Acta Pediatr. Port., 2003; Nº 3; Vol. 34: 203-216

Avaliação do Sistema de Avaliação de Alunos da Disciplina de Pediatria I - 6º Parte

LEONOR LEVY *

* *Professora Auxiliar da Faculdade de Medicina de Lisboa*

Resumo

A sexta parte do estudo é constituída pela discussão da sua metodologia e dos seus resultados e por algumas recomendações tendentes a aumentar a validade, fiabilidade, exequibilidade e reprodutibilidade do sistema de avaliação de alunos da Disciplina de Pediatria I, sugerindo que um *Objective Structured Clinical Evaluation* (OSCE) poderia proporcionar uma oportunidade de melhorar a qualidade da avaliação.

Palavras-Chave - Pediatria I, Avaliação, Validade, Fiabilidade, Exequibilidade, Reprodutibilidade, OSCE

Summary

Evaluation of the Assessment System of Paediatrics I' Students - Part VI

In this sixth part, we discuss the methodology and the results of the students' assessment in the Chair of Paediatrics I. We also propose some changes, mainly a *Objective Structured Clinical Evaluation* (OSCE), in order to improve the quality of students' assessment.

Key Words - Paediatrics I, Assessment, Validity, Reliability, Feasibility, Reproductibility, OSCE

Introdução

O I Mestrado em Educação Médica realizado em Portugal pres-supôs a frequência do Curso "Diploma on Medical Education", ministrado por docentes da Universidade de Cardiff. A obtenção do "Diploma on Medical Education", dependeu, para além da frequência do curso, da aprovação em diferentes provas efectuadas ao longo do curso.

Para a obtenção do título de Mestre em Educação Médica em Portugal foi necessária a elaboração de uma Tese de Mestrado em Educação Médica, avaliada através de uma dissertação e discussão da Tese por um Júri constituído por três Professores.

O tema escolhido para o estudo conducente a essa mesma Tese, foi a "Avaliação do sistema de avaliação de alunos da Disciplina de Pediatria I".

O estudo foi efectuado durante o ano lectivo de 1998/1999. Esta Tese de Mestrado contêm seis partes.

A sexta parte do estudo é constituída pela discussão da sua metodologia e dos seus resultados e por algumas recomendações

Discussão

Este estudo teve como objectivo a Avaliação do Sistema de Avaliação dos Alunos da Disciplina de Pediatria I.

Não se trata de um tema fácil ou cómodo, tanto mais que a autora deste estudo faz parte da equipa de docentes responsáveis pelo sistema de avaliação dos alunos da Disciplina de Pediatria I, mas é necessário e imperioso termos a coragem de avaliar aquilo que fazemos, a fim de procedermos a eventuais mudanças.

Existem algumas limitações neste estudo, como a falta de treino da fiabilidade entre os docentes e a diferença de condições entre as circunstâncias do 1º e do 2º semestre, nomeadamente o número de docentes e o *ratio* docente/discente, com a consequente diminuição do número de aulas práticas para cada aluno do 2º semestre.

Também houve diferenças entre o teste do 1º semestre e os testes do 2º semestre: o teste do 1º semestre teve um maior predomínio de perguntas de escolha múltipla e resposta única que os testes do 2º semestre, o que aumenta a probabilidade de existência de uma maior fiabilidade na correcção do teste do 1º semestre.

Optou-se por fazer primeiramente a discussão de cada um dos três tempos ou "timings" de avaliação dos alunos da Disciplina de Pediatria I, deixando para depois a discussão dos problemas comuns aos três tempos, a fim de evitar redundâncias.

Email: leonorl@mail.telepac.pt

Aspectos metodológicos

Desenho do estudo

Tratou-se de um estudo prospectivo, exploratório e não experimental com o objectivo de fazermos a avaliação dos sistemas de avaliação dos alunos da Disciplina de Pediatria I em vigência.

O estudo efectuou-se em três tempos ou "timings", um primeiro tempo no fim do 1º semestre da Disciplina de Pediatria I, em Fevereiro de 1999, um segundo tempo no fim do 2º semestre da Disciplina de Pediatria I (1ª chamada), em Julho de 1999 e um terceiro tempo no fim do mês de Julho de 1999 (2ª chamada).

Para a elaboração do texto foi especialmente importante o Manual de Pedro Serrano [1], enquanto que para o estudo estatístico, foram utilizados os Manuais de Anthony Walsh [2], de Pestana e Gageiro [3] e de Reis e Melo [4].

Instrumentos de observação

Os três exames teóricos diferiram ligeiramente uns dos outros; constatou-se uma maior predominância de perguntas de escolha múltipla e resposta única no 1º semestre, do que no 2º semestre; o teste do 1º semestre não conteve nenhuma pergunta de interpretação de casos clínicos, enquanto cada teste do 2º semestre conteve duas perguntas de interpretação de casos clínicos, o que pode ter constituído um factor de enviesamento.

Nos três testes, a maior parte das perguntas de escolha múltipla e resposta única integravam testes da mesma disciplina efectuados em anos anteriores, tendo sido apenas uma minoria de perguntas (20%) elaboradas para o efeito.

Foram introduzidas no sistema de avaliação dos alunos da Disciplina de Pediatria I algumas inovações, como a grelha de observação do exame prático e a respectiva cotação.

A fim de melhorar a avaliação da avaliação ou *"assessment's evaluation"*, foram elaborados os questionários sobre a satisfação de docentes e discentes sobre o sistema de avaliação dos alunos da Disciplina de Pediatria I, sob a forma de escalas de Likert.

Esta técnica parte do princípio de que podemos medir as atitudes através das crenças, opiniões e avaliações dos sujeitos acerca de um determinado objecto, e que a forma mais directa de acedermos a estes conteúdos cognitivos é através da autodescrição do posicionamento individual.

As escalas de Likert baseiam-se nas capacidades das pessoas que estão a ser avaliadas, para situarem as frases numa escala de intervalos iguais, considerando as avaliações como medidas ordinais.

Nas escalas de Likert, a selecção das frases que a compõem, é feita pelo investigador procurando frases que manifestem claramente apenas dois tipos de atitude: uma claramente favorável e outra claramente desfavorável em relação a um mesmo objecto.

A medição da atitude do sujeito á dada pelo seu posicionamento face ao conjunto destas frases radicais. Esta, como outras técnicas de papel e lápis, apresentam alguns problemas: a resposta do sujeito pode não corresponder à sua atitude real, mas sim a uma tentativa de agradar ao investigador; outro problema prende-se com a relevância da atitude para o sujeito; a própria linguagem em que a questão é formulada também pode constituir um problema; outro problema a considerar é o "timing" em que a escala é apresentada ao aluno [5].

A avaliação dos alunos da Disciplina de Pediatria I foi baseada em critérios e não na norma. Foram definidos dois critérios mínimos de aprovação: como critério mínimo dos testes foi definida a nota de cinco valores, correspondendo a 50% da nota máxima possível e como critério mínimo de aprovação na Disciplina de Pediatria I foi definida a nota de dez valores, correspondendo a 50% da nota máxima possível.

Uma avaliação baseada na norma (NRT) baseia-se na comparação de cada estudante com os resultados de uma curva de Gauss. Avalia-se, assim, o seu desempenho relativo. Este tipo de avaliação tem habitualmente uma percentagem fixa de alunos que conseguem ou não conseguem atingir a excelência.

O NRT é adequado quando se trata de escolher entre os candidatos a uma dada vaga. Existem algumas limitações sérias quanto ao NRT. O NRT depende da variância, o NRT não define aquilo que o estudante ser capaz de fazer no fim do programa de ensino-aprendizagem e para efeitos de padrões absolutos de sucesso e para o *feed-back* do estudante, o NRT não é adequado, pois os seus resultados dependem dos resultados de toda a classe. Por outro lado, os testes baseados no NRT tendem a concentrar-se nas raridades que têm poder discriminativo entre os melhores e os piores alunos e os princípios básicos podem nunca ser testados [6,7].

A avaliação baseada no critério (CRT), baseia-se na determinação de objectivos e no grau da sua concretização. Com este processo, os estudantes não são comparados com uns com os outros, mas consigo próprios, o CRT compara o desempenho do estudante com um critério e não com os resultados da classe. Uma crítica que se faz ao CRT é a de que o CRT não encoraja a excelência, mas apenas garante que a maior parte dos alunos atinge o nível mínimo de aceitação.

Para Lowry [6,7], a única maneira razoável de assegurar que os padrões mínimos sejam atingidos, é que esses mesmos padrões sejam definidos e assim os alunos que não atinjam esses padrões não devam ter aprovação no fim do programa de ensino-aprendizagem.

