Rev. Fac. Educ., São Paulo
8(1): 39-69, 1982

A DOCIMOLOGIA EM PERSPECTIVA

Maria José Miranda (*)

INTRODUÇÃO

A avaliação escolar é uma questão que preocupa psicólogos e educadores, porque implica variáveis psicológicas e se encontra indissoluvelmente ligada ao ensino. O controlo da realização dos objectivos pedagógicos em que assenta a acção educativa passa necessariamente pela medição dos resultados alcançados. O estudo científico dos problemas psicopedagógicos da avaliação de conhecimentos escolares em situação de exame e de concurso, foi iniciado de maneira sistemática nos princípios dos anos 20 pelo eminente psicólogo francês Henri Piéron, geralmente considerado o fundador da Docimologia.

Docimologia significa "o estudo sistemático dos exames (atribuição de notas, variabilidade interindividual e intraindividual dos examinadores, factores subjectivos, etc.)" (Piéron, 1973, p. 126), e o termo foi proposto por Piéron a partir das palavras gregas relativas aos exames: *δοκιμή* e *δοκιμασία* (prova), *δοκιμάζω* (examinar), *δοκιμαστής* (examinador) e *δοκιμαστικός* (apto para examinar) (Piéron, 1969, p. 6, nota 1). As experiências docimológicas incidindo, nomeadamente, sobre a adequação das provas aos objectivos pedagógicos, os processos de classificação, a preparação dos examinadores para a tarefa de avaliação, e o recurso a métodos objectivos de apreciação dos conhecimentos, deveriam constituir a base de uma docimástica racional (Piéron, 1969, p. 157).

Os primeiros trabalhos docimológicos puseram em evidência a instabilidade das avaliações sob os pontos de vista das diferenças interindividuais e intraindividuais, da validade e da precisão. O desenvolvimento das investigações levou a preconizar medidas de atenuação das divergências de carácter mais ou menos sistemático verificadas na avaliação escolar, medidas essas que incluem as escolas de notas, concertação entre os examinadores sobre os critérios a ter em conta na apreciação das provas, moderação das classificações, e a utilização de provas estandardizadas de correcção objectiva.

Uma outra linha mais recente de trabalhos centra a questão docimológica

(*) Professora da Faculdade de Psicologia e de Ciências da Educação da Universidade de Lisboa.

no próprio comportamento de avaliação, isto é, no seu estudo sistemático em termos das variáveis em presença e sua interacção. As investigações que têm vindo a ser realizadas em França desde os finais da década de 60, no sentido da identificação dos factores de instabilidade e sua articulação num modelo explicativo global, ampliaram o campo da Docimologia e abriram novas perspectivas de aperfeiçoamento da avaliação escolar.

O presente trabalho procura ser uma revisão actualizada da problemática da Docimologia, pondo em relevo os desenvolvimentos recentes. Na primeira parte são examinadas questões da avaliação escolar em termos do seu enquadramento no sistema educativo e formas por que se efectua. No segundo ponto são referidas as principais conclusões dos primeiros trabalhos docimológicos e as novas perspectivas de investigação deles decorrentes. Na terceira parte analisam-se vários aspectos do aperfeiçoamento dos métodos e das técnicas de avaliação escolar. Finalmente, no último ponto, é feito um balanço de resultados de investigações recentes sobre mecanismos cujo isolamento experimental revelou constituirem factores intervenientes na avaliação de provas escolares.

1. *SIGNIFICADO, MODALIDADES E LINGUAGENS DA AVALIAÇÃO ESCOLAR*

A avaliação escolar é por vezes posta em causa por argumentos outros que os de ordem psicopedagógica, nomeadamente de natureza ideológica e em termos da sua simples supressão (Lobrot, 1968) (1). Mas mesmo se a avaliação constitui "um problema que não pode deixar de se pôr, como uma tarefa que não pode deixar de ser assumida pelos educadores" (Reuchlin, 1974, p. 209), a especificação da sua legitimidade e funções no quadro do sistema educativo é um ponto que não deve ser escamoteado.

Em sentido lato, avaliar significa julgar (um acontecimento, um indivíduo, um objecto) em referência a um critério. No caso específico da avaliação de alunos, e a partir sobretudo da década de 50, o critério define-se em termos de objectivos educacionais. Num estudo recente, Besse sistematiza os objectivos educacionais, enquanto fins da acção educativa, em três grupos: os objectivos gerais, de caráter axiológico; os objectivos específicos, formulados em termos comporta-

(1) A resposta de Reuchlin publicada no mesmo volume (Reuchlin, 1968), embora frontalmente dirigida ao artigo de Lobrot, constitui uma pertinente reflexão sobre a necessária distinção dos aspectos ideológicos da análise do problema.

mentais e que estão na base dos planos de estudo; os objectivos concretos, pontuais, visados pela situação de aprendizagem escolar e que supõem processos de verificação (Besse, 1977, p. 117). A função da avaliação consiste no controlo da realização destes objectivos, e subseqüente fornecimento de informação relevante para a planificação e desenvolvimento curriculares (Astin, Panos, 1971, p. 733).

Alguns autores (Reuchlin, 1974, pp. 210-213; Noizet, Caverni, 1978, pp. 13-15) sublinham a distinção entre os aspectos de natureza social e de natureza pedagógica dos objectivos educacionais e, por consequência, entre as funções social e pedagógica da avaliação. Os objectivos sociais dizem respeito às finalidades estabelecidas, a prazo, da acção educativa nas suas sucessivas fases e tendo em vista as condições criadas pelo desenvolvimento tecnológico e científico da sociedade em que se insere; a função social da avaliação consiste no controlo da quantidade e qualidade de formação, e da "rendibilidade" do investimento, dada a absorção, pela instituição escolar, de uma parcela substancial do erário público (2). Os objectivos pedagógicos referem-se a comportamentos terminais, isto é, que se esperam no termo de um determinado nível de formação (3); a função pedagógica da avaliação põe-se simultaneamente em relação ao aluno (informação sobre o seu rendimento escolar e subseqüentes factores de decisão intervenientes na sua orientação), ao professor (informação sobre a turma e sobre a produtividade e eficácia do seu ensino) e ao sistema escolar (informação sobre o funcionamento geral do aparelho escolar).

Os objectivos pedagógicos são especificados, ao nível dos programas de ensino, sob o ponto de vista do conteúdo. A sua definição operacional e especificação em termos de comportamentos observáveis, e inserção num modelo classificativo hierárquico de categorias de comportamentos, dá lugar a uma sistemática – a uma Taxonomia (ou Taxionomia). A taxonomia dos objectivos educacionais

(2) Reuchlin chama a atenção, neste ponto, para o facto de que se a Escola prescindisse do exercício da sua competência de avaliação, esta passaria a ser executada pelas entidades empregadoras a partir de critérios que, a par de eventualmente insatisfatórios, teriam não obstante directa repercussão na própria estrutura do ensino; e isto em consequência de a duração da formação não poder constituir, isoladamente, um índice de competência (Reuchlin, 1974, p. 210).

(3) Noizet e Caverni assinalam a este propósito a importância da distinção entre "comportamento terminal" e "conhecimentos adquiridos", que não são reversíveis. A distinção liga-se ainda à necessidade de análise da adequação dos programas de ensino ao nível de desenvolvimento psicológico das populações a que se destinam (Reuchlin, 1960).

mais freqüentemente citada, e que Landsheere considera ter servido de inspiração ou de modelo para a maior parte das que se lhe seguiram (Landsheere, Landsheere, 1977, pp. 78-79), é a de Bloom e colaboradores (Bloom, Engelhart, Furst, Hill, Krathwohl, 1956; Krathwohl, Bloom, Masia, 1964) e diz respeito aos domínios cognitivo e afectivo (4). Respeita o princípio estrutural da complexidade crescente, e as categorias hierarquizadas incluem comportamentos do mais simples para o mais complexo (domínio cognitivo) e do menos interiorizado para o mais interiorizado (domínio afectivo).

Um trabalho recente (Birzea, 1979), em que também são apresentados em anexo quadros comparativos de diversas taxonomias dos domínios sensorio-motor, cognitivo e afectivo (5), põe em evidência que os diversos autores não utilizam os mesmos critérios e modelos de operacionalização, e atribui às dificuldades metodológicas e técnicas do processo a sua ainda limitada utilização na elaboração dos programas de ensino, no desenvolvimento dos instrumentos de avaliação, e na escolha de métodos pedagógicos. Campos situa as categorias taxonómicas na análise de tarefas e sublinha a importância da definição funcional dos objectivos pedagógicos (Campos, 1974, pp. 75-81).

A preocupação pela operacionalização não é independente de um recrudescimento de interesse, a partir dos anos 60, pelas teorias da aprendizagem com vista ao desenvolvimento de modelos de aprendizagem. Nesta linha se situam os estudos de Gagné sobre as condições de aprendizagem e aquisição de conhecimentos (Gagné, 1970), o ensino programado (Guglielmi, 1970), e o ensino assistido por computador, de que diversas experiências vêm sendo realizadas na Europa como na América do Norte (Baker, 1971, pp. 231-232; Glaser, Nitko, 1971, p. 647 e p. 664; Bacher, 1973, pp. 81-82). A importância conferida ao ritmo de aquisição por parte dos alunos no modelo de aprendizagem escolar proposto por Carroll em 1963 levou a uma concepção de Educação fundada no *mastery learning* (ou *learning for mastery*) desenvolvida por Bloom e colaboradores (Bloom, Hastings, Madaus, 1971, pp. 43-56), que implica a verificação pontual das aquisições por etapas sucessivas do processo de aprendizagem, e subsequente adopção de medidas compensatórias diferenciadas (6).

(4) O domínio psicomotor, inicialmente previsto, não chegou a ser tratado por Bloom e colaboradores. A taxonomia deste domínio mais desenvolvida deve-se a outro autor (Harrow, 1972).

(5) Anexos I, II, III (Birzea, 1979, pp. 199-210). No aspecto da comparação sistemática tem interesse o estudo de Vandervelde e Vander Elst sobre as taxonomias de Bloom e de Guilford (Vandervelde, Vander Elst, 1979).

(6) Refira-se a propósito o projecto de "investigação-acção" iniciado no ano lectivo de 1978/79 e com a duração prevista de seis anos, sobre 150 crianças do ensino primário de três escolas de Genebra, entre os 9 e os 12 anos e frequentando os ní-

A operacionalização dos objectivos inclui à partida a possibilidade de avaliação. Alguns autores, entre os quais Besse, alertam para o perigo de serem eventualmente descurados no ensino aspectos que não se ligam imediatamente a comportamentos susceptíveis de medição (Besse, 1977, pp. 132-133). Para Landsheere esses comportamentos, no entanto, não traduzem necessariamente micro-objectivos (Landsheere, Landsheere, 1977, pp. 246-247).

