

# Avaliação e Classificação da Aprendizagem em Química

– alguns aspectos técnicos

M. Arminda Pedrosa <sup>a</sup>



Arminda Pedrosa

Nasceu a 25 de Março de 1950 em Vilar de Figos, Barcelos. Completou na Universidade de Coimbra a licenciatura em Química, ramo científico (Química-Física) em 1974 e o Mestrado em Química-Física em 1985. Obteve o PhD na Universidade de East Anglia (Reino Unido) em Julho, de 1988, com a tese «The Use of Oral Assessment in Chemistry»; foi-lhe concedida a equivalência ao grau de Doutor em Química, especialidade de Educação em Química em Março de 1989. Exerce funções docentes na área de Química desde 1974, fundamentalmente na Universidade de Coimbra, tendo de 1979 a 1981 desenvolvido essa actividade na mesma área na Escola de formação de Professores para o Ensino Secundário, na República de Cabo Verde. É presentemente Professora Auxiliar no Departamento de Química da Faculdade de Ciências e Tecnologia da Universidade de Coimbra. Tem orientado acções de formação de professores dos Ensinos Básico e Secundário, apresentado comunicações em conferências nacionais e internacionais e é autora/co-autora de artigos em publicações nacionais e estrangeiras.

*Neste artigo referem-se sumariamente os objectivos educacionais da avaliação da aprendizagem e a utilização das classificações dos estudantes com fins diversos dos estritamente educacionais. Considera-se as classificações dos estudantes, em termos da sua aprendizagem no sistema de ensino formal, como uma medida do seu grau de sucesso na consecução dos objectivos gerais e específicos dos diferentes programas de ensino de disciplinas de química integradas em programas de ensino desta disciplina nos vários níveis do sistema de ensino em Portugal, básico, secundário e superior. Discute-se as componentes sumativa e formativa da avaliação da aprendizagem e os seus pesos relativos em sistemas de ensino em que os fins sociais da avaliação têm uma importância considerável. Definem-se e discutem-se as características de rigor a que devem obedecer as classificações e referem-se alguns métodos de estatística descritiva aplicáveis no tratamento da informação relativa a resultados de testes e exames de grupos de estudantes - turmas, anos de escolaridade, etc. Faz-se uma reflexão sobre as origens dos erros das classificações atribuídas aos estudantes em testes de avaliação da aprendizagem e sugerem-se procedimentos a adoptar com o objectivo de diminuir a componente de erro das classificações. Definem-se o erro de medida padrão das classificações obtidas por um grupo particular de estudantes num dado teste e o erro de medida padrão para cada classificação individual. Apresenta-se informação que permite determinar em que condições é que duas classificações diferindo mais do que o erro de medida padrão correspondem de facto a classificações verdadeiras diferentes ou se pelo contrário se pode considerar que a diferença surge por acaso.*

## Introdução

Muitos dos conceitos que aqui se introduzem e discutem aplicam-se a situações de avaliação da aprendizagem decorrente do ensino formal das diversas disciplinas. Contudo, sempre que possível utilizam-se contextos e exemplos do domínio da Química. A natureza deste trabalho impõe a introdução de conceitos fundamentais ao tratamento da informação resultante das classificações dos testes na perspectiva do rigor necessário na classificação dos estudantes, mas não permite uma discussão tão aprofundada quanto se desejaria pelo que se aconselha com esse propósito a bibliografia pertinente utilizada e indicada.

<sup>a</sup> Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade de Coimbra, 3000 Coimbra - Portugal.