Os testes que utilizam o CRT são difíceis de elaborar, consumidores de tempo e difíceis de obter fiabilidade. Estes testes avaliam sobretudo o conhecimento no seu nível mais baixo e não em níveis mais elevados como o da interpretação e resolução de problemas [8].

Para cada pergunta de escolha múltipla e resposta única dos três testes foram observados os Índices de Discriminação e de Facilidade.

Índice de Discriminação

O Índice de Discriminação é a medida do poder que tem cada pergunta do teste para discriminar entre os mais capazes (ou os que têm mais conhecimentos) e os menos capazes. Um poder discriminativo menor que 0.2, faz com que as perguntas sejam retiradas e entre 0.2 e 0.3, sugere que as perguntas sejam revistas [9].

O Índice de Discriminação varia entre +1 e -1; um Índice de Discriminação negativo significa que a pergunta está mal feita. Para um *item* ser encarado como um bom discriminador, o seu poder de discriminação deverá ser positivo e igual ou superior a 0,3.

Pode haver um enviesamento possível no que respeita ao Índice de Discriminação, se uma dada pergunta for formulada de uma maneira ambígua ou mal construída [10].

Índice de Facilidade

O Índice de Facilidade define a proporção dos candidatos que responderam de maneira acertada a uma dada pergunta. O Índice de Facilidade varia de 0 a 1.

Um Índice de Facilidade igual a 1 significa que todos os candidatos acertaram na pergunta ou seja que a pergunta era muito fácil; um Índice de Facilidade igual a 0 significa que nenhum candidato acertou na pergunta.

Um alto Índice de Facilidade (pergunta muito fácil) ou baixo Índice de Facilidade (pergunta muito difícil) depende de vários factores, incluindo a maneira como a pergunta é formulada. Numa pergunta de escolha múltipla e resposta única, o Índice de Facilidade é muito influenciado pelo *distractores.*

Uma mesma pergunta poderá ter como resultado diferentes desempenhos, consoante a plausibilidade dos *distractores* e a maneira como a resposta correcta está escondida no meio desses *distractores.* [9].

Tratamento estatístico

A complexidade do projecto levou à criação de diferentes bases de dados em SPSS 8 para Windows, que pudessem conter todas as informações colhidas ao longo do estudo.

O tratamento estatístico foi da responsabilidade da autora do estudo e contou com a ajuda do Professor Doutor António Gouveia de Oliveira.

Para este efeito, a autora do projecto frequentou vários cursos de introdução à investigação e cursos de estatística.

O Curso de Estatística em Investigação Médica. Aplicação SPSS a Windows, efectuado em Outubro de 1998 foi especialmente útil.

A análise estatística iniciou-se com a apresentação das variáveis nominais, ordinais e intervaladas em termos de estatística descritiva [2-4].

As associações de variáveis foram estudadas através do teste t e dos testes de Mann-Whitney e de Kruskall-Wallis e ainda correlações bivariadas.

O teste t pressupõe que as variáveis estudadas sejam pelo menos de nível intervalado, que as populações sejam normalmente distribuídas e que as variâncias das populações sejam iguais.

Quando os pressupostos para a aplicação do teste t não se verificaram, utilizou-se o teste de Mann-Whitney, que é o teste paramétrico alternativo ao teste t para duas amostras independentes. O teste t compara as médias de duas amostras independentes, enquanto o teste de Mann-Whitney compara o centro de localização de duas amostras, como forma de detectar diferenças entre as duas amostras correspondentes [2-4].

Assim como teste t pode ser generalizado para mais do que dois grupos através da Anova, também o teste de Mann-Whitney pode ser generalizado para mais de dois grupos através do teste de Kruskall-Wallis [2-4].

O teste de Kruskall-Wallis constitui numa alternativa não paramétrica ao teste Anova utilizado, quando não se encontram reunidos os pressupostos da normalidade ou da igualdade de variâncias. Este teste é utilizado para testar a hipótese de igualdade em localização [2-4].

Avaliação dos alunos da Disciplina de Pediatria I 1º semestre

A variável *nota total* é o somatório das quatro notas parcelares dos alunos, reflectindo assim, as circunstâncias ligadas aos diferentes métodos do sistema de avaliação na Disciplina de Pediatria I [11].

Os resultados da *nota total*, assim como as suas representações gráficas, bem como o teste de Kolmogorov-Smirnov, sugerem que a curva da nota total não é normal e tem um enviesamento à direita, traduzindo provavelmente uma benevolência dos examinadores ou uma grande facilidade dos sistemas de avaliação.

Os resultados do *exame teórico* e as suas representações gráficas, bem como o teste de Kolmogorov-Smirnov, mostram uma curva normal, com uma média de 7,92 para uma nota máxima de 10 valores; realce-se um desvio padrão pequeno e a existência de alguns *outliers,* quer no diagrama de *caule e folhas*, quer na *caixa de bigodes*, que são a expressão de valores extremos.

O teste escrito tem consistência interna, avaliada através do alpha de Cronbach; ao avaliarmos o Índice de Discriminação, verificamos que apenas 4 das 44 perguntas de escolha múltipla e resposta única têm poder de discriminar entre os alunos mais capazes (ou mais sabedores) e os menos capazes.

Verifica-se que 6 das perguntas de escolha múltipla e resposta única têm um poder de discriminação igual a 0 e

que uma das perguntas tem um poder discriminativo negativo. Verifica-se ainda que 10 perguntas de escolha múltipla e resposta única têm um poder discriminativo que se situa entre o 0,2 e o 0,3.

Assim, apenas 10% destas perguntas serão de manter futuramente num exame, 10 perguntas poderão ser reformuladas a fim de aumentar o seu poder de discriminação para, pelo menos 0,3 e todas as outras perguntas deverão ser retiradas [9].

Poderão existir algumas explicações para um Índice de Discriminação tão baixo; as perguntas poderão ser mal formuladas ou induzir a resposta certa ou os *distractores* poderão não ser adequados [9].

Ao observarmos o Índice de Facilidade, verificamos que trinta e duas das 44 perguntas de escolha múltipla e resposta única foram perguntas aparentemente muito fáceis: 75% dos alunos responderam acertadamente.

A nota do teste escrito reflecte também a cotação das perguntas de resposta curta e de interpretação de casos clínicos. A este propósito recorde-se que cada docente fez a correcção dos testes dos seus próprios alunos e a não existência de uma estruturação da cotação das respostas, o que pode originar uma falta de fiabilidade entre os docentes.

Os resultados do *exame prático*, assim como as suas representações gráficas e o teste de Kolmogorov-Smirnov, sugerem que a curva do *exame prático* não é normal e tem um enviesamento à direita, traduzindo provavelmente uma benevolência dos examinadores ou uma grande facilidade deste sistema de avaliação. Saliente-se que 12 alunos atingiram a nota máxima (5 valores) no exame prático, enquanto que 23 alunos tiveram uma nota igual ou superior a 4,5 valores, não tendo havido nenhuma nota inferior a 3 valores.

A observação directa do aluno e dos seus desempenhos numa situação real ou simulada tem uma inegável validade, mas a sua fiabilidade é menor, a não ser que se recorra a uma grelha de observação estruturada [12-14].

Apesar das grelhas de observação utilizadas para avaliação dos desempenhos dos alunos no exame prático, as notas sugerem a necessidade de proceder ao treino da fiabilidade inerente a um melhor sistema de avaliação.

Ao analisarmos os resultados da variável *trabalho de campo*, verificamos que a sua curva não é normal e que as suas representações gráficas sugerem um enviesamento para a direita. Verificamos ainda que houve um erro na cotação do trabalho de campo em dois casos, em que a nota dada foi respectivamente de 3 e 3,25 valores, quando a nota máxima possível era de 2,5 valores.

Isto sugere que os docentes em causa avaliaram conjuntamente o relatório do trabalho de campo e a avaliação do desempenho dos alunos nas aulas práticas, com a nota máxima de 5 valores para o conjunto dos dois tipos de avaliação.

De realçar o número considerável de alunos (13 alunos) que tiveram a nota máxima (2,5 valores), sugerindo que é necessária uma estruturação na avaliação do relatório do trabalho de campo, a fim de aumentar a fiabilidade entre os docentes.

Ao analisarmos os resultados da variável *avaliação contínua*, verificamos que a sua curva não é normal e que as suas representações gráficas, sugerem um enviesamento para a direita (Gráfico 1) [11].

