Como Construir Boas Questões?
Uma introdução à Teoria Clássica dos Testes

Ana Rita Mota e J. M. B. Lopes dos Santos

CF/UM/UP, Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto

Resumo

Uma parte muito significativa do tempo dos professores é ocupada por tarefas relacionadas com a avaliação. Apesar da diversidade de meios e instrumentos, o teste ou questionário continua a ser uma componente essencial do processo de avaliação. Mas que confiança temos na nota que atribuímos ao resultado do teste de um estudante? Será uma medida fiável das suas competências e conhecimentos? Ainda que tenhamos elaborado o teste com cuidado e sentido crítico, não seria útil dispor de critérios que nos permitissem aferir a qualidade e utilidade das perguntas?

Este artigo tem como objetivo divulgar um processo de análise de testes, desenvolvido no contexto das Ciências Sociais - a Teoria Clássica dos Testes - que pode ser aplicado com vantagem ao contexto educativo. A análise clássica de um teste é feita à posteriori, com base nos resultados obtidos pelos estudantes e aborda quer a questão da fiabilidade global do teste quer a da qualidade e utilidade de cada uma das questões (itens). Pode, pois, constituir um complemento ao processo de elaboração de um teste e promover a melhoria contínua da qualidade da avaliação.

1 Introdução

Indicadores internacionais são unânimes ao considerar que a avaliação é um dos fatores que mais condiciona o processo de aprendizagem [1,2]. Apesar da importância dos testes de avaliação sumativos, sabemos hoje que, no contexto ensino-aprendizagem, a grande utilidade da avaliação se prende mais com o seu caráter formativo, isto é, com a capacidade de se tornar num instrumento de aprendizagem. Diferentes tipos de avaliação, com feedback permanente, permitem desenvolver diferentes tipos de competências num menor espaço de tempo, para além de estimularem os estudantes a ser mentalmente mais ativos na sala de aula e proporcionarem uma aprendizagem significativa [2]. Um rápido e sistemático feedback é a chave da avaliação pedagógica, porque permite que os alunos saibam em que patamar estão e o que é preciso fazer para atingir os seus objetivos.

Qualquer metodologia de ensino que se pretenda implementar deverá ser pensada e construída com base num modelo de avaliação sólido, contínuo e diversificado [1]. Esta mudança de paradigma carece de um pensamento estruturado por parte dos professores, tornando-se necessário um esforço coletivo para pensar e preparar, com rigor, novos instrumentos de avaliação. Desta reflexão surgem, inevitavelmente, algumas inquietações: dado o elevado número de alunos numa turma e consequente trabalho de correção, como operacionalizar um modelo de avaliação formativa com feedback rápido? Neste contexto, os testes de escolha múltipla aparecem como uma alternativa atraente. Mas, como sabemos se estamos a construir bons itens?

Quando medimos o comprimento de uma mesa ou o período de oscilação de um pêndulo, temos uma ideia clara da natureza da grandeza que medimos e da adequação dos instrumentos que usamos. Mas, se pretendemos medir um coeficiente de inteligência, um traço de personalidade, ou, até, o grau de conhecimento em Física Newtoniana, a situação é muito diferente. No contexto das Ciências Sociais ou no contexto educativo, as variáveis que pretendemos medir são habitualmente designadas por latentes, para dar ênfase à sua inacessibilidade. A sua própria validade conceptual é reforçada pela qualidade do processo de medição. Por isso, os investigadores em ciências sociais e os psicólogos, em particular, sempre dedicaram muita atenção à análise dos processos de medida, que, em geral, tomam a forma de administração de testes. A Teoria Clássica dos Testes (TCT) é uma das respostas a esta problemática; apesar de existirem alternativas mais recentes e complexas, que beneficiam do aumento do poder computacional, a TCT continua a ser largamente utilizada [3,4].

A análise de um teste em TCT tem duas componentes fundamentais: (a) avaliação da fiabilidade do teste como um todo; (b) avaliação da qualidade individual de cada pergunta (item), usando dois índices, um de dificuldade e outro de poder de discriminação. Faremos uma apresentação muito sucinta destas componentes com alguns exemplos. Para uma exposição mais detalhada, recomendamos um dos textos de referência da área,

com o título sugestivo "*Introduction to Measurement Theory*" [5]. Começaremos pela análise de itens.

