UNIVERSIDADE DA BEIRA INTERIOR
Ciências Sociais e Humanas

Avaliação sumativa em matemática no Ensino Superior com recurso a questões de escolha-múltipla: uma abordagem utilizando a metodologia investigação-ação

José Manuel Monteiro Lopes de Azevedo

Tese para obtenção do Grau de Doutor em
Educação
(3.º ciclo de estudos)

Orientadora: Professora Doutora Ema Patrícia de Lima Oliveira
Coorientadora: Professora Doutora Patrícia Damas Beites

Covilhã, maio de 2017

Dedicatória

À minha MÃE, pela força, trabalho e teimosia. Conseguiu, apesar das dificuldades, levar a bom porto toda a tripulação! Bem-haja!

À memória de meu Pai.

Agradecimentos

Destaco o apoio que sempre me foi dado ao longo deste trabalho pela minha orientadora, Professora Doutora Ema Oliveira, e pela minha coorientadora, Professora Doutora Patrícia Beites. Agradeço, em especial, a paciência que sempre tiveram comigo e as pertinentes correções. Estar-lhes-ei eternamente grato.

Agradeço o apoio incondicional do Professor Doutor António Pedrosa, que, no ISCAP, me permitiu realizar este trabalho. Sem ele esta tese não tinha sido possível.

Agradeço à Professora Doutora Luísa Branco, por me ter incentivado, desde o primeiro contacto para avançar com esta investigação.

Ao Presidente do ISCAP, o meu obrigado pela autorização que me deu para poder usar os dados que foram fundamentais para a tese.

Aos meus colegas que trabalharam comigo neste projeto, o meu muito obrigado. Também a alguns colegas do Departamento, que de uma ou de outra forma também me incentivaram a continuar ou deram algumas sugestões; a minha gratidão.

A todos os professores e alunos que participaram nas entrevistas e, em especial, aos alunos que ajudaram a melhorar as questões participando no teste de validação, estou-lhes muito agradecido.

À Maria do Céu pela amizade.

À minha família não tenho palavras para dizer o quanto lhes agradeço o apoio dado.

Nuvem de palavras

RESUMO

A implementação do Processo de Bolonha tem vindo a colocar vários desafios às instituições de Ensino Superior europeias, impondo uma mudança de paradigma. Esta mudança implica a alteração da forma de avaliar os alunos, com um foco nas avaliações contínua e formativa, catalisadas pelo *e-assessment*. Esta tese apresenta o processo de implementação de uma estratégia de *e-assessment utilizando* testes com questões de escolha múltipla (QEM), implementados no *Moodle*. Esta estratégia foi executada de modo a permitir a utilização de avaliação contínua sumativa em unidades curriculares (UC) de Matemática numa instituição de Ensino Superior, em turmas com um número elevado de alunos, sendo por isso difícil de implementar. Foram objetivos deste trabalho: perceber como o *e-assessment* pode influenciar o processo de ensino-aprendizagem; definir boas práticas para o desenvolvimento de QEM na área da Matemática; e descobrir formas adequadas de análise das QEM de modo a fomentar uma avaliação tão justa quanto possível. O estudo foi conduzido entre 2008 e 2014, utilizando como metodologia a Investigação–Ação (IA), a qual incluiu três ciclos. O 1.º ciclo consistiu, principalmente, no desenvolvimento de um banco de QEM utilizando o *Moodle*, com a implementação de testes para avaliação formativa por meio de trabalhos de casa. O 2.º ciclo de IA consistiu na implementação de testes com QEM para avaliação contínua sumativa, fora do ambiente de sala de aula. O 3.º ciclo consistiu na extensão da implementação dos testes com QEM para avaliação contínua sumativa em ambiente de sala de aula. As mudanças nas práticas educativas foram identificadas utilizando entrevistas aos docentes e um questionário aos alunos. A qualidade das questões foi aferida com recurso a técnicas específicas. Verificou-se, em termos gerais, uma melhoria nas classificações académicas dos estudantes nas UC em estudo ao longo da investigação, assim como um aumento na sua assiduidade às aulas e na valorização dada a um estudo regular. Quanto aos docentes, além da promoção do trabalho em equipa, foi referida a melhoria na forma como elaboram QEM e maior atenção sobre a forma como lecionam. Assim, a implementação desta estratégia de *e-assessment* pode ser considerada um sucesso, nomeadamente por ter permitido dar uma resposta adequada às necessidades inicialmente identificadas, de implementar avaliação contínua sumativa de forma adequada com um número elevado de estudantes.

Palavras-chave

e-assessment, Questões de Escolha Múltipla, Matemática, Ensino Superior, avaliação formativa, avaliação contínua, avaliação sumativa, Investigação-Ação.

ABSTRACT

With the implementation of the Bologna Process several challenges have been posed to higher education institutions in Europe, imposing a paradigm shift. This shift implies the change of the way to assess students in higher education institutions. Continuous and formative assessments emerged as the focus, catalyzed by *e-assessment*. This thesis presents the process of implementation of an *e-assessment* strategy using tests with multiple-choice questions implemented in *Moodle*. This strategy was executed in order to allow the use of continuous summative assessment in mathematics courses in a higher education institution, in classes with large number of students. The objectives of this work were: to see how e-assessment can affect the teaching and learning processes; define the best practices for the development of QEM in mathematics; and find proper ways of analyzing QEM in order to foster an assessment as fair as possible. The research was conducted between 2008 and 2014 using the Action-Research methodology. Three cycles of Action-Research were identified. The first cycle consisted mainly in the development of a bank of multiple choice questions, using *Moodle*, and in the implementation of tests, with those multiple choice questions, as homework, mainly for formative assessment. The second cycle of Action-Research consisted in the implementation of tests with multiple choice questions for continuous summative assessment, but out of class environment. The third cycle of Action-Research consisted in extending the implementation of tests with multiple choice questions for continuous summative assessment in a class environment. Changes in educational practices were ascertained using interviews to teachers and a questionnaire to students. The quality of the questions was analyzed with specific techniques. It was found, in general, an improvement in academic achievement of students in the course studied throughout this research, as well as an increase in their class attendance and a better appreciation of regular study. As for the teachers, besides the promotion of teamwork, an improvement was reported in the way they create the questions and more attention on how they teach. Thus, the implementation of this *e-assessment* strategy can be considered a success, namely because it allowed an adequate response to the main needs identified initially, of adequately implementing continuous summative assessment with a large number of students.

Keywords

E-assessment, Multiple Choice Questions, Mathematics, Higher Education, Formative assessment, Continuous assessment, Summative assessment, Action-Research.