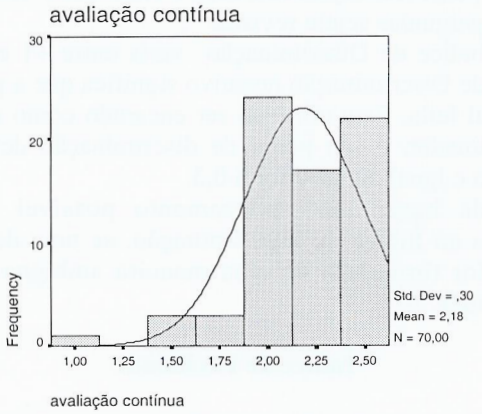

Gráfico 1

De realçar a quantidade de alunos (20 alunos) que tiveram a nota máxima (2,5 valores), sugerindo a benevolência dos docentes ou a impossibilidade de fazer uma boa avaliação dos alunos dado o *ratio* docente/discente (1/7) existente nas aulas práticas [11].

Recorde-se que estes resultados se referem ao 1º semestre da Disciplina de Pediatria I, em que havia 5 docentes livres que deixaram de colaborar na Disciplina de Pediatria I no 2º semestre.

Estes resultados sugerem a urgência em estruturar a avaliação contínua, quanto à assiduidade, pontualidade, interesse e intervenção nas aulas práticas, a fim de aumentar a fiabilidade entre os docentes, prática essa, no entanto, de muito difícil execução.

Quanto às correlações entre as notas obtidas no 1º semestre, podemos observar que não existem correlações entre o exame teórico e qualquer um dos outros métodos de avaliação dos alunos.

Também não existe uma correlação entre a avaliação contínua e o trabalho de campo.

Verificamos existirem correlações entre o exame prático e a avaliação contínua e ainda entre o exame prático e o trabalho de campo.

Poderemos especular se poderão existir alguns factores de enviesamento na avalição dos alunos, consistindo num efeito de "aura" de alguns alunos ou ainda na falta de fiabilidade entre os docentes, que se poderão comportar como "pombas" ou como "águias".

Quanto ao grau de satisfação dos docentes sobre o sistema de avaliação na Disciplina de Pediatria I,

saliente-se que apenas 6 dos 11 docentes preencheram e entregaram respectivo questionário; este facto poderá ter diferentes explicações, como o do desinteresse dos docentes, ou a descrença numa melhoria do sistema ou ainda grandes responsabilidades assistenciais ou falta de tempo.

Sendo o questionário preenchido anonimamente, não podemos fazer a distinção entre as respostas dos assistentes contratados ou dos assistentes livres.

A satisfação global é distribuída pelas cotações de 2 a 4, representando um grau de satisfação médio.

A validade do teste e do exame prático obtêm respostas que traduzem resultados médio ou acima da média, não existindo respostas que contemplem o grau mais elevado.

Quanto à validade do trabalho de campo, a maior parte dos docentes atribui-lhe um valor acima da média, existindo um resultado correspondendo à média e outro abaixo da média.

A validade da avaliação contínua é cotada como sendo média ou inferior à média, não existindo qualquer valor acima da média.

A maior parte dos docentes atribui ao teste uma fiabilidade superior à média.

Metade dos docentes atribui ao exame prático uma fiabilidade superior à média, e a outra metade reconhece-lhe apenas uma fiabilidade média ou inferior.

A fiabilidade do trabalho de campo é avaliada desde o grau 2, resultado inferior à média, até 5, que corresponde ao grau máximo.

A maior parte dos docentes atribui à fiabilidade da avaliação contínua uma fiabilidade inferior à média, existindo duas respostas que a consideram como média.

A exequibilidade do teste obteve respostas desde o grau 2 (inferior à média) até ao grau 5, correspondendo ao grau máximo.

A maior parte dos docentes avalia a exequibilidade da avaliação contínua como inferior à média, traduzindo esta resposta uma insatisfação ligada às dificuldades encontradas nas aulas práticas, realçando o problema existente do *ratio* docente/discente.

A exequibilidade do trabalho de campo é avaliada como sendo superior à média pela maior parte dos assistentes.

Quanto à reprodutibilidade do teste, os resultados parecem revelar uma boa avaliação da parte dos docentes, pois os resultados contemplam os graus 4 e 5 da escala de Likert utilizada.

A reprodutibilidade do exame prático é avaliada por metade dos docentes como acima da média, sugerindo a utilidade das grelhas de observação e da estruturação da avaliação do exame prático.

A reprodutibilidade da avaliação contínua é avaliada como média ou inferior, traduzindo as dificuldades existentes neste campo.

A reprodutibilidade do trabalho de campo é avaliada nos graus 2, 3 e 4, sugerindo a necessidade de fazer uma estruturação desse mesmo trabalho assim como da sua avaliação.

Assim, é de realçar sobretudo os problemas ligados à avaliação contínua.

Grau de satisfação dos alunos sobre o sistema de avaliação no 1º semestre

O grau de satisfação global foi claramente positivo, pois a maior parte das respostas contemplam os graus 3 e 4, ou seja, o grau médio e o imediatamente superior.

Também foram positivas as opiniões os alunos quanto à correspondência entre o teste e os objectivos da Disciplina de Pediatria I, pois a maior parte das respostas contemplam os graus 3 e 4, ou seja, o grau médio e o imediatamente superior.

A adequação dos textos de apoio, distribui-se pelas cinco categorias, sendo a categoria 3, a mais prevalente, traduzindo a necessidade de melhorar os textos de apoio.

Também são positivas as opiniões dos alunos quanto ao equilíbrio da matéria no teste final, pois a maior parte das respostas contemplam os graus 3 e 4, ou seja o grau médio e o imediatamente superior.

Quando interrogados se o exame prático tinha cor--respondido às expectativas, a maior parte dos alunos responde também muito positivamente, pois as respostas estão distribuídas quase equitativamente pelos graus 3, 4 e 5, tal como acontece para a avaliação do grau de cor--respondência entre a matéria dada nas aulas práticas e o exame prático.

Também a maior parte dos alunos avalia a satisfação por existir avaliação contínua nos graus 4 e 5 da escala de Likert utilizada, o que parece corresponder a um grau de satisfação elevado, assim como também parece existir um grau de satisfação elevado associado à existência do trabalho de campo, noção esta que corrobora o feedback que os docentes têm por parte dos alunos.

Os resultados obtidos contrariam o sentimento geral de que os alunos não estão contentes com o sistema de avaliação da Disciplina de Pediatria I, podendo-se especular que este sistema de avaliação se adapta às suas necessidades, pois mesmo não sendo discriminativo, lhes atribui quase uniformemente notas muito altas.

Avaliação dos alunos da Disciplina de Pediatria I -2º semestre - 1ª chamada

No 2º semestre da Disciplina de Pediatria I só existiram 6 docentes para 76 alunos, que teve como consequência que os alunos só terem tido 50% das aulas práticas a que tinham direito, a fim de manter um *ratio* docente/discente de 1/6, razoável quando se trata de aulas

práticas, mas ainda assim longe um de *ratio* docente/discente ideal [15].

Os resultados da variável *nota total*, são sobreponíveis aos resultados obtidos no 1º semestre. Com efeito, quer os resultados, quer as representações gráficas desta variável, sugerem que a curva obtida não é normal, existindo um enviesamento à direita, traduzindo uma benevolência dos docentes ou falhas no sistema de avaliação dos alunos.

Quanto à variável *exame teórico*, embora o teste de Kolmogorov-Smirnov não permita rejeitar a hipótese de a curva ser normal, a análise da sua simetria está no valor limite para a aceitação da curva e as suas representações gráficas, sugerem um enviesamento para a direita, havendo ainda a existência de outliers que correspondem a valores extremos [15].

Podemos ainda observar que o teste escrito tem consistência interna, avaliada pelo alpha de Cronbach. Ao observarmos o índice de Discriminação, verificamos que apenas 3 das 27 perguntas de escolha múltipla e resposta única têm poder de discriminação entre os alunos mais capazes e os menos capazes; verificamos ainda que 10 das perguntas têm um Índice de Discriminação igual a 0, enquanto que 5 *items* têm um poder discriminativo entre 0,2 e 0,3.

Isto significa que para testes a efectuar futuramente, apenas poderemos aproveitar 3 perguntas, embora existam 5 *items* passíveis de reformulação.

Ao observarmos o Índice de Facilidade verificamos que dezoito das 27 perguntas de escolha múltipla e resposta única foram perguntas aparentemente muito fáceis, pois 75% dos alunos responderam acertadamente.