Independentemente da maior ou menor formalidade das avaliações da aprendizagem e da responsabilidade pela sua concepção e execução, estas existem sob formas e designações diversas: participação nas aulas, testes escritos administrados regularmente ao longo do ano lectivo e exames. Importa afinal recolher dados que permitam classificar cada estudante numa escala de valores, actualmente de 1-5 ou 0-20 dependendo do nível de ensino, conforme o seu desempenho em tarefas requeridas nos instrumentos de avaliação da aprendizagem, de ora em diante designados simplesmente por testes. Assim o carácter sumativo da avaliação domina sobre o formativo e a classificação dos estudantes é um objectivo perseguido pelos inúmeros testes de química aplicados por cada professor no decurso de um ano lectivo. A posição proeminente dos exames e a importância atribuída em geral às classificações obtidas pelos estudantes constituem indicadores da preponderância da componente sumativa da avaliação da aprendizagem sobre a formativa. Este estado de coisas resulta, com certeza, de a componente sumativa não só responder a preocupações educacionais como também servir propósitos diversos resultantes do papel social associado à maioria dos exames públicos e profissionais. O uso de exames e seus resultados na definição de instrumentos de selecção no acesso ao Ensino Superior, entrada no mercado de trabalho e admissão em certas carreiras profissionais, ilustra bem este aspecto social associado aos exames e às classificações dos estudantes nas várias disciplinas nomeadamente em Química.

De facto, surgem pontos de vista porventura polémicos relativamente a esta temática havendo quem defenda que, seja qual for o nível de ensino formal, os exames são, para a maioria das pessoas, exercícios de seriação dos estudantes de modo a ordená-los numa escala de valores; a posição na escala exprime o nível de desempenho que é apenas revelador da capacidade de reproduzir conhecimento em condições artificiais e dentro de uma gama estreita de critérios [10] e não é revelador de capacidades de desempenho na vida activa como profissionais. O mesmo autor considera que a maioria das pessoas entende e espera que os exames respondam à necessidade que os sistemas de educação formal têm de arranjar uma base para selecção e de fornecer alguma indicação de potencial para futuros desempenhos, embora considere, baseado em resultados de investigação, que não existe evidência que apoie o ponto de vista de que os resultados dos exames constituem bons indicadores de desempenhos profissionais.

Entende-se, contudo, que dos vários testes feitos pelos professores de Química ao longo de um programa de ensino resultam aspectos importantes de avaliação formativa, para além da sumativa já referida, dos quais se destacam informações relativas ao ritmo imprimido às aulas e possíveis dificuldades de acompanhamento na incorporação e assimilação de novos conhecimentos pelos estudantes, a áreas de conteúdo de difícil compreensão, a concepções alternativas dos estudantes [6], [2] e [7], aspectos estes que podem constituir barreiras ao desejável progresso da aprendizagem. Todos os aspectos mencionados podem ser considerados o «feedback» indispensável ao professor para proceder às necessárias alterações, que para professores dos ensinos básico e secundário se traduzirão em modificações da planificação do ensino,

nomeadamente no que respeita a estratégias de ensino e materiais utilizados e aconselhados aos estudantes; para professores do ensino superior, a quem é concedida maior autonomia pedagógica, as alterações resultantes do «feedback» das classificações dos testes poderão incluir modificações ao programa da disciplina e ao desenvolvimento inicialmente previstos, para além das já mencionadas para professores de outros níveis de ensino.

Igualmente importante para uma aprendizagem eficiente e activa é a avaliação e os seus resultados como fonte de «feedback» para os estudantes que podem assim alterar os meios e as estratégias de aprendizagem no sentido de melhor responderem às exigências do ensino formal, por forma a obterem o sucesso desejado de acordo com as capacidades, aspirações e ambições individuais. Assim, os testes desempenham um papel preponderante na definição do que é mais ou menos importante apreender e aprender para o sucesso escolar. Na verdade, o trabalho individual dos estudantes no seu esforço de aprendizagem fora das aulas é mais orientado pelas tarefas incluídas em testes da mesma disciplina utilizados para avaliação da aprendizagem em anos anteriores do que pelos objectivos específicos das disciplinas de química, seus conteúdos programáticos ou sumários das aulas.

Tendo em conta a importância de testes e exames de Química, nas dimensões educacional e social bem como nas dimensões formativa e sumativa, importa que a concepção de instrumentos de avaliação seja cuidada e os resultados da sua administração aos estudantes sejam tratados com rigor.