2. Análise de itens

A TCT [3,4] realça as propriedades psicométricas dos itens de uma prova que correspondem aos parâmetros índice de dificuldade e índice de discriminação.

2.1 Índice de dificuldade

O índice de dificuldade, com valores entre 0 e 1, é definido pela fração de acerto nesse item, ou seja, é a razão entre a soma das pontuações obtidas pelos estudantes e a pontuação máxima que poderiam obter (pontuação máxima do item vezes o número de estudantes). Um item é tanto mais difícil quanto menor é o seu índice de dificuldade. Não parece uma nomenclatura feliz, mas está consagrada na literatura. Para que um instrumento de avaliação tenha um nível ideal de dificuldade, Pasquali [6] recomenda uma distribuição de níveis de dificuldades de itens de um teste dentro de uma curva normal, conforme a Tabela 1.

Quantidade ideal de itens na avaliação (em %)	Índice de dificuldade do item	Classificação do item em relação ao índice de dificuldade
10%	Superior a 0,9	Muito fáceis
20%	De 0,7 a 0,9	Fáceis
40%	De 0,3 a 0,7	Medianos
20%	De 0,1 a 0,3	Difíceis
10%	Inferior a 0,1	Muito difíceis

Tabela 1 - Distribuição sugerida de itens por índice de dificuldade ([6])

A Tabela 2 apresenta as cotações de 10 alunos a duas questões, Q1 e Q2, e mostra como se calcula o índice de dificuldade de cada item. A questão Q1 é de escolha múltipla (0 ou 6 pontos) e a questão Q2, com maior índice de dificuldade, é uma resposta restrita (máximo 10 pontos).

	A	B	C	D
1				
2			Q1	Q2
3		Aluno 1	0	8
4		Aluno 2	6	10
5		Aluno 3	6	6
6		Aluno 4	6	3
7		Aluno 5	6	4
8		Aluno 6	6	10
9		Aluno 7	0	6
10		Aluno 8	0	0
11		Aluno 9	0	8
12		Aluno 10	0	4
13		Índice de dificuldade	0,50	0,59
14			SOMA((C3:C12)/(10*6)	SOMA((D3:D12)/(10*10)

Tabela 2 - Cáculo do índice de dificuldade. O fundo amarelo refere-se às fórmulas

2.2 Índice de discriminação

O índice de discriminação de um item compara o sucesso obtido nesse item de dois grupos de estudantes: os que obtiveram maior classificação na prova com os que obtiveram menor classificação, isto é, mede a capacidade do item diferenciar os alunos de acordo com o seu desempenho global no teste. É habitualmente definido como a diferença entre as frações de acerto[1] de dois grupos de estudantes: os 27 % respondentes com pontuações mais altas (grupo A) e os 27 % respondentes com pontuações mais baixas (grupo C). O item é tanto mais discriminativo quanto maior for o valor do índice de discriminação. Este pode assumir qualquer valor entre -1, caso todos os estudantes do grupo C tenham cotação máxima e os do grupo A cotação nula e, +1, caso todos os estudantes do grupo A tenham cotação máxima e os do grupo C cotação nula. Quando as percentagens de respostas certas no grupo A e C são muito próximas, o índice de discriminação será próximo de zero. Itens com índice de discriminação superior a 0,3 são habitualmente considerados discriminativos [7,8] (Tabela 3).

[1] Soma das cotações obtidas pelos estudantes sobre a cotação máxima que poderiam obter.

Valor do índice de discriminação	Classificação
Inferior a 0,2	Item deficiente, deve ser rejeitado
De 0,2 a 0,3	Item marginal, sujeito a reformulação
De 0,3 a 0,4	Item bom, sujeito a aprimoramento
Superior a 0,4	Item bom

Tabela 3 - Classificação dos itens com base no índice de discriminação ([7])

Consideremos novamente as respostas de 10 alunos às questões Q1 e Q2, à qual acrescentámos uma terceira coluna com a cotação total no teste (admitindo que o teste é constituído apenas pelas questões Q1 e Q2). O primeiro passo é identificar o percentil 27 e o percentil 73 (Tabela 4, coluna C-16 e C-17).