Tal como no 1º semestre, os resultados do *exame prático*, assim como as suas representações gráficas e o teste de Kolmogorov-Smirnov, sugerem que a curva do *exame prático* não é normal e tem um enviesamento à direita, traduzindo provavelmente uma benevolência dos examinadores ou uma grande facilidade deste sistema de avaliação. É de realçar a quantidade de alunos que atingiu a nota de 4,5 ou superior (22 alunos) (Gráfico 2) [15]

Exame prático Stem-and-Leaf Plot

Frequency	Stem & Leaf	
4,00	Extremes	(=<3,5)
1,00	3 .	8
8,00	4 .	02233344
19,00	4 .	5556666777778888899
3,00	5 .	000
Stem width:	1,00	
Each leaf:	1 case(s)	

Gráfico 2

Também a variável *trabalho de campo* não tem uma curva normal e as suas representações gráficas sugerem um enviesamento à direita. Realce-se que 11 alunos receberam a nota máxima (2,5 valores) (Gráfico 3) [15].

Trabalho de campo Stem-and-Leaf Plot

Frequency	Stem & Leaf	
1,00	Extremes	(=<1,3)
1,00	1 .	5
2,00	1 .	77
,00	1 .	
8,00	2 .	00000011
11,00	2 .	22222233333
12,00	2 .	455555555555
Stem width:	1,00	

Gráfico 3

A variável *avaliação contínua* também não tem uma curva normal e as suas representações gráficas mostram uma enorme assimetria e um enviesamento à direita.

Saliente-se que 24 em 35 alunos receberam uma nota de 2,4 ou 2,5 valores, que é a nota máxima admitida neste tipo de avaliação.

Correlações entre as notas obtidas no 2º semestre - 1ª chamada:

Verificou-se haver correlações entre quase todos os sistemas de avaliação apenas não foi encontrada qualquer correlação entre as notas do exame prático e do trabalho de campo.

Avaliação dos alunos da Disciplina de Pediatria I 2º semestre - 2ª chamada

Apesar dos resultados da variável nota total e das suas representações gráficas sugerirem um enviesamento à direita, o teste de Kolmogorov-Smirnov não permite rejeitar a normalidade da curva. Estes resultados são diferentes aos resultados nos primeiros dois tempos ou "timings" de avaliação [16].

Quanto aos resultados da variável *exame teórico* embora as suas representações gráficas sugiram um ligeiro enviesamento à direita, o resultado do teste de Kolmogorov-Smirnov não permite rejeitar a normalidade da curva, tal como se observou nos dois "timigs" anteriores.

O teste tem consistência interna, avaliada pelo alpha de Cronbach.

Ao observarmos o Índice de Discriminação verificamos que apenas 5 das 28 perguntas de escolha múltipla e resposta única têm poder de discriminar entre os alunos mais capazes e os menos capazes; verifica-se ainda que apenas 4 perguntas têm um poder discriminativo entre 0,2 e 0,3, sugerindo que são susceptíveis de serem reaproveitados em testes futuros, se forem sujeitas a reformulações.

Ao observarmos o Índice de Facilidade verificamos que quinze das 28 perguntas de escolha múltipla e resposta única foram perguntas aparentemente muito fáceis, pois 75% dos alunos responderam acertadamente.

Tal como verificámos nos "timings" anteriores, os resultados do *exame prático*, assim como as suas representações gráficas) e o teste de Kolmogorov, sugerem que a curva do *exame prático* não é normal e tem um enviesamento à direita, traduzindo provavelmente uma benevolência dos examinadores ou uma grande facilidade deste sistema de avaliação.

De realçar que embora nenhum aluno tenha obtido a nota máxima, se verifica a existência de 17 alunos que obtiveram 4,5 valores ou mais na prova prática.

Tal como observámos anteriormente, a variável *trabalho de campo* não tem uma curva normal e as suas representações gráficas sugerem um enviesamento à direita.

Realce-se o número de alunos (15 alunos) que recebeu a nota máxima (2,5 valores). A observação de que existem muitos alunos que receberam ou 2,0 ou 2,5 valores, sugere que a pontuação desta prova tenha uma amplitude de 0 a 20 valores, entrando depois esta nota na devida proporção para a obtenção da nota total.

Tal como verificámos anteriormente, a variável *avaliação contínua* também não tem uma curva normal e as suas representações gráficas mostram uma enorme assimetria e um enviesamento à direita (Gráfico 4) [16].

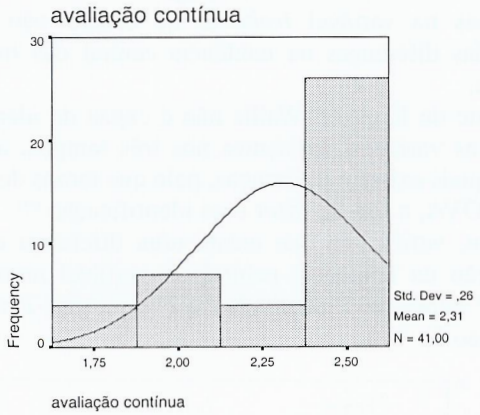

Gráfico 4

Saliente-se que 22 em 41 alunos receberam uma nota de 2,5 valores, que é a nota máxima admitida neste tipo de avaliação.

Quanto às correlações entre as notas obtidas neste "timing", verifica-se existirem correlações entre o exame teórico e o trabalho de campo entre o exame prático e a avaliação contínua e entre a avaliação contínua e o trabalho de campo; não se observam correlações entre o exame teórico e o exame prático entre o exame teórico e a avaliação contínua e entre o exame prático e o trabalho de campo.

Grau de satisfação dos docentes sobre o sistema de avaliação na Disciplina de Pediatria I - 2º semestre

Quanto ao grau de satisfação dos docentes sobre o sistema de avaliação na Disciplina de Pediatria I- 2º semestre, saliente-se que apenas metade dos docentes preencheram e entregaram o respectivo questionário.

Este facto, assim como os resultados obtidos poderão reflectir algum descontentamento da parte dos docentes, pelas dificuldades crescentes ligadas ao ensino desta Disciplina, dada a desistência de colaborar no ensino de Pediatria I por parte dos assistentes livres, sobrecarregando enormemente os seis docentes remanescentes.

O grau de satisfação global é de grau médio ou inferior à media, repartindo-se pelas categorias 1, 2 e 3.

A validade atribuída ao exame prático correspondeu aos graus 2, 3 e 4 e a validade atribuída ao trabalho de campo foi de 2 ou 4.

Já a validade atribuída à avaliação contínua contemplou apenas o grau mais baixo da escala de Likert utilizada, reflectindo as dificuldades ligadas a esta actividade, dado o pequeno número de aulas práticas e o *ratio* docente/discente de 1/7.

A fiabilidade atribuída ao teste contemplou apenas o grau 4, correspondendo a uma categoria superior ao grau médio.

A fiabilidade atribuída ao exame prático foi de grau médio ou superior e a do trabalho de campo inferior ou superior à categoria média.

Mais uma vez observamos a existência de problemas ligados à avaliação contínua, desta vez relacionados com a fiabilidade; os docentes foram unânimes em atribuir à fiabilidade da avaliação contínua a categoria 1, ou seja o grau mais baixo da escala de Likert, realçando mais uma vez as dificuldades que os docentes no ensino prático da Disciplina de Pediatria I.

Quanto à exequibilidade do teste, as respostas recaíram nas categorias 3 e 5, enquanto que a exequibilidade do exame prático se distribui por uma resposta por cada uma das categorias 2, 3 e 4. A avaliação da exequibilidade do trabalho de campo distribuiu-se pelas categorias 3 e 4.

Já para a exequibilidade da avaliação contínua, os docentes foram unânimes em considerá-la ao nível 1, ou seja o grau mais baixo da escala de Likert.

A reprodutibilidade de cada um dos seguintes métodos de avaliação, ou seja, o teste, o exame prático e o trabalho de campo, contemplou o grau médio ou os superiores da escala de Likert, ou seja os graus 3, 4 e 5; mais uma vez a reprodutibilidade da avaliação contínua foi unanimemente considerada como pertencendo ao grau mais baixo da escala de Likert, o grau 1.

É de realçar o problema existente neste semestre em todos os *items* referentes à avaliação contínua, que condiz com os resultados obtidos nas avaliação da variável avalia-

ção contínua dos alunos, pondo em causa este sistema de avaliação de alunos.

Grau de satisfação dos alunos sobre o sistema de avaliação na Disciplina de Pediatria I - 2º semestre.

Quanto ao grau de satisfação dos alunos sobre o sistema de avaliação na Disciplina de Pediatria I, durante o 2º semestre, curiosamente foi claramente positivo, não parecendo reflectir as dificuldades existentes durante este semestre.

