro instrumento ou de uma escala de proficiência já ligados ao QECR, ligar directamente ao QECR. No caso presente, uma vez que não existe qualquer instrumento de aferição ligado ao QECR no contexto escolar, a ligação terá de ser feita de forma directa.

A prática actual é recorrer a um painel de especialistas que definirá o padrão de desempenho considerado como operacionalização mínima dos descritores do QECR. Dois caminhos são possíveis, ambos idealmente deveriam ser aplicados na organização do ensino de PLNM. O primeiro é a análise de produções linguísticas de uma amostra de alunos representativa da população escolar. Estas produções podem ser relacionadas com os descritores do QECR e, quando se atinja um bom grau de consenso entre os membros do painel sobre quais as produções que melhor representam cada nível, uma

colecção de exemplos ilustrativos e comentados deve ser criada para treino de professores. Uma vez demonstrada a capacidade destes de aplicarem de forma coerente os critérios de classificação aos exemplos calibrados, poder-se-á passar ao treino da aplicação dos critérios de classificação a produções dos próprios alunos. Este processo é um garante de comparabilidade e, se seguido com rigor e de acordo com as boas práticas recomendadas pelo Conselho da Europa, levará a uma maior consistência no sistema de ensino e potenciará futuras investigações sobre a validade e o impacto das práticas educativas.

O segundo caminho prende-se com os próprios testes a utilizar. Independentemente do formato de teste escolhido na fase de definição da forma e conteúdo (cf. Cap. 4 deste texto), há que decidir que classificação mínima corresponde a cada nível QECR. Os textos do Conselho da Europa (2003 e 2004) têm descrições detalhadas de como este processo deve ser conduzido e documentado, pelo que não é necessário descrever o seu conteúdo aqui. Contudo, de notar que o manual do Conselho da Europa (2003 e 2004) foi pensado primeiramente para aferição com uso certificativo. No caso presente, o padrão de desempenho em testes tem de ser calculado, possivelmente, para dois formatos de teste e os mesmos têm de demonstrar alguma correlação. A concepção de teste 'diagnóstico' tem grande influência neste processo. Se se seguir a abordagem de Pascoal e Oliveira (2006), de se usar o mesmo teste para fins de colocação em GNP e de certificação das aprendizagens, então apenas será necessário definir um padrão de desempenho por teste/nível de proficiência, uma vez que apenas existe uma operacionalização do nível. No entanto, se se optar por formatos distintos de testes então um padrão de desempenho terá de ser calculado para cada um, assumindo-se que o formato de cada teste se mantém estável ao longo do tempo, ainda que cada item possa ter versões diferentes.

Deste modo, parece que uma real ligação ao QECR apenas será possível se se estiver na posse de instrumentos de aferição comuns a todo o sistema educativo, uma vez que a complexidade e morosidade do processo torna quase impraticável que em cada escola se proceda a uma ligação testes/QECR.

Um procedimento económico seria a definição de uma escala comum de proficiência em cada uma das capacidades 'tradicionais' (ler, escrever, ouvir, falar) e do funcionamento da língua, ligando-se a esta uma colecção de itens. Estes, uma vez calibrados, poderiam ser usados na criação de diferentes instrumentos de aferição (de colocação em nível, de progressão) e até mesmo de materiais didácticos.

No entanto, no caso em análise duas dificuldades acrescem à questão de definição e operacionalização inequívoca de padrões de desempenho. Essas dificuldades são 1) a instabilidade do programa de ensino, 2) a possibilidade de a inclusão/exclusão dos alunos em GNP se poder fazer com assimetria de desenvolvimento de cada capacidade.

Quanto à primeira dificuldade, esta traz implicações para a definição do padrão de desempenho considerado mínimo para a inclusão no GNP. Se o programa fosse único e aplicado de forma estável, o padrão de desempenho no teste de colocação seria aquele valor que se correlaciona com o sucesso de todos os alunos no fim do programa de estudos: chegar a um patamar de proficiência que permite progredir para o GNP seguinte e que seja função do programa de ensino a que o aluno foi sujeito. É por este motivo que alguns autores, como Bachman (1990:253), defendem que no caso de um teste de colocação em nível (afinal, o principal uso do teste 'diagnóstico' em estudo) não é essencial recorrer a uma teoria explicativa da competência, apenas demonstrar um elevado grau de capacidade de previsão do desempenho futuro do aluno no contexto de ensino/aprendizagem de PLNM. No entanto, uma vez que não existe um programa único definido, antes se admite que com base nas orientações programáticas os professores

organizem livremente o seu trabalho e em qualquer momento procedam à progressão dos alunos para o GNP seguinte, não é possível usar o teste para prever a progressão do aluno ao longo do programa de estudos. Resta conhecer as implicações desta orientação: um estudo de validação mais alargado deveria investigar como os professores reagem aos resultados dos testes 'diagnósticos' e se de facto há flexibilidade na execução do programa e na progressão dos alunos, ou se, pelo contrário, as práticas dos professores se manterão essencialmente inalteradas e os alunos só poderão progredir em momentos pré-definidos (por exemplo, o fim de cada período ou ano de escolaridade) ao invés de qualquer momento em que demonstrem ter atingido um novo padrão de desempenho. Desta forma, a orientação de Ribeiro e Ribeiro (1989) e Pascoal e Oliveira (2006) para a definição do conteúdo do teste 'diagnóstico' parece ser a mais adequada: um teste que permita saber se o aluno já atingiu ou não um patamar de desempenho que lhe permita iniciar novas aprendizagens, recorrendo-se a itens que avaliam os pré-requisitos (neste caso, o padrão de desempenho do nível anterior) e algumas das aprendizagens futuras.

A segunda dificuldade sobre as decisões a tomar com base em padrões de desempenho prende-se com a indefinição sobre os critérios de decisão para a progressão entre GNP. De facto, embora se organize todo o programa de PLNM em torno de uma competência lata, a competência comunicativa, reconhece-se que a mesma pode ser desenvolvida em graus diferentes por cada capacidade. Admite-se que um aluno seja incluído num GNP com diferentes graus de desenvolvimento, mas não se define quão diferentes podem ser ou, tão-pouco, se existirá alguma precedência entre uns e outros. Por exemplo, é possível incluir no nível Avançado um aluno que demonstre ter um padrão de desempenho B2 em leitura, compreensão do oral e vocabulário, mas seja apenas B1 em expressão escrita? Ou vice-versa? Caso estes critérios de decisão não sejam explicitados a própria definição de padrões de desempenho pode tornar-se inútil: as

decisões tomadas com base na aferição do nível de desenvolvimento de competência serão incomparáveis, não porque baseadas em instrumentos de aferição pouco fiáveis, mas porque baseadas em critérios de decisão distintos e porventura operacionalizados também eles de forma divergente. O resultado será uma sistemática invalidade de todo o programa de PLNM.

