Revista Brasileira de Ensino de Física, vol. 40, nº 2, e2401 (2018)
www.scielo.br/rbef
DOI: http://dx.doi.org/10.1590/1806-9126-RBEF-2017-0232

Pesquisa em Ensino de Física

Efeito da penalização em itens dicotômicos no ensino de Física

Effect of penalizing wrong answers in true/false physics tests

Fábio F. Monteiro*[1], Cecília B. Alves[2,3], Bernardo A. Mello[1]

[1]Instituto de Física, Universidade de Brasília, Caixa Postal 04455, 70919-970, Brasília, DF, Brasil
[2]Programa de Pós-Graduação em Desenvolvimento, Sociedade e Cooperação Internacional, Universidade de Brasília, Brasília, DF, Brasil
[3]Medical Council of Canada, Ottawa, ON, Canada

Recebido em 27 de Julho, 2017. Aceito em 28 de Agosto, 2017.

As vantagens e desvantagens do uso de questões de julgamento em avaliações de aprendizagem, bem como a inserção da penalização de respostas erradas como mecanismo de ajuste sobre o acerto casual, já foram apresentadas e discutidas por diversos autores no contexto da avaliação de aprendizagem. No entanto, pouco estudo tem sido feito no Brasil no sentido de entender a extensão da influência do uso da penalização de respostas erradas na medida da proficiência do aluno no domínio avaliado. Neste sentido, este artigo apresenta um estudo realizado na Universidade de Brasília, com os alunos das disciplinas básicas de Física-1 e Física-2 oriundos de 20 cursos diferentes, no modelo das Disciplinas Unificadas da Física, que foram submetidos alternadamente a avaliações com penalidade e sem penalidade. O efeito da redução da confiabilidade do escore devido à penalidade é confrontado estatisticamente com o efeito do aumento do poder discriminativo da prova devido à redução do "chute". Uma nova grandeza chamada ganho na qualidade γ é proposta para interpretar os resultados desses efeitos combinados. Ao final, fica demonstrado que o aumento do poder discriminativo da prova supera a redução da confiabilidade do escore quando se aplica a penalização de respostas erradas em itens de julgamento.
Palavras-chave: disciplinas de massa, itens dicotômicos, penalização de respostas, confiabilidade do escore, poder discriminativo da prova.

The advantages and disadvantages of using true/false tests in learning assessments, as well as the insertion of penalization (negative marking) of wrong answers as mechanism of adjustment on the guessing success, have already been presented and discussed by several authors in the context of learning assessment. However, few studies has been done in Brazil to understand the extent of the influence of using penalization of wrong answers in the measure of student proficiency in the evaluated domain. Hence, this article presents a study carried out at the University of Brasília, with the students of the basic disciplines of Physics-1 and Physics-2 coming from 20 different courses, in the unified Physics disciplines model. The students were alternately submitted to assessments with penalty and without penalty. The effect of reducing the reliability of the score due to the penalty is statistically compared with the effect of increasing the discriminatory power of the test due to the reduction of the guessworking. A new magnitude called "statistical yield" is proposed to interpret the results of these combined effects. To conclude, it is demonstrated that the increase in the discriminatory power of the test outweighs the reduction of the reliability of the score when applying penalization of wrong answers in true/false tests.
Keywords: Mass disciplines, true/false tests, penalization of answers, reliability of the score, discriminatory power of the test

1. Introdução

O principal objetivo das avaliações de aprendizagem é fornecer informações sobre o nível de domínio de alguém em um determinado assunto, a partir de suas respostas a testes. Em outras palavras, busca-se avaliar a verdadeira proficiência do indivíduo no domínio avaliado. Para tanto, o escore atribuído a ele deve corresponder, o máximo possível, à sua proficiência real e não a variáveis diversas, tais como a (má) qualidade das questões, falhas na aplicação dos testes, ou os métodos de atribuição de pontos. Crescentemente, a avaliação tem sido vista como parte vital do processo de ensino e aprendizagem, uma vez que auxilia na compreensão sobre se os objetivos da aprendizagem tem sido ou não alcançados [1].

Dietel, Herman e Knuth (1991) [2] definiram avaliação como qualquer método utilizado para melhor compreender o conhecimento de um estudante, em um dado momento. As avaliações de aprendizagem no Brasil utilizam-se tanto de questões abertas quanto de questões objetivas para a aferição de conhecimentos e habilidades.

*Endereço de correspondência: fmonteiro@unb.br.

Nas questões abertas (também chamadas discursivas), os respondentes utilizam suas próprias palavras no desenvolvimento das respostas, sem o benefício de sugestões ou alternativas dadas a priori. Embora esse tipo de questão permita que o aluno manifeste sua capacidade de expressão, criatividade e nível de compreensão acerca de assuntos variados, desde simples a complexos, as questões discursivas tem limitadores sérios a sua ampla utilização: sua correção é sujeita à maior influência de aspectos subjetivos do avaliador, exigem muito mais tempo para a correção e, por conter usualmente poucas questões, permitem a avaliação de uma gama bem mais restrita da lista de conhecimentos e habilidades desenvolvidos em sala de aula.