O grau de satisfação global foi bastante positivo, pois a maior parte das respostas contemplam a categoria 3 e 4.

Também foram positivas as opiniões os alunos quanto à correspondência entre o teste e os objectivos da Disciplina de Pediatria I, pois a maior parte das respostas contemplam o grau 4 (18 alunos), ou seja um grau superior ao grau médio.

A adequação dos textos de apoio, distribui-se pelas cinco categorias, sendo a categoria 3, a mais prevalente, traduzindo a necessidade de melhorar os textos de apoio.

Também são positivas as opiniões dos alunos quanto ao equilíbrio da matéria no teste final, pois a maior parte das respostas contemplam os graus 3 e 4, ou seja o grau médio e o imediatamente superior.

Quando interrogados se o exame prático tinha correspondido às expectativas, a maior parte dos alunos escolhe a categoria 4, havendo alguns alunos que escolhem as categorias 2, 3 e 5; para a avaliação do grau de correspondência entre a matéria dada nas aulas práticas e o exame prático, as categorias com maior número de respostas são as categorias 4 e 5, ou seja as categorias superiores da escala de Likert utilizada.

Também a maior parte dos alunos avalia a satisfação por existir avaliação contínua nos graus 4 e 5 da escala de Likert utilizada, o que parece corresponder a um grau de satisfação elevado, assim como também parece existir um grau de satisfação elevado associado à existência do trabalho de campo, noção esta que corrobora o *feedback* que os docentes têm por parte dos alunos.

Estes resultados não reflectem as dificuldades sentidas durante o 2º semestre da Disciplina I. Poder-se-á especular que os docentes, à custa de um enorme esforço, conseguiram ultrapassar as dificuldades existentes e motivar os alunos para uma auto-aprendizagem ou que, pelo contrário, a falta de condições com a consequente benevolência dos docentes, nomeadamente no que respeita à avaliação do *exame prático, avaliação contínua e trabalho de campo*, servem os propósitos dos alunos que poderão ser de, prioritariamente, alcançarem notas altas.

Comparação dos resultados das notas obtidas nos dois semestres da Disciplina de Pediatria I

Decidiu-se fazer a comparação entre as notas obtidas pelos alunos da Disciplina de Pediatria I, comparando-as de duas maneiras: comparando as notas obtidas nos dois semestres e comparando as notas obtidas nos três "timings" de avaliação da Disciplina de Pediatria I, ou seja, 1º semestre, 2º semestre -1ª chamada e 2º semestre -2ª chamada [18].

A comparação dos resultados das notas obtidas nos dois semestres da Disciplina de Pediatria I foi feita através do teste não paramétrico de Mann-Whitney.

Ao fazermos a comparação dos dois semestres entre si, verifica-se que não existem diferenças nas distribuições da *nota total* em termos de distribuição central.

Já no *exame teórico*, pode-se verificar que existe uma diferença nas tendências centrais das duas distribuições, sendo mais elevados os resultados obtidos no 1º semestre.

Nas variáveis *exame prático e trabalho de campo*, também não se verificam diferenças na tendência central das respectivas distribuições nos dois semestres.

Para a variável *avaliação contínua*, verifica-se a existência de uma diferença entre as tendências centrais das duas distribuições, sendo mais elevados os resultados no 2º semestre; pode-se especular que os docentes não quiseram penalizar os alunos pelas dificuldades existentes no 2º semestre, atribuindo-lhes notas mais elevadas na avaliação contínua.

Comparação entre os resultados das notas obtidas nas três épocas da Disciplina de Pediatria I

Comparámos ainda as notas obtidas nas três épocas ou "timings" da Disciplina de Pediatria I, através do teste não paramétrico de Kruskall-Wallis.

Verifica-se uma diferença na tendência central de quatro variáveis (nota total, exame teórico, exame prático e avaliação contínua) nos três "timings".

Apenas na variável *trabalho de campo* não foram encontradas diferenças na tendência central das três distribuições.

O teste de Kruskall-Wallis não é capaz de identificar de entre as variáveis existentes nos três tempos, aquelas entre as quais existem diferenças, pelo que temos de recorrer à ANOVA, a fim de fazer essa identificação [18].

Assim, verifica-se que existe uma diferença entre a distribuição na tendência central da variável *nota total,* entre o 2º semestre -1ª chamada e o 2º semestre -2ª chamada (Gráfico 5) [18].

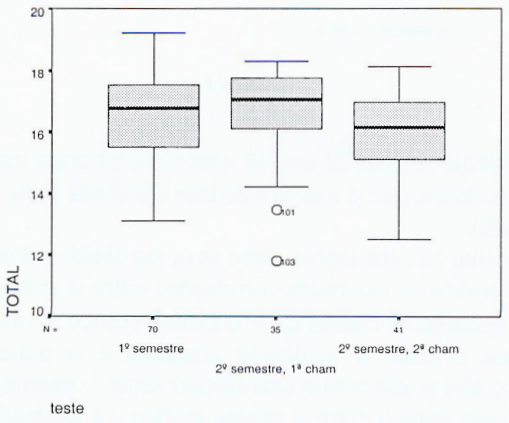

Gráfico 5

Verifica-se ainda a existência de uma diferença entre a distribuição na tendência central da variável *exame teórico*, entre o 1º semestre e o 2º semestre -2ª chamada (Gráfico 6) [18].

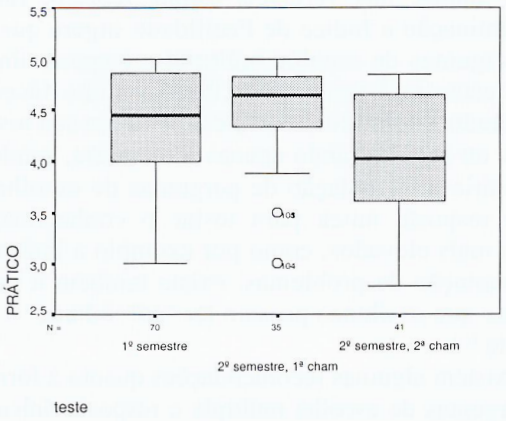

Gráfico 6

Também se verifica que existe uma diferença entre a distribuição na tendência central da variável *exame prático*, entre o 2º semestre -1ª chamada e o 2º semestre 7) -2ª chamada (Gráfico 7)[18].

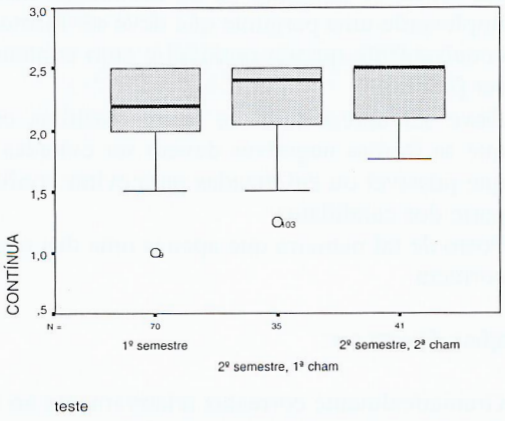

Gráfico 7

Verifica-se ainda a existência de uma diferença entre a distribuição na tendência central da variável *avaliação contínua*, entre o 1º semestre e o 2º semestre -2ª chamada (Gráfico 8) [18].

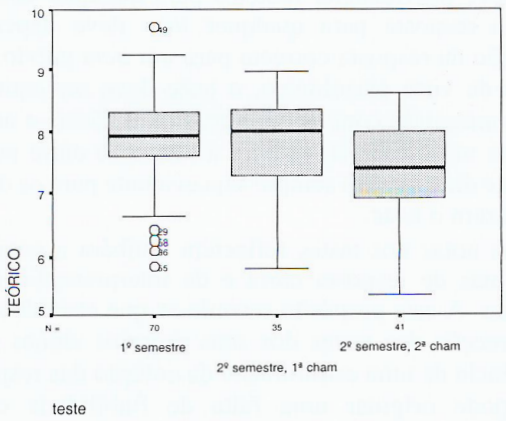

Gráfico 8

Comparação do grau de satisfação dos docentes nos dois semestres da Disciplina de Pediatria I

A comparação do grau de satisfação dos docentes nos dois semestres da Disciplina de Pediatria I foi feita através do teste t.

Apenas se fez a comparação ao nível da *satisfação total*, ou seja, o somatório de todos os resultados obtidos pelo preenchimento das escalas de Likert do questionário, dado que a variável assim obtida é uma variável de nível intervalado, susceptível de tratamento estatístico pelo teste t.