A Escola, perspectivada do ângulo da sua intervenção sistemática, desempenha um tríplice papel de instrução, de formação e orientação, e a avaliação recorre a instrumentos de verificação, de diagnóstico e de prognóstico da aprendizagem (Bonboir, 1972, p. 32). O contexto em que se realiza traduz-se em modalidades de avaliação (contínua ou pontual, externa ou interna, exame ou concurso). As técnicas utilizadas são o exame tradicional (oral ou escrito) e o teste de conhecimentos. No quadro das modalidades de avaliação, merece referência especial a distinção entre avaliação formativa e somativa.

A avaliação contínua consiste na apreciação regular, na situação de turma e pelo professor que conhece os seus alunos, das produções escolares destes. Uma avaliação deste tipo, não sendo imune a efeitos de estereotipia e de *halo* (Bonboir, 1972, p. 20; Landsheere, 1974, pp. 33-36; Perrenoud, 1979, p. 37), permite uma regulação interactiva no decurso da aprendizagem (Allal, 1979, p. 142). A avaliação pontual define-se por um "veredicto final" (Bonboir, 1972, p. 56) que incide sobre uma produção relativa a um tema ou temas extraídos de uma gama de temas possíveis, e liga-se a uma regulação retroactiva. Em termos de finalidade trata-se, no primeiro caso, de um acompanhamento da própria aprendizagem (permitindo um registro de progressos, de retrocessos, de dificuldades, dentro ou fora do quadro de uma "pedagogia correctiva"), no segundo de um balanço de aquisições efectuadas num período mais ou menos longo. As duas formas de avaliação podem ser combinadas, verificando-se então uma contaminação recíproca: a avaliação pontual pode realizar-se em intermitência, e a avaliação contínua ser expressa pela acumulação de avaliações pontuais; ou, na avaliação final ser tomada em consideração a informação registrada ao longo do período de aprendizagem.

A diferenciação entre avaliação externa e interna reside em ser efectuada ou não pelo responsável da acção pedagógica. A avaliação pontual é predominan-

veis escolares da 3ª à 6ª classes. O projecto tem o nome de R-A-P-S-O-D-I-E: *Recherche-Action sur les Prérequis Scolaires, les Objectifs, la Différenciation et l'Individualisation de l'Enseignement*. (Allal, Cardinet, Perrenoud, 1979, pp. 68-108).

temente externa, a contínua predominantemente interna. A avaliação externa reúne, relativamente à interna, algumas condições mais favoráveis sob o ponto de vista da objectividade, como situação relativamente estandardizada (em termos de tarefa e critérios de correcção), e ausência de informação sobre os candidatos para além da produção sobre a qual incide o juízo de avaliação; contém, por outro lado, os riscos inerentes às diferenças interindividuais e intraindividuais geradoras de divergências de apreciação. No que respeita a finalidade, tem carácter selectivo, em termos de admissão-rejeição.

A situação de exame difere da de concurso fundamentalmente nos fins da avaliação: no primeiro caso pretende-se verificar um nível de aquisições em função de um padrão ou norma de referência, no segundo, seleccionar, a partir de uma tarefa, os mais capazes ou mais aptos de um grupo de candidatos, e a norma é extrínseca e pré-fixada (número de vagas). A distinção é menos nítida no plano prático; pense-se, por exemplo e no caso português actual, nos resultados do ano propedêutico como condição de acesso à Universidade: a concepção de exame é contaminada pela de concurso na ordenação final cumulativa dos candidatos, a qual determina o seu ingresso ou não ingresso; na fase dita de "repescagem" ocorre de novo contaminação, agora no sentido inverso.

A distinção entre exame tradicional e teste de conhecimentos reside basicamente nas diferenças quanto à métrica de construção de cada um dos tipos de provas. Sendo um teste uma medida objectiva e estandardizada de uma amostra de comportamentos (Anastasi, 1976, p. 23), um teste de conhecimentos supõe (a) a definição clara de um conteúdo programático e seus objectivos, em relação ao qual é feita uma amostragem de questões, (b) a satisfação de princípios psicométricos no que respeita a sensibilidade, a precisão de vários aspectos da validade do instrumento, (c) normas, (d) condições estandardizadas de aplicação relativamente a instruções, cronometragem, cotação. A utilização generalizada de testes de conhecimentos na avaliação escolar nalguns países (com normas para diferentes grupos de referência, análise das distribuições de sucessos nas diversas questões encontradas nesses mesmos grupos) tem-se revelado de grande utilidade sob o ponto de vista pedagógico e para a orientação, e simultaneamente tem permitido o aperfeiçoamento progressivo dos instrumentos (com a organização de "bancos de itens", por exemplo) e o esclarecimento sobre as suas vantagens e inconvenientes, permitindo uma planificação mais fundamentada das práticas de avaliação (Bacher, 1973, pp. 57-58).

A distinção entre avaliação formativa e somativa deve-se a Scriven que, num artigo publicado em 1967, propôs os termos para designar, respectivamente, a avaliação no decurso do desenvolvimento da unidade didáctica, e a avaliação vi-

sando a apreciação de um produto terminal (Bloom, Hastings, Madaus, 1971, p. 262). A diferença entre as duas formas de avaliação é independente da extensão da unidade didáctica, e respeita tão só o alcance da avaliação: se tem por fim determinar o grau de competência, isto é, até que ponto o aluno domina os vários elementos de uma ordenação hierárquica e subsequente identificação dos pontos fracos, de molde a permitir tomada de decisões sobre o prosseguimento do ensino, é formativa (ID., p. 28, p. 61); se tem por fim uma apreciação de progressos, classificar os alunos uns relativamente aos outros ou determinar a realização de objectivos de uma unidade didáctica, é somativa (ID., p. 61, p. 117). A segunda tem efeitos diferenciadores, a primeira efeitos de homogeneização (Noizet, Caverni, 1978, p. 20) (7). Enquanto a avaliação somativa se liga a um ensino indiferenciado e a uma distribuição das aquisições isomorfas da distribuição das aptidões, a avaliação formativa liga-se a um ensino diferenciado (ou individualizado) (Cardinet, 1977, pp. 13-19; Allal, 1979) e a uma configuração em "j" da curva das aquisições (Landsheere, 1974, pp. 182 e 186-189).

A avaliação escolar, exprime-se por notas (apreciação sintética) e por resultados (cotação por soma ou subtracção de pontos, segundo regras fixas) (Landsheere, 1974, p. 16). Trata-se em ambos os casos de utilização de linguagens específicas, devidamente codificadas, tendo a avaliação, não obstante, um carácter indirecto: incide sobre produtos observáveis, que constituem sinais (Bonboir, 1972, p. 11). O sistema de resultados corresponde a uma preocupação de cotação mais objectiva das provas escolares, normalizadas ou não, sendo a sua validade preditiva superior à das notas (Bacher, 1965). O sistema de notas recorre, na prática, a escalas de medida, nomeadamente de intervalos e ordinais. As notas numéricas (0-20, 1-5) ordenam-se em escalas de intervalos, enquanto as categorias (A-B-C) e as apreciações qualitativas, em escalas ordinais. O índice de tendência central que resume a informação é, no primeiro caso, a média, no segundo a mediana. A amplitude de cada um dos valores da escala, e por conseqüência o grau de discriminação que permite, depende do número total de escalões. Estudos comparativos incidindo sobre a utilização simultânea de escalas numéricas e não numéricas apontam no sentido da superioridade prática das primeiras (Bacher, 1973, pp. 36-37; Reuchlin, 1974, pp. 216-217).

(7) A avaliação formativa, uma linha de estudos muito recentes levada a cabo na Suíça de língua francesa e relativamente ao tronco comum da escolaridade obrigatória (Cardinet, 1978; Allal, Cardinet, Perrenoud, 1979), pode ainda contribuir para a democratização do ensino no sentido de igualdade de oportunidades, na medida em que permite a adopção de medidas compensatórias das diferenças individuais intervenientes no rendimento e adaptação escolares.

2. *DOS PRIMEIROS TRABALHOS DOCIMOLÓGICOS ÀS NOVAS PERSPECTIVAS DE INVESTIGAÇÃO*

A investigação docimológica teve como ponto de partida (8) um estudo realizado em França por H. Piéron, Mme. Piéron, e H. Laugier sobre o certificado da instrução primária em 1922, cujo objectivo principal consistia em verificar o valor dos testes utilizados em orientação. Uma bateria de seis testes de aptidões foi aplicada a 117 alunos no termo da escolaridade primária e repartidos por três escolas, e os resultados comparados com as classificações escolares (ao longo do ano lectivo e exame final), agrupadas as disciplinas em três grupos, conforme apelando predominantemente para aquisições mnemónicas (história, geografia), aspectos intelectuais (redacção, aritmética), e qualidades extra-intelectuais (desenho, ginástica, canto). As discrepâncias entre os resultados dos testes e as notas escolares levaram os autores a pôr em causa o valor do exame como índice de aptidão escolar, e o seu estatuto eliminatório decisivo. Cinco anos depois, H. Laugier e D. Weinberg realizaram um estudo sobre a estabilidade das notas atribuídas por dois correctores a 166 provas escritas de concurso de acesso à Escola Normal Superior (História e Geografia); as divergências encontradas entre as classificações obtidas por dupla correcção diziam respeito às notas atribuídas à mesma prova, à média das notas de cada corrector, e à ordenação das provas pelos dois correctores. Outros estudos realizados entre 1927 e 1930 e fazendo intervir, a par da múltipla correcção, outras variáveis – Como experiência dos correctores, repetição da correcção em ocasiões diversas – confirmaram as verificações iniciais. Estes primeiros trabalhos foram reunidos num fascículo do *Travail Humain*, intitulado "Etudes docimologiques sur le perfectionnement des examens et des concours" e publicado em 1935.

Em 1931 a *Carnegie Corporation* pôs à disposição do Instituto Internacional de Educação do *Teacher's College* da Universidade de Columbia (Nova Iorque) os fundos necessários à realização de um inquérito internacional subordinado ao tema "As concepções, os métodos, a técnica e o alcance pedagógico e social dos exames e concursos". A primeira conferência teve lugar em Eastbourne em 1931, reunindo diversas comissões nacionais (americana, inglesa, escocesa, francesa, alemã, finlandesa, suíça) e foi seguida de duas outras, em Folkstone (1935) e Dinard (1938), tendo nesta última sido designada uma comissão permanente para o estudo das questões levantadas. A guerra européia suspendeu o de-

(8) Assinale-se, no entanto, a existência de outros trabalhos, menos sistemáticos, realizados fora da França, nomeadamente na Grã-Bretanha, Suíça, E.U.A., sobre correcção múltipla de provas escolares e elaboração de provas objectivas de conhecimentos (Hotyat, 1962; Piéron, 1969).

senvolvimento dos trabalhos. Entre as publicações mais importantes decorrentes do empreendimento contam-se *The Examination of Examinations* (Londres, 1935) e *La correction des épreuves écrites dans les examens. Enquête expérimentale sur le baccalauréat* (Paris, 1936).