	A	B	C	D	E
1					
2			Q1	Q2	TOTAL
3		Aluno 1	0	8	8
4		Aluno 2	6	10	16
5		Aluno 3	6	6	12
6		Aluno 4	6	3	9
7		Aluno 5	6	4	10
8		Aluno 6	6	10	16
9		Aluno 7	0	6	6
10		Aluno 8	0	0	0
11		Aluno 9	0	8	8
12		Aluno 10	0	4	4
13					
14					
15					
16		Percentil 27	6,86	PERCENTIL(E3:E12;0,27)	
17		Percentil 73	11,14	PERCENTIL(E3:E12;0,73)	
18					

Tabela 4 - Cálculo do índice de descriminação - Parte I

O próximo passo é reordenar os alunos por ordem crescente (ou decrescente) de cotação total da prova para melhor identificar os alunos que pertencem a cada um dos percentis e calcular o índice de discriminação (Tabela 5). Neste caso, verificamos que a questão Q1 é a mais discriminadora.

	C	D	E	F
		Q1	Q2	TOTAL
	Aluno 8	0	0	0
	Aluno 10	0	4	4
	Aluno 7	0	6	6
	Aluno 1	0	8	8
	Aluno 9	0	8	8
	Aluno 4	6	3	9
	Aluno 5	6	4	10
	Aluno 3	6	6	12
	Aluno 2	6	10	16
	Aluno 6	6	10	16
	PERCENTIL 27	0,000	0,333	
		SOMA(D4:D6)/(CONTAR(D4:D6)*6)	SOMA(E4:E6)/(CONTAR(E4:E6)*10)	
	PERCENTIL 73	1,000	0,867	
		SOMA(D11:D13)/(CONTAR(D11:D13)*6)	SOMA(E11:E13)/(CONTAR(E11:E13)*10)	
	Índice de discriminação	1,000	0,533	
		D17-D14	E17-E14	

Tabela 5 - Cálculo do índice de descriminação - Parte II

3. Fiabilidade

3.1 O conceito de fiabilidade

Uma das características mais relevante de um teste é sua fiabilidade [9]. A fiabilidade é uma avaliação da reprodutibilidade dos resultados de um teste. Ao contrário do que acontece em medidas físicas, em que podemos realmente repetir em condições invariantes, não podemos, na prática, repetir a administração de um teste, e, mesmo que o pudéssemos fazer, é duvidoso que o pudéssemos considerar como uma nova medida, independente da primeira. Como a medida não pode ser repetida, fica claro que a tarefa avaliar a fiabilidade de um teste tem de ser conseguida com os resultados de uma única medida, com base em hipóteses sobre os erros. A discussão do conceito de fiabilidade requer o recurso a alguns conceitos de estatística. Para tornar a apresentação mais acessível, usaremos, preferencialmente, exemplos concretos, em vez de formulações mais gerais.

O ponto de partida é um questionário (teste) administrado a um grupo de $m = 6$ respondentes, cujos resultados estão reproduzidos na Tabela 6. O investigador (professor) apura a pontuação (score) de cada respondente no teste, expresso por um número X_α, em que $\alpha = 1,...,6$ é um índice que percorre todo os sujeitos do teste. A média das pontuações na amostra dos respondentes no teste está calculada na célula B8:

$$\overline{X} := \frac{1}{6}(X_1 + X_2 + X_3 + X_4 + X_5 + X_6) = 12{,}17 \qquad (1)$$

	A	B	C	D
1	Estudante	Pontuação, X_α	$X_\alpha - \overline{X}$	$(X_\alpha - \overline{X})^2$
2	A1	8,00	-4,17	17,36
3	A2	9,00	-3,17	10,03
4	A3	13,00	0,83	0,69
5	A4	17,00	4,83	23,36
6	A5	13,00	0,83	0,69
7	A6	13,00	0,83	0,69
8		Média, $\overline{X}$= **12,17**	Variância, ΔX^2 =	**8.81**
9		**=MEDIA(B2:B7)**		**=MEDIA(D2:D7)**