5.2 Criar argumentos em sustentação da validade

5.2.1. Validade interna

Estabelecidos padrões de desempenho claros, o passo seguinte na demonstração da validade é a investigação da sua componente interna nos testes utilizados, ou seja, o valor intrínseco do teste: a qualidade dos itens (valores de dificuldade e discriminação, ausência de parcialidade), a adequação dos itens/testes ao nível proposto, a fiabilidade dos itens e dos testes, a representatividade do constructo (Conselho da Europa, 2003, Cap. 6.2). Qualquer uma destas investigações ultrapassa a capacidade de um professor isolado, ou até mesmo de uma escola/agrupamento de escolas. Por um lado, dado o volume de informação que é necessário recolher e tratar, por outro, porque é necessário demonstrar a validade do instrumento de aferição para além do contexto limitado de uma escola. Recorde-se que o objectivo final de todo o programa é a integração dos alunos, presume-se na sociedade portuguesa e não apenas no contexto escolar, e que os objectivos são os mesmos para todo o sistema de ensino português. Logo, qualquer evidência que se recolha tem de ser generalizável a todo o sistema de ensino, bem como o impacto de cada acção individual tem de ser comparada com o mais lato contexto do impacto do sistema de ensino sobre a sociedade.

No que concerne à qualidade dos itens utilizados nos testes, há que demonstrar, antes de mais, 1) que cada item afere apenas aquilo que supostamente deve aferir, 2) que a dificuldade estimada do item é aquela que se verifica empiricamente. A segunda

condição é verificável estatisticamente, recorrendo, por exemplo, a outras medidas da mesma competência (o que poderá ser difícil no caso presente, dada a ausência de quaisquer instrumentos de aferição) ou à pilotagem dos itens/testes numa amostra representativa da população a que se destinam. No caso presente, este segunda abordagem é a única possível nesta fase e seria recomendável que a aplicação de itens/testes fosse feita a uma amostra representativa da população (ou, se possível, a toda) durante a fase de recolha de produções linguísticas a utilizar na definição de padrões de desempenho para a expressão escrita e oral. Assim, seria possível na fase de definição de classificações mínimas do teste cotejar as prestações do mesmo grupo de alunos em instrumentos de aferição distintos, usando os dados originados na calibração de itens/testes e na formação de padrões de desempenho para as capacidades produtivas.

A primeira condição, usualmente, é também atalhada recorrendo ao julgamento de especialistas, muito à semelhança do que acontece com a definição de padrões de desempenho. Várias críticas são feitas a este procedimento: *grosso modo*, podemos reduzi-las à noção que, só porque um conjunto de especialistas concorda sobre o que um item afere ou a que nível se destina, isso não quer dizer que o item afira realmente essa competência, pode apenas significar que os especialistas arranjaram uma forma de se pôr de acordo. No entanto, deixar apenas ao critério de professores separados dos seus pares, ou de especialistas não identificados que trabalham em circuito fechado, o ónus (ou dir-se-ia a prerrogativa) de decidir se as suas asserções são ou não verdadeiras, leva-nos a uma situação de reificação: a coisa é o que cremos ser porque cremos que o é (ou alguém crê por nós e aceitamos esse julgamento acriticamente por via de um argumento de autoridade). Claro que o julgamento de especialistas pode simplesmente limitar-se a trocar uma reificação por outra, no entanto tal julgamento, se bem documentado

e publicitado, está sempre sujeito a desafios e a novas confirmações ou infirmações, logo encerra maior potencial de validade.

Outra crítica que se pode levantar ao julgamento por especialistas é que estes não conseguem abarcar todas as *nuances* das necessidades de informação daqueles que são, em última análise, os verdadeiros utilizadores dos testes, os professores, e que portanto o seu julgamento é menos válido porque menos pleno de significado prático. Assim, faria mais sentido deixar à consideração dos professores a criação e avaliação dos instrumentos de aferição, uma vez que ninguém melhor do que eles pode saber da sua utilidade. Este argumento é particularmente sedutor em situações de avaliação formativa, em que as acções a tomar com base na interpretação dos resultados dos testes são altamente contextuais. Todavia, tal abordagem limita-se simplesmente a redefinir o conceito de especialista: troca-se o académico informado por pesquisa bibliográfica e investigação empírica por uma espécie de leigo bem treinado e melhor intencionado, o único 'realmente' consciente das necessidades.

A abordagem mais sensata, e recomendada pelo Conselho da Europa (2003, 2004) é o recurso a painéis de especialistas que conheçam bem o QECR, reflictam diferentes competências, sensibilidades e concepções sobre competência comunicativa e aprendizagem mas que sejam igualmente reconhecidos como tendo a capacidade de contribuir para um processo de investigação que é socialmente construído: saber o que é que afinal se está a aferir com um teste e o que fazer em função dos resultados que se obtém.

Deste modo, seria recomendável a programação de um procedimento de definição de padrões de desempenho que também incluísse uma investigação sobre a perspectiva de elaboradores e utilizadores dos testes sobre a representatividade do constructo a aferir e a natureza dos itens utilizados. Numa fase posterior, recolhidos dados empíricos,

seria possível saber da adequação dos itens utilizados às inferências que se pretende fazer. Por exemplo, se se incluiu um determinado item no teste com o fim de aferir o domínio do vocabulário mas os resultados da pilotagem sugerem que os resultados do mesmo se correlacionam fortemente com os resultados em itens de compreensão da leitura, o que é que isto nos diz sobre a validade do constructo do teste? Se os elaboradores do teste incluem itens destinados a aferir a capacidade de fazer inferências mas o painel de especialistas considera que esses itens visam aferir a capacidade de identificar a ideia principal do texto, é o teste válido?

Vários procedimentos de análise qualitativa e quantitativa devem ser ensaiados em conjunto (Cf. Conselho da Europa 2003, Cap. 6), no entanto será sempre necessário recorrer a uma teoria psicométrica que permita contextualizar os dados e orientar as decisões sobre a construção e aplicação dos testes. O uso da Teoria da Resposta ao Item (TIR), embora requerendo elevada competência técnica, permitiria o desenvolvimento de uma escala única de proficiência sobre a qual se colocassem itens, testes e alunos, facilitando assim a comparabilidade e a criação de bancos de itens que poderiam ser utilizados em momentos de aferição e ensino. Outra vantagem de se desenvolver uma escala única é que, uma vez demonstrada a sua relação com o QECR, relacionar novos testes ou itens será apenas uma questão de demonstrar a relação destes com a escala desenvolvida (uma forma de ligação indirecta). É cada vez mais claro que um bom procedimento de desenvolvimento de padrões de desempenho, alargado e complementado por investigações sobre a aplicação dos testes e o seu impacto sobre as aprendizagens, teria um importante papel na produtividade e eficácia do ensino de PLNM.