Porque classificar estudantes em Química (ou outra disciplina) se traduz, para cada estudante da população submetida a testes, na atribuição de um número numa escala de valores, pode considerar-se que se trata de um acto de medir. Esta medida é com certeza de natureza diversa e requer a utilização de processos diferentes dos das medidas de grandezas físicas com que a população de professores de Química está familiarizada e habituada a designar deste modo, como sejam medidas de massa, de tempo, de quantidade de substância, de concentração, etc. No entanto, podem estabelecer-se analogias entre os dois tipos de medidas. Assim, exactidão e precisão são requisitos exigidos às medidas de grandezas físicas; a exactidão refere-se à proximidade de uma dada medida de um padrão aceite ou imposto externamente e a precisão refere-se à consistência interna de uma série de medidas, ou seja, refere-se à reproducibilidade dos dados obtidos nos actos de medir. Como exemplo considere-se os valores da concentração de uma solução obtidos nos vários ensaios numa titulação volumétrica [3]. A concordância dos valores das concentrações calculadas para as várias titulações realizadas constitui evidência da reproducibilidade ou precisão dos resultados das várias titulações da mesma solução; contudo, para que os valores da concentração obtidos sejam considerados exactos require-se que, para além de serem concordantes, os volumes registados sejam exactos em termos da unidade de volume, que, neste caso, depende da calibração da bureta.

As características de rigor exigidas a medidas educacionais são a validade e a confiança correspondentes respectivamente às características de exactidão e precisão exigidas a medidas de grandezas físicas.



## Validade de testes de avaliação da aprendizagem em Química

A validade segundo Lindquist [4] é a exactidão com que se mede aquilo que se pretende medir, i.e., o grau de proximidade da infalibilidade ao medir o que se tem intenção de medir. A utilização de grelhas de especificações [3, 1] na concepção e elaboração de testes de avaliação da aprendizagem constituem um meio de reflexão precioso que conduz à categorização das diferentes tarefas que constituem o teste, em termos das capacidades requeridas e das actividades de conteúdo programático em que se pretende ver demonstradas essas capacidades. É nosso entender que para disciplinas de química, de um modo geral, será suficiente considerar-se as seguintes capacidades: memorização, compreensão, aplicação, análise e síntese e resolução de problemas. Esta classificação das tarefas em categorias definidas nas dimensões de capacidades e actividades de conteúdo é importante para ajuizar sobretudo da validade de conteúdo e da validade de construção («construct») de testes administrados regularmente como meio de obter resultados para classificar estudantes. Outros tipos de validade, como a de previsão e a concorrente são de considerar noutras situações mas não a testes considerados isoladamente. Na situação a que nos referimos de testes normalmente concebidos para avaliação periódica da aprendizagem em disciplinas de química, a determinação das validades de conteúdo e de construção pode fazer-se recorrendo à análise e julgamento dos itens do teste por colegas (de grupo nas escolas preparatórias e secundárias) - mecanismo de validade directa. Os itens do teste serão classificados por colegas docentes de Química no que diz respeito ao grau com que as tarefas requeridas nos diferentes itens reflectem os objectivos gerais e específicos da disciplina - validade de conteúdo. O mesmo grupo de colegas pode, simultaneamente, manifestar a sua concordância ou discordância da classificação dos itens expressa na grelha de especificações no que concerne às capacidades requeridas - validade de construção.

## Confiança nas classificações dos testes

Confiança exprime a concordância ou consistência de uma série de classificações - «notas» - atribuídas a uma qualidade particular normalmente escolhida com base em critérios de importância para o desenvolvimento de capacidades requeridas como critérios de competência na disciplina. Em Química a compreensão de leis e modelos explicativos de fenómenos assim como a resolução de problemas são certamente qualidades julgadas como relevantes para classificar a competência na disciplina nos vários níveis de ensino formal, particularmente no ensino secundário e superior.

