Equiparação de Pontuações de Testes feitos em Computador por pequenos Turnos de Estudantes

Moura D.[1], Amaral M.[2], Severo M.[1]
[1] Faculdade de Medicina, Universidade do Porto, Portugal
[2] GATIUP, Universidade do Porto, Porto, Portugal

Identificação da unidade curricular

Nome: Farmacologia I (incluindo Farmacologia Geral)
Faculdade: Faculdade de Medicina da Universidade do Porto
Ano/Semestre: 3º ano/1º semestre
Plataforma: Moodle UP
Nº de Estudantes: 300

1. Introdução

Esta unidade curricular pratica há vários anos uma forma de ensino presencial apoiado por um componente digital à distância construído nas plataformas da Universidade do Porto, inicialmente do modelo WebCT e atualmente do modelo Moodle [1]. As componentes do portal da unidade curricular pertencem ao padrão geral destas plataformas de telemática educativa e, por isso, não serão objeto deste trabalho. O nosso propósito é apresentar uma componente nova com a qual se procurou resolver um problema específico no uso das tecnologias de informação ao serviço das tarefas de estudantes e professores. Trata-se especificamente da preparação e realização de testes de escolha múltipla através do computador.

Há já um núcleo robusto de docentes na Universidade do Porto que utiliza itens de escolha múltipla para a avaliação dos seus estudantes. A construção dos enunciados e as correções das provas são já correntemente feitas através de tecnologias digitais mas a realização das provas no computador enfrenta uma limitação logística muito importante. O número de salas e de máquinas disponíveis em condições adequadas a provas de avaliação sumativa é pequeno para a dimensão de muitos cursos.

Uma solução para este constrangimento é a distribuição dos estudantes por turnos ajustados à dimensão das salas e aos computadores disponíveis. Este processo tem sido bem sucedido em várias unidades orgânicas no que respeita à logística e à adaptação da plataforma Moodle UP. Há, portanto, na UP condições de gestão de espaços e de recursos informáticos. Falta agora resolver o problema da equidade entre as provas dos diferentes turnos.

Por razões óbvias não se pode repetir a mesma prova a estudantes que a vão realizar em momentos separados. Por outro lado, a introdução de diferentes perguntas em cada um dos turnos cria a necessidade de se garantir a equidade entre as diferentes provas ou de corrigir desvios que possam ocorrer.

O problema é bem conhecido da docimologia e a solução denomina-se equiparação ou equivalência (*equating*) entre formas diferentes da mesma prova [2]. No caso presente há, porém, uma limitação séria: as limitações físicas não permitem realizar mais do que 30 a 50 provas de cada vez. Ora, os métodos de equiparação mais robustos e experimentados baseiam-se na teoria da resposta a itens que, no entanto, só são válidos para grandes amostras.

Recentemente, a equiparação de pontuações para amostras pequenas tem vindo a ser analisada em múltiplos estudos que dão orientações muito úteis para o tipo de problema com que esta unidade curricular se confronta: cursos de mais de 300 estudantes para os quais está definido um calendário de provas muito rígido e incompatível com um espaçamento calmo de períodos de teste e, em simultâneo, uma capacidade de equipamento informático e de espaços que não permitem mais do que 30 a 50 exames ao mesmo tempo [3].

O objetivo deste trabalho foi assim preparar e testar, em condições reais de exames sumativos da unidade curricular, um procedimento de testes por turnos com aplicação de um processo de equiparação docimologicamente fundamentado.

2. Resultados de um ensaio piloto

2.1. Método

Nas épocas de recurso há menos estudantes do que nas épocas normais. Este contingente menos numeroso de estudantes oferece uma boa amostra para que se possa realizar um ensaio, em condições reais da prática docente e com respeito pelas legítimas preocupações dos estudantes, das ferramentas de análise e correção de desvios na equidade entre as diferentes formas do teste.

