Avaliação formativa — uma experiência no 7º ano

J. C. David Vieira
Universidade de Aveiro

Introdução

A experiência, objecto desta comunicação, teve origem numa preocupação de avaliação no ensino superior: **como avaliar em disciplinas com mais de 1000 alunos sem que os docentes envolvidos nessa disciplina sejam completamente absorvidos pelo trabalho de fazer e corrigir testes ?**

Uma ideia, não original, foi a de recorrer à informática. Pensou-se na construção de uma Base de Dados (**BD**) onde seriam gerados testes por unidades temáticas e globais, com vários níveis, a que os alunos teriam acesso.

A passagem de um teste de nível **i** a um de nível **i+1** [**0<i<t**, **t** fixado] estaria condicionada à obtenção de uma certa pontuação no teste de nível **i**.

A realização com sucesso de toda uma bateria de testes permitiria a apresentação do aluno a exame final. Aquela avaliação seria dominantemente formativa, apesar de proporcionar a obtenção de uma nota mínima de acesso à prova de avaliação final, e seria acompanhada por um docente da disciplina (acção tutorial).

Tendo em conta os fins em vista impunha-se um tipo de teste **objectivo**, de **fácil correcção** e que permitisse ao aluno a detecção e superação das suas principais dificuldades e eventuais

bloqueios. A estes testes estariam associados mecanismos de rectroacção.

A escassez de meios materiais e humanos para concretizar aquele trabalho em curto prazo (o grupo criado para o efeito procede actualmente à recolha de material) levou-me a pensar numa experiência de avaliação formativa por computador a nível básico.

Aproveitava o facto de coordenar o projecto "Matemática-Ensino" (PM/E) no Dep. de Matemática da U.A. e de orientar o seminário do 5º ano da lic. em Ensino da Matemática e um núcleo de estágio na Escola Secundária nº1 de Aveiro.

A experiência - Fase 1

O objectivo

O objectivo imediato era o de construir uma **BD** para avaliação formativa (AF) no 2º CEB. A opção por este tipo de avaliação tem a ver, por um lado, com a origem da experiência e por outro, com a importância que, no meu entender, este tipo de avaliação deveria ter na acção pedagógica. Com efeito esta avaliação ocorre em geral no decurso de uma aprendizagem e permite verificar o grau de progresso do aluno e detectar eventuais dificuldades no processo de aquisição de conhecimentos. Permite ainda ao professor controlar o andamento do seu programa pedagógico. À AF devem, pois, estar associados dois tipos de rectroacção - um sobre o aluno e outro sobre o professor. Enfim, pretendia-se construir um instrumento de avaliação que fosse interessante para os alunos e que permitisse a auto-regulação das suas aprendizagens bem como o controle do processo por parte do professor.

O tema

A decisão de desenvolver a experiência no âmbito do seminário do 5º ano da lic. em Ensino da Matemática foi tomada pelo facto de a maior parte dos alunos inscritos no seminário serem ao mesmo tempo estagiários numa Escola Secundária o que permitia testar o trabalho produzido.

O tema "**Equações numéricas do 1º grau em Q**" foi escolhido pelos alunos. Inicialmente o tema incluia **problemas,** mas posteriormente esta componente foi abandonada por se concluir que com o formato pretendido e com o tempo disponível os objectivos da resolução de problemas dificilmente seriam atingidos.

Foi decidido que o teste sobre o tema escolhido seria precedido de um teste de pré-requisitos envolvendo quase toda a matéria tratada antes do tema seleccionado.

A equipa

A equipa-base foi constituida por 4 elementos do PM/E e foi dividida em dois grupos: 1) científico-pedagógico (cp) 2) técnico (tec). O grupo científico-pedagógico teve a colaboração dos estagiários e alunos do seminário e da orientadora (Escola) do núcleo de estágio.

O grupo técnico teve o apoio durante cerca de um mês de um investigador do Dep. de Electrónica e Telecomunicações, ao tempo monitor no Dep. de Matemática.

Os testes

O formato. Em algumas sessões semanais foram apresentados e discutidos os objectivos da experiência e o formato do teste-modelo.

O tipo genérico estava pre-determinado - deveria ser um teste com questões fechadas, objectivo e de correcção simples. Propus uma variante do chamado "**teste americano**" composta apenas por uma parte de questões do tipo **falso/verdadeiro** e outra parte com questões de **resposta múltipla**. As duas partes deveriam ter algumas questões com objectivos análogos (questões emparelhadas), embora com formulações diferentes.