Nas questões de múltipla escolha, os respondentes devem escolher a opção correta dentre, em geral, quatro ou cinco opções de resposta. Outro tipo de questão bastante utilizada é a de julgamento, ou resposta dicotômica, em que o aluno assinala se uma afirmação é verdadeira ou falsa (V ou F). Questões objetivas têm sido criticadas por incentivar a mentalidade [de que há somente] uma-resposta-correta [3], forçar respondentes a darem resposta simplista a assuntos complexos [4], que não conseguir medir níveis cognitivos elevados, tais como síntese, avaliação e pensamento criativo [5] e levar a uma maior probabilidade de acerto casual [6]. Contudo, várias vantagens na utilização de desse tipo de questão também têm sido identificadas. Por exemplo, questões objetivas são reconhecidas por: (1) permitir a cobertura de uma faixa mais ampla de conhecimentos e habilidades almejados, favorecendo uma amostragem mais representativa do conteúdo dado em sala de aula; (2) Eliminar o subjetivismo na atribuição das notas; (3) Permitir correção simples e rápida, propiciando um rápido feedback acerca do desempenho dos alunos [7].

Entretanto, elaboradores de testes, professores, pedagogos, administradores educacionais etc, têm opiniões diversas sobre a utilização de questões objetivas para aferir o conhecimento dos alunos. A utilização das questões de julgamento e sobre a forma de pontuar tais questões, atribuindo pontos negativos para respostas erradas ou não [8], tem sido posto em debate. Também tem sido discutido aspectos na administração dos testes, como por exemplo, se os examinandos deveriam ser incentivados a responder todas as questões, mesmo que isso signifique, por vezes, chutar cegamente. Acerca do chute cego, Richard Burton é enfático ao afirmar em seu artigo "Multiple-choice and true/false tests: myths and misapprehensions" (2005), que as respostas dos alunos raramente são completamente desinformas ou aleatórias (cegas). Ao contrário, geralmente, os respondentes são guiados por um conhecimento parcial do assunto. Também afirma que os chutes serão menos frequentemente cegos se existirem pistas, erros ou distratores implausíveis nas questões.

Essa questão acerca dos acertos ao acaso, e a consequente elevação do escore final do respondente, mesmo quando este sabe pouco sobre o assunto, é uma preocupação bastante frequente entre pesquisadores [6, 9, 10]. O "chute" exerce influência sobre a fidedignidade das provas, que, por sua vez, está intimamente ligada à atribuição de escores que representem acuradamente a proficiência do respondente no domínio avaliado [11–13] .

Em estudo realizado por Espinosa e Gardeazabal (2010) [14], encontrou-se que a quantidade de itens deixados em branco, ao se utilizar a correção com apenação, depende da pena utilizada: quanto maior a pena, maior o número de omissões. Além disso, segundo esses autores, respondentes mais aversos ao risco tendem a deixar mais itens em branco do que respondentes com a mesma proficiência no assunto, mas que sejam mais inclinados ao risco.

Apesar de as críticas acerca da utilização de questões objetivas serem antigas, sendo a principal delas o aumento da probabilidade de chute [15], poucos estudos, em especial no Brasil, tem se sido conduzidos no sentido de investigar em que extensão a utilização de itens de julgamento e a utilização da penalidade são prejudiciais para a obtenção dos escores nas provas que representem, de fato, a proficiência do aluno no domínio avaliado.

Esse artigo tem por objetivo contribuir para o entendimento de como a utilização da penalidade (marcação negativa) afeta os escores obtidos pelos alunos, e a consequente interpretação de sua proficiência, por meio de uma prova com itens de julgamento, itens de múltipla escolha, e itens numéricos. Para tanto, foi realizado um estudo de delineamento quase-experimental, comparando dois grupos de alunos universitários, que receberam, alternadamente, provas com e sem penalidade.

2. Metodologia

No segundo semestre de 2015 os alunos da Universidade de Brasília, matriculados nas disciplinas básicas de Física-1 e Física-2, oriundos de 20 cursos diferentes, no modelo de Disciplinas Unificadas da Física [16], foram submetidos, simultaneamente às provas unificadas P2 e P3. Os alunos foram divididos em dois grupos em função da primeira letra do seu nome, o primeiro grupo com letras A a J e o segundo com letras de K a Z, resultando em quantidade muito próxima de alunos nos dois grupos, como pode ser visto da Tabela 1.

O modelo de prova aplicada foi do tipo avaliativo-objetivo composto de 20 itens distribuídos em: 12 questões de julgamento; 4 questões numéricas, subdivididas em 3 itens cada uma; e 4 questões de múltipla escolha. As questões de julgamento exploravam o domínio dos conceitos Físicos e a compreensão de fenômenos naturais. O aluno deveria demonstrar a habilidade de julgar a veracidade de afirmações relacionadas a conceitos centrais de Física, apresentados de diversas formas na linguagem científica e matemática. As questões numéricas exigiam a competência da compreensão de fenômenos naturais, e a habilidade de inter-relacionar diferentes objetos de conhecimento de Física e Matemática, além da habilidade

Tabela 1: Consolidação dos resultados das provas utilizados nesse artigo. Os coeficientes angulares foram obtidos das regressões lineares da Fig. 3 e medem o poder discriminativo das questões de julgamento. O desvio padrão $\sigma_{\text{médio}}$, calculado em cada um dos intervalos de agrupamento dos dados da Fig. 3, mede o erro atribuído às notas das questões de julgamento. O fator de qualidade da prova, Q, é definido pela Eq. 12 e o ganho de qualidade, γ, pela Eq. 13.