Deste modo, a validação interna dos testes implicaria demonstrar, entre outros critérios, se:

1. É claro para todos os utilizadores dos testes que competência(s) cada item visa aferir e que tal asserção é sustentada por dados empíricos (correlação entre medidas da mesma competência feita com recurso a itens semelhantes/métodos distintos e não correlação entre itens que aferem características diferentes do constructo);
2. Todas as dimensões importantes do constructo a medir estão devidamente representadas;
3. Os valores de dificuldade e descriminação são conhecidos com adequado grau de segurança estatística;
4. Os valores de dificuldade empiricamente conseguidos se correlacionam com as estimativas de dificuldade feitas pelo painel de especialistas;
5. Os itens descriminam entre alunos apenas em função de características do constructo (desenvolvimento da competência aferida) e não de factores externos, como língua materna, nacionalidade ou género;
6. Os processos de aplicação e classificação dos testes são comparáveis em todos os contextos e consequentemente há um grau satisfatório de fiabilidade nos resultados;
7. Diferentes formas do mesmo teste devolvem resultados comparáveis;
8. Alunos em níveis diferentes têm resultados diferentes em testes orientados para um único nível;
9. Os processos mentais usados na resolução dos itens são análogos aos que se pretende aferir e como tal os resultados se correlacionam mais com a proficiência do que com estratégias de resolução de testes;

5.2.2 Validade externa

No documento do Conselho da Europa (2003, Cap. 6.3), a validade externa é sinónimo de demonstrar a ligação entre um determinado valor num dado instrumento de aferição, ou um padrão de desempenho, e um ponto na escala do QECR. Contudo, mesmo que tal asserção seja demonstrada não segue logicamente que os padrões de desempenho estabelecidos e/ou o recurso aos níveis QECR, do modo como é feito, sejam válidos. Notem-se as palavras de Kaftandjieva (2004:19):

> *For example, as far as it concerns the CEF scales of languge proficiency there is evidence of their validity as performance standards (North, 2000, Kaftandjieva & Takala, 2002). This fact, however, does not guarantee valid interpretations of the CEF scales in any particular case of their application. Therefore, the validation effort in every linkage between language examinations and the Common European Framework for Languages (CEF) should provide enough evidence not only for the plausibility of proposed cut-off scores interpretations, but also for the validity of CEF scale interpretations as well as for the validity of the score interpretation as a whole."*

Assim, ainda que se venha a demonstrar a validade das classificações mínimas estabelecidas para os instrumentos de aferição, a validade dos padrões de desempenho para tarefas de produção/interacção, a validade do processo que levou à sua definição e a sua ligação ao QECR, resta demonstrar o aspecto mais importante da validade: que as decisões tomadas com base nestes instrumentos e nestes padrões de desempenho são adequadas ao contexto em causa. O recurso ao QECR, ainda que com forte potencial retórico num argumento de validação, não é garantia da validade das práticas educativas, mormente da selecção de determinados níveis como critério de inclusão/exclusão em programas de apoio. Essa validação tem de ser feita com base nos termos em que se

desenvolveu e aplicou o programa e os instrumentos de aferição. Em suma, é uma questão de validade do constructo.

5.3 Validade do constructo – aspectos consequenciais

O modelo integrado de Messick (1989) para a validade do constructo baseia-se em seis dimensões fundamentais: conteúdo, substância, estrutura, generalização, externalidade e consequências. Várias destas características são abordadas no processo de validação interna dos instrumentos de aferição. A dimensão de conteúdo ao se demonstrar que nada do que é importante para o constructo a aferir foi deixado de fora, assim como nenhuma dimensão supérflua foi incluída. A dimensão substantiva na investigação das causas para as classificações observadas, sobretudo à luz de um racional teórico que explique a proficiência e, sobretudo, de evidências que demonstrem que os processos cognitivos que se visa aferir – o constructo conceptualmente definido – são de facto operacionalizados na interacção com os itens de aferição (este aspecto é de particular importância para a avaliação diagnóstica e de colocação, uma vez que se pretende fazer inferências sobre determinadas competências por apelo a um raciocínio de subsidiariedade entre sub-competências operacionalizadas e macro-competências sobre o desenvolvimento das quais se pretende fazer inferências). O aspecto estrutural pretende demonstrar uma ligação entre as classificações possíveis uma teoria de desenvolvimento da complexidade do constructo. A capacidade de generalizar é um desafio à asserção de que as aferições feitas são generalizáveis entre momentos de aferição, grupos de alunos com diferentes perfis, condições de aplicação dos testes e tarefas seleccionadas. A dimensão externa reclama evidências de que diferentes métodos ou técnicas de aferição devolvem resultados comparáveis, bem como itens que versem competências consideradas distintas devolvem também resultados distintos. Num outro sentido, a validação externa é a demonstração de utilidade prática dos instrumentos (o impacto do modelo de

Bachman e Palmer, 1996) e da relevância dos critérios utilizados na definição do constructo.

Por fim, na busca das evidências que sustentem uma asserção de validade está uma ligação entre o desempenho do aluno no teste e o seu putativo desempenho em situações de uso de língua externas ao momento de aferição e para as quais, supostamente, o programa de ensino se dirige. É por este motivo que é essencial definir os domínios de uso da língua que se pretende potenciar através da instrução e é tão importante definir o conceito de 'integração' no caso em análise: as consequências que se almeja e, pelo menos raciocinando-se de modo inverso, as que se pretende obviar.

Se se entender como contexto da integração a escola, então os domínios de uso da língua a orientar a criação de programas e instrumentos de aferição serão o educativo e o público. No entanto, se se estender o contexto de integração, outros domínios de uso da língua (privado, profissional) deverão ser também incluídos. Poder-se-á argumentar que qualquer um destes domínios é condição necessária à integração, seja qual for o entendimento que se tiver do conceito (desde a assimilação à coexistência de múltiplas identidades culturais) e que portanto a questão é espúria.

Também se poderá argumentar *a contrario* que a integração é o objectivo primeiro da Escola, seja qual for a Língua Materna do aluno, e que portanto reiterar o facto nas orientações ministeriais para o PLNM é apenas repetir um propósito que não levanta objecção. De todo o modo, a necessidade de se incluir um programa de PLNM no ensino público decorre do reconhecimento de que diferentes necessidades reclamam distintas linhas de acção, sempre tendo em vista o lato objectivo da integração dos indivíduos na sociedade via instrução. Impõe-se, então, perguntar que factores (comunicativos, linguísticos) são óbice à integração de alunos de PLNM que melhor são atalhados através deste programa diferenciado, ficando o resto da 'integração' para aquilo que já é (e

será) comum ao ensino de todos os alunos: o programa de Língua Portuguesa, assim que atinjam um nível de competência B2, e os programas de todas as outras disciplinas. Resta depois demonstrar também que essa via possibilita a integração destes alunos na mesma medida que os alunos que têm o português como língua materna. Em suma, não definir o que no objectivo de integração não é específico do desenvolvimento de competência comunicativa em língua portuguesa é deixar a porta aberta, no programa e nos instrumentos de avaliação, a objectivos e actividades de utilidade discutível, possivelmente com o sacrifício de outros objectivos mais proveitosos.

Este é um problema de validade do constructo na dimensão do conteúdo: o que é que deve fazer parte do programa? Mas é também uma questão consequencial. Sem se definir, pelo menos, quais são os objectivos almejados não é possível confirmar ou infirmar o sucesso do programa e das acções desenvolvidas – avaliação incluída – por comparação com as consequências que se verifiquem. Ou seja, não há condições de sucesso ou insucesso, qualquer evidência que se verifique pode ser compreendida como confirmação ou infirmação dos pressupostos, ou até mesmo descartada como irrelevante, uma vez que não há um pressuposto teórico, uma tese, com a qual se possa confrontar.