Dificuldades técnicas e o pouco tempo disponível levaram a optar por um formato de compromisso apenas com questões de resposta múltipla. Cada questão passou a ser composta de quatro subquestões podendo ser todas verdadeiras, todas falsas ou algumas falsas e as outras verdadeiras.

Questões deste formato são designadas por alguns autores por questões do tipo **Falso-Verdadeiro Generalizado** (FVG) [4].

Foi decidido que as questões seriam geradas aleatoriamente por modelos a construir por Paula Carvalho.

A cotação. Como a experiência se iria realizar com crianças do 7º ano de escolaridade (12-13 anos) e dado que cada um teria um teste diferente, embora de nível equivalente, propus uma tabela classificatória pouco penalizante da resposta á sorte. Teve-se ainda em conta que o teste era de tipo formativo e a acção de informação directa dada aos alunos pelos professores.

Os modelos. Este assunto vai ser objecto de uma comunicação independente [2] pelo que não me alongarei. Direi apenas que este trabalho, de importância capital no desenvolvimento do projecto, foi precedido de todo um trabalho de recolha de material, seleccão, levantamento dos erros mais frequentes nas matérias em questão e graduação por níveis de dificuldade.

Geração e aplicação em sala de aula. Criados os modelos a equipa técnica estruturou a **BD** por forma a gerar testes nas condições definidas - um teste de pre-requisitos (dividido em dois TF-N1 e TF-N2 devido á dificuldade de o aplicar numa só aula de 50 minutos) e um teste de equações.

A equipa enfrentou e superou várias dificuldades devidas às limitações dos suportes informáticos utilizados - DOS e CLIPPER- e à exiguidade do tempo.

A necessidade de não haver qualquer perturbação na planificação das aulas já efectuada e a escassez de meios informáticos da Escola inviabilizaram, nesta primeira fase, a ideia original que era a do acesso do aluno ao computador para a realização da avaliação formativa sobre as unidades em causa. Os testes foram impressos e distribuidos aos alunos para resolução. Houve uma preparação prévia dos alunos para o tipo de teste que iriam fazer e para a necessidade de se evitar a resposta à sorte. Anexo a cada teste estava uma folha de instruções e a tabela de classificação [Anexo -1]. No final da unidade "equações 1º grau em Q" foi realizado o teste (formativo) sobre esta matéria.

O interesse despertado por este tipo de trabalho foi enorme, mas os resultados em termos de avaliação formativa foram diminutos.

Dadas as condições de desenvolvimento da experiência, na altura da aplicação dos testes os objectivos fundamentais estavam já

reduzidos a testar o material produzido e o interesse dos alunos. Estes objectivos foram plenamente conseguidos. As informações e os dados recolhidos são muito importantes para o prosseguimento da experiência.

Apesar dos cuidados havidos não foi possível eliminar a resposta à sorte e a troca de informações com o colega de carteira. Houve igualmente muito de lúdico e de desafio perante uma experiência que faziam pela primeira vez (nós preocupados com o tempo e no TF-N1 alunos houve que entregaram o teste ao fim de 15 minutos...!).

A análise dos dados recolhidos deve ser feita com todos os cuidados e tendo em conta as condições de realização da experiência.

A comparação destes dados com os níveis atingidos nos periodos anteriores ou com os testes sumativos relativos à mesma matéria não permite tirar grandes conclusões, dada a insuficiência da população. No entanto os relatórios referidos na bibliografia fazem já um primeiro tratamento a estes dados e fazem considerações pertinentes.

Teremos de concluir que o que se fez foi apenas tempo perdido?

De modo nenhum. O trabalho está em curso, alguns aspectos, nomeadamente a cotação, estão a ser reformulados e pretende-se instalar a tempo a BD para ser testada efectivamente no seu objectivo fundamental-ser um instrumento interessante e útil de AF. O trabalho efectuado permitiu, além disso testar, algumas das críticas feitas aos testes de resposta múltipla. Sobre este aspecto farei considerações mais adiante.