		Sem penalidade				Com Penalidade					
Prova	Disc.	Num. Alunos	Coef. Angular	$\sigma_{\text{médio}}$	Q	Num. Alunos	Coef. Angular	$\sigma_{\text{médio}}$	Q	γ	γ_r
P2	Fis. 1	363	0,292	0,124	2,35	360	0,591	0,202	2,63	1,12	12%
	Fis. 2	233	0,207	0,105	1,97	232	0,374	0,187	2.00	1,02	2%
P3	Fis. 1	322	0,351	0,146	2,40	327	0,832	0,227	3,67	1,52	52%
	Fis. 2	217	0,185	0,106	1,75	221	0,464	0,204	2,27	1,30	30%

de selecionar e aplicar métodos adequados para análise e resolução numérica de problemas. Cada questão numérica foi dividida em três subitens, interligados, e de níveis progressivos (fácil, médio, e difícil). Em cada um dos subitens o aluno deveria obter como resposta um número com dois algarismos. As questões de múltipla escolha exploravam a competência da compreensão dos fenômenos naturais e do domínio da linguagem científica e matemática. O aluno deveria demonstrar também, a habilidade de interrelacionar diferentes objetos de conhecimento de Física e Matemática, selecionar e aplicar métodos adequados para a resolução de problemas e propor uma solução algébrica (literal). Por fim, o aluno deveria ser capaz de identificar e reconhecer como resposta uma das opções apresentadas. Em ambas as provas P2 e P3, as questões numéricas e de múltipla escolha foram contabilizadas de maneira direta, i.e., sem qualquer tipo de penalidade, enquanto as questões de julgamento foram contabilizadas de dois modos diferentes: na prova P2, os alunos do primeiro grupo tiveram as questões contabilizados de maneira direta, i.e., sem qualquer penalidade, e os alunos do segundo grupo tiveram as questões contabilizadas com penalidade, onde foi considerado o cancelamento de uma questão certa para cada questão errada; Na prova P3 os grupos foram invertidos, de modo que houve penalidade no primeiro grupo e não houve penalidade no segundo grupo.

3. Formulação matemática

Para analisar o efeito da penalidade no cálculo da nota usaremos a seguinte definição:

> *Aluno ideal* é aquele cujo conhecimento do conteúdo é perfeitamente delimitado e consciente. Ao ser confrontado com uma questão ele ou sabe respondê-la corretamente ou desconhece totalmente a resposta. Adicionalmente possui plena consciência dos limites do seu conhecimento, o que exclui a possibilidade de concepções alternativas.

O aluno ideal tem 100% de chance de acerto quando responde um item de V/F que sabe. Ao marcar um item que desconhece, sua chance de acerto é de 50%. Definimos a proficiência, P, do aluno ideal como a fração do conteúdo que ele conhece. Isso significa que a fração esperada de questões conhecidas pelo aluno ideal é igual à sua proficiência.

Ao fazer uma prova com penalidade o aluno ideal poderia considerar marcar os itens que desconhece. Ele erraria, em média, metade desses itens, cancelando a metade correta. O valor esperado da sua menção não se alteraria mas ele introduziria aleatoriedade na sua nota. Para não correr riscos, o aluno ideal opta por marcar apenas os itens que sabe, resultando na fração de itens em branco

$$F_{Bp} = 1 - P. \tag{1}$$

Essa fração é a reta mostrada na Fig. 1 para o número de questões em branco na prova com cancelamento. Marcando ou não os itens que não sabe, o rendimento esperado do aluno ideal na prova com penalidade é igual à

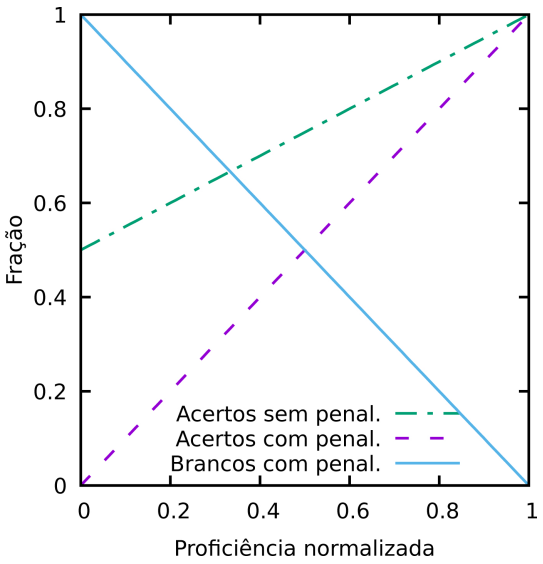

Figura 1: A reta pontilhada, "acertos com penalidade", representa a fração de acertos de itens V/F versus a proficiência do *aluno real* considerando uma prova com penalidade. Estatisticamente esta mesma reta também descreve o caso do *aluno ideal*. A reta cheia, "brancos com penalidade" representa a fração de itens V/F deixados em branco versus a proficiência do *aluno ideal* considerando uma prova com penalidade. A reta tracejada, "acertos sem penalidade" representa a fração de acertos de itens V/F versus a proficiência do *aluno real* considerando uma prova sem penalidade.