No ano letivo de 2010-2011 realizou-se, com o acordo prévio dos estudantes, o exame da época de recurso da unidade curricular aberto a todos os inscritos na época de recuso, que foram distribuídos por quatro provas no mesmo dia em horários diferentes. A prova teve o mesmo formato de todas as provas anteriores, com uma prova escrita de 70 perguntas de escolha múltipla a realizar em 2h. Os itens foram construídos de acordo com as normas aconselhadas pelo *National Board of Medical Examiner*s, que há muitos anos é responsável pelos exames de certificação em Medicina nos Estados Unidos [4]. Os itens têm 5 possibilidades de resposta e os estudantes são instruídos a selecionar a melhor delas, havendo apenas uma e sempre uma alternativa correta (*one best item*). Esta construção de testes é praticada há vários anos na unidade curricular. Os estudantes são submetidos a provas com o teste impresso em papel. Para além do enunciado é disponibilizado um impresso para onde a matriz das respostas é copiada pelo estudante e que é processado num leitor ótico. Na prova piloto reproduziu-se o formato visual das perguntas no ecran do computador e, evidentemente, prescindiu-se do formulário para o leitor ótico. O tempo de realização da prova foi mantido em 2h e procurou-se em tudo o mais manter as regras aplicadas aos exames em papel. Os estudantes foram distribuídos ao acaso por 4 turnos de 30 cada. Nem todos os lugares foram preenchidos por faltas de comparência. O primeiro turno realizou a prova das 9 às 11h, o segundo das 11 às 13h, o terceiro das 14:30 às 16:30 e o último das 16:30 às 18:00. Os estudantes dos 2 turnos da manhã e os dos 2 turnos da tarde não podiam contactar entre si porque a saída da sala dos que faziam a prova estava vedada até que chegasse a hora da entrada dos seguintes. Assim aplicaram-se dois testes diferentes, o da manhã e o da tarde, que a seguir serão designados por teste A ou teste 1 e teste B ou teste 2.

Na construção dos testes foi incluído um grupo de perguntas comuns, designadas como “âncoras”. O número de âncoras foi de 21 e o número de perguntas específicas de cada teste foi de 49. Os estudantes foram informados de todos os procedimentos analíticos, incluindo a inclusão de um número restrito de perguntas repetidas nos testes A e B. A plataforma Moodle permite a aleatorização da ordem das perguntas e dentro de cada uma dela a aleatorização das alternativas de resposta. Esta funcionalidade reduz a possibilidade de referenciação das perguntas de tal forma que se torna mais difícil a transmissão de informações entre estudantes no intervalo dos turnos.

Os resultados foram tratados estatisticamente em duas fases. Na primeira fase fez-se um estudo simples das pontuações do teste global e das perguntas-âncora através da análise de variância (ANOVA) uma vez que é um procedimento estatístico com que os estudantes estão familiarizados. Esta análise permite encontrar diferenças significativas de desempenho entre as provas A e B e ao mesmo tempo comparar os desempenhos dos diferentes grupos de estudantes perante as mesmas perguntas (parte do teste constituído pelas âncoras). Os estudantes foram informados de que se houvesse uma diferença estatisticamente significativa na média das pontuações dos testes sem que houvesse diferença estatística no desempenho das âncoras, o que daria sinal de que a proficiência dos estudantes dos diferentes turnos não era diferente, tal seria considerado um indicador de uma dificuldade diferente nos testes e, por isso, as notas seriam equiparadas às mais altas. Para a análise de variância univariada usou-se o programa STATISTICA.

A segunda fase da análise dos resultados destinou-se ao uso exclusivo dos docentes e já não se aplicou ao processo de atribuição das notas. O interesse desta segunda análise era permitir a interpretação de eventuais diferenças e, sobretudo, permitir a aplicação dos diferentes métodos descritos na literatura para equiparação de pontuações em amostras reduzidas. Nesta segunda fase usou-se a análise dos parâmetros de dificuldade e discriminação pela teoria clássica dos testes através do programa SPSS. Para as simulações e ajustes não lineares dos resultados aos diferentes modelos de equiparação para amostras reduzidas utilizaram-se os métodos de regressão não linear iterativos para ajuste à equação pré-definida utilizando-se o *software* de acesso livre “R”. Aplicaram-se os seguintes métodos: equiparação pelas médias, equiparação linear, equiparação linear encadeada de Tucker, equiparação pelo percentil, equiparação pelo percentil suavizada, equiparação pelo arco de círculo e equiparação pelo arco de círculo encadeado [5,6].

2.2. Resultados

A prova piloto foi completada por 82 estudantes, dos quais 50 realizaram o teste A e 32 o teste B. Na figura 1 estão as médias, erros-padrão da média e intervalos de confiança para 95% dos acertos totais.

Figura 1. Média, erro-padrão da média e intervalos de confiança para 95% dos acertos nos testes A e B

Verifica-se que as pontuações do teste B foram significativamente menores do que as do teste A (p=0,016). Uma vez que ambos os testes estavam ancorados a 21 perguntas procedeu-se a uma análise semelhante do desempenho dos estudantes que responderam ao teste A e ao teste B, restringindo-se o cálculo às perguntas-âncora. Os resultados estão na figura 2.

Figura 2. Média, erro-padrão da média e intervalos de confiança para 95% dos acertos nas perguntas-âncora dos testes A e B

A diferença não é estatisticamente significativa. Em seguida fez-se a análise de variância para as perguntas específicas e, tal como se esperava, a diferença é novamente significativa, ficando o desempenho no teste B abaixo do teste A (p=0,016) (figura 3).