Pretende-se saber se os graus de satisfação média total dos docentes do 1º e 2º semestres da Disciplina de Pediatria I provêm de populações com o mesmo grau de satisfação média.

Não foi encontrada uma diferença estatisticamente significativa entre o grau de satisfação dos dois grupos, ou seja, entre os dois semestres.

Repare-se, no entanto, na *caixa de bigodes*, cuja representação gráfica sugere uma maior satisfação no 1º semestre, comparado com o 2º semestre.

A exiguidade do número de docentes que responderam e entregaram os questionários torna difícil a interpretação destes resultados; pode-se especular que uma maior quantidade de respostas conduziria a resultados estatisticamente significativos.

Comparação do grau de satisfação dos alunos nos dois semestres da Disciplina de Pediatria I

A comparação do grau de satisfação dos alunos nos dois semestres da Disciplina de Pediatria I foi feita através do teste t.

Apenas se fez a comparação ao nível da *satisfação total*, ou seja, o somatório de todos os resultados obtidos pelo preenchimento das escalas de Likert do questionário, dado que a variável assim obtida é uma variável de nível intervalado, susceptível de tratamento estatístico pelo teste t.

Pretende-se saber se os graus de satisfação média total dos alunos do 1º e 2º semestres da Disciplina de Pediatria I provêm de populações com o mesmo grau de satisfação média.

Não foi encontrada uma diferença estatisticamente significativa entre o grau de satisfação dos dois grupos, ou seja, entre os dois semestres.

A representação gráfica desta comparação, sugere que a tendência central destas distribuições são semelhantes, sugerindo que as dificuldades existentes no 2º semestre, não se repercutiram na satisfação dos alunos quanto ao sistema de avaliação, o que não deixa de ser preocupante, podendo traduzir uma falta de sentido crítico dos alunos, ou o sentimento que qualquer sistema de avaliação é bom, desde que providencie boas notas, independentemente de

qualquer mudança de comportamentos, que é, afinal, a definição de Educação.

Problemas comuns aos 3 tempos de avaliação:

A avaliação dos alunos da Disciplina de Pediatria I, foi feita através de um teste escrito e de um exame prático no final de cada semestre, contribuindo ainda para a nota final a avaliação do desempenho do aluno nas aulas práticas e do relatório sobre o trabalho de campo efectuado durante o semestre.

A nota final tem um máximo de vinte valores e é dada pelo somatório das notas dos diferentes tipos de avaliação.

Existe um possível factor de enviesamento, comum aos três "timings" ou tempos de avaliação dos alunos.

Para a pontuação do teste escrito e para o exame prático é indiferente que a nota tenha um máximo de respectivamente dez e cinco valores, pois ambos os tipos de avaliação têm uma pontuação definida e no caso do exame prático existe uma estruturação do exame que deveria conduzir a uma maior fiabilidade; no entanto, o facto de a avaliação contínua para avaliar o desempenho dos alunos nas aulas práticas ter um valor até ao máximo de 2,5 valores pode constituir um enviesamento em termos de nota, pois os docentes poderão ter a tendência para darem notas mais altas do que dariam se o máximo possível fosse 20 valores, entrando depois para a constituição da nota final com 12,5%. Também o facto de a avaliação do relatório sobre o trabalho de campo ter um valor até ao máximo de 2,5 valores pode constituir um enviesamento em termos de nota, pois os docentes poderão ter a tendência para darem notas mais altas do que dariam se o máximo possível fosse 20 valores, entrando depois para a constituição da nota final com 12,5%.

Também foram observados alguns resultados comuns aos três tempos de avaliação.

A variável *nota total* revelou-se nos três tempos de avaliação, como tendo um enviesamento à direita, embora só no 2º tempo (2º semestre - 1ª chamada), através da análise da simetria, se tenha revelado assimétrica. A variável *nota total* e os seus enviesamentos traduzem os enviesamentos das variáveis que a compõem, ou seja o *exame teórico, o exame prático, a avaliação contínua e o trabalho de campo.*

A análise da variável *exame teórico* nos três tempos, revela que a análise da sua simetria não permite rejeitar a simetria das respectivas curvas, embora se observe um ligeiro enviesamento à direita nas três curvas.

Quanto à observação do Índice de Discriminação nos três tempos, os resultados revelaram que a maior parte das perguntas de escolha múltipla e resposta única não tiveram poder de discriminação entre os melhores alunos e os piores.

Também a observação do Índice de Facilidade nos três tempos revelou que a maior parte das perguntas eram muito fáceis, pois 75% dos alunos responderam acertadamente às perguntas de escolha múltipla e resposta única, sobretudo no 1º tempo.

A análise das variáveis *exame teórico*, Índice de Discriminação e Índice de Facilidade sugere que muitas das perguntas de escolha múltipla e resposta única dos testes estão mal formuladas, ou são muito fáceis, contemplando a domínio do conhecimento no seu nível mais baixo, ou seja, testando apenas a memória, sendo assim necessária a formulação de perguntas de escolha múltipla e resposta única para testar o conhecimento em níveis mais elevados, como por exemplo a interpretação e a resolução de problemas; existe também a possibilidade de que os alunos possam ter "adivinhado" a resposta certa [17].

Existem algumas recomendações quanto à formulação de perguntas de escolha múltipla e resposta única que se têm revelado úteis. São elas:

O *caule* deve ser:

- Na forma de pergunta ou de declaração incompleta implicando uma pergunta que deve ser formulada tão concisamente quanto possível e cujo conteúdo deve ser preciso.
- Deve ser apresentado de forma positiva, enquanto que as formas negativas devem ser evitadas sempre que possível ou enfatizadas para evitar confusões da parte dos candidatos
- Posto de tal maneira que apenas uma das respostas é correcta

As opções devem ser:

- Gramaticalmente correctas relativamente ao caule
- Concisas, não ambíguas e não devem ser interdependentes ou mutuamente exclusivas
- Ordenadas de tal maneira que a chave não se encontre sempre na mesma ordem relativamente aos distractores

Recorde-se ainda que dentro do mesmo teste, um *item* não deve providenciar indícios para qualquer outro *item,* nem a resposta para qualquer *item* deve depender da selecção da resposta correcta para um *item* prévio. De um ponto de vista psicológico, o teste deve ser equilibrado, não começando com as perguntas mais fáceis e acabando com as mais difíceis, embora a distinção entre perguntas fáceis e difíceis nem sempre seja evidente para os discentes que fazem o teste.

As notas dos testes reflectem também a cotação das perguntas de resposta curta e de interpretação de casos clínicos. A este propósito recorde-se que cada docente fez a correcção dos testes dos seus próprios alunos e a não existência de uma estruturação da cotação das respostas, o que pode originar uma falta de fiabilidade entre os docentes.

Wakeford [19] sugere que uma maior fiabilidade na correcção de perguntas de resposta curta seria obtida se cada discente avaliasse apenas algumas das perguntas de cada teste ou se cada teste fosse avaliado por dois discentes, podendo ainda cada aluno com uma nota muito alta ou muito baixa ser examinado por outros discentes.

Outra possibilidade seria a de transformar uma pergunta de resposta curta em perguntas de resposta curta estruturada, tendo como objectivo solicitar informações progressivas, step-by-step [20].

Ao introduzir-se perguntas de resposta curta num teste deverão ter-se as seguintes precauções: construir as perguntas de maneira precisa, preparar um esquema de avaliação estruturado, avaliar anonimamente, uma página de cada vez e de preferência cada página deve ser avaliada por examinadores diferentes, a fim de diminuir os enviesamentos [21].

A análise da variável *exame prático* nos três tempos, revelou que as curvas correspondentes não seguem curvas normais, existindo um enviesamento à direita e existindo numerosos alunos que obtiveram a nota máxima possível no 1º e 2º tempos, enquanto que no 3º tempo, nenhum aluno atingiu a nota máxima possível, havendo, no entanto 17 alunos que tiveram uma nota igual ou superior a 4,5 valores; estes resultados podem ter significados diferentes; os docentes poderão ser muito benévolos ou podem ter adaptado as suas aulas práticas ao exame prático, o exame prático pode ser muito fácil, ou os alunos poderão ser muito bons, não esquecendo ainda que não existe fiabilidade entre os docentes, o que é minorado pelo facto de cada aluno ser avaliado por dois docentes.

O facto dos alunos serem avaliados no exame prático pelo seu assistente também pode constituir um factor de enviesamento.