Tabela 6 - Resultados de 6 estudantes num teste

Os desvios da pontuação de cada estudante em relação a média, $X_\alpha - \overline{X}$, representados na coluna C, dão-nos uma ideia da variação de pontuações de cada respondente em relação à média. Por definição, a média destes desvios é nula, ou seja, a soma dos valores da coluna C é nula.

$$\overline{(X_\alpha - \overline{X})} = \frac{1}{6}\left[(X_1 - \overline{X}) + (X_2 - \overline{X}) + (X_3 - \overline{X}) + (X_4 - \overline{X}) + (X_5 - \overline{X}) + (X_6 - \overline{X})\right] = \overline{X} - \frac{1}{6}(6 \times \overline{X}) = 0 \qquad (2)$$

Na coluna D, calculámos os quadrados dos valores dos desvios $(X_\alpha - \overline{X})^2$ e, por baixo, a respetiva média, designada por variância, que, no caso presente, é 8,81; os desvios positivos e negativos contribuem agora com o mesmo sinal e o resultado é um valor positivo (ou nulo, apenas no caso em que todos as pontuações são iguais):

$$\Delta X^2 = \frac{1}{6}\left[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + (X_4 - \overline{X})^2 + (X_5 - \overline{X})^2 + (X_6 - \overline{X})^2\right] \qquad (3)$$

O desvio padrão é definido por $\Delta X := \sqrt{\Delta X^2}$.

A hipótese de base da CTC é que a pontuação X_α de um respondente ao questionário é a soma de um valor verdadeiro (*true score*), T_α, a grandeza que o teste pretende medir, com um erro aleatório,

$$X_\alpha = T_\alpha + E_\alpha, \qquad (4)$$

um conceito muito semelhante ao de uma medida física. Ao classificar o erro como aleatório, estamos a imaginar que, em hipotéticas repetições da administração do teste, em circunstâncias invariantes, a pontuação verdadeira seria a mesma, mas o erro variaria. Contudo, em cada teste só temos acesso à pontuação total, X_α, não a cada uma das suas componentes. A pontuação verdadeira T_α é um constructo hipotético, por vezes, designado por variável latente. A TCT baseia-se em algumas hipóteses acerca dos erros E_α para inferir propriedades deste processo de medição. As mais importantes são: (**a**) A média do erro para cada respondente (média numa hipotética população de testes repetidos) é nula, $\langle E_\alpha \rangle$; (**b**) A respetiva variância é a mesma para todos os respondentes, $\Delta E_\alpha^2 = \langle (E_\alpha - \langle E_\alpha \rangle)^2 \rangle = \Delta E^2$. Com algumas hipóteses adicionais sobre os erros[2] chega-se a um resultado fundamental da TCT:

$$\Delta X^2 = \Delta T^2 + \Delta E^2 \qquad (5)$$

em que ΔT^2 é a variância dos scores verdadeiros.

$$\Delta T^2 = \frac{1}{m}\sum_\alpha \left(T_\alpha - \overline{T}\right)^2 \qquad (6)$$

e $\overline{T}$ é a média das pontuações verdadeiras

$$\overline{T} := \frac{1}{m}\sum_\alpha T_\alpha \qquad (7)$$

O índice de fiabilidade de um teste, Φ, é definido, simplesmente, como a razão entre a variância das pontuações verdadeiras e a das pontuações observadas

$$\Phi = \frac{\Delta T^2}{\Delta X^2} = \frac{\Delta X^2 - \Delta E^2}{\Delta X^2} = 1 - \frac{\Delta E^2}{\Delta X^2} \qquad (8)$$

A fiabilidade toma valores no intervalo [0,1], uma vez que $\Delta X^2 \geq \Delta E^2$.

Esta definição é acompanhada de um certo desconforto: na Eq.(8) só temos acesso a ΔX^2 (Eq. 3), o que não nos permite determinar Φ. Não podemos inserir na Tabela 6, uma coluna com as verdadeiras pontuações; se as soubéssemos, o conceito de fiabilidade seria inútil.