A determinação da confiança de um instrumento de avaliação ou teste requer que se proceda a múltiplas medidas de onde se possa concluir da concordância ou não dos resultados. Na prática isto é de difícil execução, é contudo possível estimar-se a fidelidade de um teste ou exame por recurso a técnicas apropriadas. O coeficiente de confiança é o coeficiente de correlação entre uma série de classificações obtidas pelos estudantes no teste em questão e outra série de classificações obtidas pelo mesmo grupo de estudantes num teste

equivalente. A definição do coeficiente de confiança resulta do reconhecimento de que para cada teste e estudante particulares a classificação obtida é composta de duas componentes: a verdadeira (ou hipotética, porque não mensurável na prática) - obtida num teste de confiança total e o erro, que é devido ao grau de não confiança do teste. Considerando agora o conjunto das classificações obtidas pelos estudantes no teste pode fazer-se idêntico raciocínio para a variância, i.e., a variância total é a soma da variância verdadeira com a variância do erro:  $V_{\text{total}} = V_{\text{verdadeira}} + V_{\text{erro}}$ . Recorde-se que a variância é a média da soma dos quadrados dos desvios da média da amostra:

$$V = s^2 = \frac{\sum (x_i - \bar{x})^2}{N} \quad s = \text{desvio padrão.}$$

Existe sempre um erro inerente ao acto de medir, quer em medidas de grandezas físicas quer nas classificações dos estudantes em testes, consequentemente, a componente  $V_{\text{erro}}$  nunca é desprezável; tendo em atenção as características particulares das classificações em discussão aquela componente é, para o mesmo grupo de estudantes, variável de teste para teste, e para o mesmo teste variável de grupo para grupo de estudantes (e.g. turmas ou anos de escolaridade em diferentes escolas).

O coeficiente de confiança de um teste aplicado a um dado grupo de estudantes é o coeficiente de correlação entre uma série de classificações obtidas pelos estudantes no teste e uma série de classificações obtidas pelos mesmos estudantes num teste equivalente. O coeficiente de correlação,  $r$ , mede a extensão da concordância entre as classificações reais obtidas num teste particular e as classificações verdadeiras (classificações hipotéticas que seriam obtidas se o mesmo teste fosse de confiança total). O coeficiente de correlação [8],  $r = V_{\text{verdadeira}} / V_{\text{total}}$ , pode assumir quaisquer valores entre -1,0 e 1,0;  $r=0$  significa que não há qualquer relação entre as duas séries de classificações,  $r=1,0$  indica um acordo perfeito entre as duas séries de classificações e valores negativos de  $r$  ( $-1,0 < r < 0,0$ ) indicam a existência de uma relação inversa entre as duas séries de classificações.

Deve notar-se que a estimativa da confiança de um teste, quando aplicado a um grupo particular de estudantes requer a disponibilização de resultados de pelo menos duas administrações de testes equivalentes a esse grupo de estudantes. Este requisito é difícil de parantir porque a preparação de versões equivalentes de um teste é morosa e difícil e a duplicação da administração de testes é também um processo moroso e apresenta problemas de confiança resultantes das diferentes condições de aprendizagem no momento das duas administrações. A resolução deste problema é possível por recurso a métodos diferenciados cada um deles com as suas limitações próprias decorrentes dos processos utilizados característicos de cada um. Os métodos descritos na literatura [3] e [8] são: método das versões equivalentes do teste (referido na definição de  $r$ ), método de testar e retestar, método da divisão do teste em duas partes equivalentes e método da consistência interna.

O método das versões equivalentes presume, como o próprio nome indica, a preparação de duas versões paralelas e equivalentes do mesmo teste, a duplicação da sua administração ao mesmo grupo de estudantes e a classificação das tarefas

desempenhadas pelos estudantes das duas administrações das versões do teste. A qualidade das estimativas de confiança obtidas por este método depende sobretudo da capacidade de quem prepara o teste para produzir duas versões genuinamente equivalentes do mesmo teste.

