Revista Portuguesa de Educação

Avaliação da qualidade dos exames de Biologia e Geologia do ensino secundário português

Teresa Lopes[i]
Universidade do Minho, Portugal

José Precioso[ii]
Universidade do Minho, Portugal

RESUMO

Os resultados dos alunos portugueses no exame de Biologia e Geologia têm revelado uma situação grave de insucesso ao longo dos anos, com classificações médias muito baixas e taxas de reprovação excessivamente elevadas. Na opinião dos professores, as principais causas de insucesso relacionam-se com o elevado grau de complexidade do exame. Assim, é importante analisar a validade e a qualidade técnica desses exames para aferir se esses instrumentos de avaliação estão a contribuir negativamente para o insucesso dos alunos. Neste contexto, realizou-se esta investigação qualitativa, de análise documental com recurso a análise de conteúdo. Analisaram-se dois exames de Biologia e Geologia (critério de seleção: provas com os melhores e os piores resultados) nas dimensões: Área disciplinar por ano (o exame avalia as áreas de Biologia e Geologia); Tipo de questão; Dimensões do Ensino das Ciências; Taxonomia de Bloom: Dimensão do Processo Cognitivo e Dimensão do Conhecimento. Conclui-se que os exames são cognitivamente exigentes, sendo que a maioria das questões avalia categorias superiores do processo cognitivo (aplicação e análise) de conhecimento conceptual. Esta investigação demonstra a falta de validade e fiabilidade dos exames de Biologia e Geologia, assim como vários problemas de qualidade técnica.

Palavras-chave: Biologia e Geologia; Educação em ciências; Avaliação; Avaliação externa; Exames nacionais

1. INTRODUÇÃO

Em Portugal, a escolaridade obrigatória estende-se até ao final do ensino secundário (12º ano ou 18 anos de idade). A avaliação sumativa dos alunos no ensino secundário compreende a avaliação sumativa interna, cuja responsabilidade é dos professores e dos órgãos de gestão pedagógica da escola, e a avaliação sumativa externa, através da realização de exames nacionais da responsabilidade do Ministério da Educação.

A disciplina de Biologia e Geologia é lecionada nos 10º e 11º anos do curso científico-humanístico (oferta educativa do ensino secundário vocacionada para o prosseguimento de estudos no ensino superior) de Ciências

e Tecnologias e o respetivo exame é realizado no final do 11º ano. Para efeitos de certificação, o peso da classificação obtida no exame nacional é de 30% da classificação final dos alunos na disciplina, tendo a avaliação interna um peso de 70% da classificação final. No entanto, para efeito de seleção dos alunos no acesso ao ensino superior, se a disciplina for obrigatória para ingresso num determinado curso, a nota da classificação externa pode chegar a ter um peso de 50% e os alunos têm obrigatoriamente de atingir uma classificação mínima de 9,5 valores no exame para se poderem candidatar a esse curso.

Os resultados dos alunos no exame de Biologia e Geologia têm revelado uma situação grave de insucesso ao longo dos anos, com médias de classificações muito baixas e taxas de reprovação excessivamente elevadas (Lopes & Precioso, 2018). No que diz respeito à média de classificações, o valor máximo alcançado foi de 10,02 valores, em 2014 e 2016, e a classificação mínima foi de 8,21 valores, obtida em 2013 (Lopes, 2020). No que se refere às taxas de reprovação, os valores têm-se situado entre o 1/3 e os 2/3 dos alunos que vão a exame. O pior registo verificou-se no ano de 2013, em que a taxa de reprovação atingiu os 64,4%, o que significa que dos 76.501 exames realizados, 49.235 revelaram uma classificação abaixo dos 9,5 valores, números que mostram bem a dimensão do problema (Lopes, 2020).

Na opinião dos professores, as principais causas do insucesso dos alunos na avaliação externa da disciplina de Biologia e Geologia são: o elevado grau de complexidade do exame, o facto de o exame provocar *stress* e ansiedade nos alunos; as dificuldades dos alunos relacionadas com as competências de leitura, interpretação e comunicação; os critérios de correção e classificação do exame muito penalizadores; o elevado grau de complexidade da análise documental; o desfasamento entre o que é pedido no exame e o que é exigido pelo programa; a desadequação do exame à maturidade dos alunos e o facto de o programa da disciplina ser demasiado extenso (Lopes & Precioso, 2019).

Desta forma, é de grande importância analisar a validade e a qualidade técnica dos exames de Biologia e Geologia no sentido de avaliar se esses instrumentos de avaliação estão a contribuir negativamente para o insucesso dos alunos.

2. CONSTRUÇÃO DE TESTES DE AVALIAÇÃO: REGRAS E CONCEITOS IMPORTANTES

A avaliação das aprendizagens centra-se nos resultados dos alunos, sendo importante que as suas práticas asseverem uma recolha de informação rigorosa e consistente com as finalidades da aprendizagem (Fernandes, 2020).