sua proficiência,

$$R_{\rm p} = P. \tag{2}$$

Independente do seu conhecimento do conteúdo, o aluno que faz um prova sem cancelamento opta por marcar todas as questões uma vez que essa estratégia aumenta seu rendimento esperado sem incorrer em risco. Com isso, o aluno ideal ganhará, em média, 50% da pontuação das questões que não sabe. Ou seja, o valor esperado de sua nota na prova sem penalidade será

$$\begin{aligned} R_{\rm s} &= P + \frac{1}{2}(1-P) \\ &= \frac{1}{2} + \frac{1}{2}P. \end{aligned} \tag{3}$$

As frações esperadas de questões que o aluno ideal acerta em uma prova com e sem penalidade são vistas na Fig. 1. A menor inclinação da reta na prova sem penalidade, ligada ao coeficiente 1/2 na Eq. 3, reduz seu poder discriminativo, em comparação à prova com penalidade, cujo coeficiente é 1 na Eq. 2.

Além do *aluno ideal*, também usamos a seguinte definição para *aluno real*:

> *Aluno real* é aquele que não tem plena consciência dos limites do seu conhecimento. Ele sabe uma fração de uma parcela do conteúdo e possui concepções alternativas de outra parcela do conteúdo.

Ao responder uma questão, o *aluno real* tem 100% de chance de acertar um conteúdo que conhece e 50% de chance de acertar um conteúdo que desconhece. Ele tem uma probabilidade maior que 50% de acertar um conteúdo que conhece parcialmente e uma probabilidade menor que 50% de acertar um conteúdo sobre a qual possui concepção alternativa.

Pelos mesmos motivos do *aluno ideal*, o *aluno real* marca todas as questões de uma prova sem penalidade e não marca as questões que desconhece em uma prova com penalidade. Além disso, na prova com penalidade ele marca todos os itens que domina e aqueles nos quais tem conhecimento alternativo.

O *aluno real* pode optar por marcar os itens nos quais tem conhecimento parcial, uma vez que sua chance de acerto é maior que a de erro. Neste caso, elas contribuem positivamente para o valor esperado da sua nota. Caso o *aluno real* não marque essas questões, seu rendimento esperado será menor que sua proficiência.

A pontuação negativa atribuída às respostas erradas nas prova com cancelamento penaliza o aluno com concepção alternativa, permitindo distingui-lo do aluno consciente dos limites do seu conhecimento, cujo rendimento médio será maior. Nas questões nas quais o aluno possui conhecimento parcial, o cancelamento faz com que o rendimento esperado seja zero para o aluno cujo conhecimento do conteúdo é insignificante. Sem o cancelamento, o rendimento esperado nessas questões seria de 50%.

Concluindo, a retas que relacionam a fração de itens certos com a proficiência, mostradas na Fig. 1, e expressos na Eq. 2, são as mesmas para o *aluno ideal* e para o *aluno real*, desde que o *aluno real* marque as questões nas quais possui conhecimento parcial. Em comparação com o *aluno ideal* de mesma proficiência, o *aluno real* marca mais questões, deixando menos questões em branco. O número de questões em branco do *aluno real* é uma curva entre o eixo horizontal e a reta da questões em branco da Fig. 1.

Na prova com penalidade a nota do *aluno ideal* provém integralmente de questões nas quais sua chance de acertar é 100%. Já o *aluno real* obtém parte de seu rendimento de questões que acerta com chance menor que 100%. A probabilidade diferente de 1 aumenta o erro na estimativa da proficiência do *aluno real* em relação ao *aluno ideal*.

A presente investigação busca entender os efeitos da prova com e sem penalidade sobre o aproveitamento do aluno nas questões de julgamento e sua relação com a proficiência do aluno. Para o tratamento das questões de julgamento, foram definidas as frações relativas de itens certos (F_{Cp}, F_{Cs}), errados (F_{Ep}, F_{Es}) e em branco (F_{Bp}, F_{Bs}), onde os índices p e s referem-se às provas com e sem penalidade, respectivamente. Assim, para os alunos que fizeram a prova com penalidade,

$$1 = F_{Cp} + F_{Ep} + F_{Bp}, \tag{4}$$

e, para os alunos que fizeram a prova sem penalidade,

$$1 = F_{Cs} + F_{Es}, \tag{5}$$

onde,

$$F_{Bs} \approx 0. \tag{6}$$

Esta última igualdade foi usada porque, neste caso, os alunos não deixaram praticamente nenhum item em branco.