ÍNDICE

Lista de Figuras

Lista de Tabelas

Lista de Equações

Lista de Acrónimos

APM	Associação Portuguesa de Matemática
CAA	*Computer Assisted Assessment* – Avaliação Assistida por Computador
CBA	*Computer Based Assessment* – Avaliação Baseada em Computador
CCI	Curva Característica do Item
CTT	*Classical Test Theory*
EM	Escolha Múltipla
ETS	*Educational Testing Service*
GAIE	Gabinete de Apoio à Inovação em Educação
GTI	Grupo de Investigação em Matemática
IA	Investigação–Ação
IAP	Investigação-Ação Participada
ICC	*Item Characteristic Curve*
IDif	Índice de Dificuldade de uma questão
IDisc	Índice de Discriminação de uma questão
IP	*Internet Protocol* – Protocolo de Internet
IPP	Instituto Politécnico do Porto
IRT	*Item Response Theory*
ISCAP	Instituto Superior de Contabilidade e Administração do Porto
JISC	*Joint Information Systems Committe*
KR20	Índice 20 de *Kuder-Richardson*
LCA	Licenciatura em Contabilidade e Administração
LCI	Licenciatura em Comércio Internacional
LMS	*Learning Management Systems* – Sistemas de Gestão da Aprendizagem
QEM	Questão/ões de Escolha-Múltipla
PAR	*Participatory Action Research*
SOLO	*Structure of Observed Learning Outcomes*
TCT	Teoria Clássica dos Testes
TIC	Tecnologias de Informação e Comunicação
TRI	Teoria da Resposta ao Item
UC	Unidade(s) Curricular(es)

INTRODUÇÃO

O Processo de Bolonha arrancou oficialmente em Junho de 1999, com a Declaração de Bolonha. Esta define um conjunto de etapas a seguir pelos sistemas de Ensino Superior europeus, no sentido de construir um espaço europeu de Ensino Superior globalmente harmonizado. Pretende-se que a harmonização das estruturas do Ensino Superior conduza, por sua vez, a uma Europa da ciência e do conhecimento e, mais concretamente ainda, a um espaço comum europeu de ciência e de Ensino Superior, com capacidade de atração à escala europeia e intercontinental. Passados alguns anos da sua implementação em Portugal, após a publicação do Decreto-Lei nº 74/2006, verifica-se que o Processo de Bolonha constituiu uma enorme oportunidade para a reorganização do Ensino Superior em Portugal, tendo as instituições de Ensino Superior, politécnicas e universitárias, enfrentado grandes desafios. O surgimento de um novo paradigma, valorizando o aluno como sujeito central na construção da sua aprendizagem, exige novas abordagens pedagógicas que favoreçam as condições de ensino/aprendizagem. No entanto, de acordo com Redecker e Johannessen (2013), as mudanças nas práticas pedagógicas e nos processos de aprendizagem, apenas podem acontecer quando mudar também a avaliação. Tradicionalmente a avaliação no Ensino Superior baseava-se, na sua generalidade, na realização de um exame a cada uma das unidades curriculares, consistindo apenas numa avaliação final, do tipo sumativo, contrastando com o que é apontado pelo Processo de Bolonha, que aponta não só para uma avaliação contínua ao longo do(s) semestre(s), mas também englobando metodologias diversificadas (Boticki & Milasinovic, 2008; Llamas-Nistal, Fernández-Iglesias, González-Tato, & Mikic-Fonte, 2013; Mora, Sancho-Bru, Iserte, & Sánchez, 2012; Rod, Eiksund, & Fjaer, 2010).

As Tecnologias de Informação e Comunicação (TIC) levantam desafios e ao mesmo tempo oferecem aos professores ferramentas que permitem criar oportunidades de aprendizagem diferenciadas para os alunos. O seu uso é recomendado por várias organizações europeias, como por exemplo o Parlamento Europeu (Redecker, 2013; Redecker & Johannessen, 2013). A utilização das TIC no processo de avaliação torna-se assim numa mais-valia, através do formato eletrónico ou do *e-assessment*[1]. Neste caso, as TIC são utilizadas em todo o processo de avaliação desde o desenho dos testes até ao armazenamento dos resultados (Stödberg, 2012). Uma possível abordagem consiste no desenvolvimento de ambientes específicos para esse fim (Boticki & Milasinovic, 2008; Dascalu & Bodea, 2010; Llamas-Nistal et al., 2013). Outra abordagem consiste na utilização dos chamados Sistemas de Gestão da Aprendizagem ("Learning Management Systems", LMS) (Burrow, Evdorides, Hallam, & Freer-hewish, 2005; Salas-Morera, Cubero-Atienza, Redel-Macías, Arauzo-Azofra, & García-Hernández, 2012). Os LMS têm a vantagem de fornecerem um vasto conjunto de ferramentas especificamente desenhadas para a implementação de *e-assessment*. Entre essas ferramentas salientamos os

[1] De forma análoga ao que acontece com o termo *e-learning*, para o qual não é feita qualquer tradução, optamos por utilizar o termo *e-assessment*.

testes, os quais podem englobar vários tipos de questões, tais como, escolha-múltipla, verdadeiro/falso, correspondência de itens, resposta curta, entre outros.

Tendo em conta a sua finalidade, a avaliação poderá ser formativa e/ou sumativa, ou diagnóstica (Jacob, Issac, & Sebastian, 2006; Redecker & Johannessen, 2013; Stödberg, 2012). Podemos considerar que a avaliação sumativa reflete o paradigma do "Aprender para Avaliar" e que as avaliações diagnósticas e formativas refletem o paradigma do "Avaliar para Aprender" (Jacob et al., 2006). Podemos afirmar que o primeiro paradigma é o mais comum na avaliação que tradicionalmente se faz no Ensino Superior, que consiste na realização de uma ou mais provas de exame, previamente calendarizadas (Flores, Simão, Barros, & Pereira, 2015, p. 1525). O *e-assessment* tem vindo a servir como catalisador para uma mudança deste primeiro paradigma para o segundo, visto que em estudos científicos relevantes sobre este assunto, a utilização da avaliação formativa ou de ambos os tipos, formativa e sumativa, em simultâneo é mais comum do que a utilização da avaliação sumativa (Stödberg, 2012).

Outro aspeto importante tem a ver com o tipo de tarefa que é realizada no *e-assessment*. Uma classificação com a qual nos identificamos, visto resultar de uma cuidadosa revisão de literatura em algumas das mais importantes revistas científicas da área e dado que corresponde à nossa prática como docente, é apresentada por Stödberg (2012), onde o autor enumera cinco categorias: i) questões de resposta fechada, tais como questões de escolha-múltipla ou de correspondência; ii) questões de desenvolvimento; iii) portfólios; iv) produtos, tais como programas informáticos, e; v) discussões entre os alunos. Ainda no mesmo estudo, verifica-se que as questões de resposta fechada são as mais utilizadas no *e-assessment*. De entre as questões de resposta fechada, as questões de escolha-múltipla (QEM) apresentam particular relevância e revestem-se de algumas particularidades, apresentando vantagens e limitações. Estes vários aspetos, bem como a comparação deste tipo de avaliação com outros, têm sido objeto de estudo na investigação científica nesta área (Bible, Simkin, & Kuechler, 2008; Bush, 2015; Haladyna, Downing, & Rodriguez, 2002; Lee, Liu, & Linn, 2011; Liu, Lee, & Linn, 2011; Rod et al., 2010; Torres, Lopes, Babo, & Azevedo, 2011).