O objetivo da avaliação da aprendizagem é oferecer uma oportunidade justa aos alunos para demonstrarem o que aprenderam (Airasian & Abrams, 2003) e, nesse sentido, é necessário que as informações recolhidas para a avaliação sejam de qualidade. Segundo Airasian & Abrams (2003), essa qualidade das informações recolhidas para a avaliação é influenciada por três fatores: por um lado, pelas condições sob as quais as informações são recolhidas, sendo que os alunos devem ter a oportunidade para mostrar o seu desempenho típico ou o melhor; por outro lado, pela qualidade dos

instrumentos de avaliação usados, que está relacionada com a clareza dos itens de teste, os critérios de correção adequados e a adequação do nível de linguagem aos alunos, entre outros; e por fim pela objetividade da informação, que pode ser posta em causa pelo enviesamento. Segundo os mesmos autores, bons instrumentos de avaliação apresentam três características fundamentais para que as informações coletadas tenham uma base fiável e válida: (i) o que é avaliado é o que foi ensinado; (ii) os exercícios, tarefas ou questões incluem uma amostra representativa dos objetivos ou metas que se tinham estabelecido previamente para as aprendizagens dos alunos; (iii) as questões de avaliação, as instruções e os procedimentos de correção e cotação são claros, inequívocos e apropriados (Airasian & Abrams, 2003).

Fernandes (2020) faz várias recomendações que devem merecer a atenção dos professores e avaliadores, nos processos de recolha de informação, quando elaboram instrumentos de avaliação (normalmente testes) e formulam questões cujos resultados servirão para atribuir uma classificação aos alunos. São elas:

1) As questões devem ser consistentes com o que foi ensinado, isto é, não deverão ser formuladas questões cujo conteúdo não foi devidamente trabalhado com os alunos.
2) Relativamente a um determinado conteúdo, devem ser formuladas questões com graus diferenciados de dificuldade.
3) Deve haver uma congruência entre o nível de dificuldade das questões formuladas e o nível de dificuldade que foi abordado durante o processo de ensino.
4) Não devem ser formuladas questões que exijam dos alunos a mobilização de conhecimentos, capacidades ou procedimentos que não foram devidamente tratados nas aulas.
5) As perguntas devem ser escritas de forma muito clara, assegurando que todos os alunos compreendem o que se pretende.
6) As questões formuladas não podem ser ambíguas, ou seja, os alunos deverão compreender exatamente o que se pretende.
7) Deve poder garantir-se que o que se pergunta permite avaliar as aprendizagens que realmente se pretendem avaliar.
8) Devem ser utilizadas diferentes tipologias de perguntas (por exemplo, perguntas de escolha múltipla; perguntas de ordenação; perguntas de associação; perguntas de verdadeiro/falso; perguntas de resposta curta; perguntas de resposta longa) (Fernandes, 2020, p. 6).

A avaliação deve refletir o alinhamento entre o currículo definido previamente, o que foi ensinado e o que é avaliado, tanto no que diz respeito aos conteúdos, como no que se refere aos processos cognitivos. A Taxonomia de Bloom refere-se a essa coerência como alinhamento (Anderson et al., 2001; Bloom et al., 1956) que considera essencial para a eficácia do processo educativo.

A Taxonomia de Bloom – *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain* – (Bloom et al., 1956), pretendeu desenvolver um método de classificação para processos cognitivos que seriam importantes nos processos de aprendizagem.

Mais tarde, foi revista e atualizada por uma equipa de investigadores pela necessidade de incorporar novas práticas e novos conhecimentos sobre a aprendizagem que se foram construindo e que foram evoluindo (Anderson et al., 2001), sendo publicada, em 2001, sob o título A *Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* (Anderson et al., 2001). Esta taxonomia foi sendo amplamente usada a nível internacional para caracterizar objetivos educacionais em documentos oficiais, em currículos, em avaliações do desempenho dos alunos baseadas em objetivos e para alinhar os currículos com a avaliação (Lee et al, 2017). E continua a ser utilizada (Forehand, 2010), constituindo uma base para a determinação de objetivos curriculares e para a implementação de metas gerais de ensino (Krathwohl, 2002), tendo uma influência significativa e duradoura no processo de ensino e de aprendizagem em todos os níveis de ensino até aos dias de hoje (Adams, 2015). E, por isso, é utilizada neste estudo para análise das questões relativamente às duas dimensões que pressupõe na construção da aprendizagem: a dimensão do conhecimento e a dimensão dos processos cognitivos.

Tendo os resultados nos exames tanta influência na vida dos alunos, é fundamental que alunos, professores, encarregados de educação e sociedade sintam confiança nesses instrumentos e nos resultados por eles fornecidos. Para De Ketele e Gerard (2005), as condições para se ter um teste como um instrumento de avaliação de qualidade são a relevância, a validade e a fiabilidade, que, embora sejam dimensões teoricamente independentes, se interrelacionam. Para Black (1998), a confiança dos resultados depende da fiabilidade, da validade e da ausência de enviesamento. Conceitos que importa clarificar, já que, na opinião dos autores referidos, são amplamente negligenciados.

A relevância é o grau de apropriação do teste em relação aos objetivos, ou seja, é o grau de compatibilidade do instrumento de avaliação com os outros elementos do sistema ao qual pertence (De Ketele & Gerard, 2005).

A fiabilidade está relacionada com o facto de os resultados serem reproduzíveis noutras condições: com corretores diferentes, num outro período de tempo, com outros aplicadores, com conjuntos de perguntas diferentes, entre outros (Black, 1998; De Ketele & Gerard, 2005). Ou seja, está relacionada com o grau de congruência entre a nota obtida e a nota verdadeira, a que reflete verdadeiramente o que o aluno sabe e consegue fazer, não esquecendo que a nota verdadeira é uma abstração, um ponto de convergência desejado, independente dos avaliadores e das circunstâncias (De Ketele & Gerard, 2005).