Se se considerar a validação como medir a mesma coisa com métodos maximamente distintos, como Bachman (1990:240) sugere, então deve considerar-se a validação do teste diagnóstico em paralelo com o teste de progressão de nível e de certificação de competências no fim do Ensino Secundário. Por exemplo, a certificação de competências em Língua Portuguesa no final do Ensino Secundário é feita com base num exame nacional especialmente criado para cada nível. Se os procedimentos de colocação forem fiáveis e externamente válidos (ligados ao QECR) e o mesmo se afirmar dos exames de certificação, então será de esperar uma taxa de sucesso de 100% no exame

nacional. Se a classificação mínima para progressão (10 valores) corresponde a um ponto na escala do QECR imediatamente abaixo do patamar mínimo do nível, então todos os alunos colocados têm de ser capazes de alcançar essa classificação, uma vez que só foram colocados nesse GNP por já terem ultrapassado esse limiar. Caso não se verifique esta condição é necessário investigar a fonte da discrepância, no processo de colocação dos alunos, no instrumento de certificação ou em ambos.

No entanto, a concepção do que deve ser o constructo do teste diagnóstico influencia todo o processo. Caso se entenda que o teste diagnóstico deve ter formato e usos 'verdadeiramente' diagnósticos, seguindo a concepção de Alderson (2005), então há que demonstrar uma correlação entre a prestação dos alunos em testes com itens mais 'linguísticos' ou orientados para sub-capacidades, utilizados nos processos de colocação e organização do ensino, e em momentos de certificação de competência/decisão de progressão, com recurso a testes que se ancilam em itens que visam aferir as capacidades mais de mais elevada ordem[2] (recepção, produção e interacção orais e escritas). Fica em aberto a questão de saber se as variações que se verificam na proficiência são função de uma progressão 'natural', ou se pelo contrário foram potenciadas pela escolha de um plano de estudos individualizado. Para tal, é necessário recolher evidências sobre a forma como os professores utilizaram os dados obtidos com o diagnóstico, se de facto houve diferenças substanciais no seu comportamento na planificação e execução do trabalho, se tais diferenças se correlacionam fortemente com o desenvolvimento das competências dos alunos, ou se, pelo contrário, resultados distintos implicam sempre a adopção das mesmas linhas de acção e estratégias de ensino.

Instrumento de aferição antes da instrução, acções (pretendidas e efectivamente executadas) durante a instrução e prestação em instrumento de aferição com uso certifi-

[2] Emprega-se a expressão 'capacidade de elevada ordem' como equivalente a 'macro-skill', Sub-capacidade como equivalente a 'sub-skill'. Cf. Alderson (2005: 128-129, 184-189).

cativo/sumativo são, então, três elos de uma cadeia de certificação que têm de ser considerados em conjunto. De qualquer forma, a relação entre a aferição diagnóstica e certificativa não se resume apenas a uma questão estatística. Pelo contrário, a evidência estatística é um teste ao pressuposto teórico que conforma a definição do constructo e, em consequência, das especificações do teste. Por exemplo um teste de leitura, construído recorrendo a tarefas, deve correlacionar-se mais fortemente com testes diagnósticos, construídos com itens discretos, que visam sub-componentes da competência de leitura. Pelo contrário, os resultados no mesmo teste devem correlacionar-se menos com resultados em testes diagnósticos de compreensão do oral. Ou seja, aspectos consequenciais, tanto intencionais como inesperados devem ser integrados num argumento de validação e cotejados com os pressupostos que levaram a adopção de determinadas linhas de acção.

Se, por outro lado, se entender o teste diagnóstico numa acepção mais próxima de teste de colocação em nível, como simples indicador de variação da competência em função da sujeição a um plano de estudos essencialmente definido e sujeito a pouca ou irrelevante modificação, então a validação de ambos os instrumentos dependerá em larga medida de outros aspectos consequenciais: todos ou quase todos os alunos que tenham a mesma língua materna e classificações comparáveis no teste de colocação terão uma prestação semelhante no teste de certificação de competências, uma vez que a instrução é constante e a competência semelhante (como determinada pelo teste de colocação).

Resulta claro que a multiplicidade de variáveis no contexto actual – ausência de testes e outros procedimentos de aferição calibrados, ausência de padrões de desempenho, indefinição do programa a seguir – é um óbice a qualquer investigação de validade, sobretudo se o processo for resumido a alguma informalidade limitada a contextos iso-

lados. No entanto, a validade, nas suas diferentes dimensões, nunca é uma estimativa perfeita de uma característica unívoca e inequívoca. As evidências são sempre incompletas, porventura esparsas, nem sempre partilham o mesmo grau de fiabilidade e podem até ser contraditórias. Ainda assim, evidências imperfeitas são melhores que evidências nenhumas e um processo político e administrativo que se sujeita a sindicância, ainda que aproximativa, encerra sempre maior possibilidade de justiça e equidade.

6. Conclusões

A criação do Quadro Europeu Comum de Referência serve um fim que é antes político que técnico. Esse fim é o de intercompreensão e reconhecimento entre diferentes órgãos de educação e avaliação/certificação na Europa com o propósito de promover a mobilidade dos cidadãos. Tal fim, tão lato como ambicioso, implica que o QECR tem de ser suficientemente abrangente para acomodar diferentes tradições educativas e avaliativas e ao mesmo tempo suficientemente específico de forma a comportar algum significado prático. O QECR é, então, um ponto de chegada para todos aqueles que queiram referir-se, com recurso a uma linguagem comum, a padrões que reflectem necessidades contextuais, tipificadas e por vezes irrepetíveis.

É por este motivo que autores como Alderson (2005:121) consideram que o QECR não é particularmente útil como fonte de um constructo que guie a construção de instrumentos de aferição. O QECR não liga domínios de uso e tipos de texto a níveis, nem tão-pouco descreve ou explica o processo que leva à evolução da competência comunicativa, à passagem de um nível para o outro. Por este mesmo motivo, causa alguma estranheza que se estabeleçam níveis de proficiência com base no QECR, e se tomem decisões de elevado impacto com base nesses mesmo níveis, sem que antes se tenha feito qualquer definição do seu conteúdo, domínios de uso de língua, tipologias de texto ou se tenham definidos critérios operacionais (classificações em determinados testes) que permitam caracterizar o comportamento comunicativo dos alunos e, com base nele, tomar decisões.

A simples asserção de ligação de um programa de ensino a um qualquer nível do QECR, ainda que por analogia com a prática em outros contextos educativos, não é garante suficiente da validade das decisões tomadas. Particularmente, é preciso demonstrar a adequação de se considerar que um aluno com nível de proficiência no patamar

inferior de B2 não necessita de um programa de ensino distinto do dos alunos nativos. Esta asserção é ainda mais difícil de confirmar (e de infirmar) visto que em documento algum se define o que são o nível B2 e C1 no contexto do ensino de PLNM no Ensino Secundário, ao contrário do que já acontece, ainda que de forma lata, para os outros níveis. Não são definidos conteúdos para este nível ou se prevê a criação de instrumentos de aferição que permitam dar conta da evolução da competência comunicativa do aluno a partir do momento que é integrado na disciplina de Língua Portuguesa. Assim, não é possível afirmar (ou infirmar) que: 1) a proficiência do aluno continua a desenvolver-se a partir do momento em que é integrado numa turma de Língua Portuguesa, 2) o facto de estar no nível B2 lhe dá todos os requisitos comunicativos necessários a seguir o currículo nacional em igualdade de circunstâncias com os seus pares.