Coloquemos de lado, por um momento, a questão de como estimar Φ e investiguemos o seu significado, supondo que sabemos o seu valor. Da Eq. (8) tiramos $\Delta E = \sqrt{1 - \Phi}\,\Delta X$.

[2] Para uma listagem detalhada das hipóteses da TCT, ver [5].

Tal como acontece numa medida física, quando indicamos, além do valor medido, a incerteza na forma de mais ou menos, um desvio padrão, podemos expressar a pontuação do sujeito como

$$S_\alpha = X_\alpha \pm \Delta E = X_\alpha \pm \sqrt{1-\Phi}\Delta X. \quad (9)$$

Por exemplo, para $\Phi = 0{,}5, S_\alpha = X_\alpha \pm 0{,}71\Delta X$, o erro de medida é cerca de 70 % do desvio padrão da população.

Para percebermos a pouca utilidade de um tal teste, tomemos o exemplo de um exame nacional com distribuição normal de classificações, em que o score médio é $\overline{X} = 11$ valores (percentil 50) e o desvio padrão é $\sigma = \Delta X = 3{,}5$. Um cálculo simples mostra um erro nas classificações de $\pm 2{,}5$; um estudante com 13 valores teria uma classificação verdadeira entre 10,5, que é o percentil 44 da distribuição de classificações, e 15,5 que é o percentil 90! Sabemos realmente pouco sobre a competência deste estudante como resultado da administração do teste. A Tabela 7 ilustra o significado de diferentes valores de fiabilidade.

Φ	$\Delta E/\Delta X = \sqrt{1-\Phi}$
1	0
0,90	0,32
0,80	0,45
0,70	0,55
0,60	0,63
0,50	0,72

Tabela 7 - Valores de fiabilidade e razão entre o erro da medida da pontuação de cada estudante (desvio padrão, ΔE) e o desvio padrão do conjunto de pontuações de todos os estudantes (ΔX).

3.2 Estimar Φ

Suponhamos que temos a possibilidade de fazer duas administrações de testes paralelos, *A* e *B* e conhecemos as pontuações de cada respondente nos dois testes, $X_\alpha(A)$ e $X_\alpha(B)$. Os valores de $X_\alpha(A)$ e $X_\alpha(B)$ não são iguais, devido ao erro de medida, mas estão naturalmente correlacionados, uma vez que, por hipótese, cada sujeito do teste tem a mesma pontuação verdadeira nos dois testes. Uma representação gráfica da pontuação do segundo teste em função da do primeiro ajuda-nos a visualizar esta relação.

Se os erros fossem nulos ($\Delta E = 0$), teríamos os pontos $(x,y) = (X_\alpha(A)$ e $X_\alpha(B))$ sobre uma reta de declive 1, pois $X_\alpha(B) = X_\alpha(B) = T_\alpha$. Se usarmos a relação $\Delta E = \sqrt{1-\Phi}\Delta X = \sqrt{(1-\Phi)/\Phi}\,\Delta T$, podemos simular os resultados de dois testes.

A Figura 1 mostra a simulação do resultado de dois testes para diferentes valores de fiabilidade. A amostra de 30 pontuações verdadeiras é gerada de uma distribuição normal com média $\overline{T} = 11$ e desvio padrão $\Delta T = 3$. Os erros são gerados de uma distribuição de média nula e desvio padrão $\Delta E = \sqrt{(1-\Phi)/\Phi}\,\Delta T$. Esta figura mostra dois resultados possíveis, para dois valores de Φ. Em cada caso, usamos os mesmos valores de pontuações verdadeiras nos dois testes, com média $\overline{T} = 11$ e desvio padrão $\Delta T = 3$; para cada teste geramos erros aleatórios $E_\alpha(A)$ e $E_\alpha(B)$ e com desvios padrões $\Delta E = \sqrt{(1-\Phi)/\Phi}\,\Delta T$.