Em Portugal, em 2018, foram pedidas 6.822 reapreciações de provas de exame do secundário apenas na 1ª fase: 75% dessas reapreciações resultaram em subida da nota; 9% em descida da nota; e apenas 16% em manutenção da nota (Júri Nacional de Exames, 2019). No caso específico de Biologia e Geologia, foram pedidas 1.220 reapreciações de provas de exame: 784 (64%) resultaram em subida da nota; 117 (10%) resultaram em descida da nota; e 319 (26%) resultaram em manutenção da nota (Júri Nacional de Exames, 2019). Estes números demonstram bem a reduzida fiabilidade das provas nacionais, o que ainda é agravado pelo facto de que, para pedir a reapreciação da prova,

é necessário pagar com antecedência 25 euros (que, no caso de subida da nota, são devolvidos), o que limita, com base em critérios económicos, o acesso ao pedido de reapreciação, levantando questões éticas de igualdade de acesso a um direito dos alunos que realizam o exame.

Para minimizar estes efeitos, cada vez mais, se tem normalizado a aplicação da prova, os critérios de avaliação e os procedimentos de classificação pelos professores classificadores (Fernandes, 2005; Kellaghan & Madaus, 2003), o que acaba por condicionar o tipo de questões incluídas no exame, levando a que este seja sobretudo constituído por itens de seleção.

A validade é um conceito complexo (Black, 1998) que está relacionado com o grau de adequação entre o que se declara avaliar e o que realmente se avalia, ou seja, o que um instrumento pretende avaliar e o que realmente avalia (De Ketele & Gerard, 2005). A validade de conteúdo refere-se ao grau de adequação de um instrumento de avaliação ao currículo que se pretende avaliar, ou seja, se o teste avalia ou não o currículo de uma disciplina ou de um curso (Black, 1998). A validade de constructo relaciona-se com o facto de as questões avaliarem as competências e os processos cognitivos que se pretendem avaliar (Black, 1998). Deste modo, a validade de um teste está relacionada também com as inferências que se fazem a partir dos resultados dos alunos nesse teste (Black, 1998). Por outro lado, se os exames, pela sua existência, induzem um modelo de aprendizagem inadequado, então, para Black (1998), é inevitável concluir que tal implica falta de validade. Ora, isso é o que tem vindo a ser descrito pela investigação relativamente aos exames em Portugal, que induzem o *teaching to the test*, o que põe em causa a validade das provas nacionais. Vários estudos realizados no nosso país (Lopes, 2013; Madureira, 2011; Marques et al., 2015; Raposo & Freire, 2008; Rosário, 2007) mostram essa influência da realização de exames nacionais nas práticas pedagógicas e avaliativas dos professores. Os docentes orientam as suas práticas para "treinar" os alunos para o que é pedido no exame, utilizando práticas que não pensam ser as de maior qualidade para a aprendizagem; elaboram os seus testes com a estrutura semelhante à dos exames e com os mesmos critérios de classificação das questões, mesmo não concordando com eles e considerando-os penalizadores para os alunos; e mostram grande preocupação em abordar todo o programa das disciplinas, mesmo sabendo que o ritmo que impõem para o lecionar não proporciona uma aprendizagem de qualidade (Lopes, 2013).

Segundo Alves (2014), os exames do ensino secundário português carecem de validade por várias razões: avaliam o que não foi ensinado, sobrevalorizam alguns conteúdos, sem que haja coerência com o programa das disciplinas avaliadas, não levam em consideração as condições em que se deu o processo de ensino, nem o processo de aprendizagem e contemplam questões ambíguas.

Fiabilidade e validade são conceitos que estão muitas vezes correlacionados e um não faz sentido sem o outro, como explica Black (1998): um teste que tenha condições de classificação e correção bem definidas e reproduzíveis em condições diferentes e com corretores diferentes, ou seja, um teste fiável, perde todo o valor se não avaliar o currículo que se pretende avaliar; por outro lado, um teste que avalia um currículo e as competências

que se propõe avaliar, ou seja que é válido, mas que apresenta problemas de correção e classificação, fica desprovido de significância. Muitas vezes, o que acontece é que estes dois conceitos entram em conflito, já que, para aumentar a fiabilidade na correção e classificação, poder-se-á diminuir a validade, reduzindo o alcance que o teste possa ter. Por outro lado, quando aumentamos esse alcance com a introdução de questões de resposta aberta e de desenvolvimento de raciocínio, diminui-se a fiabilidade.

Este conflito de que Black (1998) fala é bem visível no caso dos exames nacionais portugueses. A maioria das questões é de resposta fechada, ou de seleção, e os critérios de correção e classificação das questões de resposta aberta ou de construção são considerados pelos professores muito limitadores da liberdade de resposta dos alunos (Lopes, 2013), em nome da fiabilidade, pondo em causa a validade do instrumento de avaliação, já que os professores relatam casos em que, para cumprir os critérios de correção impostos, atribuem zero valores a respostas que consideram pelo menos parcialmente corretas, mas que não cumprem os requisitos dos critérios de classificação.

Fernandes (2008) considera legítimo questionar a validade e fiabilidade atribuídas às avaliações externas. O autor argumenta que os conceitos base da psicometria são muito usados nas avaliações externas, mas a verdade é que, apesar das orientações emergentes da investigação em avaliação das aprendizagens se inserirem no paradigma cognitivista, não se têm desenvolvido outros conceitos mais adequados segundo as perspetivas mais atuais de avaliação e, portanto, esta é uma área que exige mais investigação (Fernandes, 2008).