No estudo apresentado em Torres e colaboradores (2011) apresentam-se algumas vantagens dos testes formados por QEM, tais como: i) podem ser utilizados em conteúdos diversificados; ii) podem medir uma grande variedade de objetivos educativos e de aprendizagem; iii) são adaptáveis a vários níveis de capacidades cognitivas; iv) são muito úteis para avaliação de turmas numerosas; v) os testes podem ser automaticamente corrigidos e avaliados e podem ser facilmente realizadas análises estatísticas, utilizando sistemas informáticos, como é o caso dos LMS, e; vi) fornecem o formato mais útil para comparações ao longo do tempo, devido à objetividade na correção. Quanto às limitações dos testes formados por QEM, o mesmo estudo apresenta as seguintes: i) podem ser difíceis de construir para níveis de capacidades cognitivas mais elevadas; ii) requerem boas capacidades de escrita por parte dos professores, de modo a que as questões sejam claras; iii) requerem boas capacidades de leitura por parte dos alunos,

de modo a interpretarem corretamente as questões; iv) não conseguem medir alguns tipos de objetivos de aprendizagem, como por exemplo, a capacidade de comunicar; v) muitas vezes é difícil encontrar bons "distratores", que correspondem às opções não corretas, e; vi) os alunos podem tentar adivinhar a resposta.

No que concerne à investigação científica relacionada com QEM, um dos principais tópicos de investigação está relacionado com a forma como se devem elaborar as QEM. Num estudo importante apresentado em Haladyna, Downing e Rodriguez (2002) é apresentado um conjunto de linhas de orientação para a construção de questões de escolha-múltipla. Este estudo apresenta 31 linhas de orientação divididas em 5 categorias, nomeadamente: i) preocupações com o conteúdo, por exemplo, evitar questões com rasteiras; ii) preocupações com o formato, por exemplo, formatar as questões na vertical e não na horizontal; iii) preocupações com o estilo, por exemplo, editar e rever as questões; iv) escrita do tema, por exemplo, assegurar que as orientações no tema são claras, e; v) escrita das opções, por exemplo, tornar plausíveis todos os "distratores".

Um outro tópico de investigação relaciona-se com o facto de os alunos poderem tentar adivinhar a resposta sem terem conhecimento dos tópicos avaliados e da forma como se poderá minimizar este fenómeno que é, sem sombra de dúvida, uma das grandes desvantagens dos testes QEM. Uma metodologia habitualmente utilizada há já muitos anos, consiste na definição de penalizações para o caso de o aluno selecionar uma resposta errada, isto é, atribuir-lhe uma cotação negativa. Ainda assim, o problema continua presente e muitos alunos continuam a tentar adivinhar a resposta, especialmente aqueles que não têm grandes expectativas em obter boas classificações (Bush, 2015). De modo a tentar minimizar este problema, foram sendo desenvolvidas várias metodologias de classificação dos testes que envolvem a utilização de questões de escolha-múltipla. Por exemplo, Triantis e Ventouras (2012) apresentam uma interessante metodologia, a que foi atribuído o nome de "Classificação com questões emparelhadas". Uma abordagem diferente, que consiste na utilização de testes de escolha-múltipla com formatos mais sofisticados, é apresentada por Bush (2015). Neste trabalho são comparados alguns desses formatos, sendo apontadas as vantagens e limitações de cada um deles. Um novo formato para as opções em testes de escolha-múltipla, chamado de "Explanation Multiple-Choice Items" (Opções de Escolha Múltipla Explanatórias), é apresentado por Liu e colaboradores (2011) e por Lee e colaboradores (2011). Este novo formato é comparado com outros, nomeadamente os de escolha-múltipla tradicional e os de resposta aberta.

Um outro tópico de interesse para a investigação neste domínio concerne à avaliação da qualidade das QEM a partir da análise das respostas dadas pelos alunos a essas questões. Existem duas grandes teorias que permitem realizar esta abordagem, a saber, a Teoria Clássica dos Testes (TCT, em inglês "Classical Test Theory" – CTT) e a Teoria da Resposta ao Item (TRI, em inglês "Item Response Theory" – IRT). A TCT remonta ao início do século XX e baseia-se

maioritariamente no cálculo de dois índices para cada uma das questões em análise, chamados Índice de Dificuldade e Índice de Discriminação. O Índice de Dificuldade de uma questão relaciona-se com a proporção de alunos que consegue responder acertadamente a essa questão e o Índice de Discriminação tem a ver com a capacidade que uma questão tem para distinguir os alunos/examinandos melhores, dos alunos/examinandos piores. A partir do cálculo e da análise dos valores do Índice de Dificuldade e do Índice de Discriminação, verificando se esses valores se encontram dentro de uma gama de valores aceite como adequada, pode-se aferir a qualidade das questões. Quanto à TRI, foi originalmente desenvolvida na década de 1940 e baseia-se na determinação do quanto um aluno/examinando possui de uma característica não observável ou latente – a sua capacidade. A cada nível de capacidade associa-se a probabilidade de o aluno/examinando responder acertadamente a uma determinada questão, através do ajuste de uma função logística cumulativa, a que se chama Curva Característica do Item (CCI, em inglês "Item Characteristic Curve" – ICC). Esta função relaciona a probabilidade de sucesso numa questão com a capacidade do aluno/examinando e com as características dessa questão. As características da questão chamam-se parâmetros e há três modelos básicos que se distinguem através do número de parâmetros utilizados para descrever a questão, a saber, modelo logístico de 1–parâmetro, modelo logístico de 2–parâmetros e modelo logístico de 3-parâmetros. Os parâmetros são a dificuldade, a discriminação e o nível de acerto casual. Todos estes parâmetros devem pertencer a uma determinada gama de valores para que a questão em análise tenha a qualidade desejada.

Problema e objetivos da investigação

Com a adequação dos diferentes cursos ao Processo de Bolonha, passaram a ingressar no Instituto Superior de Contabilidade e Administração do Porto (ISCAP) alunos que não frequentaram a disciplina de Matemática no ensino secundário. Entre estes, conta-se um número significativo de alunos que ingressam através do Acesso a Maiores de 23 Anos, verificando-se que muitos deles já não estudam há alguns anos, daí que apresentem ainda mais dificuldades (designados habitualmente como "estudantes não tradicionais"). Por outro lado, com a reestruturação dos cursos, a carga horária semanal destinada às Unidades Curriculares (UC) da Área Científica de Matemática foi bastante reduzida, as turmas tornaram-se numerosas e também se tornou necessário articular os programas de Matemática com os das outras UC, de forma a proporcionar, em tempo útil, as bases matemáticas necessárias. Assim sendo, houve uma necessidade de implementar novas estratégias e metodologias de apoio aos alunos. Além disso, decorrente do Processo de Bolonha, passou a estar presente a necessidade de implementar processos de avaliação contínua.

Tendo em conta todos estes aspetos, começou a desenvolver-se um processo de avaliação contínua através de *e-assessment*, com a utilização de QEM, implementadas na plataforma *Moodle*. A implementação de um projeto de nome MatActiva (Azevedo, Torres, Lopes, & Babo, 2009; Babo, Azevedo, & Lopes, 2008; Babo, Azevedo, Torres, & Lopes, 2010a, 2010b; Lopes,

Babo, & Azevedo, 2008; Lopes, Babo, Azevedo, & Torres, 2010, 2011; Torres, Lopes, Babo, & Azevedo, 2009; Torres et al., 2011), no qual o autor desta tese participou ativamente e cujo objetivo geral era ajudar os alunos a melhorar o seu desempenho na Matemática utilizando as características de interatividade do *Moodle*, serviu à altura como catalisador para a implementação deste tipo de avaliação em algumas UC da área científica de Matemática. Pretendia-se que os testes fossem gerados aleatoriamente pelo *Moodle*, permitindo que a cada aluno fosse apresentado um teste diferente. Para isso, foi desenvolvido um banco de QEM divididas por categorias, definidas de modo a permitir que todos os testes avaliassem os mesmos resultados de aprendizagem para todos os alunos que os estivessem a realizar. As QEM foram analisadas utilizando técnicas adequadas nomeadamente a teoria clássica dos testes e a teoria de resposta ao item.