A Docimologia surgiu pois como uma crítica aos métodos tradicionais utilizados, com fins de selecção, nos exames e nos concursos (9). Esta perspectiva da Docimologia, que se pode designar por "negativa" ou "clássica", situa-se no plano da verificação e análise das divergências de avaliação, em situação natural ou provocada. Essas divergências, sobre as quais se encontram publicadas exposições detalhadas (Vernon, 1957; Hotyat, 1962; Piéron, 1969) ocorrem, entre outros casos, relativamente às percentagens de admissões verificadas em diferentes júris de concurso e média e dispersão das notas atribuídas por esses mesmos júris (Noizet, 1961; Piéron, Reuchlin, Bacher, 1962; Piéron, Reuchlin, Bacher, Démangeon, 1962); às notas do mesmo aluno em diferentes disciplinas e na mesma disciplina na escrita e na oral (Piéron, Reuchlin, Bacher, 1962), às notas dadas à mesma prova pelo mesmo examinador em ocasiões diferidas (Wiseman, 1949; Finlayson, 1951; Nisbet, 1955; Bacher, 1965) e à classificação de uma prova por diferentes correctores (Nisbet, 1955; Penfold, 1956) (10); à escala de notas utilizadas pelos vários examinadores e peso real de cada nota na ponderação global em função da dispersão (Guerbet-Sceaux, Reuchlin, 1958); às apreciações numa escala verbal de natureza qualitativa (Démangeon, Larcebeau, 1958; Rémondino, 1965).

Do plano da verificação das divergências, que constituem um facto adquirido e sobre o qual a evidência acumulada é considerável, emerge a perspectiva positiva. A docimologia positiva contém duas grandes linhas fundamentais: por um lado, a que diz respeito ao aperfeiçoamento da avaliação; por outro, (Noizet, Bonniol, 1969), a "docimologia experimental".

O aperfeiçoamento da avaliação inclui o desenvolvimento das técnicas de construção dos instrumentos de avaliação – provas de exame e provas normalizadas de conhecimentos, análise das suas qualidades métricas, estudo da adequabi-

(9) Piéron chama a atenção para a importância acordada a nível oficial em França já na década de 40, à investigação docimológica: uma circular do Ministro da Educação em 1946 sobre o *baccalauréat* refere explicitamente "a desigualdade das provas" e "a heterogeneidade dos júris" evidenciados pela análise estatística dos resultados e que conferem "por vezes ao exame [características] de lotaria" e recomenda que os examinadores sejam "professores experimentados" (Piéron, 1969, p. 39).

(10) Neste tipo de análise o índice estatístico utilizado é o coeficiente de correlação.

lidade de um e outro tipo de provas a situações específicas, e métodos de moderação das classificações (Bacher, 1969, 1973; Bonboir, 1972; Reuchlin, 1974; Landsheere, 1974).

A "docimologia experimental" centra-se na avaliação como um comportamento, no sentidc de resposta global a uma situação, com componentes perceptivos e cognitivos (Noizet, Bonniol, 1969; Bonniol, 1974; Amigues, Bonniol, Caverni, Fabre, Noizet, 1975; Noizet, Caverni, 1978, pp. 63-146). Procura determinar experimentalmente os mecanismos intervenientes na decisão avaliativa e factores de distorsão desses mesmos mecanismos. A docimologia experimental é, assim, uma "docimonomia" (Noizet, Bonniol, 1969, p. 787).

Sob o ponto de vista metodológico, a investigação em Docimologia utiliza o método experimental, em sentido amplo e em sentido estrito. O estudo do comportamento de avaliação pode pôr o ênfase na análise das diferenças individuais (inter- e intraindividuais) entre os examinadores (perspectiva diferencial) ou na análise da tarefa (perspectiva experimental em sentido estrito).

O problema das divergências de avaliação começou por ser formulado em termos de variações aleatórias: as medições pontuais são erros de estimação, flutuações em torno do valor "verdadeiro", constituído pela média de um número elevado de classificações independentes (11). A análise estatística evidenciou, no entanto, que a grandeza das diferenças entre as médias em correcção múltipla, por exemplo, leva claramente à rejeição da hipótese nula quanto a essas diferenças. O paradigma das variações aleatórias deu lugar ao de variações sistemáticas (subavaliações, sobreavaliações etc), cuja verificação implica o estudo dos critérios utilizados na avaliação e/ou uma tipologia de avaliadores (elaborada a partir de escalas de severidade-indulgência, flutuação-constância, julgamento analítico-sintético), já avançada por Piéron. A análise das variações interindividuais ou intraindividuais decorre, porém, do conhecimento prévio do mecanismo de variação, sem o qual há perda de informação: a recorrência a planos de experiência do método experimental em sentido estrito surge assim como um corolário da própria metodologia experimental, visando a formulação, se não imediatamente de um modelo explicativo global, pelo menos de um quadro conceptual de referência contemplando a interacção entre as características pessoais ou profissionais do avaliador e as componentes da tarefa de avaliação (Bonniol, 1974, p. 202; Noizet, Caverni, 1978, pp. 64-66).

(11) No quadro do inquérito francês sobre o *baccalauréat* chegou-se mesmo a proceder ao cálculo do seu número para cada disciplina, fixando o coeficiente de precisão no valor $r = 0{,}99$ e utilizando a fórmula de profecia de Spearman-Brown (Piéron, 1969, p. 23).

3. *O APERFEIÇOAMENTO DOS PROCESSOS DE AVALIAÇÃO ESCOLAR*

Os trabalhos docimológicos realizados ao longo dos últimos cinquenta anos puseram em evidência as divergências de avaliação mediante comparações sistemáticas de, por um lado, classificações dadas em exames reais, por outro, de avaliações atribuídas em situações especialmente provocadas. Da convergência, elevada, das conclusões desses trabalhos, resultaram necessariamente estratégias conducentes à redução das divergências. Essas estratégias dizem respeito às práticas de avaliação efectuada por meio de provas de tipo tradicional, e ao desenvolvimento de outras técnicas de medição das aquisições escolares.

Bacher situa os erros de avaliação em relação a três grandes ordens de factores, que dizem respeito ao avaliador, às provas e aos alunos (Bacher, 1969). O avaliador aprecia as produções escolares sob uma determinada óptica e utiliza na avaliação uma certa margem de notas; acontece, assim, que as distribuições de notas relativas a um mesmo conjunto de provas diferem quanto à média e à dispersão, e até mesmo a ordenação das provas relativamente umas às outras, em termos de mérito relativo e a partir das notas dos diferentes examinadores, não é coincidente. No que se refere ao conteúdo das provas, o carácter aleatório da escolha dos temas propostos constitui uma base insuficiente de generalização dos conhecimentos adquiridos ao longo de um período de aprendizagem. Finalmente, relativamente ao aluno, flutuações ocasionais determinam variações de desempenho que são tomadas como indicadores de rendimento. Bonboir introduz ainda outra fonte de erro, constituída pelo aspecto dinâmico das mudanças, estáveis ou temporárias, que ocorrem no decurso de um período de tempo e cuja repercussão se estende ao avaliador, à matéria e ao aluno (Bonboir, 1972, p. 118).

Vários tipos de métodos no sentido de providenciar, por um lado, a redução das variações interindividuais (e intraindividuais) de apreciação, por outro de aumentar a validade e precisão das avaliações, têm resultado da investigação docimológica. Esses métodos empíricos de moderação das divergências visam tornar comparáveis as classificações dos exames, isto é, fazer com que elas exprimam as diferenças individuais entre os alunos mas não também entre os examinadores (Noizet, Caverni, 1978, p. 47). O seu contributo em termos de prossecução de uma maior objectividade não é idêntico, o que leva autores como Reuchlin e Bacher a falar de "soluções aparentes", ou "paliativos", e de processos mais objectivos.

Cabem no primeiro grupo tentativas do tipo substituição de escalas de notas com elevado número de escalões por outras com menor número (classificação de 1 a 5 em vez de 0 a 20, por exemplo) e eventual utilização de escalas ordinais

de tipo qualitativo em vez de escalas numéricas de intervalos. A redução do número de escalões, se diminui a possibilidade de dois correctores classificarem diferentemente a mesma prova, confere ao erro de avaliação um maior peso. Reuchlin considera que, por ora e no estado actual de desenvolvimento da investigação, se afigura mais prudente utilizar escalas com número elevado de escalões (e categorias por consequência pouco amplas), do que escalas com menor número de escalões e categorias amplas (Reuchlin, 1974, p. 216). Um estudo de Kaufmann (1975), utilizando o método psicofísico de medição das sensações e a noção de limiar diferencial sobre dados de Laugier e Weinberg recolhidos no quadro do inquérito sobre o *baccalauréat* (1936), sugere o uso de escalas com número de categorias diverso em função da disciplina (maior número na redacção do que na matemática, no caso desses dados). A substituição de notas numéricas por apreciações qualitativas, à luz da experiência docimológica actual, não só não diminui como antes agrava o desacordo entre correctores, e ainda o significado das categorias qualitativas não é unívoco (Reuchlin, 1974, p. 217).

Outro tipo de medidas no sentido de atenuar as divergências de avaliação consiste na intervenção na composição dos júris de exames e da fixação de parâmetros comuns de apreciação.

A intervenção ao nível da composição dos júris significa uma equilibração dos mesmos em termos de severidade-indulgência, por exemplo. A solução é mais aparente do que real, na medida em que os alunos são desigualmente afectados nas diversas disciplinas (Bacher, 1973, p. 39).

A fixação de parâmetros de apreciação pode assumir a forma de estabelecimento de "bitolas" por disciplina, através de discussão e concertação prévia entre os examinadores sobre a importância relativa dos vários aspectos a ter em conta na avaliação, atendendo a que aquela não é um dado da situação. Uma experiência de Bonniol, Caverni e Noizet utilizando grupo controlo e experimental e introduzindo a variável concertação neste, revelou um acordo significativamente maior, quanto às classificações como quanto à ordenação das provas, no grupo experimental (Noizet, Caverni, 1978, pp. 55-57). Um estudo de análise factorial sobre correcção de provas de redacção (Rémondino, 1959) evidenciou que as dimensões a ter em conta na avaliação, estabelecidas por concertação, se agrupavam em quatro categorias distintas (apresentação gráfica, aspectos linguísticos, qualidades de expressão, características pessoais), não sendo consideradas isoladamente nem tendo a mesma importância relativa para os diversos examinadores. A bitola pode ainda consistir num conjunto de regras extrínsecas de cotação fornecidas aos examinadores, sobre a decomposição das notas em sub-notas por questão, ou listagem de ocorrências sujeitas a penalização ou bonificação (Noizet, Caverni, 1978, pp. 53-54).

Os processos até aqui enunciados (escalas, constituição dos júris, bitolas de classificação) precedem imediatamente a avaliação. Outros intervêm noutros momentos, como o ajustamento das distribuições e a multiplicação do número de apreciações relativas a cada aluno.