Por outro lado, se a prova pretende seriar alunos, um outro conceito importante a ter em conta é o de sensibilidade, a capacidade da prova de diferenciar os alunos, ou seja, de discriminar níveis distintos de qualidade das aprendizagens dos alunos (Almeida, 2012). A sensibilidade de uma prova será tão maior, quanto maior for a sua capacidade de discriminar níveis de desempenho dos discentes (Almeida, 2012).

O enviesamento ou funcionamento diferencial de itens pode surgir entre sexos, grupos étnicos ou classes sociais diferentes, entre outras situações (Black, 1998) e está relacionado com os fatores que levam as questões a ter um impacto desigual em alunos diferentes, criando, à partida, injustiça. Há uma grande quantidade de condições que podem levar ao funcionamento diferencial das questões, tais como: contextos, género, classe social, preferência cognitiva, estilo de aprendizagem e temperamento, entre outros. Os problemas de enviesamento muitas vezes não são levados em consideração, frequentemente são muito difíceis de detetar, até porque reiteradamente não são conhecidos nem estudados, embora constituam ameaças à equidade (Black, 1998).

Torna-se assim claro que a justiça pretendida num teste de grande impacto no percurso académico dos alunos será muito difícil de alcançar com um instrumento de avaliação como o exame.

3. OBJETIVOS

Estudos prévios (Lopes & Precioso, 2018) evidenciam uma situação de insucesso escolar grave, muito prevalente e persistente, no exame de Biologia e Geologia, já que os alunos e as alunas têm vindo a obter classificações médias preocupantemente baixas (entre 8 e 11 valores) e taxas de reprovação demasiadamente altas (entre 45% e 65%). Importa então analisar a qualidade desses exames, no sentido de averiguar se o instrumento de avaliação poderá estar a contribuir para esse insucesso.

Assim, os objetivos deste estudo são: (1) analisar a validade e a qualidade técnica dos exames de Biologia e Geologia; (2) determinar se os exames de Biologia e Geologia avaliam as finalidades da disciplina previstas no Programa; (3) determinar se os exames de Biologia e Geologia avaliam a consecução dos objetivos educacionais propostos pela Taxonomia de Bloom.

4. METODOLOGIA

A população deste estudo é constituída pelos vários exames nacionais de Biologia e Geologia, da 1ª e 2ª fases, que foram realizados nos anos letivos entre 2012/2013, último ano em que se aplicaram alterações nas provas nacionais, a 2017/2018, último ano com dados fornecidos de forma completa pelo Júri Nacional de Exames (JNE). A população é, portanto, constituída por dez exames. Selecionou-se uma amostra não probabilística de conveniência, sendo o critério os extremos dos resultados dos alunos, tanto em termos de classificação média como de taxa de aprovação. Assim, a amostra é composta por dois exames: exame de Biologia e Geologia de 2014, 1ª fase, a prova em que os alunos apresentaram os melhores resultados; e exame de Biologia e Geologia de 2014, 2ª fase, a prova em que os alunos apresentaram os piores resultados. Optou-se por uma investigação qualitativa de análise documental e análise de conteúdo. Foram construídos os instrumentos de recolha de dados que consistiram em grelhas para analisar as questões dos exames de forma a permitir obter dados sobre as seguintes dimensões: Área disciplinar por ano; Tipo de questão; Dimensões do ensino das ciências; Taxonomia de Bloom: dimensão do processo Cognitivo; Taxonomia de Bloom: dimensão do conhecimento. As grelhas de análise foram construídas com a especificação das dimensões, respetivas categorias e subcategorias baseadas na literatura da especialidade, sendo submetidas a validação *a priori* por especialistas em Educação e em Educação em Ciências. Os exames nacionais sujeitos a análise e os respetivos critérios de correção foram retirados do *site* do Instituto de Avaliação Educativa (IAVE).

No tratamento de dados, realizaram-se a análise documental e a análise de conteúdo, cada uma das questões de cada um dos exames foi analisada qualitativamente, fazendo a respetiva categorização para cada uma das dimensões analisadas (presença/ausência da categoria e cotação). Por fim, fez-se uma análise de abordagem quantitativa (registo quantitativo de cada categoria e cotação correspondente), no sentido de permitir reduzir os dados de forma

a facilitar a elaboração de conclusões, registando-se os resultados em tabelas para comparar as duas provas.

5. RESULTADOS

Os exames de Biologia e Geologia são compostos por questões de vários tipos distribuídas por quatro grupos. Cada grupo apresenta sempre documentos/fontes de informação, que podem ser textos, gráficos, imagens, tabelas, mapas, cuja análise é pré-requisito para responder às questões.

Através da análise da Tabela 1, constata-se que há uma preocupação em produzir provas equivalentes, em termos de estrutura, dentro dos mesmos parâmetros. Nas duas provas são abordados conteúdos de Biologia de 10º e 11º anos e de Geologia de 10º e 11º anos, com, sensivelmente, a mesma distribuição de questões. Nas duas provas, surgem quatro questões do domínio procedimental de Biologia, não se verificando questões do domínio procedimental de Geologia.

Relativamente aos tipos de questões, as duas provas apresentam 25 itens de seleção e cinco itens de construção, o que revela um grande desequilíbrio. Ambas são compostas por 22 questões de escolha múltipla, correspondendo a 110 pontos; uma questão de associação, correspondendo a 10 pontos; duas questões de ordenação, correspondendo a 20 pontos; e cinco questões de resposta restrita, correspondendo a 60 pontos. Quer isto dizer que há manutenção da estrutura da prova, no que diz respeito a tipos de questões e respetiva cotação, mas não há uma distribuição equitativa do tipo de questões.