Parece que o recurso ao QECR acaba por ter um efeito diametralmente oposto àquele que se pretendia combater com a sua criação: a redução da justificação política de testes (e das acções neles baseadas) a aspectos de validade aparente – o reconhecimento de legitimidade pelo público em função de um julgamento superficial da forma e conteúdo; a redução da justificação a argumentos de autoridade, o mais sublime dos quais a asserção de 'qualidade técnica assegurada' via consulta do especialista. Uma decisão feita com base no resultado num teste não é válida porque é equacionada (ainda que venha a ser tecnicamente bem equacionada) com o QECR. Uma decisão é válida se demonstrar a adequação do que se faz em função dos dados de que se dispõe, da forma como se os obteve e das consequências do seu uso. O QECR é um ponto de chegada, se, onde e quando seja necessário ligar uma qualquer prática (de ensino, avaliativa) à escala comum com fins de intercompreensão, nunca o ponto de partida para a criação de testes e programas de estudo. Muito menos guia para a tomada de decisões que afectam de forma indelével a vida de milhares de alunos no nosso sistema de ensino. O QECR não

substitui o trabalho de organizadores de cursos e de elaboradores de instrumentos de aferição: definir os domínios de uso da língua, identificar e tipificar tarefas comunicativas relevantes, seleccionar e organizar conteúdos, estabelecer padrões de desempenho considerado mínimos aceitáveis.

Mas não é só o QECR que parece ser reduzido a uma figura de retórica. A própria concepção de avaliação começa a ser uma fonte de justificação de linhas de acção política e de certas opções técnicas, mesmo que não se cuide dos mais elementares aspectos metodológicos, mormente psicométricos.

A um procedimento de avaliação diagnóstica não é indiferente um modelo subjacente de conhecimento e uso da língua, assim como da sua aprendizagem. A opção por um determinado modelo de avaliação aplicado à educação implica, ainda que implicitamente, a adesão a um modelo de ensino/aprendizagem, sob pena de uma invalidade sistemática dos resultados que o instrumento de avaliação devolve[1]. No caso presente, embora os documentos do ME pareçam apontar numa determinada perspectiva da Didáctica das Línguas, deixam margem de manobra suficiente para que instituições/professores com distintas práticas e formações possam aplicar livremente as suas competências com vista ao alcance dos fins traçados.

Resta o problema da economia de meios. Uma qualquer abordagem da didáctica das línguas requer um instrumento de avaliação coeso, contudo o desenvolvimento de instrumentos de avaliação em número e qualidade (validade, fiabilidade, exequibilidade, justeza e impacto) suficientes é de difícil concretização. Não é possível garantir uma qualidade comparável entre todos os instrumentos de avaliação e tão-pouco será económico demonstrar a equivalência entre eles (ao nível do constructo que as Orientações Programáticas propõem e da ligação ao QECR).

[1] Cf. Bachman (1990:242).

Delegar em especialistas externos a elaboração de instrumentos de avaliação é uma solução que permite contornar tal obstáculo. Diferentes instrumentos são construídos, tendo em mente os mesmos critérios de qualidade e os objectivos gerais que se pretende alcançar. No entanto, há que garantir, através de formação e treino, que a aplicação dos mesmos é feita de forma consistente ao longo do sistema, de forma a garantir, antes de mais, a fiabilidade dos resultados e, mais tarde, a validade dos usos dos mesmos.

É precário o equilíbrio entre a autonomia de quem ensina (e toma decisões com distintos graus de impacte sobre a vida dos alunos) e os objectivos impostos a nível da tutela. De facto, o problema de saber quem e em que circunstâncias avalia não se resume apenas a uma questão de economia de meios ou de validade em sentido estrito. É também uma questão política, na medida em que denota a forma como atribuições e competências são distribuídas a diferentes actores do sistema educativo. A responsabilidade de seriar, agrupar e educar implica a capacidade de avaliar (medir e decidir), ordenar o trabalho, executá-lo, julgá-lo e ser julgado pelos seus (do trabalho) resultados. Mas não basta uma capacidade *de jure*, é imperativo que se verifique uma capacidade *de facto*, sob pena de se criar um híbrido irracional: um sistema educativo que não impõe soluções únicas e padronizadas para todas as situações, mas que tão-pouco cuida dos efectivos ganhos de produtividade (a existirem) resultantes de uma 'delegação de poderes' feita sem garantias de capacidade material.

Não se pense que a opção é dicotómica e extremada. De um lado teríamos um modelo imposto, 'cego', feito em patamares superiores que, apenas por via administrativa, se propaga pela hierarquia. Do outro teríamos um modelo atomizado, em que cada escola, ou até mesmo professor isolado, usaria quaisquer meios e conhecimentos ao seu dispor para aplicar orientações tão latas que qualquer interpretação poderia delas ser

derivada. Se a primeira opção ignora a impossibilidade de, por via de instrumentos jurídico-administrativos, garantir a coesão de interpretação de referenciais (note-se o caso das contradições da definição de 'diagnóstico' entre os ciclos básico e secundário), a segunda despreza a necessidade de se garantir que um sistema justo tem de ter um grau mínimo de fiabilidade e comparabilidade dos resultados, mormente aqueles que são expressos através de sumários quantitativos – as notas. A primeira opção sugere que há decisões que são demasiado importantes para serem tomadas pelos professores, ignorando que a capacidade de utilizar instrumentos com proveito varia em grau de contexto para contexto. A segunda opção reduz a noção de validade ao juízo idiossincrático, e portanto indemonstrável, de um actor isolado.

Antes, é uma questão de saber qual o conteúdo do papel de cada um dos intervenientes no processo educativo e das relações que estabelecem entre si. Saber se se pretende uma tutela que simultaneamente põe objectivos, dirige, e avalia (chegando-se a um movimento pendular entre a micro-gestão e a quase ausência), ou uma tutela que estabelece princípios gerais de actuação e se limita a apoiar, onde, quando e como assim lhe seja requerido por outros intervenientes, ainda que cuidando sempre da exequibilidade das tarefas delegadas, da capacidade efectiva de as executar e da possibilidade de comparabilidade entre resultados.

O programa de PLNM é o candidato ideal a uma outra maneira de pensar as relações entre níveis superiores e inferiores na hierarquia do sistema de ensino português. A complexidade da elaboração de programas, ainda que prototípicos, de instrumentos de aferição e da investigação da sua validade científica transcende as competências das escolas e é feita com maior economia de meios a um nível superior. Contudo, esse trabalho só terá significado, só será realmente válido, quando se demonstrar que é

proveitoso para aqueles a quem se destina: os professores, enquanto agentes educativos, os alunos enquanto objectos da educação.

Neste debate a questão da avaliação não é indiferente, antes pelo contrário. A opção por um ou outro modelo implicará sempre a existência de procedimentos de avaliação que, antes de mais, mereçam a confiança pública. Mesmo que possa parecer contra-intuitivo, a confiança pública em instrumentos de avaliação não decorre necessariamente da qualidade técnica dos mesmos, antes daquilo que alguns autores apelidam como validade aparente (*face validity*)[2]. O perigo para o sistema educativo (e para todos quantos dele dependem, directa ou indirectamente) é que tal preocupação reduza a avaliação a uma espécie de recurso retórico, dificultando a decisores políticos, intervenientes no processo educativo e ao público em geral o cabal recorte do real impacto e alcance que os instrumentos de aferição e os usos que deles se faz efectivamente têm na vida de todos. O mesmo é válido para qualquer outra linha de acção política que se manifeste em orientações ou instrumentos educativos.