No método de testar e retestar o mesmo teste é administrado ao mesmo grupo de estudantes em dois tempos diferentes suficientemente separados para reduzir o efeito de memorização das tarefas da primeira para a segunda administração. O coeficiente de confiança é dado pela correlação entre as classificações obtidas nas duas administrações do mesmo teste. Note-se que, neste método, a estimativa da confiança pode ser inflacionada pela memorização das tarefas e por alteração dos conhecimentos dos estudantes da primeira para a segunda administração do teste.

No método de divisão do teste em duas partes equivalentes o coeficiente de confiança é calculado a partir das classificações obtidas nas duas partes julgadas equivalentes do teste. A dificuldade na preparação de um teste deste tipo é idêntica à apontada para preparar duas versões equivalentes do mesmo teste. O coeficiente de confiança calculado com as classificações das metades do teste,  $r_m$ , tem que ser corrigido para se obter o coeficiente de confiança do teste na sua globalidade,  $r = 2r_m / (1 + r_m)$ .

O método de consistência interna destina-se a examinar a consistência interna ou equivalência dos itens que constituem um teste. Este método permite estimar a homogeneidade de um teste e aplica-se somente a testes homogêneos, i. e. testes em que todos os itens, em número elevado, estimam a mesma característica da aprendizagem. Por esta razão é um método utilizado em investigação educacional mas não em situações correntes de avaliação da aprendizagem.

### Origem dos erros das classificações de testes de avaliação da aprendizagem

A componente de erro nas classificações atribuídas aos estudantes pode atribuir-se a três origens distintas [3]:

1. Aos estudantes: alterações de motivação para o desempenho das tarefas requeridas no teste, de velocidade de execução das tarefas, de atenção, etc., relativamente a situações de vida escolar corrente em que a componente avaliação sumativa não é percebida pelos estudantes.
2. Às tarefas incluídas no teste: diferenças nas tarefas julgadas como equivalentes (p. ex. em termos das capacidades que testam) ou escolhidas pelos estudantes, a formulação das perguntas ou o conteúdo versado para a demonstração de capacidades particulares pode também constituir uma fonte de erro.
3. Ao professor que classifica as respostas dos estudantes: diferenças resultantes de flutuações nos padrões e objectividade utilizados nas classificações dos testes.

### Procedimentos a adoptar para aumentar a confiança das classificações de testes

As medidas tendentes a aumentar a confiança em testes são aquelas que tendem a diminuir os erros das classificações atribuídas aos estudantes em testes de avaliação da aprendizagem. Considerando as origens dos erros destas classifica-

ções explicitadas anteriormente, podem fazer-se algumas recomendações simples que conduzirão ao aumento do coeficiente de confiança, ou seja, à diminuição da componente de erro das classificações de testes de avaliação da aprendizagem [3]:

1. Dar tempo suficiente para que todos os alunos completem o teste. Deste modo reduz-se a ansiedade e evita-se que diferentes estudantes optem pela execução de tarefas distintas muito provavelmente não equivalentes.
2. Certificar-se que o teste tem um número suficiente de tarefas destinadas a medir uma qualidade ou característica particular - aumento de confiança por aumento da extensão. Deve notar-se que embora de um modo geral o coeficiente de confiança aumente com o aumento da extensão do teste, a relação entre estas duas variações não é linear; quanto mais elevado for o coeficiente de confiança de um teste menor será o aumento deste coeficiente provocado pelo aumento do número de itens do teste.
3. Eliminar perguntas de opção e evitar ambiguidades na formulação das tarefas.
4. Utilizar grelhas de especificações [3] e [1] e esquemas de classificação previamente estabelecidos; evitar a classificação por impressão ou palpite.