Também no que concerne às dimensões do ensino das ciências, verifica-se constância de uma prova para a outra, visto que as duas apresentam 26 questões (175 pontos) inseridas na categoria "Aprender ciência" e 4 questões (25 pontos) inseridas na categoria "Aprender a fazer ciência". As categorias "Aprender acerca da ciência" e "Aprender pela ciência" não se encontram representadas, embora se apresentem, nos documentos normativos da disciplina, como finalidades importantes. As questões de resposta restrita poderiam ser consideradas na categoria "Aprender pela ciência" pela argumentação e fundamentação científicas. No entanto, não foram inseridas nessa categoria porque, na realidade, não avaliam essas competências, visto que apenas aceitam uma única resposta correta.

Compreende-se que o exame incida mais sobre "Aprender ciência", ou seja, nos conteúdos; no entanto, se o programa da disciplina considera como importantes finalidades da disciplina as outras categorias, não se compreende que duas delas não estejam sequer representadas, já que o exame, como instrumento de avaliação da disciplina, deve avaliar a consecução das suas finalidades.

Relativamente à dimensão do processo cognitivo da Taxonomia de Bloom revista, surgem algumas diferenças que podem explicar as disparidades nas classificações dos alunos nas duas provas. Embora os dois exames se centrem nas categorias de aplicação e análise, há diferenças.

Tabela 1

Análise Comparada dos Rexames 2014, 1.ª Fase e 2014, 2.ª Fase

Avaliação da qualidade dos exames de Biologia e Geologia			Exame: 2014 1ª fase		Exame: 2014 2ª fase	
			n	cotação	n	cotação
Área disci-plinar por ano	Biologia	10º ano	3	20	4	25
		11º ano	8	55	7	50
		Procedimental	4	25	4	25
	Geologia	10º ano	7	45	7	45
		11º ano	8	55	8	55
		Procedimental	0	0	0	0
Tipo de questão	Questões de seleção	Escolha múltipla	22	110	22	110
		Associação	1	10	1	10
		Ordenação	2	20	2	20
		Verdadeiro/Falso	0	0	0	0
		Completamento	0	0	0	0
	Questões de construção	Completamento	0	0	0	0
		Resposta curta	0	0	0	0
		Resposta restrita	5	60	5	60
		Resposta extensa	0	0	0	0
Dimensões do Ensino das Ciências	Aprender ciência		26	175	26	175
	Aprender a fazer ciência		4	25	4	25
	Aprender sobre a ciência		0	0	0	0
	Aprender pela ciência		0	0	0	0
Taxonomia de Bloom: Dimensão do Processo Cognitivo	Lembrar		0	0	0	0
	Compreender		5	25	3	15
	Aplicar		14	75	15	80
	Analisar		10	90	11	90
	Avaliar		1	10	1	15
	Criar		0	0	0	0
Taxonomia de Bloom: Dimensão do Conhecimento	Conhecimento factual		0	0	0	0
	Conhecimento conceptual		26	175	26	175
	Conhecimento processual		4	25	4	25
	Conhecimento metacognitivo		0	0	0	0

No caso da 1ª fase, prova em que os alunos tiveram melhores resultados, há cinco questões na categoria da compreensão, correspondendo a 25 pontos, enquanto na 2ª fase, prova em que os alunos tiveram piores resultados, há apenas três questões de compreensão, correspondendo a apenas 15 pontos.

Na categoria "Aplicar", na 1ª fase, há 14 questões, correspondendo a 75 pontos, enquanto na 2ª fase, há 15 questões, correspondendo a 80 pontos. Na categoria da análise, na 1ª fase há 10 questões e na 2ª fase há 11 questões, mas correspondem igualmente a 90 pontos. Na categoria "Avaliar", a mais elevada presente nas provas, há apenas uma questão em cada prova, mas na 1ª fase, corresponde a 10 pontos e, na 2ª fase, corresponde a 15 pontos. Em nenhuma das provas surgem itens que configurem as categorias de "Lembrar", a menos exigente, e de "Criar", a mais exigente. Portanto, a 2ª fase é uma prova mais exigente cognitivamente. Por outro lado, dentro da mesma categoria, as questões podem ter graus de dificuldade diferentes, seja pelos conteúdos que abordam, em termos de quantidade e qualidade, seja pelo nível e quantidade de relações e inferências que o aluno tem de fazer para responder às questões.

No que concerne à dimensão do conhecimento da Taxonomia de Bloom revista, as provas mostram-se equiparadas, pois ambas apresentam 26 itens dentro da categoria de conhecimento conceptual, perfazendo 175 pontos, e quatro itens dentro da categoria de conhecimento processual, somando 25 pontos.

Em síntese, demonstra-se que os exames são provas cognitivamente exigentes, em que a maioria das questões avalia categorias superiores do processo cognitivo, tais como a aplicação e análise, de conhecimento conceptual. Não há questões pertencentes à categoria "Lembrar" e as questões de compreensão são raras, mas também não há questões da categoria "Criar". Há um privilégio claro das perguntas de escolha múltipla. A principal finalidade do ensino das ciências avaliada é "Aprender Ciência", embora também surja um grupo em que há questões que avaliam a finalidade "Aprender a fazer ciência", avaliando conhecimento processual. É de salientar que, por vezes, o grau de dificuldade surge com artificialismos e formulações menos corretas das questões que causam dúvidas e inseguranças aos alunos. Por fim, considera-se, pela análise das duas provas, que o exame da 2ª fase apresenta uma complexidade de análise mais elevada, sobretudo ao nível interpretativo, incidindo em categorias cognitivas mais elevas e conteúdos e conceitos mais específicos e menos familiares dos alunos, o que eleva o nível de dificuldade geral da prova, tornando-a cognitivamente mais exigente.