A falta de treino dos docentes em termos de fiabilidade é um dos defeitos apontados a um exame prático. Pode ainda ser observado um efeito de "aura" nalguns alunos que influencie os docentes ou ainda outro problema, ou seja, o de alguns discentes se poderem comportar como *pombas* e outros como *águias*, o que implica um treino da fiabilidade [12-14].

A fim de tentar melhorar a fiabilidade entre os docentes, relembre-se que no exame prático da Disciplina de Pediatria I utilizámos grelhas de observação para a avaliação de conhecimentos, atitudes e gestos, que foram construídas para o efeito [22].

A análise da variável *avaliação contínua* obtida nos três tempos, revela que as respectivas curvas não são normais, havendo um acentuado enviesamento à direita e uma enorme assimetria; realce-se que muitos alunos tiveram a nota máxima possível deste tipo de avaliação.

Estes resultados podem sugerir diferentes interpretações: o elevado *ratio* docente/discente pode não permitir uma boa avaliação dos alunos; no 2º semestre, as condições existentes pioraram, dado que cada aluno só teve 50% das aulas práticas a que tinha direito e os docentes poderão não ter querido penalizar os alunos por este facto, tendo atribuído notas altas na avaliação contínua.

Realce-se, no entanto, que os alunos parecem ter comparecido e cooperado nas aulas práticas, o que pode ter dado origem às altas notas a eles atribuídas, facto este que, no entanto, não invalida a falta de fiabilidade que parece existir e que é sentida pelos docentes da Disciplina de Pediatria I.

Estes resultados sugerem a necessidade de estabelecer uma forma mais justa e fiável da fazer a avaliação contínua dos alunos; seria possível a elaboração de uma grelha de avaliação dos alunos para cada aula prática, mas a sua exequibilidade é duvidosa.

A forma de cotação da avaliação contínua também pode enviesar os resultados; uma cotação de 0 a 20 talvez melhorasse este tipo de avaliação.

A análise da variável *trabalho de campo* obtida nos três tempos, revela que as curvas respectivas não seguem uma curva normal, havendo um enviesamento à direita e grande número de alunos que obtiveram a nota máxima possível.

Estes resultados podem ter várias explicações, como a benevolência dos docentes, ou a forma de cotação atribuída a este tipo de avaliação, sugerindo que uma cotação de 0 a 20 talvez contribuísse para aumentar a fiabilidade entre os docentes. Outra hipótese possível para aumentar a fiabilidade, seria a de que os relatórios do trabalho de campo, fossem avaliados anonimamente por dois docentes.

A falta de correlação existente entre os diferentes tipos de avaliação nos três tempos de avaliação é preocupante; afinal, apesar dos diferentes tipos de avaliação avaliarem dimensões diferentes do desempenho dos alunos, para o exercício da profissão médica é essencial a integração harmoniosa das diferentes dimensões, para alcançar a excelência no exercício da Arte e Ciência da Medicina.

Foi avaliado o sistema de avaliação da Disciplina de Pediatria I.

Em conclusão, pode dizer-se que o sistema de avaliação dos alunos da Disciplina de Pediatria I não é suficientemente bom, não discrimina entre os alunos mais capazes e menos capazes, sofrendo ainda de falta de fiabilidade.

O sistema de avaliação dos alunos da Disciplina de Pediatria I contempla os três primeiros níveis da aprendizagem, com predominância para a avaliação dos conhecimentos ao seu nível mais baixo e não é satisfatório.

Seria necessário um sistema de avaliação que avaliasse conhecimentos, atitudes e gestos e ainda a competência em comunicação interpessoal, sistema esse que, para além de contemplar preferencialmente o terceiro nível da aprendizagem, tivesse validade, fiabilidade, exequibilidade e

fosse reprodutível (Gráfico 7).

Recorde-se que o nível um da aprendizagem segundo Miller [9], o **"sabe"** pode ser avaliado através de perguntas de escolha múltipla e resposta única, de perguntas de resposta curta ou através de uma prova oral. O **"sabe como"** deve ser avaliado através da interpretação e resolução de problemas, através do "modified essay question" ou de uma prova oral.

Já o **"mostra como"** precisa de ser avaliado através de uma prova prática. O **"faz"** necessita da observação da actuação na vida real.

O **Faz**, ou seja , a competência ou o desempenho clínico é definido como aquilo que o aluno ou o médico fazem realmente na sua prática clínica real [21].

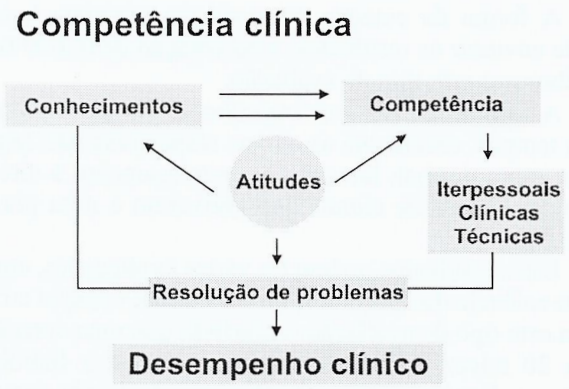

Gráfico 9

Tradicionalmente, em educação médica, a ênfase era posta na aquisição de conhecimentos essenciais em cada conteúdo programático, mas nos dias de hoje, cada vez é dada mais importância aos desempenhos e atitudes que se pensa serem características de "um bom médico" [6,7, 23,24].

Não existe nenhum tipo de avaliação que possa avaliar todas estas características.

Os discentes têm de identificar quais os aspectos que serão sujeitos a uma avaliação e decidir então os tipos de avaliação adequados.

A maior parte dos peritos em educação médica concordam com a necessidade de serem utilizados diferentes métodos para a avaliação dos desempenhos.

Pensamos que um *OSCE* (Objective Structured Clinical Evaluation) proporcionaria uma boa oportunidade de fazer uma melhor avaliação dos alunos da Disciplina de Pediatria I.

O *OSCE* é caro e administrativamente pesado, mas é um meio excelente de avaliar os desempenhos que outros métodos mais tradicionais são incapazes de avaliar [25,6,7,17].

O *OSCE* não é um método de avaliação, mas uma estrutura administrativa flexível, incorporando uma grande variedade de métodos de avaliação, permitindo a avaliação de conhecimentos, atitudes e gestos e ainda as competências de comunicação interpessoal de maneira mais objectiva [26-29].

O *OSCE* permite avaliar as diferentes áreas da competência clínica, como a colheita de história, o exame físico, desempenhos, o diagnóstico e o tratamento, etc.

O *OSCE* permite ainda avaliar no domínio dos conhecimentos, não só a sua recordação, mas a interpretação e resolução de problemas e a aplicação de conhecimentos ou desempenhos num novo contexto, a demonstração de desempenhos e as atitudes, para além das competências de comunicação interpessoais, podendo ser utilizado para avaliação formativa [30,25,31].

Classicamente o *OSCE* é constituído por diferentes "estações" com uma duração fixa de cinco minutos, existindo dois tipos diferentes de "estações", ou seja "estações" de desempenho e outras de questionário, havendo alternância dessas mesmas estações.

O *OSCE* tem como princípios o aumento do número de observações, a fragmentação da competência, a focalização em pontos críticos e a estruturação da observação. As vantagens do *OSCE* são o aumento da validade, fiabilidade e reprodutibilidade.

As desvantagens do *OSCE* são a impossibilidade de fazer uma avaliação global, a dificuldade de selecção das estações e das grelhas de observação e o tempo consumido na sua preparação. As estações de questionário do *OSCE* não precisam de observador, enquanto que as estações de desempenho precisam de um observador.

Nas estações de desempenho pode recorrer-se a doentes verdadeiros ou então a doentes simulados [14]. Estes doentes, também chamados de padronizados são pessoas normais treinados para simularem uma doença de uma maneira padronizada; estes doentes são utilizados quer na pré-graduação, quer na pós-graduação no ensino e avaliação de competências a nível de consultas e de observação [32,33].

Para Yelland [33], estes doentes padronizados preenchem um vazio importante na avaliação de competências em comunicação, aumentando a validade da avaliação e uma melhor avaliação da competência clínica do que os métodos indirectos como as perguntas de resposta curta modificada e os exames orais.

Outra vantagem dos doentes padronizados é a da descoberta de que as competências em comunicação enfatizadas pelos docentes não reflectem as competências que os doentes mais valorizam [34].

A praticabilidade do *OSCE* é um problema que enfrentamos, mas poder-se-ia ultrapassar esta dificuldade se pudéssemos recorrer a *OSCEs* sequenciais ou a *OSCEs* apenas para avaliação de desempenhos, mantendo os testes para avaliação dos conhecimentos [35].