Figura 3. Média, erro-padrão da média e intervalos de confiança para 95% dos acertos nas perguntas específicas dos testes A e B

Procedeu-se em seguida à análise dos itens de acordo com a teoria clássica dos testes. Os valores do coeficiente alfa de Cronbach para ambos os testes foi elevado (0,87 e 0,94, para os testes A e B, respetivamente), o que mostra que a fiabilidade dos testes é boa [7]. A análise do grau médio de dificuldade dos itens mostrou que, para as perguntas-âncora foi semelhante (0,54 e 0,50 para os estudantes do grupo A e B, respetivamente). Em contraste, o grau de dificuldade médio dos itens específicos do teste A foi significativamente menor do que os do teste B (0,64 e 0,55, respetivamente).

Verificando-se que o teste 2 foi mais difícil do que o teste 1 simularam-se os acertos através de diferentes métodos de equiparação (figura 4).

Figura 4. Relação entre os acertos (expressos em percentagem) nos testes A (teste 1) e B (teste 2). A linha contínua é a linha de identidade. As restantes linhas representam a estimativa de equiparação pelos diferentes métodos: equiparação da média (*mean*) equiparação linear (*linear*), equiparação dos percentis equivalentes (*Equip*.) e dos percentis equivalentes com suavização (*Equip. Smoothing*) e equiparação pelo arco do círculo (*circ*)

A linha contínua representaria a identidade entre os dois testes. Como não foi essa a situação em estudo procedeu-se à equiparação das pontuações, o que se traduz num desvio para a esquerda já que o teste 2 (B) é mais difícil. Todas as formas de equiparação cumpriram este requisito. No entanto a equiparação não é a mesma. Verifica-se que os métodos da equiparação linear, dos percentis equivalentes e dos percentis equivalentes com suavização dão diferenças maiores para as pontuações baixas e diferenças menores para as pontuações altas. Com o método da equiparação pela média há diferenças uniformes ao longo de toda a escala. O método da equiparação pelo arco do círculo dá diferenças maiores para as classificações médias e diferenças menores para as classificações extremas.

A figura 5 mostra os erros-padrão da equiparação. O método da equiparação pela média tem um erro-padrão invariável com a pontuação. O método linear tem um erro-padrão menor no centro e maior nos extremos da pontuação. O método da equiparação pelos percentis é o que mostra mais aleatoriedade no erro-padrão, que diminui após suavização. Finalmente o erro-padrão para o método do arco de círculo é maior no centro do que nas caudas. O erro menor de todos os testes é o do método do arco de círculo nas zonas das caudas das pontuações.

Figura 5. Erro padrão (SEE) de equiparação pelos diferentes métodos: equiparação da média (*m*) equiparação linear (*l*), equiparação dos percentis equivalentes (*e*) e dos percentis equivalentes com suavização (*es*) e equiparação pelo arco do círculo (*circ*)

Tirando partido da existência de perguntas comuns nos dois testes, refizeram-se os cálculos quer dos ajustes quer dos erros-padrão. Os testes passaram por isso a uma transformação de encadeamento (*chained*). Os resultados do ajuste estão expressos na figura 6.

Figura 6. Relação entre os acertos (expressos em percentagem) nos testes A e B. A linha contínua é a linha de identidade. As restantes linhas representam a estimativa de equiparação pelos diferentes métodos encadeados: da média (*mean chained*), linear (*linear chained* e *linear tucker*), dos percentis equivalentes com suavização (*Equip. Smoothing*) e do arco do círculo (*circle chained*)

Verifica-se, agora, que todos os testes se aproximam mais de paralelas à linha de identidade. No entanto se observarmos o erro das estimativas dos métodos de equiparação por encadeamento a partir das âncoras (figura 7) verifica-se que há dois métodos mais precisos do que quaisquer outros: os métodos do encadeamento pelas médias e o método do encadeamento pelo arco de círculo. A diferença está na forma da curva ao longo da escala

de pontuações. O método do encadeamento das médias é ligeiramente melhor do que o do encadeamento do arco de círculo no centro, mas os erros aumentam muito para valores de pontuação abaixo dos 25% de acertos e para valores acima dos 60% de acertos.