### Erro de medida padrão

Na teoria da confiança este conceito será o mais útil para o professor. Partindo da expressão de coeficiente de confiança,  $r$ , definida anteriormente e tendo em conta que  $V_{\text{verdadeira}} = V_{\text{total}} - V_{\text{erro}}$ , obtém-se outra expressão para  $r = 1 - V_{\text{erro}} / V_{\text{total}}$ .

O erro de medida padrão [8] é o desvio padrão do erro e pode calcular-se a partir da expressão anterior  $V_{\text{erro}} = V_{\text{total}}(1 - r)$ . Tendo em conta a relação entre a variância e o desvio padrão resulta  $s_{\text{erro}} = s_{\text{total}} \sqrt{1 - r}$ . Se os desvios padrão das classificações obtidas pelo mesmo grupo de estudantes em dois testes diferentes forem aproximadamente iguais então o teste com

coeficiente de confiança mais elevado constitui uma estimativa melhor da classificação verdadeira de um estudante. Pode estimar-se a classificação verdadeira de um estudante  $i$  desde que se conheça a média das classificações do grupo de estudantes e o coeficiente de confiança do teste aplicado a esse grupo:  $C_{\text{verdadeira}} = \bar{x} + r(x_i - \bar{x})$ .

$C_{\text{verdadeira}}$  - representa a classificação verdadeira do estudante  $i$ .

$x_i$  - representa a classificação obtida (resultante da aplicação dos critérios de correcção/classificação dos itens do teste) pelo mesmo estudante  $i$  no teste aplicado ao grupo de estudantes (e.g. turma, ano de escolaridade numa escola em particular, etc.).

$\bar{x}$  - representa a média das classificações obtidas pelo grupo de estudantes.

$r$  - representa o coeficiente de confiança do teste aplicado àquele grupo de estudantes.

Pode provar-se que o valor do erro de medida padrão varia para diferentes pontos ao longo da curva de distribuição das classificações obtidas por um dado grupo de estudantes num teste. Se um professor pretender tomar decisões acerca dos intervalos de confiança de uma dada classificação pode calcular o erro de medida padrão para cada classificação individual [5]:



$$(s_{\text{erro}})_i = \sqrt{\frac{1}{n-1} x_i (n - x_i)}$$

$(s_{\text{erro}})_i$  - representa o erro padrão para a classificação do estudante i.

$x_i$  - representa a classificação obtida pelo estudante i.

$n$  - representa o número de itens do teste.

Conhecidos os processos de cálculo dos erros de medida padrão, coloca-se a questão de saber se duas classificações individuais diferentes corresponderão ou não a classificações verdadeiras diferentes, ou seja, será essa diferença entre as duas classificações suficientemente grande para se poder considerar uma diferença real ou poderá considerar-se que tal diferença surge por acaso? Considere-se por exemplo um teste de Química administrado a 120 estudantes (p. ex. quatro turmas de um mesmo ano de escolaridade), em que a média das classificações foi de 40%, o desvio padrão 8 e o coeficiente de confiança 0.89; supondo que os estudantes A e B obtiveram as classificações 42% e 46%; pretende-se responder à pergunta: estas duas classificações corresponderão a classificações verdadeiras distintas?

A resposta a esta pergunta presume um raciocínio composto por vários passos:

1 - Cálculo do erro de medida padrão:  $EMP = s \sqrt{1-r} = 8 \sqrt{1-0.89} = 2.65$ .

2 - O valor estimado do desvio padrão que seria de esperar nas classificações dos estudantes A e B se estes realizassem o mesmo teste muitas vezes. É necessário saber qual seria então o desvio padrão das diferenças entre os pares de classificações obtidas nas diferentes aplicações do teste - uma para o estudante A e outra para o estudante B, ou seja é necessário calcular o erro padrão da diferença (EPD):  $EPD = \sqrt{2EMP^2}$ , que neste caso é  $EPD = \sqrt{2 \times 2.65^2} = 3.75$ .