O ajustamento das distribuições é susceptível de ser efectuado mediante informação aos diversos examinadores de uma mesma disciplina sobre a distribuição global de notas nessa disciplina, atribuídas pelo conjunto dos examinadores, de molde a permitir a aproximação das médias das notas dos vários examinadores à média da distribuição global. O ajustamento consiste na modificação das notas sem alteração das posições de ordem (12). A rectificação das classificações é uma subida ou descida de pontos, a partir das diferenças entre a média geral das notas e a média da distribuição a corrigir (Bacher, 1973, p. 38; Noizet, Caverni, 1978, p. 47).

A multiplicação do número de apreciações relativas a um mesmo aluno é um processo utilizado principalmente na Grã-Bretanha. Comparativamente com os outros processos corresponde a um aumento considerável de maleabilidade por parte do aluno e por parte do examinador (Bacher, 1973, p. 63). A multiplicação do número de apreciações diz respeito aos temas propostos e à multicorrecção. A produção do aluno incide sobre temas por ele seleccionados, entre os vários propostos. Finlayson (1951) mostrou que os resultados dos mesmos alunos em temas diferentes são menos concordantes entre si do que as apreciações dos examinadores sobre as suas produções num mesmo tema (coeficiente de correlação superior no segundo caso). Por outro lado, a produção sobre mais do que um tema e sobre temas escolhidos, permite fundamentar a avaliação sobre um espectro mais amplo e consequentemente mais representativo. A multicorrecção corrige enviesamentos das notas atribuídas por um mesmo examinador, pela média das notas dadas pelos diferentes examinadores à mesma prova; segundo Wiseman (1949) evita a concertação prévia e torna desnecessária a apreciação analítica: a rectificação incide sobre a apreciação global.

Os processos de moderação não se limitam a intervenções pontuais, e foram desenvolvidos e são praticados em países como a Suécia, Grã-Bretanha e

(12) No ajustamento das distribuições de notas pode ser tomada em linha de conta não só a tendência central, mas também a dispersão das distribuições; neste caso o procedimento passa pela redução das notas, isto é, a transformação das distribuições empíricas em distribuições de notas z. Este método, que mereceu a atenção de autores como Noizet, não é, porém, prática corrente (Noizet, Caverni, 1978, p. 50 e p. 104).

Suíça sistemas de moderação mais completos e sofisticados. A implementação dos sistemas supõe o estabelecimento prévio de acordo, em grande linhas, sobre os objectivos pedagógicos e sobre os programas (Bacher, 1973, p. 73).

Landsheere descreve com algum pormenor os sistemas sueco (de que o suíço é uma adaptação) e inglês (Landsheere, 1974, pp. 153-175). O sistema sueco utiliza uma prova normalizada de conhecimentos por disciplina aplicada a todos os alunos, aferida sobre uma amostra nacional ou regional; a distribuição de notas na disciplina é ajustada à distribuição dos resultados dos mesmos alunos na prova normalizada, sendo as notas nos dois tipos de provas expressas na mesma escala (1 a 7). Mais completo na opinião de Landsheere, o sistema inglês visa uma maior objectividade e comparabilidade das notas dos exames internos, e inclui a colaboração inter-escolas (assegurada pela designação, em cada uma, de um moderador encarregado de coordenar os exames na sua escola e participar numa comissão inter-escolas), na preparação e cotação das provas, no ajustamento das classificações e na atribuição da informação final. O sistema inglês introduz ainda testes de inteligência e de aptidões. O relatório de Vernon sobre a admissão no ensino secundário constitui uma análise exaustiva das diferentes práticas e métodos utilizados na moderação, suas vantagens e inconvenientes sob os pontos de vista psicológico e pedagógico, e sua inserção no sistema escolar a nível regional e nacional (Vernon, 1957).

Na prossecução de uma maior objectividade na avaliação escolar, ocupam um lugar de relevo aos testes de conhecimentos, ou "provas normalizadas de conhecimentos" (Bacher, 1973, pp. 40-41; Reuchlin, 1974, pp. 221). Um teste de conhecimentos é um instrumento de medida das aquisições efectuadas em situação estruturada de aprendizagem (Brown, 1971, p. 95), isto é, dos efeitos de um programa específico de instrução ou de treino (Anastasi, 1976, p. 398). Os testes de conhecimentos escolares referem-se a conteúdos programáticos específicos e diferenciam-se formalmente dos testes de inteligência e de aptidões não só quanto ao contexto da sua utilização como quanto à natureza da experiência anterior sobre que incidem (Anastasi, 1976, pp. 398-399). A sua construção parte da análise dos programas e dos objectivos do ensino, em termos de sistematização das unidades de informação e modalidades do seu processamento. A técnica mais frequentemente recomendada nos manuais recentes consiste na elaboração de um diagrama de dupla entrada (*blue print*) organizado segundo duas dimensões, os conteúdos e os processos, e subsequente selecção das questões a incluir – que cabem nas diversas casas do diagrama bidimensional – em função do peso relativo atribuído a cada rubrica do programa e a cada objectivo (Lewis, 1974, p. 62). Estudos realizados por diferentes especialistas sobre a metodologia de construção de testes de conhecimentos escolares e selecção das questões, a partir de diagramas utilizando na dimensão objectivos a taxonomia de Bloom, foram incluídas

numa obra deste autor; esses estudos inserem os testes de conhecimentos escolares no quadro da avaliação formativa como somativa, sobre matérias diferenciadas quanto ao conteúdo (das ciências da natureza à arte) e nível de ensino (da escolaridade pré-primária ao ensino profissional) (Bloom, Hastings, Madaus, 1971, pp. 283-905).

As questões que compõem os testes de conhecimento são geralmente formuladas sob a forma de itens de resposta por escolha múltipla. A resposta, no caso mais simples, consiste na escolha da alternativa correcta de entre as várias apresentadas, constituindo-se as alternativas incorrectas "distractores"; a dificuldade do item depende da natureza da questão como da natureza dos distractores, induzindo em maior ou menor grau ao erro (Noizet, Caverni, 1978, pp. 156-157). A sua resolução implica predominantemente uma actividade de reconhecimento por parte do aluno, e não de evocação de conhecimentos adquiridos, o que é geralmente apontado como um inconveniente. Formas mais elaboradas da apresentação das alternativas têm sido desenvolvidas, como, por exemplo, a selecção da resposta correcta implicar a realização de uma operação por parte do aluno (combinação, classificação, descoberta de uma relação), ou ainda a tarefa constituir na detecção e correcção de erros (Bacher, 1973, pp. 47-48). Um estudo de Bujas propõe a apresentação de alternativas cuja distância relativamente à resposta correcta é variável, sendo a sinalização de qualquer delas ou de nenhuma cotada por ponderação ou por penalização; a validade do instrumento é, segundo o autor, melhorada (Bujas, 1965). O aspecto comum a estas formas de apresentação da questão consiste em a resposta correcta estar antecipadamente fixada, donde a actividade do aluno ser de avaliação (escolha de uma de várias possibilidades) e não de produção, que implica elaboração original da resposta (Noizet, Caverni, 1978, p. 150).

Têm sido também realizadas, e com êxito, experiências no sentido de construir testes de conhecimentos escolares em que as questões são abertas, isto é, a resposta é livre; mais próximo da situação pedagógica, verificou-se que o acordo obtido em múltipla correcção é mais elevado do que nas provas tradicionais (Cambon, 1961; Bacher, 1973, pp. 48-49).

Relativamente às provas tradicionais, os testes de conhecimentos respondem melhor a exigências de objectividade de correcção e, embora a escolha das questões dependa do construtor do teste, a sua variedade permite uma amostragem mais satisfatória sob os pontos de vista do conteúdo e dos objectivos (13). A

(13) Landsheere apresenta, a título exemplificativo, o estudo comparativo de uma mesma questão formulada sob a forma tradicional e sob a forma de resposta por escolha múltipla. No primeiro caso estabelece bitolas de correcção para a atribuição da

organização de "bancos" de itens estandardizados pode conduzir a que as provas normalizadas de conhecimentos sejam de quando em quando renovadas (Bacher, 1973, p. 55; Landsheere, 1974, p. 157). A forma das questões permite ainda o emprego de processos de correcção automatizados (Baker, 1971).

O facto de os testes não incluírem aspectos ligados às qualidades de expressão levou Vernon a sugerir a sua utilização principalmente para a apreciação de conhecimentos factuais, reservando as provas de tipo tradicional para a apreciação de aspectos de expressão original (Vernon, 1965, p. 209). A mesma opinião é expressa por Coffman, que aponta, no entanto, a necessidade de uma formulação cuidada das questões e desenvolvimento de processos de cotação das provas tradicionais de molde a reduzir ao mínimo as variações, sistemáticas como aleatórias, das classificações (Coffman, 1971, pp. 285-296). Este tipo de considerações vai também ao encontro do problema de que o uso exclusivo dos testes de conhecimentos na avaliação escolar poderia, eventualmente, conduzir a negligenciar objectivos importantes do ensino, por parte do professor e por parte do aluno, tais como o significado global da matéria, a elaboração dos conhecimentos etc.

As características técnicas mais importantes dos testes de conhecimentos escolares, enquanto instrumentos de medição, são a validade, a precisão e a sensibilidade. A sensibilidade depende do universo programático de que as questões do teste constituem uma amostra; assim, ele pode cobrir totalmente um programa ou partes mais ou menos amplas, permitindo neste caso um controlo mais analítico dos conhecimentos adquiridos (Bacher, 1973, p. 51). A consistência interna do teste liga-se à natureza e número de itens; um teste de conhecimentos escolares é necessariamente heterogéneo (pela diversidade dos objectivos e conteúdos que estão na base da sua construção) e o método mais apropriado de análise da precisão é a utilização de formas paralelas (Stanley, 1971, pp. 404-405; Bacher, 1973, p. 51; Landsheere, 1974, p. 140). Entre os processos de aumentar a precisão contam-se a clarificação dos itens e das instruções, maior número de itens, e mais elevado índice discriminativo de cada um (Bloom, Hastings, Madaus, 1971, p. 81). No que respeita à validade do teste, importa considerar a validade de conteúdo (que remete para a especificação das rubricas do programa a incluir no diagrama bidimensional), a validade de construção (que se liga à medição dos comportamentos enquanto indicadores dos objectivos) e validade preditiva (confirmação subsequente das predições efectuadas) (Ferreira Marques,

nota, e no segundo indica diferentes modalidades consoante o grau de conhecimentos requeridos (Landsheere, 1974, anexo I, pp. 199-204).

1971; Cronbach, 1971, p. 446; Tourneur, 1974, pp. 41-43; Messick, 1975). Bacher assinala que a validade preditiva dos testes de conhecimentos escolares põe um problema, em virtude de as diferenças individuais relativamente ao critério dependerem também de aspectos não cognitivos (Bacher, 1973, p. 52).

Um outro aspecto métrico igualmente relevante é o da configuração da curva dos resultados. A distribuição normal dos resultados dos alunos em testes escolares tem sido contestada por alguns autores (Landsheere, 1974, pp. 179--183) e outras formas foram sugeridas nomeadamente a curva tipo III de Pearson (Gardner, 1952, p. 273; Angoff, 1971, p. 521). Bacher chama no entanto a atenção para o facto de que uma situação estruturada de aprendizagem não conduz necessariamente a um sucesso global, e que as diferenças individuais subsistem se bem que a diferenciação na base do programa propriamente dito seja atenuada (Bacher, 1973, p. 50).