Figura 7. Erro padrão (SEE) de equiparação pelos diferentes métodos encadeados: da média (*mean chained*), linear (*linear chained* e *linear tucker*), dos percentis equivalentes com suavização (*Equip. Smoothing*) e do arco do círculo (*circle chained*)

2.3. Discussão

O estudo piloto de equiparação entre duas formas de um mesmo teste realizado em turnos diferentes através do computador mostrou que as duas formas não eram equivalentes. Não havendo deficiências na fiabilidade de nenhum dos testes, como se mostrou pela análise do coeficiente alfa de Cronbach [7], é legítimo pensar-se que um dos testes (teste B) era mais difícil. Uma explicação alternativa seria a de que os testes eram equivalentes e que o desempenho inferior dos estudantes expostos ao teste B se deveria a uma proficiência menor deste grupo de estudantes. Os estudantes estariam pior preparados do que os estudantes expostos ao teste A e, por isso, perante testes de dificuldade equivalente, teriam pior desempenho.

A introdução das perguntas-âncora revelou-se preciosa para distinguir as duas interpretações [8]. Assim, se os estudantes das duas séries tivessem proficiências diferentes o desempenho nas perguntas âncora e a média do grau de dificuldade seria diferente, o que não foi o caso. Assim, como os estudantes que realizaram a prova B tiveram a mesma percentagem de acertos dos que realizaram a prova A na parte comum do teste (21 perguntas-âncora) e tiveram um desempenho menor nas perguntas específicas, a conclusão é a de que o grau de dificuldade do teste B era maior. Houve, portanto, que estabelecer uma equivalência de pontuações para que os resultados dos estudantes submetidos à prova B fossem calibrados para cima e se ajustassem aos resultados da prova mais fácil (A).

A decisão pragmática a tomar na altura da realização do teste não poderia esperar pela análise cuidada que a seguir se fez pela equiparação com os diferentes métodos propostos na literatura para as amostras pequenas. Como a diferença das médias entre as notas do teste A e do teste B era de aproximadamente 1 valor na escala oficial de 0-20, aumentou-se 1 valor aos resultados das provas B.

A simulação das equiparações com o emprego dos métodos não encadeados (sem perguntas âncora) e encadeados (com perguntas âncora) mostrou que a ancoragem introduz uma grande melhoria no paralelismo da equiparação, na consistência da precisão e no valor absoluto do erro. Comparando os diferentes métodos de equiparação encadeada verifica-se que o encadeamento linear pelo método de Tucker e o encadeamento do arco de círculo são manifestamente os melhores. A escolha entre eles obriga a uma ponderação entre o melhor desempenho no centro do método linear e o melhor desempenho do arco de círculo nas zonas abaixo de 25% e a partir dos 60%. A nossa ponderação é favorável ao método do arco de círculo. A sua menor precisão no centro é pequena, enquanto que a precisão nas caudas e muito maior. Por outro lado, as regras de pontuação do teste colocam o ponto de corte junto dos 60% de acertos, já que a probabilidade de acerto ao acaso em perguntas de 5 opções é de 20%. Nesse ponto os dois processos têm a mesma precisão. Para cima há uma degradação da precisão do método linear, enquanto que o método do arco de círculo tem uma aumento crescente de precisão.

2. Nota conclusiva

O nosso estudo de equiparação de diferentes formas do mesmo teste mostrou que o procedimento deteta eventuais situações de não equivalência da dificuldade das provas, o que dá uma garantia adicional de equidade da avaliação. Por outro lado permite, que, nos casos em que houver uma não equivalência, uma correção docimologicamente robusta. Além disso, permitiu-nos fazer a seleção do processo de equiparação a selecionar: introdução de perguntas-âncora e realização do método de equiparação encadeada pelo arco de círculo. Na próxima avaliação em situação real o método será aplicado na prova da época normal para uma população de cerca de 300 estudantes.

3. Bibliografia

[1] Amaral M, Moura D, Soares C, Godinho A (2008) Development of resources for computer-based testing in campus-wide IT systems. Proceedings of the IASK meeting. www.iask-web.org
[2] Pasquali L (2003) Psicometria: Teoria dos Testes na Psicologia e na Educação. Editora Vozes, Petrópolis, RJ, Brasil
[3] Livingston SA (1993) Small-sample equating with log-linear smoothing. Journal of Educational Measurement 30:23-29

[4] Case SM, Swanson DB (2002) Constructing written test questions for the basic and clinical sciences. 3rd editon revised. NBME, Philadelphia, Penn, USA.
[5] von Davier AA (2007) Potential solutions to practical equating issues. In: Dorans NJ, Pommerich M, Holland PW (eds) Linking and aligning scores and scales. Springer, New York, USA
[6] Kim S. von Davier AA, Haberman S (2008) Small-sample equating using a synthetic linking function. Journal of Educational Measurement 4:325-342
[7] Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16:297-334
[8] Livingston SA, Kim S (2009) The circle-arc method for equating in small samples. Journal of Educational Measurement 46:330-343