A avaliação contínua implementada através de *e-assessment* foi um processo moroso e delicado que levou vários anos. O estudo exposto nesta tese apresenta a implementação do mesmo. Assim, o objetivo geral do estudo é refletir sobre o processo de avaliação da aprendizagem dos alunos em UC de Matemática, utilizando *e-assessment* com testes contendo QEM. Como objetivos mais específicos pretende-se:

- perceber como o *e-assessment* pode influenciar o processo de ensino-aprendizagem por parte dos alunos;
- perceber como o *e-assessment* pode influenciar o processo de ensino-aprendizagem por parte dos docentes;
- definir boas práticas para o desenvolvimento de QEM na área da Matemática;
- descobrir formas adequadas de análise das QEM de modo a fomentar uma avaliação tão justa quanto possível para os alunos.

Metodologia de investigação

Atendendo à natureza do problema, a metodologia de investigação selecionada para conduzir este estudo foi a Investigação-Ação (IA). Nesta metodologia de investigação é dada particular ênfase à adopção por parte do investigador de um papel ativo na ação e na colaboração com os restantes participantes no estudo, provocando mudanças que têm como objetivo atingir melhorias nas práticas implementadas no contexto onde foi desenvolvido o estudo (Hughes, 2008; Sousa & Baptista, 2011; Yin, 2011). Pretende-se com a Investigação-Ação que os professores adquiram uma maior consciência e autoconfiança, levando-os a aprender e a mudar as suas práticas. No caso deste estudo, foram identificados 3 ciclos de IA, os quais se descrevem sucintamente de seguida.

O 1.º Ciclo de IA consistiu no início da implementação de uma estratégia de *e-assessment* com utilização de QEM, para utilização em avaliação contínua, quer para avaliação formativa quer sumativa, nesta fase inicial apenas como trabalhos de casa. Foi definida uma estratégia de

avaliação, implementado um banco de questões, testado um processo de revisão das questões e dos testes e foram analisadas linhas de orientação para a elaboração das QEM. Um aspeto importante é que os testes foram implementados na plataforma *Moodle*, sendo que a cada aluno era apresentado um teste diferente, gerado aleatoriamente pelo *Moodle* através da seleção das questões contidas no banco de questões.

O 2.º ciclo de IA consistiu na utilização das QEM, desenvolvidas durante o 1.º ciclo de IA e em mais algumas desenvolvidas durante este ciclo, para avaliação sumativa presencial, através de testes realizados em período letivo, fora do ambiente de sala de aula. Em primeiro lugar, tiveram de ser criadas as condições tecnológicas necessárias para a realização dos testes dado que havia falta de equipamentos para os alunos o poderem realizar, o nível de segurança da rede da escola e da plataforma *Moodle* não oferecia as garantias necessárias, bem como a capacidade dos servidores que alojavam a plataforma *Moodle* não era suficiente para responder a um nível de serviço elevado como aquele que era necessário para a realização dos testes. Em colaboração com o gabinete técnico de informática e com os responsáveis administrativos da plataforma *Moodle*, foram criadas as condições tecnológicas necessárias para garantir a realização dos testes. Neste 2.º ciclo de IA houve também preocupação com questões de âmbito científico-pedagógico. As respostas dadas pelos alunos nos vários testes foram analisadas no que diz respeito à consistência, ao nível de dificuldade e ao nível de discriminação. Atendendo às análises anteriores, as QEM foram agrupadas de acordo com as suas características.

O 3.º ciclo de IA consistiu na execução de melhoramentos no processo e na criação das condições necessárias para a realização dos testes durante o horário normal das turmas e em ambiente de sala de aula. Foram efetuadas entrevistas aos docentes e um questionário aos alunos. Foi feita a análise dos dados recolhidos, de modo a aferir sobre as mudanças nas práticas educativas resultantes da implementação deste processo de *e-assessment*.

Estrutura da tese

A tese divide-se em duas partes fundamentais, a saber, a Parte I, Enquadramento Teórico e a Parte II, Estudo Empírico.

Na Parte I, começamos por descrever, no capítulo 1, diversas modalidades de avaliação, incluindo o *e-assessment*, a avaliação formativa, a avaliação sumativa e a avaliação contínua. Seguidamente, no capítulo 2, abordam-se as QEM, referindo as suas vantagens e as suas limitações, os formatos das QEM, os Bancos de Questões e linhas de orientação para a elaboração de QEM. Segue-se a abordagem à análise de testes e questões, no capítulo 3, abordando a Teoria Clássica dos Testes e a Teoria da Resposta ao Item, assim como algumas considerações adicionais sobre a análise de testes e questões. A primeira parte da tese termina com uma abordagem às taxonomias de aprendizagem no capítulo 4, nomeadamente a Taxonomia de Bloom e a Taxonomia SOLO.

Na Parte II, apresentamos a metodologia de investigação no capítulo 5, principiando com um preâmbulo onde enunciamos os objetivos do estudo para, logo de seguida, abordarmos a Investigação-Ação enquanto opção metodológica. O contexto da investigação e participantes são descritos na secção seguinte, avançando depois para o desenho da investigação, onde são especificados cada um dos 3 ciclos de IA desenvolvidos. Ainda neste capítulo, descrevem-se os instrumentos de recolha de dados utilizados, mais concretamente: o banco de questões, explanando como foram definidas as categorias para as questões, como foram criados as questões e os testes, como foi efetuado o processo de revisão, caracterizando o banco de questões por ciclo, explicando como foram aplicados os testes opcionais do 1º ciclo de IA e como foram avaliadas as questões usando a Teoria Clássica dos Testes e a Teoria da Resposta ao Item; os questionários aos docentes no 1.º ciclo de IA; a entrevista aos docentes no 3.º ciclo de IA, e; o questionário aos alunos no 3.º ciclo de IA. O capítulo 6 consiste na apresentação e análise dos dados, tomando a evolução das classificações dos estudantes ao longo do processo de investigação, as respostas dos docentes ao questionário no 1.º ciclo de IA, a análise da qualidade dos testes e das questões usando a TCT e a TRI, as respostas dos docentes nas entrevistas no 3.º ciclo de IA e, por fim, as respostas dos estudantes ao questionário no 3.º ciclo de IA. No capítulo 7 são discutidos os resultados obtidos.

A tese termina com a apresentação da conclusão, onde salientamos os principais contributos do trabalho realizado, suas limitações ou constrangimentos e perspectivas para investigação futura.