Reuchlin aponta como principais críticas dirigidas aos testes de conhecimentos enquanto instrumentos de avaliação escolar as que se referem à fragmentação de conhecimentos em unidades elementares, à facilitação da tarefa por familiarização anterior com o tipo de questões, e à intervenção do acaso na escolha da resposta correcta (Reuchlin, 1974, pp. 222-223). No que diz respeito ao conteúdo das questões, os estudos realizados nos últimos anos e em diferentes níveis de escolaridade demonstraram já que os testes não têm que se cingir a conhecimentos pontuais e podem apelar para operações complexas (Bacher, 1973, p. 57). Quanto à facilitação por familiarização com a tarefa, ela pode ser reduzida por aplicação prévia de uma prova neutra visando um melhor nivelamento dessa mesma familiarização (Reuchlin, 1974, p. 222); ainda, o tipo de questões não deverá ser utilizado no ensino regular sob o risco, apontado entre outros autores por Vernon (1965, p. 205), de desenvolver nos alunos aptidão para responder a itens de escolha múltipla. Finalmente, o aspecto de adivinhação da resposta é susceptível de controlo pela recorrência a fórmulas de ponderação dos resultados (Choppin, 1975) e pelo número de alternativas propostas. Um estudo recente sobre precisão de testes compostos de itens de duas, três ou quatro alternativas revelou que o número de três é adequado no caso dos testes escolares, e recomenda o seu emprego pelas vantagens de construção (redacção dos itens, escolha dos distractores) e aplicação (inclusão de maior número de itens sem alargamento dos limites de tempo) que oferece (Straton, Catts, 1980, pp. 364-365). A redacção das instruções pode ainda ter efeitos de controlo: a descrição precisa quer da questão (assinalando a existência de uma única resposta correcta), quer da resposta (apelando para uma escolha fundamentada), quer das consequências da resposta (indicando a pontuação positiva da resposta correcta e as penalizações por resposta incorrecta ou omissão) podem também diminuir a probabilidade de intervenção do acaso (Leclercq, 1978, p. 22).

4. *O ESTUDO SISTEMÁTICO DO COMPORTAMENTO DE AVALIAÇÃO ESCOLAR*

Um artigo intitulado "Para uma docimologia experimental" (Noizet, Bonniol, 1969) assinala uma nova linha de trabalhos docimológicos, em que o estudo diferencial da avaliação escolar dá lugar à análise experimental do comportamento de avaliação. Investigações conduzidas nos últimos anos em França por psicólogos como Noizet, Amigues, Caverni, Fabre, Bonniol, entre outros, partem das condições objectivas em que se realiza a avaliação escolar: as variáveis presentes na situação são sujeitas a isolamento experimental, e analisados os seus efeitos enquanto factores de divergência entre os examinadores relativamente a uma mesma prova ou a um mesmo conjunto de provas.

O ponto de partida dessas investigações consiste em considerar a avaliação como um comportamento, no sentido de uma resposta global a uma situação, com componentes perceptivos e cognitivos (Noizet, Caverni, 1978, pp. 66-67). Nesta perspectiva, a explicação das divergências de avaliação passa necessariamente pela análise experimental da situação e implica o recurso a planos de experiência pondo em evidência as variáveis em presença e suas interacções (Bonniol, 1974, p. 202). O comportamento de avaliação é um comportamento de estimação, como tal sujeito a determinantes que cabe à investigação docimológica esclarecer; a Docimologia tem por objecto não só o estudo sistemático dos exames, mas também o estudo sistemático do comportamento de estimação em situação de exame (Noizet, Bonniol, 1969, p. 786).

A tarefa do examinador perante um lote de provas consiste na atribuição a cada uma das notas de uma escala, isto é, em situar um conjunto de estímulos (as provas) num conjunto de respostas (a escala de notas). Esta operação supõe um modelo de referência, constituído previamente à tarefa, que fixa as regras da correspondência (Amigues, Bonniol, Caverni, Fabre, Noizet, 1975, p. 794). A atribuição de notas é um comportamento de tomada de decisão, a partir de índices que o examinador extrai das provas e que traduzem os critérios de apreciação (Noizet, Caverni, 1978, p. 67). O examinador utiliza simultaneamente vários critérios, dada a natureza multidimensional do objecto da apreciação; a decisão supõe, assim, a compatibilização entre índices (Noizet, Caverni, 1978, pp. 126-127).

Nesta óptica, a avaliação é um comportamento que se inscreve na categoria dos que são objecto da psicologia geral, contendo aspectos de natureza cognitiva e de natureza perceptiva, e determinados pela interacção entre as variáveis do estímulo e as variáveis da personalidade.

A análise experimental do comportamento de avaliação emprega planos de experiência que incluem técnicas clássicas da investigação docimológica como a multicorrecção, as provas reais (produzidas efectivamente por alunos) e as provas fictícias (construídas expressamente para o efeito). A manipulação das variáveis independentes consiste na variação sistemática das condições em que a avaliação se realiza, nomeadamente no que se refere às informações fornecidas e à composição dos grupos de avaliadores de um mesmo conjunto de provas. Essas variações são analisadas tomando as variáveis independentes isoladas ou em interacção, e verificando os seus efeitos na variável dependente (as médias das notas atribuídas) (Noizet, Caverni, 1978, pp. 72-74). Para o estudo dos critérios de apreciação utilizados pelos examinadores e processos de recolha dos índices, recorre-se também à avaliação indirecta, desempenhando o experimentador o papel de intermediário (Noizet, Caverni, 1978, p. 122). Tratando-se a avaliação escolar de um comportamento que ocorre no quadro de uma actividade profissional, a questão da representatividade da situação experimental relativamente à situação real é contemplada, por um lado, em termos de composição dos grupos experimentais (futuros ou actuais professores), por outro, das precauções tomadas na construção das provas fictícias (utilizando, por exemplo, partes de provas reais) (Noizet, Caverni, 1978, pp. 74-76).

O modelo de referência de que dispõe o avaliador é um padrão subjectivo, constituído portanto previamente à tarefa de avaliação. O modelo é composto fundamentalmente por três tipos de elementos: o produto-norma, o produto-esperado, a escala de medida (Noizet, Caverni, 1978, pp. 69-70). O produto-norma é de definição mais ou menos imediata consoante a matéria sobre que incidem as provas (numa prova de matemática, por exemplo, poderia ser o acerto de todas as questões que o compõem). A partir do produto-norma é elaborado o segundo elemento, o produto esperado; este toma em linha de conta, relativamente ao primeiro, as informações de que o avaliador dispõe e que são geradoras de expectativas. O terceiro elemento, a escala de medida, traduz-se para o avaliador em escalões de classificação, de escalas de intervalos ou ordinais. O que caracteriza o modelo de referência é menos o aspecto estrutural do que o aspecto funcional: a sua dinâmica interna em termos de natureza e articulação das informações, a forma como elas determinam a operacionalidade dos critérios de avaliação (processos de selecção de índices pertinentes), e as regras de atribuição das classificações. As informações geram expectativas, que se constituem em representações que determinam atitudes; o tríptico representação-expectativa-atitude caracteriza o comportamento de avaliação como essencialmente cognitivo (Noizet, Caverni, 1978, p. 98).

As informações de que o examinador dispõe são de dois grandes tipos: apriorísticas e sequenciais. As informações apriorísticas são elementos extrínse-

cos à situação de avaliação e dizem respeito aos alunos que fizeram as provas ou certas condições em que elas se realizaram. As informações sequenciais são intrínsecas à própria tarefa. Enquanto geradoras de expectativas, os dois tipos de informações produzem efeitos de assimilação e de contraste. O efeito de assimilação traduz-se na procura de compatibilização entre o produto-real (a prova) e o produto-esperado em função da representação constituída, em termos de redução da distância (consonância) entre os dois. O efeito de contraste advém da ordem sequencial das provas e traduz-se em distorções de estimação (sobreavaliação ou subavaliação) decorrentes da ordenação das provas; inclui ainda fenómenos de ancoragem, isto é, fixação num estímulo que, porque privilegiado, assume de forma mais ou menos persistente o estatuto de norma de apreciação dos que se lhe seguem.

O modelo de referência, logicamente anterior à avaliação, não é por conseguinte estático, antes susceptível de modificações e ajustamentos no decurso da tarefa. Piéron referira já o facto de que uma prova pode ser diferentemente classificada consoante sucede a outra de nível excelente ou medíocre, e a prática da avaliação sugere uma tendência de severidade crescente na apreciação das primeiras para as últimas provas de uma série. Os efeitos de ordem de apresentação das provas são mais pregnantes do que os dos critérios de avaliação que os examinadores dizem utilizar (Bonniol, 1974, p. 203). O fenómeno perceptivo de ancoragem adquire no estudo sistemático do comportamento de avaliação um valor heurístico considerável.

Numa comunicação apresentada ao VI Congresso Internacional de Ciências da Educação (Paris, 1973) foi proposto um modelo explicativo, de tipo analógico, do comportamento de avaliação (Amigues, Bonniol, Caverni, Fabre, Noizet, 1975). O modelo é expressamente concebido como um sistema de âncoras, que estabelece uma correspondência entre produtos-esperados e valores de uma escala de medida. Essa correspondência passa por um produto-norma definido pelos objectivos pedagógicos e pelos conhecimentos escolares. Os produtos-esperados são uma amostra de produtos possíveis, determinados pelas informações apriorísticas e sequenciais, amostra essa que provém de uma operação de selecção. Uma outra operação tem lugar, a de comparação do produto-real com o sistema de âncoras. Num primeiro momento o produto-real é sujeito a uma assimilação; quando um produto-real é nitidamente superior ou inferior aos outros, ocorre uma modificação no sistema que determina a sobreavaliação ou a subavaliação dos que se seguem – trata-se da introdução de uma âncora parasita. Assim, os efeitos de contraste resultam, pelo menos em parte, do próprio jogo das assimilações. A validação do modelo implica a realização de estudos experimentais incidindo particularmente sobre os determinantes da avaliação (informações disponíveis, produto-norma, regras de cotação) e a operação de comparação (rela-

ção entre a articulação do sistema e a "leitura" do produto-real) (Amigues, Bonniol, Caverni, Fabre, Noizet, 1975, pp. 795-798).

O funcionamento dos mecanismos de assimilação e contraste em situação de avaliação escolar tem sido objecto de várias investigações experimentais. Essas investigações, reportando-se a situações simuladas, fazem intervir informações fictícias sobre os alunos e efeitos de ordem e de ancoragem na avaliação de um lote de provas por diferentes examinadores. Seguidamente se passam em revista resultados extraídos de alguns desses trabalhos.