Do trabalho realizado saíu igualmente a ideia da organização de uma competição matemática, **CM1**, com matemáticas escolares, que envolveu cerca de duzentos alunos (7 turmas) e que foi um sucesso[1]

Estrutura: vantagens e desvantagens. Como já referi, a estrutura estava praticamente pre-determinada à partida: o teste deveria ser objectivo e de correcção rápida (no nosso caso automatizada).

Testes com o formato escolhido têm vantagens e desvantagens que importa conhecer para explorar as primeiras e miminizar as segundas. É claro que grande parte das objecções levantadas tem a

[1]Foi feito um vídeo que se apresentará durante o Encontro.

ver com a utilização dos testes de escolha ou resposta múltipla para avaliação quantitativa. Há obras sobre estratégias para ter êxito em tais testes. O domínio dos conhecimentos foi relegado para segundo (só?) plano, houve uma perversão completa dos objectivos da avaliação enquanto parte inseparável do processo de ensino-aprendizagem.

Não é esse o caso tratando-se de avaliação de tipo formativo. Mesmo assim convem saber quais as principais vantagens e desvantagens habitualmente apontadas:

Desvantagens

D_1) tempo gasto na elaboração;

D_2) não permite avaliar o desenvolvimento de um raciocínio;

D_3) não permite avaliar a capacidade de expressão escrita;

D_4) permite a fraude com facilidade;

D_5) favorece a "resposta á sorte"

Vantagens

V_1) questões claras;

V_2) respostas breves;

V_3) correcção simples e objectiva;

V_4) permite a detecção (auto ou não)de dificuldades e ou bloqueios no acompanhamento da unidade didáctica.

V_5) envolve dois tipos de rectroacção-sobre o aluno e sobre o professor;

Sendo as vantagens óbvias farei apenas algumas considerações sobre as desvantagens.

Todas as desvantagens apontadas são pertinentes e levantam limitações sérias quanto à sua utilização como instrumento dominante de avaliação sumativa. O seu papel deve ser complementar relativamente a outros instrumentos de avaliação. É esta a posição de Vandevelde (1971). No que respeita à AF penso que é um importante instrumento de trabalho. Mesmo que a falta de meios informáticos de

uma escola não permita tirar o máximo proveito deste tipo de teste impedindo, por exemplo, a detecção quase imediata de certas dificuldades do andamento do processo de ensino-aprendizagem, tal detecção será feita inevitávelmente na confrontação com os resultados do correspondente teste sumativo.

A geração aleatória de testes e um sistema de cotação penalizador da resposta à sorte são meios de dificultar D_4 e D_5.

As desvantagens apontadas têm a ver com o formato do teste - **QRM**. Vejamos agora uma crítica relativa á estrutura deste tipo de testes.

Uma crítica de estrutura - SKinner e as soluções erradas.

Muitas das questões de um **QRM** são estruturadas tendo por base os erros dos alunos. Diz SKinner que "*toda a solução falsa, num teste de escolha múltipla, aumenta a probabilidade de que um estudante num dado momento reproduza a resposta errada em vez da resposta certa*"[2]

Experiências feitas por alguns investigadores-Preston (1965) e KarraKer (1967) parecem confirmar a opinião de SKinner[2].

A minha experiência de muitos anos de ensino, valorizando a acção formativa e explorando, sempre que possível, o erro com fins formativos "a aprendizagem pelo erro", leva-me a pensar que SKinner, globalmente, não tem razão. Numa fase posterior da experiência em curso espero poder apresentar dados que confirmem esta opinião.

A consideração de respostas falsas deve ser aproveitada para desenvolver o sentido crítico do aluno face a uma questão. Saber efectivamente o que um dado objecto matemático **é** implica saber o que ele **não é**. Saber **o que não é** já revela algum conhecimento sobre o objecto em causa. **Exemplo**: Considere-se a seguinte questão de um teste do tipo FVG:

Sendo $x \in [0,2]$, $-x|$ é igual a:

(i) $|x|$ (ii) $-|x|$ (iii) x (iv) $-x$

[2]Referências tiradas de [4].

Se um aluno deixa em branco (i) e (iii) e regista como falsas (ii) e (iv) ele revela conhecer a não negatividade da função módulo.

Fiabilidade. Numa experiência deste tipo impõe-se estudar o seu grau de fiabilidade. Este será medido em primeiro lugar pela fiabilidade dos testes - estabilidade e e consistência interna. Presentemente não temos ainda dados sificientes para que os resultados sejam significativos.