PARTE I. ENQUADRAMENTO TEÓRICO

CAPÍTULO 1. MODALIDADES DE AVALIAÇÃO DOS ESTUDANTES

A avaliação pode influenciar profundamente a motivação daqueles que aprendem, bem como moldar a sua perspetiva sobre a aprendizagem. Assim, a introdução de sistemas de avaliação diferentes poderá ter um impacto importante em todo o processo educativo (Boticki & Milasinovic, 2008; Brown, 2001; Bull & Danson, 2001; Frankland, 2007a; Garfield & Ben-Zvi, 2008; Holmes, 2015; Jacob et al., 2006; Jarvis, Holford, & Griffin, 2003; JISC, 2007; Redecker, 2013; Scouller, 1998; Smith et al., 1996; Stödberg, 2012; Wild, Triggs, & Pfannkuch, 1997). Neste capítulo, começamos por apresentar algumas definições e características ou princípios que devem nortear a avaliação dos estudantes, nomeadamente os que frequentam o Ensino Superior na atualidade, com a adesão das instituições ao Processo de Bolonha. Em seguida abordamos o *e-assessment*, apresentam-se aspetos relacionados com a avaliação formativa e com a avaliação sumativa e por fim, a avaliação contínua.

Diversas propostas podem ser encontradas na literatura para definir a avaliação no processo de aprendizagem dos estudantes. De forma bastante concreta e sucinta, Jarvis e colaboradores (2003) consideram que a avaliação serve basicamente para perceber se os alunos aprenderam ou não, quanto aprenderam e o quê. Numa abordagem mais profunda, Brown refere que a "avaliação define o que os alunos veem como sendo importante, como gastam o seu tempo e como se veem a eles mesmos como alunos e como diplomados" (traduzido de Brown, 2001, p. 4). Em termos formais, trata-se de um "processo de guardar informação sobre o desempenho individual dos alunos de modo a fazer julgamentos sobre o seu progresso (…) descobrindo se os objetivos de aprendizagem estão a ser atingidos" (traduzido de Llamas-Nistal et al., 2013, p. 72). De forma análoga, a avaliação é entendida como um "o processo de provar e determinar em que medida um aluno foi de encontro ou fez progressos em relação aos critérios definidos" (traduzido de JISC, 2006, p. 12), podendo, de forma mais abrangente, ser utilizada como uma medida da evolução dos processos de ensino e de aprendizagem (Wong, 2007). Assim, vários autores destacam a necessidade de a avaliação estar sempre alinhada com os objetivos de aprendizagem (Brown, 2001; Garfield & Ben-Zvi, 2008; Holmes, 2015; Rice & Campbell, 2007). Além de permitir averiguar o sucesso educativo do estudante, a avaliação serve também como forma de lhe dar um *feedback* sobre as suas forças e fraquezas, com base no desempenho por si demonstrado (Rod et al., 2010). Tomando os aspetos anteriormente mencionados, parece-nos que a definição seguinte permite, de certa forma, uma síntese integradora daqueles que mais se evidenciam: "A avaliação é tradicionalmente vista como o processo de acumulação de informação e a formação de julgamentos acerca daquilo que foi alcançado pelos alunos relativamente a conteúdos específicos (…) podendo resultar numa classificação; em *feedback*

para os alunos, professores e famílias; em motivos para ajustar as metodologias de ensino ou em planos de remediação" (traduzido de O'Toole, 2007, p. 469).

Por princípio, deve ser sempre garantido que a avaliação seja válida e fiável, ainda que seja difícil consegui-lo (Ferrão, 2010; Frankland, 2007a; Haladyna, 2004; Jarvis et al., 2003; Knight, 2001; Race, 2001; Rice & Campbell, 2007; Wong, 2007). A validade tem a ver com o facto de se avaliar o que efetivamente está definido nos objetivos de aprendizagem, o grau com que se testam as capacidades, conhecimentos ou competências que é suposto e se pretende realmente avaliar (JISC, 2006; McAlpine, 2002c).

Por outro lado, uma avaliação fiável deve ser independente de qual é o avaliador envolvido, do local e do momento em que um determinado avaliador classifica os trabalhos do estudante. A fiabilidade depende da definição de medidas objetivas, precisas, repetíveis e analiticamente sólidas e diz respeito "ao grau pelo qual os resultados de um teste são repetíveis e justos, quer de estudante para estudante quer de um momento para outro" (traduzido de JISC, 2006, p. 92). Como refere McAlpine (2002c), esta propriedade assegura resultados similares em circunstâncias idênticas. Deve ser conseguido um compromisso entre estes dois aspetos, pois a verdade é que quanto mais simplificamos o que tentamos avaliar, maior fiabilidade obtemos mas, por outro lado, simplificar pode de alguma forma comprometer a validade (Knight, 2001). Assim sendo, o avaliador deve focar-se em que a avaliação seja válida, justa e fiável (Race, 2001; Rice & Campbell, 2007).

A avaliação dos estudantes no Ensino Superior adquire particular relevância, pois dela dependerá a atribuição de um grau académico, ou seja, a sua capacidade para determinar se os alunos atingiram ou não os objetivos definidos como necessários para a atribuição desse grau. Tradicionalmente, a avaliação no Ensino Superior consistia num procedimento formal (Jarvis et al., 2003; McAlpine, 2002c), através da realização de exames efetuados no final de cada semestre letivo, em datas pré-definidas pelas instituições para cada uma das Unidades Curriculares (UC) que formam os diversos cursos. Seja qual for a modalidade de avaliação adotada, nos países da União Europeia os alunos obtêm o grau caso sejam aprovados em todas as UC.

Pode-se afirmar que o Ensino Superior pré-Bolonha estava ainda bastante centrado no professor que "debitava" os conteúdos em aulas nas quais os alunos desempenhavam um papel mais passivo, limitando-se a "absorver" os conteúdos assim transmitidos (Brito, 2012; Melo, 2012; Rod et al., 2010; Sousa, 2011). A entrada do Processo de Bolonha veio introduzir mudanças neste paradigma, na medida em que se defende que todo o processo educativo deve centrar-se no aluno, sendo este responsável pela construção da sua própria aprendizagem, apontando-se assim para uma mudança nas práticas educativas e consequentemente nas práticas de avaliação, levando a que se promova a realização da chamada avaliação contínua (Brito, 2012; Ferrão, 2010; Melo, 2012; Rod et al., 2010; Rust, 2001; Sousa, 2011). O Processo de Bolonha

visou "a criação de um espaço europeu mais competitivo, baseado no conhecimento e capaz de garantir um crescimento económico sustentável, através de sistemas curriculares centrados nos objetivos de formação" (Sousa, 2011, p. 33). Na verdade, o Processo de Bolonha não é primordialmente um assunto de avaliação, mas relaciona-se com ela (Blanco & Ginovart, 2012; Yorke, 2001), dado que a sua aplicação se centra, em grande medida, nos processos de reconhecimento de competências e diplomas conjuntos para as diversas instituições europeias de Ensino Superior e, além disso, a introdução de sistemas de avaliação diferentes têm impactos importantes em todo o processo educativo. Assim sendo, a mudança nas práticas de avaliação é primordial para a mudança nas práticas educativas.