A partir dessas variáveis foi definido o rendimento R, obtido nas questões de julgamento, de modo que, para os alunos que fizeram a prova com penalidade,

$$R_p = F_{Cp} - F_{Ep}, \tag{7}$$

e para os alunos que fizeram a prova sem penalidade,

$$R_s = F_{Cs}, \tag{8}$$

Assim, se fosse dada a possibilidade de fazer a mesma prova sem penalidade a um aluno que fez a prova com penalidade, ele acertaria, em média, metade dos itens que deixou em branco, e não seriam cancelados os itens errados. Neste caso, o novo rendimento R_{sp} poderia ser estimado por

$$R_{sp} \approx F_{Cp} + F_{Bp}/2, \tag{9}$$

Vale ressaltar que a estimativa no sentido contrário não poderia ser feita, pois não há informação sobre os itens que seriam deixados em branco.

 DOI: http://dx.doi.org/10.1590/1806-9126-RBEF-2017-0232

A nota obtida pelos alunos nas questões numéricas e de múltipla escolha, foram utilizadas como estimativa da sua proficiência P e normalizada de modo a obtermos uma proficiência máxima igual a unidade. Em seguida, os alunos foram agrupados conforme sua proficiência, em dez intervalos de 0,1, e a proficiência média de cada grupo de alunos foi calculada. Dentro de cada grupo, foram excluídos os alunos cujo rendimento estava além de dois desvios padrões do rendimento médio dos alunos do grupo.

Dentro de cada intervalo de proficiência foi calculada a média das frações e itens certos $F_C(P)$, a média das frações e itens errados $F_E(P)$ e a média das frações e itens em branco $F_B(P)$, sendo P a proficiência média dos alunos pertencentes ao intervalo. Em consonância com a Eq.(3), para uma prova sem penalidade, fazemos,

$$F_{Bs}(P) \approx 0. \tag{10}$$

e o rendimento médio estimado $R_{sp}(P)$ dos alunos pertencentes ao grupo pode ser dado por

$$R_{sp}(P) \approx F_{Cp}(P) + F_{Bp}(P)/2. \tag{11}$$

4. Resultados e discussões

4.1. Penalidade versus incidência do "chute"

Com a finalidade de investigar o efeito da penalidade sobre a incidência do "chute" nas questões de julgamento, as notas dos alunos foram agrupadas em intervalos de 0,1 conforme sua proficiência, e apresentadas na Fig. 2. A fração de itens V/F em branco foi representada por círculos e quadrados para o caso das provas sem penalidade e com penalidade, respectivamente. A quantidade de estudantes em cada intervalo foi representada pela área dos círculos e quadrados, e o comportamento esperado para a fração de itens V/F, no caso do *aluno ideal*, foi representado pela reta, conforme mencionado anteriormente na Fig. 1.

Assim, presença das figuras cheias muito próximas do eixo horizontal na Fig. 2, indica que, quando não há penalidade, quase todas as questões de julgamento são marcadas, isto é, a fração de itens V/F deixados em branco é praticamente nula, indicando uma forte presença do "chute" em todos os intervalos de proficiência.

As figuras ocas são próximas da reta referente ao *alumo ideal* para alunos de alta proficiência e afastadas dela para alunos com baixa proficiência. Esse é o comportamento esperado em virtude do conhecimento parcial em algumas questões. A presença de um máximo ou um platô à esquerda demonstra que os alunos de baixa proficiência deixam menos questões em branco que o esperado. Esse comportamento tanto pode ser resultado de concepções alternativas como de "chutes' ' deliberados.

4.2. Penalidade versus poder discriminativo das questões de julgamento

Outro fator investigado foi o efeito da penalidade sobre o poder discriminativo das questões de julgamento. Neste caso, os alunos foram agrupados em intervalos de 0,1 conforme sua proficiência, e representados no gráfico 3 por losangos cheios e quadrados cheios, associados aos casos com cancelamento e sem cancelamento, respectivamente. Novamente, a quantidade de alunos em cada intervalo foi representada pela área dos losangos e quadrados.

Em seguida, uma reta foi associada, por regressão linear, aos dados sem cancelamento (quadrados cheios) e outra aos dados com cancelamento (losangos cheios).