Um critério de fiabilidade de um teste é que a maioria das subquestões tenha um índice de discriminação (ID) alto.

Dada a insuficiência de dados já referida não farei uma análise de dados fina.

Limitar-me-ei a fazer certas considerações que julgo pertinentes a partir dos valores do índice de discriminação das subquestões q_{ij} do TF-N1 (ver tabela 1 e gráficos nas páginas seguintes).

Este estudo foi feito sobre os 114 alunos que fizeram os dois testes formativos dos pre-requisitos.

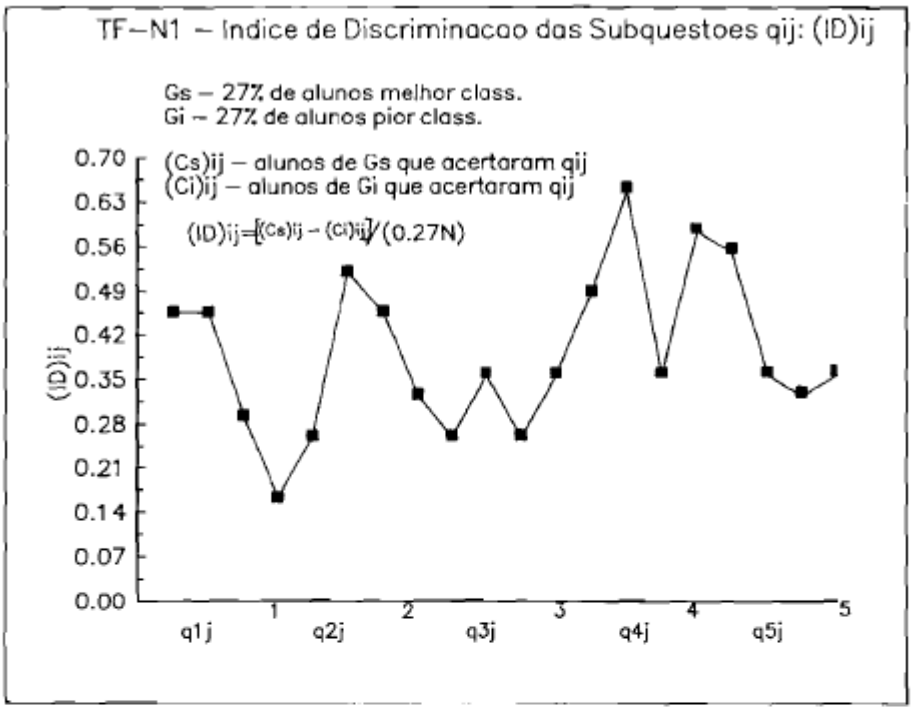
Ora para uma população de menos de 300 elementos os resultados não devem ser tomados como definitivos. Para uma população pequena o ID é muito sensível à influência do acaso. Mesmo assim, julgo que os valores encontrados para o ID nos permitem apontar para a fiabilidade do TF-N1.

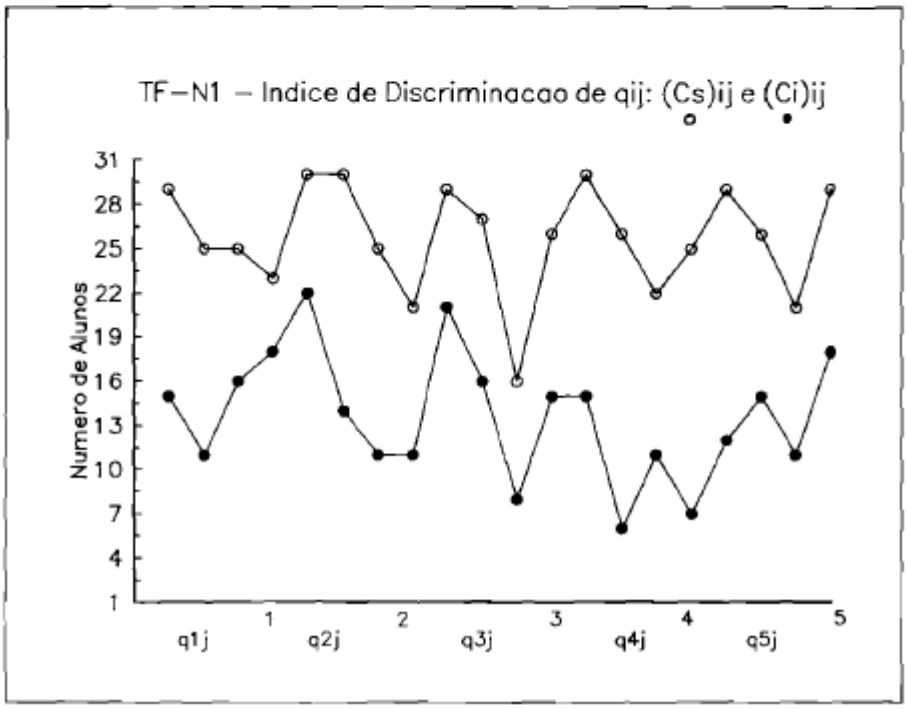
De um modo geral uma questão q_{ij} é considerada **satisfatória** ou **não satisfatória** consoante se tenha $ID_{ij}>0,40$ ou $ID_{ij}<0,20$ respectivamente.