1.1. *E-assessment*

As Tecnologias de Informação e Comunicação (TIC) trazem novos desafios aos professores e ao mesmo tempo oferecem-lhes ferramentas que lhes permitem criar oportunidades de aprendizagem diferenciadas para os alunos. A sua utilização é recomendada por várias organizações europeias, nomeadamente, pelo Parlamento Europeu, sendo que as TIC têm vindo a emergir como um tópico em crescente investimento na área da Educação (Blanco & Ginovart, 2012; Cook & Jenkins, 2010; Redecker, 2013). A utilização das TIC no processo de avaliação dos alunos encontra-se numa fase ainda mais recente de aplicação e desenvolvimento, estando associada, em grande medida, à necessidade de adequação das formas tradicionais de avaliação nos cursos ou formações realizadas em formato de *e-learning*. Várias designações têm sido utlizadas para esta forma de avaliação: *e-assessment*, avaliação assistida por computador ("Computer Assisted Assessment", CAA) ou avaliação baseada em computador ("Computer Based Assessment", CBA). Podemos afirmar que os três termos são equivalentes e representam o mesmo conceito (JISC, 2006, 2007; Jordan, 2013; Redecker, 2013). Neste documento será utilizado o termo *e-assessment*.

O *e-assessment* inclui todo o processo de avaliação, cobrindo uma vasta gama de atividades que vão desde o desenho das atividades a atribuir, até ao armazenamento dos resultados, passando pela entrega de avaliações, classificações e todos os processos de elaboração de relatórios, armazenamento e transferência de dados associados quer a avaliações internas quer externas, processo este no qual as TIC são utilizadas em qualquer uma dessas atividades (JISC, 2006, 2007; Stödberg, 2012). Bull e Danson (2001) apresentam o *e-assessment* como sendo um termo genérico, o qual diz respeito à aplicação das tecnologias informáticas no processo de avaliação. Verifica-se que a maioria das aplicações de *e-assessment* incluem as chamadas questões de resposta fechada, como por exemplo questões de escolha múltipla ou de correspondência, mas também podem ser encontrados outros tipos de tarefas de avaliação no *e-assessment* tais como portefólios ou discussões (Cook & Jenkins, 2010; Stödberg, 2012).

Na literatura encontram-se diferentes tipos de abordagem para o *e-assessment*. Uma abordagem consiste no desenvolvimento de ambientes específicos, os quais têm como principal vantagem serem desenhados e implementados de acordo com as necessidades do utilizador (Boticki & Milasinovic, 2008; Dascalu & Bodea, 2010; Gruttmann, Böhm, & Kuchen, 2008; Guo, Palmer-Brown, Lee, & Cai, 2014; Jordan, 2013; Llamas-Nistal et al., 2013; McGuire, Youngson, Korabinski, & McMillan, 2002; Vora & Shinde, 2014; Wilson, Boyd, Chen, & Jamal, 2011). Alguns autores utilizam os Sistemas de Gestão da Aprendizagem ("Learning Management Systems", LMS) os quais têm a vantagem de disponibilizarem diversas ferramentas especificamente desenhadas para a implementação de atividades de avaliação (Blanco & Ginovart, 2012; Holmes, 2015; Mora et al., 2012; Moscinska & Rutkowski, 2012; Salas-Morera et al., 2012; Sorensen, 2013). Há ainda autores que utilizam os chamados Sistemas de Avaliação, sistemas esses desenvolvidos especificamente para elaboração e apresentação aos alunos de questões destinadas a avaliação através de *e-assessment* (Burrow et al., 2005). Estes sistemas contêm bancos de questões previamente desenvolvidos e que podem ser utilizados principalmente para avaliação formativa (Hauk, Powers, & Segalla, 2015; Mathai & Olsen, 2013).

Encontram-se aplicações de *e-assessment* nas mais diversas áreas, tais como Geografia (Holmes, 2015; Rod et al., 2010; Wilson et al., 2011), Gestão (Jacob et al., 2006), Química (Sorensen, 2013), Medicina (Harris et al., 2015) ou Engenharia (Boticki & Milasinovic, 2008; Burrow et al., 2005; Jacob et al., 2006; Moscinska & Rutkowski, 2012). Também na Matemática se podem encontrar alguns exemplos de aplicação (Acosta-Gonzaga & Walet, 2013; Blanco & Ginovart, 2012; Ferrão, 2010; Gruttmann et al., 2008; Hauk et al., 2015; Mathai & Olsen, 2013). Nos estudos científicos aqui citados, combinam-se vários componentes de avaliação, sendo que pelo menos um desses componentes é um teste com Questões de Escolha Múltipla (QEM). No entanto, na maior parte dos casos, os testes com QEM ou são utilizados para avaliação formativa ou representam apenas uma percentagem muito pequena na avaliação sumativa.

Uma das grandes vantagens do *e-assessment* é a possibilidade de ser utilizado na avaliação de um elevado número de alunos, facilitando assim o trabalho do professor e permitindo poupar em termos de economia de espaço e de tempo (Blanco & Ginovart, 2012; Boticki & Milasinovic, 2008; Bull & Danson, 2001; Jordan, 2013; Mora et al., 2012; Moscinska & Rutkowski, 2012; Rust, 2001; Yorke, 2001). Na literatura podem encontrar-se muitas outras vantagens, as quais se apresentam resumidamente na Tabela 1. A ordem de apresentação não pretende definir qualquer ordem de importância às vantagens apresentadas. A maioria das vantagens foca-se no professor, havendo no entanto algumas vantagens que se focam mais no aluno. Estas últimas encontram-se a sombreado na tabela. Algumas das vantagens apontadas, nomeadamente, "Aliviar o trabalho que representa para o professor avaliar um elevado número de alunos", "Reduzir a carga associada à correção/classificação dos elementos de avaliação" ou "Obtenção quase instantânea das classificações", estão diretamente ligadas à realização *de e-assessment*

através de questões de resposta fechada, como por exemplo as QEM ou as questões de correspondência, sendo que estas são as formas mais comuns de realização de *e-assessment*.

Tabela 1: Vantagens do *e–assessment*

Vantagens
• Menor trabalho em avaliar um elevado número de alunos (Blanco & Ginovart, 2012; Boticki & Milasinovic, 2008; Bull & Danson, 2001; Jordan, 2013; Mora et al., 2012; Moscinska & Rutkowski, 2012; Rust, 2001; Yorke, 2001)
• Menor carga associada à correção/classificação dos elementos de avaliação (Bull & Danson, 2001; Jordan, 2013; Race, 2001; Redecker, 2013)
• Poupança de recursos (Bull & Danson, 2001; Gruttmann et al., 2008; Jordan, 2013; Mora et al., 2012)
• Poupança de tempo (Bull & Danson, 2001; Jordan, 2013)
• Rapidez na obtenção das classificações, podendo estas, por vezes, ser obtidas de forma instantânea/automática (Bull & Danson, 2001; Cook & Jenkins, 2010; Mora et al., 2012; Redecker, 2013)
• O meio de realização dos elementos de avaliação pode ser mais rico do que o baseado em papel, podendo incluir cores, animação, som e mesmo vídeo (Bull & Danson, 2001; Cook & Jenkins, 2010; Jordan, 2013; Mora et al., 2012; Redecker, 2013)
• Possibilidade de obter provas adaptadas a diferentes necessidades (personalização, redimensionamento) (Bull & Danson, 2001; Cook & Jenkins, 2010; JISC, 2007)
• Uma avaliação pode ser repetida várias vezes, apresentando-se de cada vez pequenas variações no conteúdo das questões (Bull & Danson, 2001; JISC, 2007)
• Maior diversidade naquilo que é testado (a nível dos conteúdos, das tarefas, das atividades e dos métodos) (Cook & Jenkins, 2010; JISC, 2007; Jordan, 2013)
• Pode constituir um meio poderoso para a realização de avaliação contínua, dado que permite um *feedback* mais rápido aos alunos e envolve, normalmente, menores custos e menos recursos (Gruttmann et al., 2008; McAlpine, 2002c)
• Melhor fiabilidade (mais objetivo e menor enviesamento nas classificações) (Cook & Jenkins, 2010; Jordan, 2013; Moscinska & Rutkowski, 2012)
• Pode ser efetuado em qualquer altura, em qualquer lugar (Cook & Jenkins, 2010)
• Facilidade de armazenar, editar, reproduzir, recombinar e reutilizar informação (Cook & Jenkins, 2010; Redecker, 2013)
• Capacidade para gerar automaticamente indicadores de qualidade para as questões (McAlpine, 2002c)
• Possibilidade de fornecer *feedback* impessoal, sem a noção de julgamento (Jordan, 2013)
• Obtenção quase instantânea das classificações (Bull & Danson, 2001; Cook & Jenkins, 2010; Mora et al., 2012; Moscinska & Rutkowski, 2012; Yorke, 2001)
• Obtenção atempada de *feedback* específico (Bull & Danson, 2001; Cook & Jenkins, 2010; McAlpine, 2002c; Mora et al., 2012; Yorke, 2001)
• Possibilidade de fomentar um maior envolvimento e motivação nos alunos (Jordan, 2013; Redecker, 2013)
• Possibilidade de facilmente rever as questões e alterar as respostas (Cook & Jenkins, 2010)