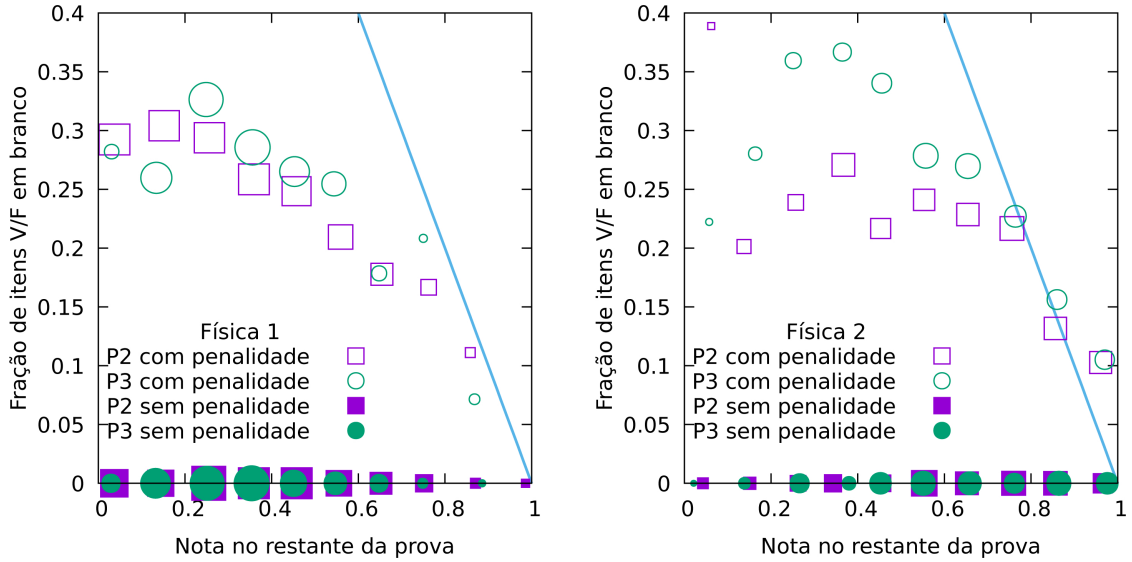

Figura 2: "Fração de itens V/F deixados em branco" versus "Nota no restante da prova". As figuras cheias referem-se à prova sem penalidade. As figuras ocas referem-se à prova com penalidade. A linha reta é uma referência ao caso do *aluno ideal* apresentado na Fig. 1 pela reta cheia "brancos com penalidade". Os quadrados referem-se à prova P2 e os círculos referem-se à prova P3.

A comparação da inclinação das referidas retas com o caso do *aluno ideal*, descrito na Fig. 1, define o poder discriminativo das questões de julgamento. Neste caso, ficou evidente que as provas com cancelamento apresentaram maior poder discriminativo (maior inclinação) que as prova sem cancelamento, ou seja, a presença da penalidade contribui positivamente para melhorar a capacidade média da prova de classificar o aluno conforme sua proficiência.

4.3. Penalidade versus erro associado às notas atribuídas às questões de julgamento

A partir das barras de erro, calculadas para cada intervalo de proficiência, apresentadas na Fig. 3 foi possível avaliar o efeito da penalidade sobre a ocorrência de erro nas notas atribuídas às questões de julgamento. A observação direta, mostra que as barras de erro são maiores nas provas com penalidade e menores nas provas sem penalidade. Isto é um indicativo de que a presença da penalidade aumenta o erro associado às notas atribuídas às questões de julgamento.

4.4. Aumento do poder discriminativo versus aumento do erro

A introdução da penalidade nas questões de julgamento apresenta dois efeitos: aumento do poder discriminativo das questões, e aumento do erro associado à nota atribuída às questões. Para saber se o aumento do poder discriminativo irá ou não suplantar o aumento do erro associado, foi definida a qualidade da medida Q como,