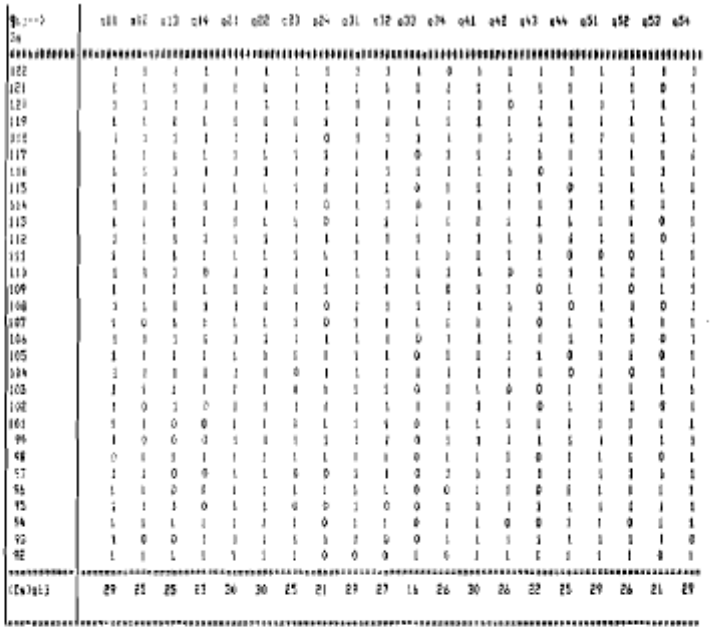
Uma questão com ID **negativo** afecta a fiabilidade do teste (indicia muitas respostas à sorte). Isto é a teoria geral, mas há que ter muito cuidado a tirar conclusões.

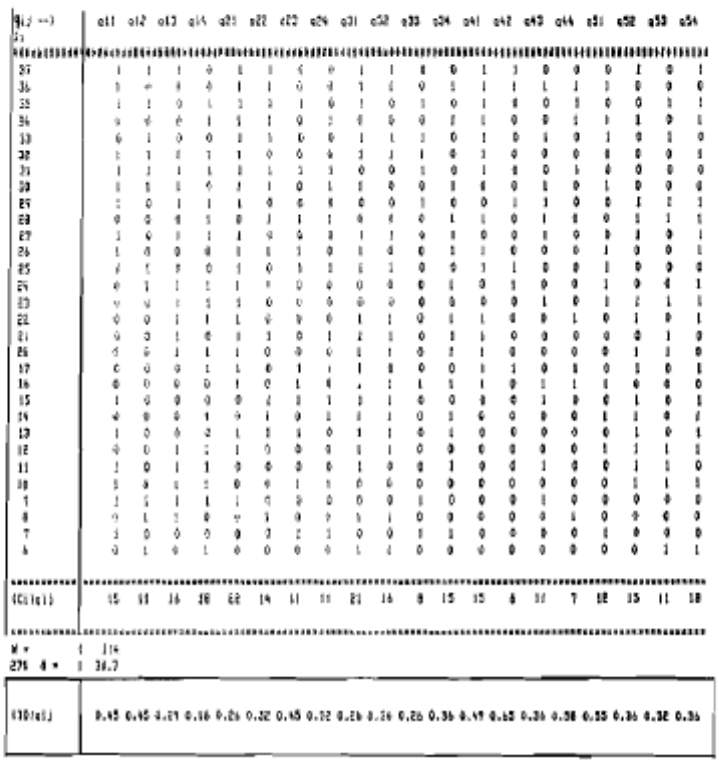
De facto um dado conhecimento básico pode ter sido adquirido por quase toda a população escolar (e por que não toda?) e nesse caso uma questão que procurasse testar a sua aquisição teria um índice de discriminação próximo do zero ou mesmo zero. Parece ser essa a interpretação para ID_{14} =0,17 . Com efeito q_{14} é uma questão gerada por