6. CONCLUSÕES E IMPLICAÇÕES

Os exames de Biologia e Geologia são provas exigentes, em que a maioria das questões avalia categorias superiores do processo cognitivo, tais como a aplicação e análise, de conhecimento conceptual.

Estes resultados vão ao encontro dos resultados de outros estudos realizados no mesmo âmbito. Preto (2008) analisou o grau de dificuldade dos itens relacionados com atividades experimentais da 1ª e 2ª fases do ano de 2006 e concluiu que as questões requerem um nível elevado de exigência

conceptual por exigirem a associação de diferentes conteúdos solicitando principalmente competências no domínio do raciocínio.

Esteves e Rodrigues (2012) analisaram, segundo o domínio cognitivo da Taxonomia de Bloom, os itens dos exames da disciplina de História, de 2006 a 2010, 1ª e 2ª fases, verificando que não existiam questões na categoria "Conhecer", mas sim em todas as outras categorias, sendo a categoria da compreensão aquela que apresentava maior número de itens e maior cotação. As autoras concluíram que as provas apresentam uma complexidade cognitiva elevada. Tal como já havia sido constatado por Esteves e Rodrigues (2012), também o exame de Biologia e Geologia apresenta fundamentalmente questões que avaliam categorias superiores do processo cognitivo.

Por outro lado, o privilégio claro que é dado às perguntas de escolha múltipla retira validade à prova, uma vez que este tipo de itens dificulta a avaliação da capacidade do aluno desenvolver um raciocínio (Black, 1998). O aluno pode estar a fazer um raciocínio pelo menos parcialmente correto e errar a questão, ou pode acertar a questão fazendo um raciocínio errado. E, portanto, não está a ser dada uma oportunidade justa ao aluno para mostrar aquilo que sabe neste domínio. Esta opção pelas questões de escolha múltipla está relacionada com a tentativa de incrementar a fiabilidade da prova, mas diminui a sua validade, o que evidencia o conflito em que entram estes dois conceitos, denunciado por Black (1998).

Relativamente às finalidades do ensino das ciências, a principal finalidade avaliada é "Aprender ciência", embora também surja um grupo em que há questões que avaliam a finalidade "Aprender a fazer ciência", avaliando conhecimento processual. Pode compreender-se que o exame incida mais sobre "Aprender ciência", ou seja, nos conteúdos, no entanto, se o programa da disciplina considera como importantes finalidades da disciplina as outras categorias, –"Aprender acerca da ciência" e "Aprender pela ciência"–, não é aceitável que estas dimensões não estejam sequer representadas, uma vez que o exame, como instrumento de avaliação, deve avaliar a consecução das finalidades da disciplina. Este facto retira validade ao exame já que este não avalia o que o programa da disciplina preconiza.

É impreterível produzir instrumentos de avaliação com melhor qualidade técnica, aperfeiçoando os seguintes aspetos na elaboração das provas: as questões não podem, de forma inequívoca, extrapolar o programa da disciplina, tanto no que se refere aos conteúdos, como no que se refere ao grau de complexidade com que são abordados; os tipos de itens devem ser mais diversificados, introduzindo por exemplo questões de resposta curta, retirando o privilégio claro dado às questões de escolha múltipla; as questões de ordenação devem ser retiradas porque, por um lado, por vezes não são claras para os alunos e, por outro, o seu critério de classificação é muito penalizador; as questões devem ser claras e conter indicações objetivas daquilo que se pretende que o aluno demonstre saber fazer; a escolha dos documentos de suporte deve ser mais criteriosa no sentido de estes se adequarem melhor à maturidade cognitiva e à idade dos alunos; a linguagem deve ser clara, do domínio dos alunos e adaptada em termos etários; não devem existir artificialismos que aumentam a complexidade na forma como é elaborada a questão, seja de linguagem, seja de interpretação da questão,

seja de pormenores dissimulados que determinam a resposta, entre outros; deve ser procurado um maior equilíbrio nos processos cognitivos avaliados, assim como das finalidades da disciplina.

Os critérios de correção e classificação devem ser repensados, tornando-os mais flexíveis, devendo haver, por parte do IAVE, uma maior consideração pelas sugestões e contribuições dos professores classificadores acerca do que deve ou não ser melhorado na primeira versão do documento dos critérios de classificação que lhes é fornecida, transferindo para os professores classificadores maior responsabilidade, o que também implica mais e melhor formação nessa área.

Uma forma de equilibrar validade e fiabilidade é as provas serem corrigidas por mais do que um corretor, independentemente, fazendo depois um balanço das classificações atribuídas pelos diferentes corretores.