A introdução de novas formas de avaliação envolve críticas que não são habitualmente consideradas no processo tradicional de avaliação (Bull & Danson, 2001). Muitas vezes, essas críticas prendem-se apenas com a habitual resistência às mudanças. Habitualmente os alunos manifestam uma opinião favorável sobre o *e-assessment* (Blanco & Ginovart, 2012; Burrow et al., 2005; Dascalu & Bodea, 2010; Douglas, Wilson, & Ennis, 2012; Ferrão, 2010; Green & Mitchell, 2009; Jacob et al., 2006; JISC, 2007; Rod et al., 2010; Sorensen, 2013; Wilson et al., 2011). Hauk e colaboradores (2015) apresentam um estudo no qual comparam as classificações obtidas pelos alunos, sendo que alguns alunos realizavam trabalhos de casa executados em papel e outros alunos realizavam trabalhos de casa executados em computador. Conclui-se que os trabalhos de casa realizados em computador são pelo menos tão efetivos como os realizados em papel, para os alunos estudados (uma Unidade Curricular de Álgebra nos Estados Unidos da América). Wilson e colaboradores (2011), em aulas do primeiro ano de um curso de Geografia, utilizaram avaliação formativa na forma de *e-assessment* ao longo do semestre. Essa avaliação era feita voluntariamente pelos alunos e verificou-se que os alunos que utilizaram esta avaliação formativa obtiveram resultados melhores na avaliação sumativa final e que maiores volumes de utilização da plataforma de e-assessment estavam associados a melhores resultados na avaliação sumativa final. De qualquer das formas, é reconhecido na literatura que o *e-assessment* apresenta algumas limitações. Na Tabela 2 encontram-se algumas dessas limitações, sendo que a ordem de apresentação não pretende definir qualquer ordem de importância às limitações apresentadas.

Tabela 2: Limitações do *e-assessment*

Limitações
• Dificuldades em garantir as condições tecnológicas de modo a que não haja alunos discriminados, isto é, de modo a que todos os alunos possam aceder à avaliação em condições idênticas (Bull & Danson, 2001)
• É mais difícil e moroso escrever as questões (Cook & Jenkins, 2010; Jordan, 2013)
• Muito tempo despendido com o início do processo de implementação (Cook & Jenkins, 2010; Green & Mitchell, 2009; Yorke, 2001)
• Pode apresentar custos elevados para a implementação (Cook & Jenkins, 2010)
• Alguns sistemas apresentam funcionalidades limitadas (Cook & Jenkins, 2010)
• Apresenta grandes desafios a nível organizacional (adequação às normas de avaliação internas e externas, gestão das infraestruturas de apoio e gestão de picos de utilização) (Cook & Jenkins, 2010; JISC, 2007)
• Pode representar custos elevados na formação dos docentes (Mora et al., 2012)

Um aspeto importante do *e-assessment* diz respeito ao tipo de tarefa que é realizada. Uma classificação dos tipos de tarefas *no e-assessment*, com a qual nos identificamos, considerando

que é o resultado de uma cuidadosa revisão de literatura em algumas das revistas científicas mais relevantes na área e porque corresponde à nossa prática como professor, é apresentada por Stödberg (2012). Nesta classificação são consideradas cinco categorias de tarefas no *e-assessment*: i) questões de resposta fechada, tais como as QEM, ou as questões de correspondência; ii) questões de resposta aberta; iii) portfólios; iv) produtos, tais como programas de computador, e; v) discussões entre os alunos.

Muitas vezes associa-se o *e-assessment* apenas às QEM. No entanto, há um grande leque de atividades diferenciadas que podem ser implementadas com *e-assessment* que não se podem implementar em formato papel (Bull & Danson, 2001). Ainda assim, a verdade é que as questões de resposta fechada continuam a ser as mais utilizadas no *e-assessment* (Stödberg, 2012). Entre este tipo de questões, as QEM apresentam particular relevância e têm algumas especificidades, apresentando algumas vantagens e também algumas limitações, para além daquelas que são apresentadas para o *e-assessment* em geral, nas Tabela 1 e 2, respetivamente. As vantagens e limitações que dizem respeito especificamente às QEM serão apresentadas numa das próximas secções desta tese.

1.2. Avaliação formativa e avaliação sumativa

Atendendo ao seu propósito, a avaliação pode ser formativa e/ou sumativa, ou diagnóstica (Jacob et al., 2006; Jarvis et al., 2003; Redecker & Johannessen, 2013; Stödberg, 2012). Alguns autores apresentam a avaliação diagnóstica como sendo um caso especial da avaliação formativa (Knight, 2001; McAlpine, 2002c; O'Toole, 2007). Na Tabela 3 encontra-se um resumo de conceitos de avaliação formativa e de avaliação sumativa que se encontram na literatura. Pode-se considerar que a avaliação sumativa reflete o paradigma de "Aprender para Avaliar" e que a avaliação formativa reflete o paradigma "Avaliar para Aprender" (Jacob et al., 2006). O primeiro paradigma era mais comum na avaliação que tradicionalmente se fazia no Ensino Superior pré-Bolonha (Frankland, 2007c), que consistia na aplicação de exames, previamente agendados, em formato papel. O *e-assessment* pode ser útil e pode trazer benefícios para ambos os tipos de avaliação, formativa e sumativa (Bull & Danson, 2001; McAlpine, 2002c). Podemos afirmar que o *e-assessment* tem vindo a servir como catalisador para uma mudança deste primeiro paradigma para o segundo, visto que em diversos estudos científicos sobre *e-assessment* se verifica que a utilização de avaliação formativa ou o uso de ambos os tipos de avaliação, formativa e sumativa, é mais comum do que o uso de apenas avaliação sumativa (Stödberg, 2012).