$(-a)^2$ é igual a : d) a^2 (**a** inteiro não nulo e **a** entre **-9** e **9**)


TF–N1 – Indice de Discriminacao das Subquestoes qij: (ID)ij
Gs – 27% de alunos melhor class.
Gi – 27% de alunos pior class.
(Cs)ij – alunos de Gs que acertaram qij
(Ci)ij – alunos de Gi que acertaram qij
(ID)ij=[(Cs)ij – (Ci)ij]/(0.27N)
(ID)ij
0.70
0.63
0.56
0.49
0.42
0.35
0.28
0.21
0.14
0.07
0.00
1
2
3
4
5
q1j
q2j
q3j
q4j
q5j


TF–N1 – Indice de Discriminacao de qij: (Cs)ij e (Ci)ij
Numero de Alunos
31
28
25
22
19
16
13
10
7
4
1
1
2
3
4
5
q1j
q2j
q3j
q4j
q5j





Ora é razoável aceitar que ao fim de quatro meses de aulas este conhecimento esteja adquirido pela maioria dos alunos. E se uma dada questão tiver ID<0? Elimina-se de imediato? Não, sem previamente analisar bem quer a questão e seus objectivos, quer as causas que determinaram que os alunos de Gi tenham respondido melhor do que os de Gs. Recordo-me de um exemplo que poderia muito bem conduzir a um resultado deste tipo referido na AMC,1988- uma percentagem razoável de alunos muito bons falharam questões iniciais muito simples que os alunos mais fracos acertaram. Razão apontada: aqueles alunos mal reflectiram naquelas questões para guardarem o tempo para as questões mais difíceis (que acertaram).

Acresce no nosso caso o facto de se tratar de testes de avaliação formativa o que permite estudar melhor este tipo de situações.

Na reavaliação dos micro-objectivos das subquestões geradas pelos modelos vamos ter em atenção entre vários instrumentos teóricos, a **teoria dos conhecimentos locais**.

Por exemplo, o teste nº 123 - Nível 1 incluia a seguinte subquestão

$|6+1|$ é igual a: c) $|6|+|1|$ d) $6+1$

O aluno respondeu c) F d) V

A resposta a c) não poderá ser devida a um conhecimento correcto - o módulo da soma nem sempre é igual á soma dos módulos? Do mesmo modo uma resposta correcta em c) não poderia ser consequência de um conhecimento errado - o módulo da soma é igual á soma dos módulos? E em d) poderemos tirar conclusões definitivas sem saber como é que o aluno responderia á questão: $|1-6|$ é igual a: d) $1-6$ [ ] ?

A teoria dos conhecimentos locais chama a atenção do professor e do investigador para a qualidade da resposta - não basta que a resposta a uma questão esteja certa, é preciso detectar se essa resposta **resulta de conhecimentos correctos**. A avaliação formativa, repito-o ainda, tem um papel insubstituível neste campo. Quando os dados recolhidos forem suficientes procederemos então a uma análise de dados fina que permita validar os modelos e dar informações para o seu aperfeiçoamento.

Experiência - Fase 2

Nesta segunda fase, que já está em curso, pretende-se:

1) continuar a recolha de dados para aferir a BD construída na Fase-1;

2) aplicar a BD nas turmas do 8º ano da Escola Secundária nº1 constituídas dominantemente por alunos que participaram na experiência no seu 7º ano; esta aplicação tem intuito diagnóstico e serve para recolha de dados a comparar com os obtidos na primeira aplicação;