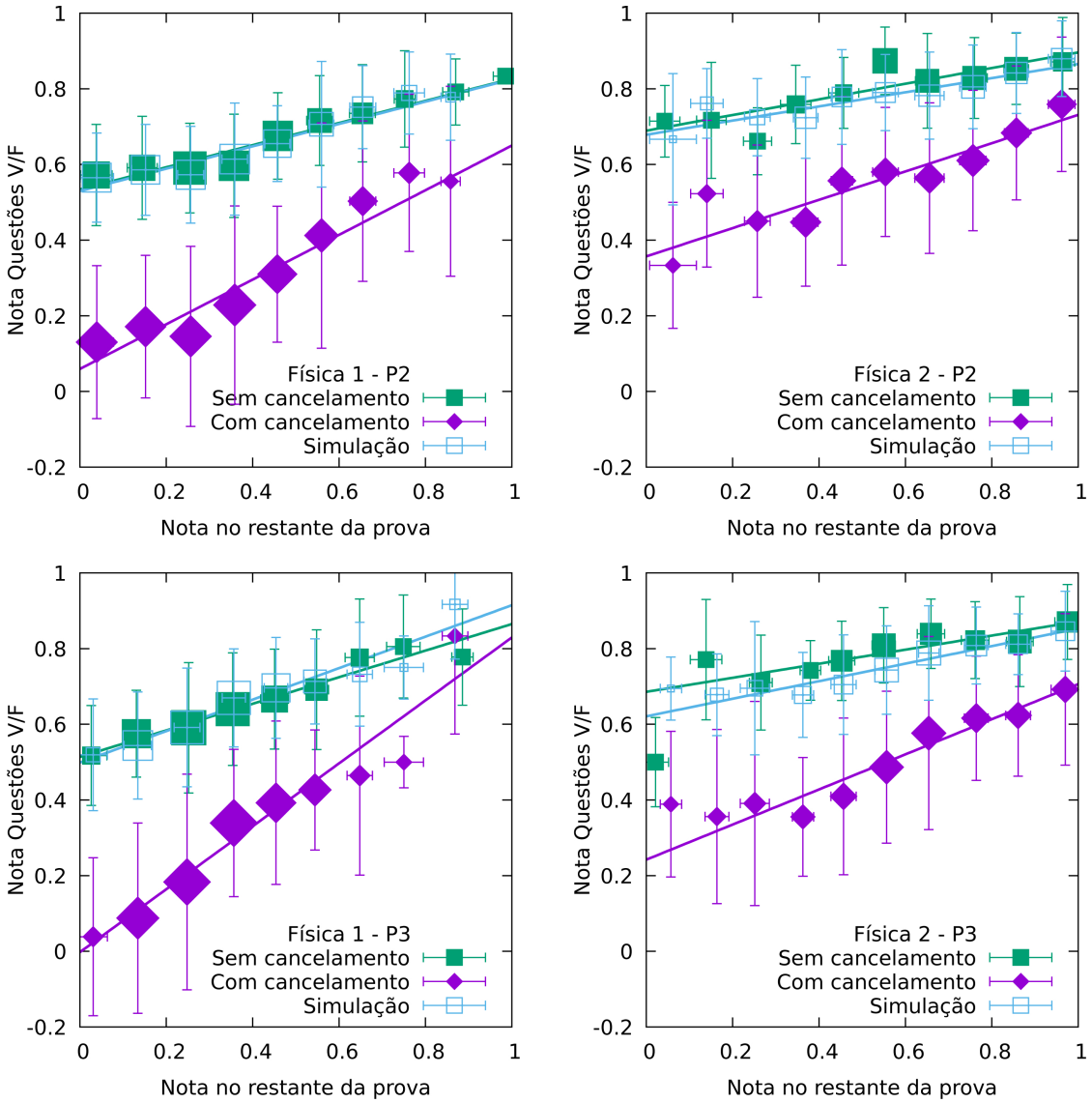

Figura 3: Nota nas questões V/F versus nota no restante da prova de Física 1 e 2. Os quadrados cheios referem-se à prova sem cancelamento. Os losangos cheios referem-se à prova com cancelamento. Os quadrados ocos referem-se à estimativa estatística da nota da prova sem cancelamento, utilizando os dados obtidos na prova com cancelamento.

 DOI: http://dx.doi.org/10.1590/1806-9126-RBEF-2017-0232

$$Q = \frac{\text{Coef. Angular}}{\sigma_{\text{médio}}}, \tag{12}$$

onde o coeficiente angular das retas da Fig. 3, que indica a inclinação das retas, foi utilizado como medida do poder discriminativo das questões de julgamento, e a média dos desvios padrões $\sigma_{\text{médio}}$, calculados em cada um dos intervalos em que as notas foram agrupadas para construir a Fig. 3, foi utilizada como medida do erro associado às notas atribuídas às questões de julgamento.

Desse modo, considerando os casos sem penalidade e com penalidade, para cada prova (P2 e P3) e para cada disciplina (Física 1 e Física 2) foi indicado na Tabela 1 o número de alunos envolvidos, o coeficiente angular, o desvio padrão σ, a qualidade da medida Q e o ganho na qualidade γ. Pontos atípicos (algo em torno de 2,2% a 4.5% dos alunos de cada grupo) foram suprimidos para evitar distorções nas medidas.

Quanto maior o coeficiente angular, maior a semelhança entre a nota dos alunos nas questões de julgamento e a medida de sua proficiência, estimada a partir da sua nota nos itens restantes da prova. Os dados da Tabela 1 indicam que as provas com penalidade apresentaram maior coeficiente angular que as provas sem penalidade, ou seja, nas questões de julgamento, as provas com penalidade apresentaram maior semelhança entre a nota dos alunos nas questões de julgamento e a medida estimada de sua proficiência.

Por outro lado, quando maior o desvio padrão médio $\sigma_{\text{médio}}$, maior o ruído na medida, ou seja, menor o grau de confiabilidade da nota atribuída às questões de julgamento.

Assim, para avaliarmos o quão melhor é a qualidade da medida Q nas provas com ou sem penalidade foi definido o ganho de qualidade γ, como

$$\gamma = \frac{Q_{\text{com penal.}}}{Q_{\text{sem penal.}}} \tag{13}$$

Os dados da Tabela 1 indicam que, na provas P2 de Física 2, o ganho de qualidade relativo γ_r foi pequeno (2%), indicando que, para este caso, o uso ou não da penalidade mostrou-se insignificante na avaliação do aluno. Contudo, na prova P2 de Física 1, o ganho de qualidade relativo γ_r mostrou-se relativamente maior (12%), indicando que, para este caso, o uso da penalidade apresentou um pequeno ganho de qualidade na avaliação do aluno. Por outro lado, nas provas P3 de Física 1 e 2, o ganho de qualidade relativo γ_r foi significantemente maior que a unidade (52% e 30%), indicando que a aplicação do fator de correção foi benéfico, isto é, que o aumento do poder discriminativo decorrente da aplicação da penalidade nos itens dicotômicos superou a redução da confiabilidade das nota decorrente do aumento do erro associado à presença da penalidade.