Um estudo de Bonniol, Caverni e Noizet (1972) põe em evidência o efeito do estatuto escolar induzido (por informação previamente fornecida) na avaliação de um conjunto de oito provas de redacção. O plano de experiência contém como variáveis independentes a experiência dos examinadores (16 professores em exercício e 16 futuros professores), o estatuto escolar fictício dos alunos (indicação sobre se pertencem a uma turma de bons ou de maus alunos), e a origem real das provas (quatro provenientes de alunos de uma turma boa, as outras quatro de alunos de uma turma fraca). A variável dependente é a classificação, numa escala de 0 a 20. As oito redacções foram escolhidas segundo um critério de provas medianas, de duas turmas diferenciadas do primeiro ano do ensino secundário. A variável estatuto escolar fictício determina diferenças de médias estatisticamente significativas, sendo mais elevada a das notas atribuídas às provas supostamente redigidas por alunos da turma superior; o cruzamento da variável estatuto com a variável origem real, evidencia diferenças de médias no mesmo sentido. O efeito é idêntico nos dois grupos de examinadores, não se verificando fenômenos de interacção. A variável experiência dos examinadores traduz-se ao nível dos resultados por uma maior severidade de avaliação por parte dos professores em exercício.

Num outro estudo (Caverni, Fabre, Noizet, 1975) a hipótese é formulada explicitamente em termos de assimilação das avaliações actuais às avaliações anteriores, independentemente do conteúdo das provas. Duas experiências reproduzem as condições de avaliação contínua e de avaliação pontual. Na experiência 1, numa primeira fase doze examinadores (professores de inglês do ensino secundário e do ensino superior) classificaram doze provas; numa segunda fase, dois meses mais tarde, aos mesmos examinadores foi pedida a classificação de um lote de seis provas, todas de nível idêntico, imaginariamente redigidas pelos mesmos alunos. Em cada uma das seis provas figurava uma nota fictícia anterior, sendo as seis notas fictícias as três mais altas e as três mais baixas efectivamente atribuídas por cada examinador na primeira fase. As médias das notas das três provas com informação anterior favorável são superiores às médias das notas com informação anterior desfavorável, sendo a diferença significativa no nível de probabilidade de

.005. Na experiência 2, dezesseis professores do ensino secundário, repartidos por quatro grupos, classificaram quatro provas de ciências naturais. Em cada uma das provas figurava uma série de notas fictícias, supostamente relativas ao rendimento escolar do mesmo aluno durante o ano lectivo. Quatro séries de cinco notas fictícias, numa escala de 0 a 20 e diferenciadas quanto à tendência central e à dispersão, foram associadas a cada uma das provas; os valores da média e do desvio-padrão comuns a cada duas séries eram, respectivamente, de 13 e de 7, de 3,8 e de 1,00. A análise das notas dos dezesseis examinadores, em termos de médias, evidencia que a média das notas atribuídas às provas é mais elevada quando a média das notas fictícias é alta do que quando é baixa, não sendo a diferença entre médias, porém, significativa num nível de probabilidade exigente. Tomando em consideração o índice de dispersão das distribuições de notas fictícias, verifica-se que a média das notas é mais alta quando a dispersão é fraca, sendo a diferença entre médias significativa no nível de probabilidade de .05.

Os resultados dos dois trabalhos apontam no sentido da dependência da avaliação actual relativamente à informação disponível, que em ambos os casos é uma avaliação escolar anterior. Essa dependência manifesta-se por uma assimilação, isto é, uma redução de distância, entre a nota virtual e notas anteriores. A redução da distância é uma procura de consonância cognitiva (Caverni, Amigues, 1977, p. 17; Noizet, Caverni, 1978, p. 100).

Outros trabalhos fazem intervir na situação experimental informações extra-escolares do tipo estatuto sócio-cultural fictício dos alunos ou origem étnica, igualmente imaginária. São descritos por Noizet e Caverni dois trabalhos em que é aduzido como factor experimental a identidade social dos examinadores (Noizet, Caverni, 1978, pp. 88-96). Em síntese, os resultados revelam uma assimilação da avaliação à categorização experimentalmente induzida, em termos de valência positiva ou negativa; ainda, da compensação de preconceitos desfavoráveis (Noizet, Caverni, 1978, p. 92 e p. 95).

Como referido antes, os efeitos de contraste provêm de informações sequenciais concomitantes à tarefa de avaliação. O contraste exprime-se por sobreavaliações ou subavaliações determinadas pela posição de ordem, absoluta ou relativa, das provas na série. O seu estudo experimental implica por consequência análise de efeitos de ordem, que inclui a identificação e a localização de provas que pela sua qualidade são "exteriores" à série e que fazem intervir fenómenos de ancoragem.

A influência da ordem de apresentação das provas na avaliação das mesmas foi posta em evidência num trabalho experimental em que metade de um grupo de 18 examinadores classifica 26 provas de tradução de inglês numa dada sequên-

cia, e a outra metade na sequência inversa (14). A análise das médias revela que as primeiras provas são sobreavaliadas relativamente às últimas, nas duas ordenações, à excepção da primeira prova que é subestimada. As diferenças entre médias verificam-se relativamente à série no seu conjunto (ordenações directa e inversa), como às partes (primeira metade/segunda metade, primeiro terço/último terço). A posição relativa (em termos de classificação) das provas varia nas duas ordenações, atingindo as diferenças nalguns casos valores elevados; considerando a ordem de qualidade absoluta das 26 provas definida pelo lugar que cada uma ocupa no conjunto tomando em conta as médias das notas atribuídas nas duas ordenações, verifica-se que a grande diferença de posição relativa de algumas provas decorre de se seguirem imediatamente, em cada ordenação, a uma prova boa ou a uma prova fraca. A sobreavaliação ou subavaliação é, assim, um efeito de contraste. Isto leva o autor do trabalho a afirmar, parafraseando Piéron, que para prever a nota de uma prova é importante não só conhecer o examinador como ainda as provas que a precedem – mais do que a prova em questão (Bonniol, 1974, p. 205). A subestimação da primeira prova poderá explicar-se pela sua comparação, na falta de informações sequenciais, ao produto-norma (na ocorrência, a tradução sem erros); assim, também por um efeito de contraste (Noizet, Caverni, 1978, p. 106).

Noizet e Caverni submetem os mesmos dados a processos matemáticos de tratamento que põem em evidência os efeitos de contraste. As distribuições de notas dos diferentes examinadores foram previamente reduzidas a uma escala comum (média 10, desvio-padrão 3), a fim de legitimar a comparação. Um dos processos consiste na determinação das médias das notas dadas às provas cujo único ponto comum é o lugar que ocupam na série: médias das notas das provas 1-2-3, etc., na ordem directa, e das provas 26-25-24, etc na ordem inversa. As médias assim obtidas são independentes do conteúdo das provas; verifica-se, no entanto, uma oscilação regular e sistemática em torno do valor 10 (a média reduzida). Um outro processo baseia-se no cálculo da diferença entre as médias das notas dadas às provas na posição n e na posição n+1. Mais uma vez se trata de notas dadas a provas diferentes, que têm em comum apenas a posição de ordem na série. São obtidos 25 valores de d ($d = M_{n+1} - M_n$), analisáveis em termos de 24 pares sucessivos. Na ausência de efeitos de contraste, o número de pares em que d é positivo seria igual ao número de pares em que d é negativo; se somente interviessem efeitos de contraste, todos os d sucessivos seriam de sinal contrário. O facto

(14) Não foi possível a consulta directa do trabalho publicado em 1965. O autor faz-lhe referência num outro artigo (Bonniol, 1974), e no livro de Noizet e Caverni os dados são reapreciados à luz de técnicas de tratamento que põem em evidência os fenómenos de contraste (Noizet, Caverni, 1978, pp. 104-106).

de em 18 dos 24 pares sucessivos d ser de sinal contrário, indica a presença de efeitos de contraste (Noizet, Caverni, 1978, pp. 106-107).

Uma forma de estudar experimentalmente os fenómenos de ancoragem é fornecer a dois grupos equivalentes de examinadores um conjunto de provas, sendo num dos casos incluída uma âncora (grupo experimental) e no outro não (grupo controlo). A âncora experimental introduzida é uma prova excepcional relativamente às outras (no sentido positivo ou negativo), um estímulo "exterior" à série, portanto. O estudo dos efeitos da âncora consiste na análise de distorções (sobrestimação ou subestimação das provas que se lhe seguem), e sua persistência (Bonniol, 1974, p. 204).

Uma experiência conduzida sobre dezassete provas de inglês e catorze de matemática distribuídas por um grupo controlo e seis grupos experimentais de entre doze a quinze examinadores cada, fez intervir como valores da variável independente a natureza da âncora (alta ou baixa), o peso da âncora (uma única prova ou três provas sucessivas do mesmo nível de qualidade), e o lugar da âncora (no fim do primeiro terço ou no início do terceiro). Os resultados evidenciaram efeitos de contraste na classificação das provas das duas disciplinas: no inglês o fenómeno repercute-se nas quatro ou cinco provas que sucedem à âncora, e é mais acentuado quando ela é constituída por três provas; na matemática verificam-se efeitos no mesmo sentido embora de menor amplitude, sobretudo quando a âncora é elevada (Bonniol, 1974, p. 204). Os efeitos de contraste são mais acentuados quando as âncoras se situam no início da série (Noizet, Caverni, 1978, p. 111).

No quadro dos estudos sistemáticos do comportamento de avaliação merecem ainda referência os trabalhos sobre a natureza dos critérios de apreciação e o modo como se efectua a recolha dos índices que lhes correspondem. Como referido antes (p. 44), esse tipo de trabalhos emprega por vezes a avaliação indirecta; outras, provas construídas segundo dimensões objectivamente definidas, e que decorrem dos critérios que os examinadores dizem utilizar nas diversas disciplinas.

Num estudo recente (Caverni, Amigues, 1977) trinta e dois futuros professores classificaram oito provas de redacção sobre o mesmo tema, atribuídas a alunos do primeiro ano do ensino secundário, e construídas segundo três dimensões, o estilo, a ordem de apresentação das ideias e o conteúdo. Nos aspectos de estilo e de ordem de apresentação das ideias, os índices experimentalmente introduzidos referiram-se, respectivamente, à correcção de sintaxe e às qualidades lógicas da exposição; no aspecto do conteúdo, permitiam induzir o meio socio-económico dos alunos. As médias das notas atribuídas às provas sem incorrecções de esti-

lo, bem como às provas seguindo uma ordem lógica de exposição, são superiores às médias das notas atribuídas às provas onde o mesmo não se verifica, sendo a diferença entre as médias estatisticamente muito significativa. No que diz respeito aos índices de conteúdo, as diferenças entre médias não são significativas; a análise dos resultados evidencia, porém, que a recolha dos índices relativos ao meio sócio-económico dos alunos foi diferentemente efectuada pelos examinadores de meio sócio-económico diferente – a interacção entre os dois factores experimentais é significativa no nível de probabilidade de $p < .001$.