Os resultados simulados dos testes são:

$$X_\alpha(A) = T_\alpha + E_\alpha(A), \quad (10)$$

$$X_\alpha(B) = T_\alpha + E_\alpha(B). \quad (11)$$

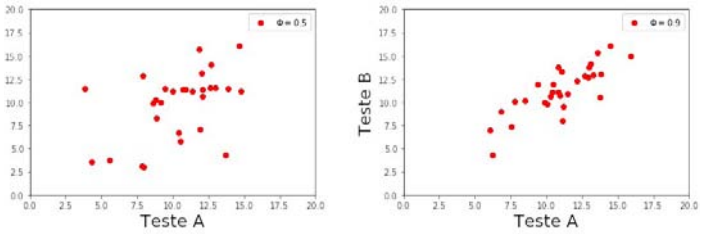

Figura 1 - Simulação de resultados de dois testes para diferentes valores de fiabilidade. A amostra de 30 pontuações verdadeiras é gerada de uma distribuição normal com média $\overline{T} = 11$ e desvio padrão $\Delta T = 3$. Os erros são gerados de uma distribuição de média nula e desvio padrão $\Delta E = \sqrt{(1-\Phi)/\Phi}\,\Delta T$.

É claro que quanto maior for Φ, maior é a correlação entre as pontuações dos dois testes. Se fizéssemos uma regressão linear entres os valores dos dois testes, o coeficiente de regressão R^2 seria mais próximo de 1 no caso de maior valor de Φ. As hipóteses da TCT permitem concluir, precisamente, que Φ é a raiz quadrada de R^2

$$\Phi = \sqrt{R^2}. \quad (12)$$

Parece que não estamos mais próximos do nosso objetivo, pois, desde o princípio, afirmámos a impossibilidade prática de administrar dois testes paralelos. Mas, suponhamos, por um momento, que os diferentes itens do teste medem a mesma competência. Se dividirmos o teste em duas metades, as pontuações verdadeiras de cada respondente em cada metade, normalizadas ao mesmo valor máximo (20 valores ou 100 pontos), devem ser iguais. Ou seja, o teste em si já consiste em dois testes paralelos com metade das perguntas, cujas respostas conhecemos.

Com base neste conceito, num artigo extremamente influente [10], L. Cronbach propôs uma estatística para estimar Φ, conhecida como o parâmetro α de Cronbach, e mostrou que ele constituía uma média da estimativa da fiabilidade sobre todas as partições possíveis do teste em duas metades.

Para um teste com k itens (perguntas), o α de Cronbach é dado por

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} \Delta p_i^2}{\Delta X^2}\right] \quad (13)$$

em que Δp_i^2 é a variância das pontuações no item i.

Vejamos um exemplo de cálculo deste parâmetro, feito numa folha de cálculo (Tabela 8). O exemplo refere-se a 10 alunos num questionário de 7 perguntas. As somas das pontuações de cada aluno estão na coluna I. Na linha 16, usamos a função VAR() para calcular as variâncias Δp_i^2 de cada questão e das pontuações totais, ΔX^2 (célula I16). A fiabilidade é estimada na célula B17 com a fórmula de Cronbach [Eq.(13)].

A estimativa da fiabilidade, através da correlação entre meios-testes, requer que estes sejam testes paralelos. O α de Cronbach só é uma boa estimativa de Φ se as respostas aos itens tiverem uma correlação elevada, ou seja, num teste homogéneo, em que todos itens, ou uma boa parte deles, tenham subjacente a mesma competência[3]. Mas convém salientar que a ausência de correlações dimimui o valor de α, e por isso, o valor de fiabilidade de um teste heterogéneo será, em geral, superior a α. Por esta razão, alguns autores escrevem

$$\Phi \geq \alpha = \frac{k}{k-1}\left[1 - \frac{\sum_i \Delta p_i^2}{\Delta X^2}\right] \qquad (14)$$

Assim, um α de valor razoável ($\alpha > 0{,}7$, por exemplo), mesmo num teste heterogéneo, é uma boa indicação da fiabilidade do teste. Um α baixo pode simplesmente traduzir a heterogeneidade do teste, cuja fiabilidade (reprodutibilidade) pode ser significativamente superior a α (para uma discussão mais detalhada consultar o artigo de Cronbach [10]).