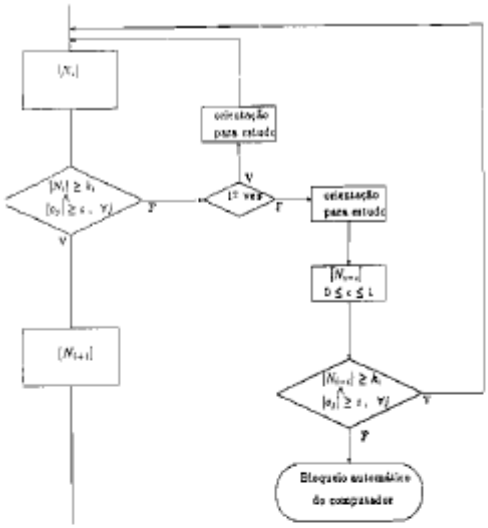
3) testar novo sistema de cotação, desta vez mais penalizante da resposta á sorte: **1** ponto para resposta certa, **0** pontos para omissão de resposta e **-1** para resposta errada por subquestão.

4) com base nos dados já recolhidos preparar a BD para ser utilizada, se possível já a partir de Janeiro próximo, como instrumento de avaliação formativa em turmas do 7º ano de escolaridade de várias escolas secundárias onde funcionam núcleos de estágio. Pretende-se que a aplicação da BD à avaliação formativa seja feita segundo o esquema da fig. 1.

O ajustamento automático[3] do nível dos testes será feito por um algorítmo "inteligente" baseado em certas funções distância a definir no espaço das respostas e em funções de selecção de questões previamente graduadas. Este trabalho de graduação das subquestões e das questões, quer globalmente, quer por objectivos, é fundamental para o êxito da experiência. A aplicação, que esperamos poder fazer já este ano, faz parte da fase experimental para aferição do modelo agora em construção.

NOTA: Durante o desenvolvimento da unidade didáctica o modelo deve ser utilizado seguindo todas as fases, do nível mais baixo ao mais elevado. Para efeitos de recapitulação de matéria deve ser permitido ao aluno iniciar a utilização da BD onde

[3] A ideia deste tratamento foi sugerida pela leitura do artigo "A nearest hyperrectangle learning method" S.Salzberg, Machine Learnig, 6 (1991). O método referido insere-se na teoria da aprendizagem por exemplos, com geração de excepções e com retenção selectiva (um exemplo já memorizado - um hiperrectangulo num espaço euclidiano n-dimensional, onde n é o número de variaveis de cada exemplo - é substituido por outro melhor).
O Programa EACH (Exemplar-Aided Constructor of Hyperrectangles) referido no artigo utiliza varias funções distância e ajusta-as consoante a situação apresentada.



Fluxograma a implementar para o controlo dinâmico do 'Feed-Back'

$[N_i]$ - classe de testes de nível N_i gerados por um modelo ou por modelos equivalentes; $|Ni|$ - cotação de um teste de nível N_i; K_i - cotação mínima exigida para a passagem de um teste N_i a um teste N_{i+1}; c - cotação mínima exigida para cada objectivo

Bibliografia

[1] Castro, L.F. , Relatório de seminário, 1991
[2] Carvalho, M.P. , Avaliação formativa por computador, 1991
[3] Landshere, G .- A investigação experimental em pedagogia, D Quixote, 1986
[4] LECLERCQ, D.- La conception des questions á choix multiples,a Ed. Labor, Buxelas, 1986
[5] Léonard, F e Sackur, C.- Connaissances locales et triple approche, une méthodologie de recherche, Rech. en did. des math., vol.10/2.3,1991
[6] Morisette, D. -La mesure et l'évaluation en enseignement, Les Presses de l'Univ. Laval,1984
[7] Noizet, G.e Caverni,J.- Psychologie de l'évaluation scolaire, PUF,1978
[8] Salzlerg,S.- A nearest hyperrectangle learning method, Machine Learning,6, 1991
[9] Vergnaud, G.- La theórie des champs conceptuels, Recherches en didactique des mathématiques, vol.10/2.3,1991

Grupo de trabalho — Avaliação formativa no 7º ano

Coordenador - João C. David Vieira, [c.p.];M. Paula Carvalho, PM/E [c.p.] [tec.]; A. Batel Anjo, PM/E [tec.]; Querubim Terra, PM/E Seminário [tec]; Amaro de Sousa - estudante de pos-graduação do Dep. ET. [tec.] (colaboração temporária): Colaboradores: M. Teresa Neto — orientadora (Escola) do núcleo de estágio — e Estagiários [núcleo de estágio da Esc. Sec. nº1 de Aveiro]