Este trabalho põe em evidência que critérios de natureza subjectiva intervêm na avaliação escolar, a par de critérios de natureza objectiva. E, consequentemente, que os examinadores utilizam na apreciação critérios não explícitos. Num trabalho anterior, não publicado, Noizet e Caverni tinham chegado a conclusões idênticas, e verificado que os critérios que os examinadores dizem utilizar são tanto quantitativos como qualitativos, e que nestes é ainda possível distinguir critérios subjectivos de outros mais objectivos; por outro lado, que os critérios que os examinadores utilizam de facto na recolha de índices não coincidem, no que se refere à sua importância relativa, com os que dizem utilizar (Noizet, Caverni, 1978, pp. 122-124).

Trabalhos realizados sobre a possível dependência entre a variedade de critérios de apreciação e as propriedades formais de escalas de notas, não sãoconclusivos (Noizet, Caverni, 1978, pp. 137-140). Apontam, no entanto, no sentido da existência de uma escala implícita, de cada examinador, a qual se caracteriza pelo número de escalões e pelo intervalo médio entre duas notas sucessivas; na situação de avaliação o examinador adapta essa escala implícita à escala de notas de que dispõe (Noizet, Caverni, 1978, p. 139).

CONCLUSÃO

A Docimologia conheceu, em pouco mais de meio século de existência, um desenvolvimento considerável, no plano da teoria como no das aplicações.

Ressalte-se, em primeiro lugar, a importância da evolução da Docimologia de uma perspectiva estritamente crítica dos processos de avaliação – pondo em evidência a instabilidade das classificações, as divergências entre os examinadores, em suma, a inadequação dos exames e dos concursos aos fins propostos – para uma perspectiva de estudo científico da problemática e das técnicas da avaliação escolar.

Saliente-se, em segundo lugar, os progressos realizados no que respeita a

questão da objectividade da avaliação escolar: por um lado, o aperfeiçoamento da avaliação tradicional, no que se refere aos aspectos técnicos a ter em consideração na construção das provas, à concertação entre os examinadores sobre os critérios de apreciação, e aos métodos de moderação das classificações; por outro, a importância conferida ao desenvolvimento de instrumentos estandardizados de medição das aquisições escolares.

Assinale-se, ainda, a diversificação dos métodos de investigação docimológica. A análise experimental da situação de avaliação permitiu identificar factores de distorção das apreciações independentemente das diferenças individuais entre os examinadores, e esclarecer aspectos da interacção entre os factores subjectivos e as variáveis da situação.

À luz dos trabalhos docimológicos é possível equacionar os problemas da avaliação escolar em termos de efeitos das diferenças interindividuais e intraindividuais dos examinadores, da diversidade dos critérios de apreciação e de classificação, das expectativas constituídas, da ordem de apresentação das provas, dos aspectos formais e das características técnicas dos instrumentos de avaliação. A Docimologia contribuiu ainda, pela pertinência das questões que constituem o objecto de estudo, para uma reformulação da problemática da avaliação escolar na perspectiva de realização dos objectivos pedagógicos e da formação dos alunos.

A análise da evolução da Docimologia sugere uma interrogação: qual o futuro próximo da investigação docimológica?

Duas grandes linhas de orientação emergem no estádio actual do desenvolvimento dos trabalhos: por um lado, o aperfeiçoamento e diversificação das técnicas de avaliação dos conhecimentos escolares, enquanto instrumentos de controlo e enquanto meios auxiliares da realização dos objectivos pedagógicos; por outro, o estudo dos determinantes do comportamento de avaliação, do seu peso relativo e da sua interdependência.

Afigura-se provável que a Docimologia se encaminhe no sentido da confluência de ambas, e que da articulação das duas metodologias resultem novas perspectivas para a avaliação escolar.

BIBLIOGRAFIA

ALLAL, L.K. – Stratégies d'évaluation formative: conceptions psychopédagogiques et mo-

dalités d'aplication in Allal. L., Cardinet, J., Perrenoud, P. – (Eds.) *L'evaluation formative dans un enseignement differencié*. Bern: Lang, 1979.

ALLAL, L. K., CARDINET, T., PERRENOUD, P. (Eds.) *L'évaluation formative dans un enseignement differencié* (Actes du colloque à l'Université de Genève, Mars 1978). Bern: Lang, 1979.

AMIGUES, R., BONNIOL, J. J., CAVERNI, J. P., FABRE, J. M., MOIZET, G. – Le comportement d'évalution de productions scolaires: à la recherche d'un modèle explicatif. *Bulletin de Psychologie*, 1975, *28*, pp. 793-799.

ANASTASI, A. (Ed.) – *Testing Problems in Perspective*. Washington: American Council on Education, 1966.

ANASTASI, A. – *Psychological Testing* (4th ed.). New York: Macmillan, 1976.

ANDRE, C., CRETIGNY, I. – Les production des élèves. *Recherches Pédagogiques*, 1974, nº 66, pp. 83-89.

ANGOFF, W. H. – Scales, norms and equivalent scores. *in* Thorndike, R. L. (Ed.) *Educational Measurement* (2nd ed.). Washington: American Council on Education, 1971, pp. 508-600.

ASTIN, A. W., PANOS, R. J. – The evaluation of educational programs *in* Thorndike, R. L. (Ed.) *Educational Measurement* (2nd ed.). Washington American Council on Education, 1971, pp. 733-751.

BACHER, F. – L'évaluation des résultats scolaires au niveau de l'école moyenne. *Travail Humanin*, 1965, *28*, pp. 219-230.

BACHER, F. – L'utilisation des tests dans les écoles en France. *B.I.N.O.P.*, 1968, *24*, pp. 13-15.

BACHER, F. – Congrès de Berlin sur les possibilités et les limites de l'application des tests dans les écoles. *B.I.N.O.P.*, 1968, *24*, pp. 42-49.

BACHER, F. – La normalisation de la notation. *B.I.N.O.P.*, 1969, *25*, pp. 75-90.

BACHER, F. – La docimologie *in* Reuchlin, M. (Ed.) *Traité de Psychologie Appliquée* (Tome 6). Paris: Presses Universitaires de France, 1973, pp. 27-87.

BACHER, F., REUCHLIN, M. – Le Cours d'Observation. Enquête sur l'ensemble des élèves d'un département (Loiret, 1960-62). *B.I.N.O.P.*, 1965, *21*, pp. 149-236.

BAKER, F. B. – Automation of test scoring, reporting and analysis *in* Thorndike, R. L. (Ed.) *Educational Measurement* (2nd ed.). Washington: American Council on Education, 1971, pp. 202-234.

BESSE, J. M. – Vers une pédagogie par objectifs? *Bulletin de la Société A. Binet et Th Simon*, nº 556, 77ème année, III, 1977, pp. 114-147.

BIRZEA, L. – *Rendre opérationnels les objectifs pédagogiques*. Paris: Presses Universitaires de France, 1979.

BLOOM, B. S., ENGELHART, M. D., FURST, E. J., HILL, W. H., KRATHWOHL, D. R. – *Taxonomy of Educational Objectives. The Classification of Educational Goals. Handbook I: Cognitive Domain*. New York: McKay, 1956.

BLOOM, B. S., HASTINGS, H. T., MADAUS, G. F. – *Handbook on Formative and Summative Evaluation of Student Learning*. New York: McGraw-Hill, 1971.

BONBOIR, A. – *La pédagogie corrective*. Paris: Presses Universitaires de France, 1970.

BONBOIR, A. – *La docimologie*. Paris: Presses Universitaires de France, 1972.

BONBOIR, A. – *La méthode des tests en Pédagogie*. Paris: Presses Universitaires de France, 1972.

BONBOIR, A. (Ed.) *Une pedagogie pour demain*. Paris: Presses Universitaires de France, 1974.

BONNIOL, J. J. – Les comportements d'estimation dans une tâche d'évaluation d'épreuves

scolaires. Étude de quelquer-uns de leurs déterminants. *Bulletin de Psychologie*, 1974, *27*, pp. 202-205.

BONNIOL, J. J., CAVERNI, J. P., NOIZET, G. – Le statut scolaire des élèves comme déterminant de l'évaluation des devoirs qu'ils produisent. *Cahiers de Psychologie*, 1972, *15*, pp. 83-92.

BONORA, D. – L'évalution des connaissances: quelques problèmes de mesure. *Pédagogie*, 1972, nº 2, pp. 158-177.

BONORA, D. – Les buts de l'éducation in Reuchlin, M. (Ed.) *Traité de Psychologie Appliquée* (Tome 6). Paris: Presses Universitaires de France, 1973, pp. 139-191.

BROWN, F. G. – *Measurement and Evaluation*. Itasca: Peacock Publishers, 1971.

BRUNELLE, L. (Ed.) *L'Université en question: Pourquoi les exames?* Paris: Société des Editions Rationalistes, 1968.

BUJAS, Z. – L'influence du mode de notation sur la validité des tests de connaissances. *Travail Humain*, 1965, *28*, pp. 197-202.

CAMBON, J. – Objéctivité de la notation des tests de connaissances à réponses libres. *B.I.N.O.P.*, 1961, *17*, pp. 329-334.

CAMPOS, B. P. – Taxonomie des objectifs. Analyse des tâches. (Evaluation pour une orientation intégrée au processus de l'Education) *in* Bonboir, A. (Ed.) *Une pédagogie pour demain*. Paris: Presses Universitaires de France, 1974, pp. 65-85.

CAMPOS, B. P. – L'observation des aptitudes en vue de l'orientation *in*: Bonboir, A. (Ed.) *Une pédagogie pour demain*. Paris: Presses Universitaires de France, 1974, pp. 105-136.

CARDINET, J. – *Objectifs éducatifs et évalution individualisée*. Neuchâtel: Documents de l'I.R.D.P., 1977.

CARDINET, J. – *L'évaluation scolaire et égalité de chances*. Documents de l'I.R.D.P., 1978.

CARDINET, J. – L'évaluation formative, problème actuel. In: Allal, L. K., Cardinet, J., Perrenoud, P. (Eds.) *L'évaluation formative dans un enseignement différencié*. Bern: Lang, 1979, pp. 10-18.

CARDINET, J., ALLAL, L.K. – *La mesure des résultats de l'enseignement* (compte rendu de l'atelier de contact européen organisé à Windsor du 13 au 18 Juin 1976 par le Conseil de l'Europe). Neuchâtel, 1976.

CASTRO, D. de – Recherches docimologiques sur quelquer matières inscrites au programme d'examens universitaires. *Travail Humain*, 1965, *28*, pp. 231-265.

CAVERNI, J.P., AMIGUES, R. – Aggrégation des critères de choix et interactions sociales dans une tâche d'évaluation réelle: la notation de production scolaires. *Cahiers de Psychologie*, 1977, *20*, pp. 15-26.

CAVERNI, J.P., FABRE, J.M., NOIZET, G. – Dépendence des évaluations scolaires par rapport à des évaluations antérieurs: études en situation simulée. *Travail Humain*, 1975, *38*, pp. 213-222.

CHOPPIN, B. – Guessing the answer on objective tests. *British Journal of Educational Psychology*, 1975, *45*, pp. 206-213.