No ano letivo 2019/2020, devido à pandemia de COVID-19 que levou a aulas não presenciais no final do 2º período e todo o 3º período, o exame de Biologia e Geologia sofreu alterações, tanto na sua estrutura, como na sua classificação. O exame foi constituído por 33 itens, dos quais apenas 10 eram obrigatórios. Dos restantes 23, contribuíram para a classificação final da prova os 15 itens em que cada aluno teve melhor pontuação. Quanto ao tipo de questões, foram introduzidas questões de resposta curta e questões de completamento de seleção (com escolha múltipla). A todas as questões foi atribuída a mesma cotação. Estas alterações deram maior oportunidade aos alunos para mostrarem o que sabem e o que são capazes de fazer e, por isso, merecem uma reflexão acerca de se deveriam manter-se nos próximos exames. As modificações no instrumento de avaliação levaram a uma melhoria significativa nos resultados dos alunos. A média dos resultados da 1ª fase (dados do JNE, sem especificar se a média se refere a alunos internos ou a todos os alunos que realizaram o exame na 1ª fase) foi de 14 valores e a taxa de reprovação ficou-se pelos 15,57%. É ainda de salientar que apenas os alunos que pretendessem usar a prova nacional como prova específica para acesso ao ensino superior tinham a obrigatoriedade de a realizar. No entanto, a diferença no número de alunos a realizar a prova não foi grande, já que se realizaram 41460 provas na 1ª fase de 2020 (na 1ª fase de 2019 realizaram-se 42848 provas). Estes resultados vêm ao encontro do que se defende nesta investigação, já que mudanças no instrumento de avaliação produziram grandes diferenças nos resultados dos alunos, e mostram que uma prova pontual não é um instrumento rigoroso incontestável que defina o que um aluno sabe e é capaz de fazer e que, portanto, não pode determinar de forma tão decisiva o percurso académico dos alunos. Por outro lado, a melhoria nos resultados dos alunos vem demonstrar que a prova não pode ser vista como um instrumento de avaliação inquestionável e infalível, mas antes que deve ser repensado e continuamente melhorado.

Por fim, salienta-se que esta investigação apresenta algumas limitações. Por um lado, pela amostra diminuta de provas analisadas e, por outro, por não se não se ter estudado aprofundadamente se o exame avalia o conteúdo programático de forma adequada.

AGRADECIMENTOS

Este trabalho foi financiado por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia no âmbito do projeto do CIEC (Centro de Investigação em Estudos da Criança da Universidade do Minho) com a referência UIDB/00317/2020; Bolsa de Doutoramento SFRH/BD/123731/2016.

REFERÊNCIAS

Adams, N. E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association, 103*(3), 152-152. https://doi.org/10.3163/1536-5050.103.3.010

Airasian, P. W., & Abrams, L. M. (2003). Classroom student svaluation. In J. T. Kellaghan & D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 533-548). Kluwer Academic Publishers.

Almeida, L. (2012). Avaliação dos alunos: Combinando as razões e os modos. In J. Karpicke, H. Sousa, L. Almeida, & Fundação Francisco Manuel dos Santos (Ed.), *A avaliação dos alunos* (pp. 73-88). Fundação Francisco Manuel dos Santos.

Alves, J. (2014). Exames: Mitos e realidades. In J. Machado & J. Alves (Orgs). *Melhorar a escola: Sucesso escolar, disciplina, motivação, direção de escolas e políticas educativas* (pp. 155-169). Universidade Católica Portuguesa.

Anderson, L. W. & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Black, P. J. (1998). *Testing, friend or foe?: The theory and practice of assessment and testing*. Psychology Press.

Bloom, B. S. (Ed.), Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). T*axonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. David McKay.

De Ketele, J. M., & Gerard, F. M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences. *Mesure et Évaluation en Éducation, 28*(3), 1-26. https://doi.org/10.7202/1087028ar

Esteves, M. & Rodrigues, A. (2012). Exames nacionais e contextualização no ensino da História. *Revista Interacções, 8*(22), 135-162. https://doi.org/10.25755/int.1539

Fernandes, D. (2005). *Avaliação das aprendizagens: Desafios às teorias, práticas e políticas*. Texto Editores.

Fernandes, D. (2008). Para uma teoria da avaliação no domínio das aprendizagens. *Estudos em Avaliação Educacional, 19*(41), 347-372. http://dx.doi.org/10.18222/eae194120082065

Fernandes, D. (2020). *Avaliação sumativa*. In Projeto MAIA – Projeto de Monitorização Acompanhamento e Investigação em Avaliação Pedagógica. https://apoioescolas.dge.mec.pt/sites/default/files/2021-02/folha_avaliacao_sumativa.pdf

Forehand, M. (2010). Bloom's taxonomy. In M. Orey (Ed.), *Emerging perspetives on learning, teaching, and technology* (pp. 41-74). https://textbookequity.org/Textbooks/Orey_Emergin_Perspectives_Learning.pdf

Júri Nacional de Exames (2019). *Processo de avaliação axterna da aprendizagem – Provas de aferição, provas finais e exames nacionais 2018*. Ministério da Educação, Direção-Geral da Educação. https://www.dge.mec.pt/sites/default/files/JNE/relatorio_anual_do_jne_2018_final_lv.pdf?fbclid=IwAR2ogq7WPSzMp-NbI5jCUpML2uv560P0QLDW7vXbyqHtI6qk5ucRCyMKeUA

Kellaghan, T. & Madaus, G. (2003). External (public) examinations. In T. Kellaghan & D. Stufflebeam (Eds.), I*nternational handbook of educational evaluation* (pp. 577-600). Kluwer Academic Publishers.

Krathwohl, D. R. (2002) A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212-218. https://doi.org/10.1207/s15430421tip4104

Lee, Y., Kim, M., Jin, O., Yoon, H., & Matsubara, K. (2017). *East-Asian primary science curricula - An overview using revised Bloom's taxonomy*. Springer.