Assim, que modelo para a avaliação diagnóstica da competência comunicativa em Português Língua Não Materna no sistema educativo português? E que forma de o executar com garantias de eficácia, eficiência, impacto e justeza? Neste trabalho procurei contribuir para responder a estas questões. A um tempo olhando para a dimensão 'diagnóstica' da avaliação, a outro, para os necessários corolários a extrair das necessidades e das condicionantes materiais e humanas à sua aplicação.

A noção de 'avaliação diagnóstica', pelo que é possível inferir da revisão bibliográfica feita, transcende em boa medida a representação tradicional (e ainda muito comum) da avaliação como processo externo e distinto ao ensino/aprendizagem. O diagnóstico não é controlo de qualidade, ou o ablativo processo de dizer se, dado um

[2] Cf. Bachman (1990:285), Bachman e Palmer (1996:42), ALTE members (1998:145, 394)

determinado padrão de desempenho (ou resultado num instrumento de aferição em comparação com um resultado mínimo arbitrariamente estabelecido), alguém deve ou não ver-lhe atribuído um determinado direito. O diagnóstico, de facto, reduz (ou idealmente reduziria) a distância entre ensino e avaliação no tempo como no espaço. Na visão de Alderson (2005), Ribeiro e Ribeiro (1989) e Pascoal e Oliveira (2006), o diagnóstico é uma de várias fases de todo um processo de ensino: a fase de recolha de informação detalhada, significativa, que leva a uma reorientação, no sentido da optimização, de práticas docentes. No entanto tal abordagem implica que haja uma prática instituída, ou pelo menos um protótipo de comportamento, cuja eficácia e eficiência sejam conhecidas e sobre a qual se considere que se pode agir de forma diferenciada com ganho. Tal não é o caso presente: não existem programas, não existem padrões de desempenho, não existem materiais didácticos, não existe um corpo profissional experiente que possa assimilar novas informações e delas extrair corolários. De facto, nesta fase, não existe nada que se possa melhorar: apenas um caminho aberto para a experimentação, seja ela sistemática e guiada, ou avulsa. É de todo o interesse que, antes que se ensaiem medidas de diferenciação mais refinadas (por exemplo, em função da LM dos alunos), se instituam práticas comuns mínimas que permitam a recolha de dados comparáveis, fiáveis e válidos que todos possam usar em benefício das aprendizagens a potenciar.

Esta visão do que são o ensino e a avaliação distingue-se fortemente da concepção vigente, que vê caminhos pré-estabelecidos (os programas, os ciclos de estudos) como garantes suficientes para o alcance de determinados objectivos: um ensino para a cidadania, um ensino para a proficiência, a 'curva em J'. Pelo contrário, esta concepção, herdeira da Didáctica das Línguas (especialmente do ensino a adultos) reconhece a necessidade (ou a imposição) de se distinguir o que é diferente e de se agir em confor-

midade. No entanto, tal distinção implica que se verifiquem, pelo menos, três condições: 1) que haja uma escala de descrição comum e inequivocamente interpretada e partilhada por todos (no caso, pretende-se que o QECR sirva esse fim, não obstante a ausência de uma ligação real), 2) o uso de instrumentos de aferição fiáveis, válidos e com impacto, cujos resultados tenham significado fora da abstracção do uso em sala de aula e, não menos importante, 3) que os processos de decisão, na concepção e aplicação destes e de outros instrumentos, sejam transparentes, documentados e sindicáveis pelo público, este entendido numa acepção lata que vai do mero 'leigo' ao 'especialista'.

Não se pode iludir o facto de a validação da avaliação tanto ter uma dimensão retórica, e portanto iminentemente política e socialmente construída, como técnica e empírica[3]. Tal facto não implica uma posição extremada entre um dos aspectos, ou até mesmo o confortável refúgio do apelo a uma solução 'equilibrada' e de 'consenso'. Antes, implica que se reconheçam o âmbito e o alcance de cada uma destas dimensões e se actue de forma consequente. A dimensão retórica implica que o debate em torno de conteúdo, técnicas e padrões mínimos de desempenho tem de ser participado tanto pelos interessados, como por especialistas e utilizadores. Não obstante, tal debate não se pode reduzir à aparência dos instrumentos, ou à necessidade de se conseguir sucesso a qualquer custo, até iludindo o que as classificações realmente representam em termos de desenvolvimento de competência, desempenho comunicativo e escolar, integração. Um debate alargado é um debate tecnicamente bem informado. Um debate sobre avaliação produtivo é um debate que se ancila no que se consegue demonstrar com evidências empíricas e não apenas na declaração de vontades, ainda que aparentemente consensuais. Por isso é tão importante cuidar da fiabilidade e comparabilidade dos resultados das aferições, sejam elas de baixo ou elevado impacto.

[3] Cf. Messick (1994:8-9)

A criação do programa de PLNM foi um primeiro passo neste sentido, contudo mais trabalho precisa de ser feito.

Um primeiro passo seria a definição clara do que se entende por ser a componente comunicativa/linguística da integração. Não é justo que se permita a associação da noção de integração, quase por sinonímia, ao desenvolvimento da competência comunicativa em PLNM. Essa será, sem dúvida, uma dimensão essencial, mas a integração é tanto um processo bidireccional como é conformado por aspectos económicos, sociais, culturais, jurídicos e, bem assim, linguísticos. Definir o que é o papel do professor de PLNM para a integração é criar as condições para uma avaliação consequente da actuação de toda a escola para a integração dos alunos.

Numa perspectiva mais técnica, duas linhas de acção se impõem urgentemente. A primeira é a criação de padrões de desempenho claros, porventura calibrados em relação ao QECR, numa primeira fase, para as capacidades de expressão/interacção escrita e oral, a médio trecho para testes de colocação em nível e de competências. Apesar da ausência da definição de conteúdos para os níveis B2 e C1, também para estes níveis devem ser definidos padrões de desempenho, para que se possa eventualmente verificar o impacto de se excluir os alunos de um programa de PLNM assim que atingem o patamar mínimo de proficiência do nível B2. Estes padrões de desempenho deverão ser usados em estudos que cotejem a correlação entre o desenvolvimento da competência comunicativa e o desempenho académico mais geral, em função de variáveis como língua materna, ocupação profissional/grau de escolaridade dos pais, nível socioeconómico, área de estudo do aluno.

A segunda, a criação de programas e instrumentos de aferição claros, exaustivos e facilmente aplicáveis pelos professores. Sendo uma área nova no contexto do ensino público português, são à partida grandes os obstáculos com que os professores se depa-

ram na planificação, execução e monitorização do seu trabalho. A abordagem seguida impõe aos professores que organizem os seus próprios programas. Esta é uma abordagem inovadora cuja aplicação em outras áreas do currículo merece investigação. No entanto, se em disciplinas com uma maior tradição os professores podem usar outras referências que não apenas orientações programáticas na planificação do seu trabalho (o currículo implícito em manuais, a prática de anos – seja do próprio professor, seja de colegas –, a referência a padrões de desempenho reflectidos em instrumentos de aferição estabelecidos e bem conhecidos), no caso do PLNM não existe, para a maior parte dos profissionais, uma referência clara que lhes permita articular o que, para muitos, será uma novidade: metodologia, programa, instrumentos didácticos e de aferição.