Um dos problemas que um *OSCE* levanta é o da escolha dos seus *items* e da adequação aos problemas que os médicos vão encontrar na sua prática diária, de modo a actuarem com a máxima eficácia possível e ainda da sua capacidade de aprendizagem individual e de auto-aperfeiçoamento e ainda de adaptação à mudança [36,37].

Uma boa maneira de seleccionar os *items* de um *OSCE*

é recorrer à opinião de vários membros membros da equipa da Disciplina de Pediatria I, pois está provado que quando vários docentes cooperam nesta actividade, conseguem identificar os *items* mais importantes de um *OSCE* e aumentar a sua fiabilidade [38]

Experiência da Disciplina de Pediatria I em 1984

No ano lectivo de 1984/1985, os docentes da Disciplina de Pediatria I levaram a cabo uma experiência pedagógica muito interessante, que consistiu na avaliação dos alunos através de um *OSCE* [39].

O *OSCE* consistiu em oito "estações", alternando as "estações" de conhecimento com as de desempenho; a "estação" nº 1 tinha como conteúdo "Crescimento", as "estações" nº 2 e nº 3 tinham como conteúdo "Desenvolvimento", as "estações" nº 4 e nº 5, tinham como conteúdo "Alimentação", a "estação" nº 6 tinha como conteúdo "Imunizações", a "estação nº 7 tinha como conteúdo "Acidentes/Intoxicações" e a nº 8, "Semiologia".

Os principais problemas encontrados foi o grande consumo de tempo na preparação do *OSCE* e o desequilíbrio entre a avaliação dos conhecimentos, atitudes e gestos.

As vantagens deste *OSCE* foram a grande poupança de tempo na sua aplicação e uma maior fiabilidade na avaliação dos alunos, apesar da curva de Gauss resultante, ter mostrado um enviesamento à direita [39].

Parece haver necessidade de uma mudança no sistema de avaliação da Disciplina de Pediatria I. Temos consciência das dificuldades em introduzir mudanças em qualquer sistema, mas temos também consciência da necessidade e da premência de uma mudança [40-43].

Já Maquiavel dizia que "é preciso considerar que não existe nada mais difícil de desenvolver, nem tão duvidoso em termos de sucesso, nem mais perigoso de manejar, como iniciar uma nova ordem das coisas".

Bibliografia

1. Serrano P. Redacção e Apresentação de Trabalhos Científicos. *Editores Relógio D'Água*, 1996.
2. A. Statistics for the Social Sciences. With Computer Applications. *New York Harper & Row, Publishers*, 1988.
3. Pestana MH, Gageiro JN. Análise de Dados para Ciências Sociais. A complementaridade do SPSS. *Edições Sílabo*, 1998.
4. Reis E, Melo P, Andrade R, Calapez T. Estatística Aplicada. Vol. 2. *Edições Sílabo* 1997.
5. Vala J, Monteiro MB. Psicologia Social. Serviço de Educação *Fundação Calouste Gulbenkian* 1993.
6. Lowry S. Assessment of students. *BMJ* 1993; 306: 51-54.
7. Lowry, S. Making change happen.*BMJ* 1993; 306: 320-322.
8. Turnbull JM. What is.....Normative versus Criterion-referenced Assessment. *Medical Teacher* 1989; 2: 145-150.
9. Sutton, R.A. (1995). An Introduction to Assessment & Evaluation Processes and Procedures. *University College Cardiff.*
10. Hobsley M. Counting apples with oranges: a limitation of the discrimination index. *Medical Education* 1999; 33: 192-196.
11. Levy L. Avaliação do Sistema de Avaliação dos Alunos da Disciplina de Pediatria I. 2ª Parte. *Acta Ped Port*, 2001; 5:325-34
12. Ludbrook J, Marshall VR. Examiner training for clinical examinations. *British Journal of Medical Education* 1971; 5: 152-155
13. Newble D, Hoare J, Sheldrake P. The selection and training for clinical examination. *Medical Education* 1980; 14: 345-349.
14. Van der Vleuten CPM, Swanson DB. Assessment of Clinical Skills With Standardized Patients: State of the Art. *Teaching and Learning in Medicine* 1990; 2: 58-76.
15. Levy L. Avaliação do Sistema de Avaliação dos Alunos da Disciplina de Pediatria I. 3ª Parte. *Acta Ped Port*, 2001; 6: 399-406
16. Levy L. Avaliação do Sistema de Avaliação dos Alunos da Disciplina de Pediatria I. 4ª Parte. *Acta Ped Port*, 2002; 1: 45-53 *Acta Ped Port*, 2001, 6:
17. Van der Vleuten CPM, Newble DI. How can we test clinical reasoning? *The Lancet* 1995; 345: 1032-1034.
18. Levy L. Avaliação do Sistema de Avaliação dos Alunos da Disciplina de Pediatria I. 5ª Parte. *Acta Ped Port*, 2002; 2: 137-43
19. Wakeford RE, Roberts S. Short answer questions in an undergraduate qualifying examination: a study of examiner variability. *Medical Education* 1984; 18: 168-173.
20. Webber RH. Structured short-answer questions: an alternative examination method. *Medical Education* 1992; 26: 58-62.
21. Newble D, Cannon R. A Handbook for medical teachers. *United Kingdom, Kluwer Academic Publishers* 1994.
22. Van Thiel J, Kraan HF, Van der Vleuten CPM. Reliability and feasibility of measurin medical interviewing skills: the revised Maastrich History-Taking and Advice Checklist. *Medical Education* 1991; 25: 224-229.
23. Mulholland H. Teaching skills. How to assess junior doctors. *Hospital Update*, 1993; 56S-58S.
24. Neufeld VR. Assessing Clinical Comptence. *An Introduction to Measurement Properties, in Neufeld and Norman (Eds.)* 1985.
25. Carpenter JL. Cost Analysis of Objective Structured Clinical Examinations. *Academic Medicine* 1995; 9: 828-833.
26. Harden RM., Stevenson M., Downie WW, Wilson GM. Assessment of Clinical Competence using Objective Structured Examination. *Medical Education* 1975; 1: 447-451.
27. Harden RM., (1979). Assess Students: An Overview. *Medical Teacher* 1979; 2: 65-70.
28. Harden RM. Twelve tips for organizing an Objective Strutured Clinical Examination (OSCE). *Medical Teacher* 1990; 12: 259-64.
29. Hodges B, Regehr G, Hanson M, Naughton, N. An objective Structured Clinical Examination for Evaluating Psychiatric Clinical Clerks. *Acad Med* 1997; 72: 715-21
30. Biran LA. Self-assessment and learning through GOSCE (group objective structured clinical examination). *Medical Education* 1991; 25: 475-479.
31. Hodder RV, Rivington RN, Calcutt LE, Hart IR. The effectiveness of immediate feedback during the Objective Structured Clinical Examination. *Medical Education* 1989; 23: 184-188.
32. Robins LS, Zweifler AJ, Alexander GL et all. Using standardized patients for teaching and assessment. *Acad Med*, 1997; 72: s91-s93
33. Yelland MJ. Standardized patients in the assessment of general practice consulting skills. *Medical Education* 1998; 32: 8-13.
34. Cooper C, Mira M. Who should assess medical students' communication skills: their academic teachers or their patients?. *Medical Education* 1998; 32: 419-421.
35. Cass A, Regehr G. Sequencial testing in the Objective Structured Clinical Examination: Selecting Itmes for the Screen. *Acad Med* 1997; 72: S25-S27.

36. Gunderman RB. The Outcome of Medical Outcomes Assessment: It's Not Necessarily Academic. *Acad Med* 1997; 72: 682-687.
37. Stone SL, Qualters DM. Course-based Assessment: Implementing Outcome Assessment in Medical Education. *Acad Med*; 73: 397-401.
38. Valentino J, Donnelly MB. The Reliability of Six Faculty Members in Identifying Important OSCE Items. *Acad Med* 1998; 73: 204-205.
39. Gomes-Pedro J, Madeira M, Gouveia R *et all*. Programa de Ensino-Aprendizagem de Pediatria I (3). Avaliação Pedagógica. *Rev. Port. Ped.* 1988; 19: 101-7.
40. Anderson PC. Obstacles to change in Medical Education. Journal of *Medical Education* 1970; 45: 139-143.
41. Barnes PC. Managing change. British Medical Journal 1995; 310: 590-592.
42. Berwick DM. A primer on leading the improvement of systems. *British Medical Journal* 1996; 312: 619-622.
43. Owen AV. Getting the best from people. *BMJ* 1995; 310: 648-651.