COFFMAN, W.E. – Essay examination *in* Thorndike, R.L. (Ed.) *Educational Measurement* (2nd ed.). Washington: American Council on Education, 1971, pp. 271-302.

COOLEY, W.W., LOHNES, P.R. – *Evaluation Research in Education*. New York: Wiley, 1976.

CRONBACH, L.J. – Test validation *in* Thorndike, R.L. (Ed.) *Educational Measurement* (2nd ed.). Washington: American Council on Education, 1971, pp. 443-507.

CRONBACH, L.J., SNOW, R.E. – *Aptitudes and Instructional Methods. A Handbook for Research on Interactions*. New York: Wiley, 1977.

DAVIS, F.B. – Estimation des décalages entre connaissances et aptitudes. *Travail Humain*, 1965, *28*, pp. 213-218.

DEBESSE, M., REUCHLIN, M. (Eds.) – *Traité des Sciences Pédagogiques. Tome 4: Psychologie de l'Education.* Paris: Presses Universitaires de France, 1974.

DEMANGEON, M., LARCEBEAU, S. – Une expérience de correction multiple. *B.I.N.O.P.*, 1958, *14*, pp. 131-156.

Evaluation I: savoir noter, noter le savoir. *Cahiers Pédagogiques*, 1978, *34*, nº 162.

Evaluation II: La mauvaise conscience. *Chaiers Pédagogiques*, 1978, *35*, nº 168.

Examens et concours. Hommage à Henri Piéron. Paris: Les Cahiers Rationalistes, 1965, nº 227.

FABRE, J. M. – *Jugement et certitude. Recherche sur l'évaluation des conaissances.* Berne: P. Lang, 1980.

FINLAYSON, D. S. – The reliability of marking essays. *British Journal of Educational Psychology*, 1951, *21*, pp. 126-134.

GAGNÉ, R.M. *Te Conditions of Learning* (2nd ed.). New York; Holt, Rinehart & Winston, 1970.

GARDNER, E. F. – The importance of reference groups in scaling procedures (1952). *In* Anastasi, A. (Ed.,) *Testing Problems in Perspective.* Washington: American Council on Education, 1966, pp. 272-280.

GLASER, R., NITKO, A.J. – Measurement in learning and instruction, *in* Thorndike, R. L. (Ed.) *Educational Measurement* (2nd. ed.). Washington: American Council on Education, 1971, pp. 625-670.

GRONLUND, N.E. *Stating behavioral objectives for classroom instruction.* New York: Macmillan, 1970.

GRONLUND, N.E. *Measurement and evaluation in teaching* (3rd ed.). New York: Macmillan, 1976.

GUEBERT-SCEAUX, A.M., REUCHLIN, M. – Etude sur l'examen d'entrée en sixième dans cinq établissements scolaires parisiens. *B.I.N.O.P.*, 1958, *14*, pp. 9-19.

GUGLIELMI, J. – *L'enseignement programmé à école.* Paris: Presses Universitaires de France, 1970.

HARROW, A.J. – *A Taxonomy of the Psychomotor Domain.* New York: McKay, 1972.

HARTOG, P., RHODES, E.C. – *An Examination of Examinations.* London: Macmillan, 1935.

HOTYAT, F. – *Les examens. Les moyens d'évaluation dans l'enseignement.* (Documents Pédagogiques Internationaux de l'Institut de l'UNESCO pour l'Education). Paris: Bourrelier, 1962.

HOTYAT, F. – L'organisation des examens. Exposé introductif. (XVi colloque de l'Association Internationale de Pédagogie Expérimentale de Langue Française). *Les Sciences de l'Education* (Nº sp. Docimologie et Education). Paris: Didier, 1969, p. 3-17.

KAUFMANN, J. – Note sur les problèmes de métrique en matière de notation scolaire. *Travail Humain*, 1975, *38*, pp. 133-148.

KRATHWOHL, D.R., BLOOM, B.S., MASIA, B.B. – *Taxonomy of Educational Objectives: the Classification of Educational Goals. Handbook 2: Affective Domain.* New York: McKay, 1964.

La correction des épreuves écrites dans les examens. Enquête expérimentale sur le baccalaréat. Paris: La Maison du Livre, 1936.

LANDSHEERE, G. de – *Evaluation continue et examens. Précis de Docimologie* (3e. ed.). Paris: F. Nathan, 1974.

LANDSHEERE, V. de LANDSHEERE, G. de – *Definir os objectivos da Educação* (2ª ed.). Lisboa: Moraes, 1977 (tradução portuguesa).

LAUGIER, H. PIÉRON, H., PIÉRON, Mme., TOULOUSE, E., WEINBERG, D. – *Etudes*

docimologiques sur le perfectionnement des examens et des concours. Publications du Travail Humain, série A, nº 3, 1935.

LECLERQ, D. – La question à choix multiple (Q.C.M.): un outil d'hier ou de demain? *Cahiers Pédagogiques*, 1978, *34*, pp. 22-24.

LEWIS, D.G. – *Assessment in Education.* London: University of London Press, 1974.

LOBROT, M. – Le problème des examens *in* Brunelle, L. *L'Université en question: pourquoi les examens?* Paris: Société des Editions Rationalistes, 1968, p. 85-112.

MARQUES, J. H. FERREIRA – *O problema da validade em Psicologia Diferencial.* Separata da Revista da Faculdade de Letras. Lisboa, 1971.

MESSICK, S. – The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*, pp. 955-966.

NGUYEN-XUAN, A. – Recherches anglaises sur la notation des devoirs de rédaction. *B.I.N.O.P.*, 1963, *19*, pp. 254-266.

NISBET, J. D. – English composition in secondary school selection. *British Journal of Educational Psychology*, 1955, 25, pp. 51-54.

NOIZET, G. – Etude docimologique sur la correction de lécrit du baccalauréat. *B.I.N.O.P.*, 1961, *17*, (257-267).

NOIZET, G., BONNIOL, J.J. – Pour une docimologie expérimentale. *Bulletin de Psychologie*, 1969, *22*, pp. 782-787.

NOIZET, G., CAVERNI, J.P. – *La psychologie de l'évaluation scolaire.* Paris: Presses Universitaires de France, 1978.

PELNARD, J. – Mission d'information sur les méthodes utilisées en Grande Bretagne dans les examens, et sur différentes recherches psychopédagogiques liées aux problèmes de mesure. *B.I.N.O.P., 24*, pp. 332-334.

PELNARD-CONSIDÈRE, J. – Travaux docimologiques sur les examens en Faculté de Médecine. *B.I.N.O.P., 25*, pp. 241-256.

PENFOLD, D.M. EDWARDS – Essay marking experiments: shorter and longer essays. *British Journal of Educational Psychology*, 1956, *26*, pp. 128-136.

PERRENOUD, P. – Les différences culturelles aux inégalités scolaires: l'évaluation et la norme dans un enseignement indifférencié. *In*: Allal, L.K., Cardinet, J., Perrenoud, P. (Eds.) *L'évaluation formative dans un enseignement différencié.* Bern: Lang, 1979, p. 20-55.

PIÉRON, H. – *Examens et docimologie* (2e. ed.). Paris: Presses Universitaires de France, 1969.

PIÉRON, H. – *Vocabulaire de Psychologie* (5e. ed.). Paris: Presses Universitaires de France. 1973.

PIÉRON, H., REUCHLIN, B., BACHER, F. – Une recherche expérimentale de docimologie sur les examens oraux de Physique au niveau du baccalauréat de Mathématiques. *Biotypologie*, 1962, *23*, pp. 48-73.

PIÉRON, H. REUCHLIN, M., BACHER, F., DEMANGEON, M. – Analyse des corrélations entre notations à une session du baccalauréat. *Biotypologie*, 1962, *23*, pp. 17-47.

Rapsodie, Groupe-Prévenir les inégalités scolaires par une pédagogie différenciée: à propos d'une recherche-action dans l'enseignement primaire genevois. *In*: Allal, L.K., Cardinet, J., Perrenoud, P. (Eds.) *L'évaluation formative dans un enseignement différencié.* Bern: Lang, 1979, p. 68-108.

REMONDINO, C. – Etude factorielle sur la notation des compositions scolaires portant sur la langue maternelle. *Travail Humain*, 1959, *22*, pp. 27-40.

REMONDINO, C. – Recherche sur les systèmes numériques d'évaluation scolaire. *Travail Humain*, 1965, *28*, pp. 263-265.

REUCHLIN, M. – Applications de la psychologie génétique et de la psychologie différen-

tielle au cours du Cours d'Observation. *B.I.N.O.P.* 1960, *16*, pp. 59-104.
REUCHLIN, M. – Le symposium "Les problèmes méthodologiques de l'évaluation des connaissances scolaires" au XVe Congrès de l'Association Internationale de Psychologie Appliquée (Ljubliana, 2-8 Août 1964). *Travail Humain*, 1965, *28*, pp. 193-195.
REUCHLIN, M. – Les examens: problèmes vrais et solutions fausses *in* Brunelle, L. (Ed.) *L'Université en question: Pourquoi les examens?* Paris: Société des Editions Rationalistes, 1968, p. 113.127.
REUCHLIN, M. (Ed.) – *Traité de Psychologie Appliquée. Tome 6: l'éducation, la psychologie et les institutions educatives.* Paris: Presses Universitaires de France, 1973.
REUCHLIN, M. – Problemes d'évaluation *in* Debesse. M., Mialaret, G. (Eds.) *Traité de Sciences Pédagogiques*, (tome 4). Paris: Presses Universitaires de France, 1974, pp. 207-236.
STANLEY, J.C. – Reliability *in* Thorndike, R.L. (Ed.) *Educational Measurement* (2nd ed.) Washington: American Council on Education, 1971, pp. 356-442.
STRATON, R.G., CATTS, R.M. – A comparison of two, three and four-choice item tests given a fixed total number of choices. *Educational and Psychological Measurement*, 1980. *40*, pp. 357-366.
THORNIDIKE, R.L. (Ed.) *Educational Measurement* (2nd ed.). Washington: American Council on Education, 1971.
TOURNEUR, Y. – Analyse des objectifs dans l'évaluation de maîtrise – étude de quelques indices. *Les Sciences de l'Education*, 1974, nº 1, pp. 36-56.
VANDERVELDE, L., VANDER ELST, P. – *Os objectivos em educação: será possível defini-los com precisão?* Coimbra: Almedina, 1979 (tradução portuguesa).
VERNON, P.E. – *Secondary School Selection – A British Psychological Inquiry*. London: Methuen, 1957.
VERNON, P.E. – Evaluation objective des résultats obtenus dans les études de niveau élevé. *Travail Humain*, 1965, *28*, pp. 203-212.
VERNON, P.E. – *Intelligence and Attainment Tests*. London: University of London Press, 1972.
VERNON, P.E., MILLIGAN, G.D. – Futher study of the reliability of English essays. *British Journal of Statistical Psychology*, 1954, 7, pp. 65-74.
WISEMAN, S. – The marking of English composition for Grammar School Selection. *British Journal of Educational Psychology*, 1949, *19*, pp. 200-209.