Será de todo o interesse que numa fase inicial a aplicação do programa de PLNM se articule de muito perto com projectos de investigação universitária e de formação de docentes. Um programa de criação de níveis de desempenho seria uma hipótese a explorar para o treino de professores e para a sua familiarização com o QECR. Recorde-se que, embora as orientações ministeriais incitem os professores ao uso do QECR, este ainda é um documento não muito divulgado entre a maioria dos profissionais e cuja real utilidade, bem como o seu alcance, nem sempre é entendida da melhor forma.

De igual modo, a criação de um banco de itens, com a colaboração de especialistas da área do PLNM já familiarizados com o QECR e dos professores que os utilizarão, seria também uma linha de acção a explorar. Tal banco, com itens calibrados com o QECR, facilitaria a criação de materiais didácticos, assim como de testes, e a sedimentação de padrões de desempenho claros que dêem uma "semântica" aos resumos de competência comummente usados: as notas. O processo de familiarização com o QECR e o treino para a construção de itens seria também uma forma de simultaneamente

garantir a fiabilidade e comparabilidade das práticas dos professores, assim como partilhar experiências e dar um novo sentido prático à investigação que a academia produz.

Algumas palavras finais sobre a noção de avaliação diagnóstica. Seguindo o raciocínio de Alderson (2005), podemos propor que a avaliação diagnóstica é um processo contínuo, que se assemelha ao método de experimentação científica (colocar hipótese, testar hipótese, analisar os resultados e agir em conformidade). Reduz a distância entre avaliação e ensino e tem o potencial de tornar o professor num agente activo e reflexivo das suas práticas. Neste sentido, é possível que consuma em si o conceito de avaliação formativa, em vez de o integrar ou dele se distinguir. No modelo proposto por Alderson (2005) a avaliação diagnóstica não tem necessariamente de se ancilar em aferição diagnóstica, embora sugira que, a par do que acontece em outras áreas, os instrumentos de aferição diagnóstica deveriam basear-se em itens discretos, teorica e empiricamente relacionados com aspectos de elevada ordem do uso da língua com fins comunicativos. Claramente mais pesquisa é necessária, na definição de uma teoria da competência comunicativa em LNM e do seu desenvolvimento. Contudo, talvez fosse de considerar uma outra abordagem à questão da optimização do ensino, enquanto a linguística teórica não nos oferece modelos mais completos e satisfatórios do desenvolvimento da competência comunicativa numa língua estrangeira e a linguística aplicada outros instrumentos de ensino e aferição. Essa abordagem é que, mais do que o aspecto técnico da constituição de testes diagnósticos, o essencial é saber como os professores procuram, conseguem e manipulam informação nas suas tarefas do dia-a-dia. Em princípio, quanto mais lato for o reportório de estratégias de um professor, maior será a sua necessidade de informação e a sua capacidade de a integrar. O caminho poderia ser o inverso: em lugar de se investigar que item afere que aspecto de um sub-constructo da competência comunicativa, saber que tipo de informação um professor procura, como a utiliza e que

fontes, sobretudo aquelas que o professor já conhece e compreende, podem ser usadas com proveito na sua criação. Um teste diagnóstico, num sistema de ensino ideal, é um instrumento deliberadamente criado, com um fim específico, que é confirmar ou infirmar uma hipótese que orientará uma prática futura. Esse teste requer um professor consciente, activo, reflexivo e pronto a mudar. Esse professor requer as condições, materiais, jurídicas e humanas, para fazer aquilo em que ninguém o pode substituir com proveito: ajudar alguém a aprender.

Referências Bibliográficas

Abrantes, Paulo (2002) "Introdução: A avaliação das aprendizagens no ensino básico", *in* Abrantes, Paulo; Araújo, Filomena (coords.) (2002), *Reorganização Curricular do Ensino Básico,* Lisboa: Ministério da Educação.

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Alderson, J. C. (2005), *Diagnosing foreign language proficiency,* London: Continuum International Publishing Group.

Alderson, J.C.; Clapham, Caroline; Wall, Diane (1995) *Language Test Construction and Evaluation*, Cambridge: Cambridge University Press.

Alonso, Luísa (2002) "Integração Currículo-Avaliação: Que significados? Que constrangimentos? Que implicações?", *in* Abrantes, Paulo; Araújo, Filomena (coords.) (2002), *Reorganização Curricular do Ensino Básico,* Lisboa: Ministério da Educação.

ALTE members (1998), *Multilingual glossary of language testing terms*, Cambridge: Cambridge University Press.

Bachman, Lyle F. (1990), *Fundamental considerations in language testing,* 7.ª impressão, Oxford, Oxford University Press.

Bachman, Lyle F. (2004), *Statistical analyses for language assessment,* Cambridge: Cambridge University Press.

Bachman, Lyle F., Palmer, Adrian S. (1996), *Language testing in practice,* Oxford: Oxford University Press.

Buck, G. (2001). *Assessing Listening.* Cambridge: Cambridge University Press.

Casteleiro, J. M., Meira, A. & Pascoal, J. (1988). *Nível Limiar.* Lisboa: Instituto de Cultura e Língua Portuguesa.

Conselho da Europa (2001) *Common European Framework of Reference for languages: Learning, Teaching, Assessment*, Cambridge, Cambridge University Press, trad. port. Rosário, Maria e Soares, Nuno, (2001) Quadro Europeu Comum de Referência para as línguas – Aprendizagem, ensino, avaliação, Porto: Edições Asa.

Conselho da Europa (2003) *Relating language examinations to the Common European Framework of reference for Languages: Learning, Teaching, Assessment (CEF) - Manual Preliminary Pilot Version*, Estrasburgo: Conselho da Europa.

Conselho da Europa (2004) *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of reference for Languages: learning, teaching, assessment*, Estrasburgo: Conselho da Europa.

Dionízio, Sandra (coord.) (2005) Análise do Inquérito no Âmbito do Conhecimento da Situação Escolar dos Alunos cuja Língua Materna não é o Português – Relatório Final, Lisboa: IESE, retirado de: http://sitio.dgidc.min-wdu.pt/linguaportuguesa/Documents/portLNMRelatorioFinal.pdf

Fernandez, Sonsoles (2003) *Propuesta curricular y Marco Comum Europeo d Referencia. Desarrollo por Tareas*, Madrid: Edinumen.

Fulcher, Glenn (2004) "Are Europe's Tests being built on an 'unsafe' framework?" *in* Guardian Weekly, 18 de Março, retirado de: http://www.guardian.co.uk/education/2004/mar/18/tefl2

Kaftandjieva, Felianka (2004) "Standard Setting" *in* Conselho da Europa (2004) *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of reference for Languages: learning, teaching, assessment*, Estrasburgo: Conselho da Europa.

Leiria, Isabel; Queiroga, Maria João; Soares, Nuno Verdial (2006) "Perfis Linguísticos" *in* DGIDC-ME (2006), *Português Língua Não Materna no Currículo Nacional – Perfis Linguísticos – Teste Diagnóstico*, Lisboa: Direcção-Geral de Inovação e Desenvolvimento Curricular – Ministério da Educação.