	A	B	C	D	E	F	G	H	I	J
1										
2										
3		GRUPO IV							Score	
4	ALUNO	Q3	Q4	Q5	Q6	Q7	Q8	Q9		
5	1	6	10	6	10	6	6	10	54	SOMA(B5:H5)
6	2	6	5	0	0	0	6	0	17	
7	3	6	8	6	10	6	6	10	52	
8	4	6	8	6	10	6	0	5	41	
9	5	6	5	6	10	6	6	8	47	
10	6	6	8	6	10	6	6	8	50	
11	7	6	8	6	10	6	6	10	52	
12	8	6	0	6	0	0	0	8	20	
13	9	6	7	0	10	6	6	10	45	
14	10	0	8	0	10	6	0	0	24	
15										
16	Variância	3,600	7,789	8,400	17,778	6,400	8,400	15,656	204,84	VAR(I5:I14)
17	Fiabilidade	0,78	7/6*(1-SOMA(B16:H16)/I16)							
18	CorrMedia	0,37	B17/(B17+7*(1-B17))							

Tabela 8 - Exemplo de cálculo de Fiabilidade. As linhas 5 a 14 contêm a pontuação de cada estudante nas 7 questões, somadas na coluna I. A linha 16 calcula as variâncias de cada coluna e na célula B17 é calculada a fiabilidade. A célula B18 calcula a correlação média entre itens.

4. Exemplo de aplicação da TCT

Aplicou-se a TCT a um teste de avaliação de Física e Química, 11.º ano, realizado numa escola portuguesa, em 2018. A prova teve a duração de 120 minutos, a que acresceu a tolerância de 30 minutos. Foi constituída por 28 questões, incluindo itens de seleção (por exemplo, escolha múltipla) e itens de construção (por exemplo, resposta curta e resposta restrita). O teste abordou conteúdos de Física e Química e foi aplicado a 100 alunos no 3.º período.

As respostas aos itens foram tratadas como sendo dicotómicas (0 ou 6 pontos) e politómicas (pontuando também as respostas classificadas com códigos intermédios de 0 a 10 pontos), num total de 20 itens dicotómicos e 8 politómicos. No gráfico da figura 2, os itens da prova, dicotómicos (azul) e politómicos (vermelho), estão registados em função do seu índice de dificuldade e de discriminação. O gráfico mostra que 7 % dos itens são muito fáceis, 50 % fáceis e 43 % são medianos.

A Figura 2 revela, ainda, que 18 itens (cerca de 64 %) têm um índice de discriminação superior a 0,3 pelo que são considerados discriminativos, isto é, bons itens, sendo que cerca de 18 %, precisam de melhorias. Apenas um item deve ser rejeitado e a média dos alunos nesta prova foi de 14,0 valores, o que confirma o elevado número de itens de baixo grau de dificuldade. Contudo, o facto de os itens mais fáceis serem os menos discriminativos confirma a qualidade dos mesmos. A estimativa de fiabilidade, através do alfa de Cronbach, revelou $\alpha = 0{,}85$, ou seja, uma boa fiabilidade.

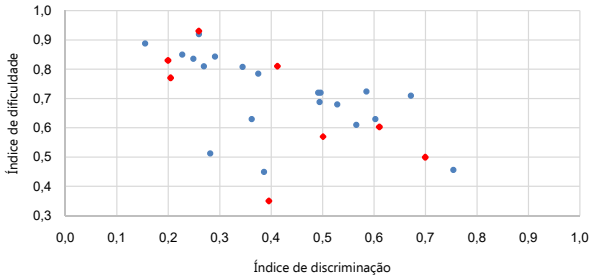

Figura 2 - Índice de dificuldade vs índice de discriminação num teste de 28 questões; a azul as questões dicotómicas e a vermelho as politómicas."

4. Notas finais

A avaliação é, pela sua natureza, subjetiva, mas isso não impede que produza resultados úteis, rigorosos e com significado. O objetivo deste artigo foi apresentar alguns dos conceitos fundamentais da TCT e mostrar que esta abordagem, rápida e simples, pode ser útil no aperfeiçoamento de instrumentos de avaliação, durante a prática letiva no ensino básico, secundário e universitário.