4.5. Penalidade versus fator emocional gerador de queda de rendimento

Um argumento bastante presente entre os alunos é o de que as provas com penalidade provocam um fator emocional que contribui para a queda rendimento nas provas. Para investigar este fato, foi conduzida uma simulação estatística (quadrados ocos),na Fig. 3, de quais seriam as notas dos alunos que fizeram prova com penalidade (losangos cheios), caso eles pudessem fazer a mesma prova sem penalidade. A técnica empregada foi a de atribuir respostas aleatórias aos itens deixados em branco pelos alunos que fizeram a prova com penalidade. Utilizando essas respostas aleatórias, foi calculado a nota média e o desvio padrão, apresentados nas Fig. 3 por meio dos quadrados ocos.

A observação direta do gráfico indica um alto nível de coincidência entre as notas obtidas nas provas sem penalidade (quadrados cheios) e a simulação (quadrados ocos) obtida por meio estatístico a partir da notas obtidas nas provas com penalidade (losangos cheios). Se o "efeito emocional" mencionado pelos alunos fosse significativo, não seria possível, a partir de considerações puramente matemáticas, reproduzir as notas obtidas numa prova sem penalidade a partir das notas obtidas numa prova com penalidade. O alto nível de coincidência é um indicativo de que se o "efeito emocional" existe, ele não afeta significativamente as notas dos alunos como um todo.

4.6. Marcação dos itens com conhecimento parcial

Como mencionado anteriormente, a marcação dos itens com conhecimento parcial é importante na avaliação do aluno real. Caso o aluno não marcasse esses itens na prova com penalidade, seu rendimento estimado na prova sem penalidade resultaria menor que a nota obtida pelos alunos que fizeram a prova sem penalidade. A inexistência dessa diferença indica que o aluno real efetivamente marca as questões nas quais possui conhecimento parcial. Graças a isso, não fica prejudica a estimativa da sua proficiência fornecida pelo rendimento na prova.

5. Conclusão

As vantagens e desvantagens do uso da penalização foram abordadas do ponto de vista estatístico para investigar o comportamento do poder discriminativo dos itens de julgamento, o efeito sobre o erro do valor atribuído à nota, a ocorrência do "chute" nas provas com e sem penalidade e a possibilidade de efeitos emocionais associados à penalização.

A distribuição estatística dos itens deixados em branco na presença de penalização deixou claro que os alunos menos proficientes são mais propensos ao "chute" que os alunos mais proficientes, ao passo que a ausência de penalização estimula o "chute" em todos os níveis de proficiência.

A comparação dos efeitos combinados do aumento do poder discriminativo com o aumento do erro associado às notas dos itens dicotômicos, por efeito da penalização, evidenciou que a aplicação da penalização nos itens de julgamento contribui positivamente para a qualidade da medida da proficiência do aluno.

As notas obtidas nas provas sem penalidade puderam ser estimadas, por argumentos puramente estatísticos, a partir das notas obtidas nas provas com penalidade, permitindo-se concluir que a presença de "efeitos emocionais" decorrentes do uso da penalização não apresenta efeito estatístico significativo sobre a medida da proficiência do aluno.

Foi verificado também que o *aluno real* marca os itens nos quais possui conhecimento parcial, permitindo a medida não viesada da sua proficiência.

Referências

[1] H.L. Jackman, *Early Education Curriculum: A Child's Connection to the World* (Delmar, Albany, 2001).

[2] R.J. Dietel, J.L. Herman and R.A. Knuth, *What Does Research Say About Assessment?* (North Central Regional Education Laboratory, Oak Brook, 1991).

[3] R.K. Hambleton and E. Murphy. Applied Measurement in Education **5**, 1 (1992).

[4] G.F. Vito, J.C. Kunselman and R. Tewsbury, *Introduction to Criminal Justice Research Methods: An Applied Approach* (Charles C. Thomas, Springfield, 2014).

[5] K. Burke, *How to Assess Authentic Learning* (Corwin Press, Thousand Oaks, 2009).

[6] J. Diamond and W. Evans, Review of Educational Research **43**, 181 (1973).

[7] T.S. Roberts, `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.6175&rep=rep1&type=pdf`, acesso em 11/07/2017.

[8] R.F. Burton, Medical Education **36**, 805 (2002).

[9] R.B. Frary, Educational Measurement: Issues and Practice **7**, 33 (1988).

[10] D. Budescu and M. Bar-Hillel, Journal of Educational Measurement **30**, 227 (1993).

[11] R.F. Burton and D.J. Miller, Assessment & Evaluation in Higher Education **24**, 399 (1999).

[12] R.F. Burton, Assessment and Evaluation in Higher Education **29**, 587 (2004).

[13] R.F. Burton, Assessment and Evaluation in Higher Education **30**, 65 (2005).

[14] M.P. Espinosa and J. Gardeazabal , Journal of Mathematical Psychology **54**, 415 (2010).

[15] R.L. Ebel, *Essentials of Educational Measurement* (Prentice-Hall, Englewood Cliffs, 1979).

[16] B.A. Mello, Revista Brasileira de Ensino de Física **37**, 3503 (2015).