Leiria, Isabel (coord.) (2008) *Orientações Programáticas de Português Língua Não materna (PLNM) Ensino Secundário, s/l,* Direcção-Geral de Inovação e Desenvolvimento Curricular – Ministério da Educação, retirado de: http://sitio.dgidc.min-edu.pt/linguaportuguesa/Documents/OrientProgramatPLNMVersaoFinalAbril08.pdf

Luoma, S. (2004). *Assessing Speaking.* Cambridge: Cambridge University Press.

McKay, P. (2006). *Assessing Young Learners.* Cambridge: Cambridge University Press.

Messick, S. (1989). "Validity", *in* Linn, R. L. (Ed.), *Educational Measurement,* New York, NY: Macmillan Publishing Company.

Messick, Samuel (1992). *The Interplay of Evidence and Consequences in the Validation of Performance Assessments. Research Report*, Princeton, NJ: Educational Testing Service.

Messick, Samuel (1994) *Standards-Based Score interpretation: Establishing Valid Grounds for Valid Inferences. Research*, Princeton, NJ: Educational Testing.

Nascimento, Maria Fernanda Bacelar (1984) *Português Fundamental*, 3 vols., Lisboa: Instituto Nacional de Investigação Científica.

North, Brian (2004) "Europe's Framework promotes language discussion, not directives" *in* Guardian Weekly, 15 de Abril, retirado de: http://www.guardian.co.uk/education/2004/apr/15/tefl6

Pascoal, José (1992) *Contribuição para um estudo docimológico aplicado ao português como língua estrangeira: o CELP e o DILP*. Dissertação de Mestrado em Linguística Portuguesa Descritiva apresentada à Faculdade de Letras da Universidade de Lisboa.

Pascoal, José; Oliveira, Teresa; (2006) "Teste Diagnóstico" *in* DGIDC-ME (2006), *Português Língua Não Materna no Currículo Nacional – Perfis Linguísticos – Teste Diagnóstico*, Lisboa: Direcção-Geral de Inovação e Desenvolvimento Curricular – Ministério da Educação.

Peralta, Maria Helena (2002) "Como avaliar competências? – Algumas considerações" *in* Abrantes, Paulo; Araújo, Filomena (coords.) (2002), *Reorganização Curricular do Ensino Básico,* Lisboa: Ministério da Educação.

Pinto, Jorge; Santos, Leonor (2006) *Modelos de Avaliação das Aprendizagens*, Lisboa: Universidade Aberta.

Purpura, J. E. (2004). *Assessing Grammar.* Cambridge: Cambridge University Press.

Read, J. (2000). *Assessing Vocabulary.* Cambridge: Cambridge University Press.

Ribeiro, António Carrilho; Ribeiro, Lucie Carrilho (1989) *Planificação e Avaliação do Ensino-Aprendizagem*, Lisboa: Universidade Aberta.

Roldão, Maria do Céu (2006), *Gestão do Currículo e Avaliação de Competências,* 4ª edição, Queluz de Baixo: Editorial Presença.

Santos, Leonor (2007), *Dilemas e desafios da avaliação reguladora,* retirado de http://www.educ.fc.ul.pt/docentes/msantos/avaliacao.htm no dia 4 de Janeiro de 2008.

Soares, António *et al.,* (2005) *Português Língua Não Materna no Currículo Nacional – Documento Orientador*, Lisboa: Direcção-Geral de Inovação e Desenvolvimento Curricular – Ministério da Educação

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Legislação e outros actos normativos

Despacho Normativo 30/2001, de 19 de Julho (Estabelece princípios e orientações a observar na avaliação das aprendizagens no Ensino Básico).

Decreto-Lei 74/2004, de 26 de Março (Estabelece os princípios orientadores na organização e gestão do currículo, bem como da avaliação das aprendizagens).

Despacho Normativo 1/2005, de 5 de Janeiro (Estabelece princípios e orientações a observar na avaliação das aprendizagens no Ensino Básico, revoga o Despacho Normativo 30/2001).

Despacho Normativo 30/2007, de 10 de Agosto (Estabelece os princípios orientadores para o ensino de PLNM no Ensino Secundário).

Anexo I

Princípios e procedimentos na avaliação das aprendizagens nos ensinos Básico e Secundário

DN 30/2001 (Básico)	DN 1/2005 (Básico)	DL 74/2004 (Secundário)
Art.º 17º — A avaliação formativa inclui uma vertente de diagnóstico tendo em vista a elaboração e adequação do projecto curricular de turma e conduzindo à adopção de estratégias de diferenciação pedagógica.	Art.º 18º — A avaliação diagnóstica conduz à adopção de estratégias de diferenciação pedagógica e contribui para elaborar, adequar e reformular o projecto curricular de turma, facilitando a integração escolar do aluno, apoiando a orientação escolar e vocacional. Pode ocorrer em qualquer momento do ano lectivo quando articulada com a avaliação formativa.	
Art.º 16º — A avaliação formativa é a principal modalidade de avaliação do ensino básico, assume carácter contínuo e sistemático e visa a regulação do ensino e da aprendizagem, recorrendo a uma variedade de instrumentos de recolha de informação, de acordo com a natureza das aprendizagens e dos contextos em que ocorrem. Art.º 18º — A avaliação formativa fornece ao professor, ao aluno, ao encarregado de educação e aos restantes intervenientes informação sobre o desenvolvimento das aprendizagens e competências, de modo a permitir rever e melhorar os processos de trabalho.	Art.º 19º — A avaliação formativa é a principal modalidade de avaliação do ensino básico, assume carácter contínuo e sistemático e visa a regulação do ensino e da aprendizagem, recorrendo a uma variedade de instrumentos de recolha de informação, de acordo com a natureza das aprendizagens e dos contextos em que ocorrem. Art.º 20º — A avaliação formativa fornece ao professor, ao aluno, ao encarregado de educação e aos restantes intervenientes informação sobre o desenvolvimento das aprendizagens e competências, de modo a permitir rever e melhorar os processos de trabalho.	Art.º 11º, Nº 2 — A avaliação formativa é contínua e sistemática e tem função diagnóstica, permitindo ao professor, ao aluno, ao encarregado de educação e a outras pessoas ou entidades legalmente autorizadas obter informação sobre o desenvolvimento das aprendizagens, com vista ao ajustamento de processos e estratégias.
Art.º 22º — A avaliação sumativa consiste na formulação de uma síntese das informações recolhidas sobre o desenvolvimento das aprendizagens e competências definidas para cada área curricular e disciplina, no quadro do projecto curricular de turma respectivo, dando uma atenção especial à evolução do conjunto dessas aprendizagens e competências.	Art.º 24º — A avaliação sumativa consiste na formulação de um juízo globalizante sobre o desenvolvimento das aprendizagens do aluno e das competências definidas para cada disciplina e área curricular.	Art.º 11º, Nº 3 — A avaliação sumativa consiste na formulação de um juízo globalizante, tem como objectivos a classificação e a certificação e inclui: a) A avaliação sumativa interna, da responsabilidade dos professores e dos órgãos de gestão pedagógica da escola; b) A avaliação sumativa externa, da responsabilidade dos competentes serviços centrais do Ministério da Educação, concretizada na realização de exames finais nacionais.