Apesar da Teoria Clássica dos testes ainda ser muito utilizada para analisar a qualidade dos itens, a Teoria de Resposta ao Item (TRI) [3] é a metodologia mais usada, quando se trabalha em larga escala, como, por exemplo, no tratamento estatístico de provas internacionais como o PISA, o TIMMS ou, até, o ENEM (Brasil), uma vez que permite também avaliar o "perfil" do candidato. Por outras palavras, a Teoria Clássica dos testes não permite que o desempenho de estudantes, que fizeram testes diferentes, seja comparado com igualdade, já que a dificuldade da prova poderá interferir diretamente na pontuação de cada um. A TRI, com um modelo mais sofisticado, permite contornar essa dificuldade.

[3] Uma nota de cautela: para a mesma correlação média entre itens, o valor de fiabilidade aumenta com o número de itens e testes com mais itens tenderão a ter maior fiabilidade.

Referências

[1] MICHAELSON, Larry K.; SWEET, Michael; PARMELEE, Dean X. Team-Based Learning: small group learning's next big step. New York: Wiley Periodicals, 2008.

[2] MOTA, Ana Rita; ROSA, C. W. Ensaio Sobre Metodologias Ativas: Reflexões E Propostas. Revista Espaço Pedagógico, Vol. 25, n.º 2, p. 261-76, 2018.

[3] SARTES, L. M. A.; SOUZA-FORMINGONI, M. L. O. Avanços na psicometria: da Teoria Clássica dos Testes à Teoria de Resposta ao Item. Psicologia: Reflexão e Crítica, v. 26, n.º 2, p. 241-250, 2013.

[4] BORGATTO, Adriano Ferreti; ANDRADE, Dalton Francisco de. Análise Clássica de Testes com diferentes graus de dificuldade. Estudos em Avaliação Educacional, São Paulo, v. 23, n.º 52, p. 146-156, 2012.

[5] ALLEN, M.J.; YEN, W.M. Introduction to Measurement Theory. Waveland Press, 2001.

[6] PASQUALI, L. Psicometria: Teoria dos Testes na Psicologia e na Educação. Petrópolis: Vozes, 2003.

[7] RABELO, M. Avaliação Educacional: Fundamentos, Metodologia e Aplicações no Contexto Brasileiro. Rio de Janeiro: SBM, 2013

[8] COSTA, P.; OLIVEIRA, P.; FERRÃO, M. E.. Statistical issues on multiple choice tests in engineering assessment. SEFI 37th Annual Conference, 2009.

[9] MAROCO, João; GARCIA-MARQUES, Teresa. Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? v. 4, n.º 1, pg. 65-90, 2013.

[10] CRONBACH, Lee J.. Coefficient alpha and the internal structure of tests. Psychometrika, v. 16, n.º 3, p. 297-334,1951.

Ana Rita Mota, licenciou-se em Ensino da Física e Química na Universidade de Aveiro e obteve o doutoramento em Física na Faculdade de Ciências da Universidade do Porto, sob orientação do professor João Lopes dos Santos. Em 2015/2016 integrou, na Universidade de Harvard, o grupo de investigação de Eric Mazur. Foi sua professora assistente no curso *Applied Physics*, conhecido pela metodologia *Team & Project-based approach*. Em Harvard, foi também professora assistente do físico David Keith. É autora de artigos publicados em revistas nacionais e internacionais e tem como principal interesse de investigação as áreas de ensino colaborativo, avaliação e metacognição. Atualmente é investigadora pós-doc no Departamento de Física e Astronomia da FCUP. Paralelamente, leciona Física e Química no colégio de Nossa Senhora do Rosário, Porto.

João Lopes dos Santos, é professor de Física na Faculdade de Ciências da Universidade do Porto. Doutorou-se em Física Teórica da Matéria Condensada no Imperial College, e desenvolve o seu trabalho científico nesta área. O ensino da Física e a divulgação científica tiveram sempre um papel importante na sua carreira de professor. Foi coordenador do Projecto Faraday de intervenção no ensino secundário, promovido pela Fundação Calouste Gulbenkian. Considera que ser Físico é uma das melhores ocupações para quem tem uma curiosidade que atravessa muitas fronteiras.