3 - A razão entre a diferença das duas classificações e o EPD dá o que se designa por razão t:  $t = (x_1 - x_2) / EPD = (46 - 42) / 3.75 = 1.07$ . Note-se que, quanto maior for a diferença entre as duas classificações relativamente ao EPD, mais provável é que a diferença entre as classificações não tenha surgido por acaso.

4 - Se o valor de t igualar ou exceder o(s) valor(es) críticos tabelados para níveis de significância diversos [8, 9], pode inferir-se que a probabilidade de a diferença ser real é, em percentagem igual ao nível de significância x 100. Consultando a tabela de valores críticos de t para um grupo de 120, verifica-se que o valor 1.07 obtido é menor que o valor crítico para o nível de significância mais elevado 0.20, ou seja, a diferença de 4 pontos nas duas classificações consideradas pode ter ocorrido por acaso mais do que 20% das vezes (0.20 x 100). Não há por isso razões para crer que os estudantes A e B diferem nas suas capacidades em Química avaliadas por este teste.

Pode considerar-se agora a questão de saber qual deve ser a diferença mínima entre duas classificações num teste de

avaliação da aprendizagem em Química para se poder considerar que tais classificações correspondem a capacidades distintas. De acordo com as tabelas de valores críticos de t pode considerar-se que para que a diferença a calcular tenha ocorrido por acaso somente 5 vezes em 100 - nível de significância de 0.05 - o valor de t deve ser maior ou igual a 1.98. Logo

$$1.98 = \frac{\text{diferença das classificações}}{3.75}$$

de onde resulta a diferença das classificações = 7.25.

Note-se que nas aplicações consideradas utilizou-se sempre o mesmo valor de EMP para as classificações do teste embora, como ficou já expresso, este valor varie ao longo da curva de distribuição de classificações. Contudo, esta opção deriva de se considerar que este é o procedimento mais provavelmente adoptado por qualquer professor preocupado com a problemática das classificações atribuídas aos seus alunos e o rigor possível na interpretação de tais classificações.

Deve referir-se que o tratamento das classificações dos estudantes que se discutiu pode fazer-se por recurso a meios informáticos ficando então extremamente facilitado.

Muitas outras aplicações de estatística podem ser utilizadas pelos professores na sua tarefa de avaliadores da aprendizagem e das capacidades dos seus alunos em Química e noutras disciplinas. As opções feitas na elaboração deste artigo resultaram da intenção de divulgar alguns conceitos fundamentais em avaliação sumativa e procedimentos julgados importantes no tratamento e interpretação dos resultados das classificações de testes.

#### Referências

- [1] Domingos, Ana Maria, Isabel Pestana Neves e Luísa Galhardo, Uma forma de estruturar o Ensino e a Aprendizagem (3.ª Ed.), Livros Horizonte, 1987.
- [2] Driver, Rosalind, Edith Guesne and Andrée Tiberghien, Open University Press, 1985.
- [3] Kempa, Richard, Assessment in Science, Cambridge Science Education Series, Cambridge University Press, 1986.
- [4] Lindquist, E. F., A First Course in Statistics, Houghton Mifflin Co., Boston Mass, 1942. - Citado por Richard Kempa.
- [5] Lord, F. M., Do Tests of the same Length have the same Standard Error of Measurement?, *Educational and Psychological Measurement*, 17, 510-21, 1957.
- [6] Osborne, Roger and Peter Freyberg, Learning in Science - The implications of children's science, Heinemann, 1985.
- [7] Santos, Maria Eduarda, Mudança Conceptual na Sala de Aula - Um desafio Pedagógico, Livros Horizonte, 1991.
- [8] Satterley, David, Assessment in Schools - (Theory and Practice in Education, no. 1), Basil Blackwell Ltd., 1981.
- [9] Snodgrass, Joan Gay, The Numbers Game, Oxford University Press, 1977.
- [10] Tolley, George, Learning and Assessment, em Patricia Murphy e Bob Moon at the Open University (Ed.), Developments in Learning and Assessment, 1989.