Lopes, T. (2013). *Perceções de professores, alunos e encarregados de educação sobre o (in)sucesso na disciplina de Biologia e Geologia* [Dissertação de Mestrado, Universidade do Minho]. Repositório Institucional da Universidade do Minho. https://hdl.handle.net/1822/28923

Lopes, T. (2020). *Insucesso escolar na disciplina e no exame de Biologia e Geologia e fatores associados* [Tese de Doutoramento, Universidade do Minho]. Repositório Institucional da Universidade do Minho. https://hdl.handle.net/1822/77124

Lopes, T. & Precioso, P. (2018). Evolução do insucesso escolar nos exames nacionais do ensino secundário, por sexo, em Portugal. *Revista Iberoamericana de Evaluación Educativa, 11*(2), p. 53-69. https://doi.org/10.15366/riee2018.11.2.003

Lopes, T. & Precioso, J. (2019). *O que pensam os professores portugueses sobre as causas do insucesso dos alunos no exame nacional da disciplina de Biologia e Geologia?* In XV Congresso Internacional Galego-Português de Psicopedagogia. Universidade da Coruña.

Madureira, M. B. (2011). *A influência dos exames nacionais de Física e Química A e respetivos resultados nas práticas de ensino e de avaliação dos professores* [Dissertação de Mestrado, Universidade do Minho]. Repositório Institucional da Universidade do Minho. https://hdl.handle.net/1822/19112

Marques, M.; Sousa, J.; Costa, N.; & Pacheco, J. (2015). Efeitos da avaliação externa das aprendizagens no desenvolvimento profissional de professores de Matemática do ensino básico em Portugal. *Meta: Avaliação, 7*(19), 58-84. https://doi.org/10.22347/2175-2753v7i19.781

Preto, A. (2008). *Ensino da Biologia e Geologia no Ensino Secundário: Exames e trabalho experimental* [Dissertação de Mestrado, Universidade de Lisboa]. Repositório da Universidade de Lisboa. http://hdl.handle.net/10451/1312

Raposo, P. & Freire, A. (2008). Avaliação das aprendizagens: Perspectivas de professores de Física e Química. *Revista da Educação, XVI*(1), 97-127. https://www.academia.edu/50102733/Avalia%C3%A7%C3%A3o_das_Aprendizagens_Perspectivas_de_Professores_de_F%C3%ADsica_e_Qu%C3%ADmica

Rosário, M. A. (2007). *Influência do exame nacional do 9º ano de escolaridade nas práticas de ensino e de avaliação em Matemática* [Dissertação de Mestrado, Universidade do Minho]. Repositório Institucional da Universidade do Minho. https://hdl.handle.net/1822/7180

i Instituto de Educação, Universidade do Minho, Portugal.
https://orcid.org/0000-0001-8361-1429

ii Instituto de Educação, Universidade do Minho, Portugal.
https://orcid.org/0000-0002-7889-8290

Toda a correspondência relativa a este artigo deve ser enviada para:

Teresa Lopes
Instituto de Educação (IE), Universidade do Minho
Campus de Gualtar
4710-057 Braga
teresaflopes@netcabo.pt

Recebido em 10 de dezembro de 2020
Aceite para publicação em 18 de fevereiro de 2022

Quality assessment of the Portuguese secondary school Biology and Geology exams

ABSTRACT

The results of Portuguese students in the Biology and Geology exam have revealed a serious situation of failure over the years, with very low average scores and excessively high failure rates. In teachers' opinion, the main causes of failure are related to the high complexity of the exam. Thus, it is important to analyze the validity and technical quality of these exams to assess whether these assessment instruments are negatively contributing to the students' failure. In this context, this qualitative research was carried out, using document analysis and content analysis. Two Biology and Geology exams were analyzed (selection criteria: exams with the best and worst results) in the dimensions: Disciplinary area per year (the exam evaluates the areas of Biology and Geology); Question type; Dimensions of science education; Bloom's Taxonomy: cognitive process dimension and knowledge dimension. It is concluded that the exams are cognitively demanding, and most questions evaluate higher categories of the cognitive process (application and analysis) of conceptual knowledge. This investigation demonstrates the lack of validity and reliability of the Biology and Geology exams, as well as several technical quality problems.

Keywords: Biology and Geology, Science education, Evaluation, External assessment, National exams.

Evaluación de la calidad de los exámenes de Biología y Geología en la educación secundaria portuguesa

RESUMEN

Los resultados de los estudiantes portugueses en el examen de Biología y Geología han revelado una grave situación de reprobación a lo largo de los años, con puntuaciones medias muy bajas y tasas de fracaso excesivamente altas. En opinión de los profesores, las principales causas de reprobación están relacionadas con el alto grado de complejidad del examen. Por tanto, es importante analizar la validez y calidad técnica de estos exámenes para evaluar si estos instrumentos de evaluación están contribuyendo negativamente al fracaso de los estudiantes. En este contexto, se llevó a cabo esta investigación cualitativa, utilizando análisis de documentos mediante análisis de contenido. Se analizaron dos exámenes de Biología y Geología (criterio de selección: pruebas con mejor y peor resultado) en las dimensiones: Área disciplinaria por año (el examen evalúa las áreas de Biología y Geología); Tipo de pregunta; Dimensiones de la enseñanza de las ciencias; Taxonomía de Bloom: dimensión del proceso cognitivo y dimensión del conocimiento. Se concluye que los exámenes son cognitivamente exigentes y la mayoría de las preguntas evalúan categorías superiores del proceso cognitivo (aplicación y análisis) del conocimiento conceptual. Esta investigación demuestra la falta de validez y confiabilidad de los exámenes de Biología y Geología, así como varios problemas de calidad técnica.

Palabras clave: Biología y Geología, Educación en ciencias, Evaluación, Evaluación externa, Exámenes nacionales.