Tabela 3: Noções de avaliação formativa e sumativa

Avaliação formativa	Avaliação sumativa	Traduzido de:
"fornece *feedback* aos alunos durante o curso, de modo a que tenham oportunidade de melhorar"	"contribui para as classificações finais de um módulo, nível ou grau"	(Brown, 2001, p. 6)
"permite aos alunos e seus professores aferir quanto é que foi aprendido, identificar áreas que necessitam de mais trabalho e ajudar os alunos a reforçar a sua aprendizagem"	"tem como objetivo a medição da aprendizagem do aluno, habitualmente no final do programa de estudos"	(Cook & Jenkins, 2010, p. 8)
"refere-se ao *feedback* fornecido durante a aprendizagem de modo a que os alunos e os professores saibam como o ensino e aprendizagem estão a prosseguir e como podem ser melhorados"	------	(Frankland, 2007c, p. 70)
"é levada a cabo para ajudar a planear como o ensino ou a aprendizagem devem ter lugar, ou para alterar o ensino ou a aprendizagem enquanto estes prosseguem"	"apenas nos diz o que foi aprendido no final do processo de aprendizagem ou de ensino"	(Jarvis et al., 2003, p. 159)
"é aquela que fornece *feedback* sobre o desenvolvimento de um aluno (e talvez também do professor) sobre um item, um grupo de itens ou sobre tópicos que com eles se relacionem, de modo a ajustar os seus planos para a aprendizagem que se segue.	"aquela que geralmente é feita no final de uma atividade ou um programa de aprendizagem e que é usada para fazer um juízo sobre os progressos globais conseguidos pelos alunos"	(JISC, 2006, p. 52 e 101)
"é aquela que fornece *feedback* sobre o desenvolvimento de um aluno sobre os seus entendimentos e competências. Pode também ser descrita como avaliação para aprender"	"a avaliação final do desempenho de um aluno, levando geralmente a uma qualificação ou certificação formal de uma competência. Também referida como avaliação da aprendizagem"	(JISC, 2007, p. 6)
"destina-se a informar os alunos sobre como podem fazer melhor"	"fornece um resultado, na forma de uma prova de desempenho ou competência (por exemplo, um certificado), e na forma de informação que pode ser usada como indicador de desempenho para avaliar o trabalho de professores, departamentos, escolas e o sistema nacional de educação"	(Knight, 2001, p. 3 e 7)
"é desenhada para apoiar o processo de aprendizagem fornecendo *feedback* ao aluno, o que pode ser utilizado para evidenciar áreas que necessitem mais estudo e portanto melhorar o desempenho futuro"	"destina-se à progressão e/ou análise externa, dada no final de um curso e concebida para julgar o desempenho global dos alunos "	(McAlpine, 2002c, p. 6)
"procura guardar evidências sobre a proficiência dos alunos com o objetivo de influenciar os métodos e as prioridades do ensino"	"é utilizada para determinar o que os alunos conseguiram alcançar no final de um programa de trabalho"	(Redecker & Johannessen, 2013, p. 79)
"é levada a cabo durante o processo de aprendizagem numa disciplina ou curso e pretende fornecer *feedback* aos alunos sobre os seus progressos de modo a apoiar a sua aprendizagem"	"pretende sumariar aquilo que os alunos conseguiram alcançar, através da realização de uma apreciação ou determinação de uma classificação"	(Stödberg, 2012, p. 595)

Uma Unidade Curricular engloba habitualmente avaliação formativa e avaliação sumativa (Brown, 2001). A avaliação sumativa deve sempre fornecer algum *feedback* aos alunos, apresentando assim valor formativo. A uma tarefa de avaliação formativa não deve ser atribuído um valor sumativo, porque se pode perder a sua essência, visto que na avaliação formativa espera-se que o aluno manifeste abertamente as suas dificuldades, enquanto que na avaliação sumativa pode tentar escondê-las para obter melhor classificação (Hernández, 2007; Knight, 2001). No entanto, em alguns casos opta-se por atribuir algum valor sumativo, embora pequeno, à avaliação formativa, de modo a aumentar o compromisso dos alunos (Holmes, 2015). Um dos elementos-chave da avaliação formativa é o *feedback* fornecido aos alunos (Llamas-Nistal et al., 2013; Oldham, Freeman, Chamberlain, & Ricketts, 2007). Outro aspeto importante é a chamada avaliação pelos pares, que é uma das formas de operacionalizar os princípios da avaliação formativa (Frankland, 2007b).

1.3. Avaliação contínua

Tal como já foi referido anteriormente, o Processo de Bolonha aponta para diversas formas de avaliação implementadas durante o semestre/ano académico - num sistema de avaliação contínua-, ao passo que a avaliação que tradicionalmente se utilizava no Ensino Superior consistia na aplicação de exames numa única avaliação final. "A avaliação contínua pode ser definida como aquela que utiliza testes ao longo de uma unidade de aprendizagem, e a acumulação dos resultados numa classificação final" (traduzido de Holmes, 2015, p. 2). A avaliação contínua, normalmente, propicia mais a avaliação formativa do que a avaliação sumativa. No entanto, há também muitas vezes avaliação contínua sumativa, da qual é exemplo a avaliação que é feita, na maior parte dos casos, no Ensino Superior no período pós-Bolonha.

A avaliação contínua apresenta várias vantagens, entre as quais se salientam (Borba & Penteado, 2001): i) favorece a existência de itinerários de aprendizagem alternativos; ii) permite uma aprendizagem mais dinâmica; iii) estimula e apoia os progressos, dando prioridade aos elementos positivos, em vez de dar prioridade aos elementos negativos; iv) desenvolve a capacidade de reflexão, visto que dá a oportunidade aos alunos para se organizarem e entenderem os seus procedimentos e progressões. Por seu turno, o "*e-assessment* pode fornecer um poderoso meio para a realização de avaliação contínua, fornecendo aos alunos e professores *feedback* rápido e detalhado sobre o processo de aprendizagem" (traduzido de McAlpine, 2002c, p. 8). Na literatura podemos encontrar alguns exemplos de utilização de *e-assessment* em ambientes de avaliação contínua. Apresentamos, por ordem cronológica, alguns que consideramos relevantes:

- Boticki e Milasinovic (2008) desenvolveram um sistema de *e-assessment* baseado na Web, através do qual os alunos de Engenharia, como complemento a testes intermédios

e exames finais, desenvolveram ao longo do semestre trabalhos de casa de programação, avaliados automaticamente pelo sistema, e testes com questões de escolha múltipla. Esta avaliação era sumativa, sendo que o *e-assessment* representava um peso de 30% na classificação final;

- Rod e colaboradores (2010) desenvolveram uma estratégia de avaliação contínua que envolveu a utilização de um conjunto de QEM implementadas no LMS da organização, as quais eram utilizadas essencialmente para avaliação formativa, embora tivessem um pequeno peso de 8% na avaliação sumativa;
- Mora e colaboradores (2012) utilizaram *e-assessment* na avaliação contínua sumativa e formativa de alunos de Engenharia. A componente de *e-assessment* envolveu a utilização de testes periódicos com QEM, valendo 20% da avaliação sumativa;
- Llamas-Nistal e colaboradores (2013) desenvolveram uma ferramenta de *e-assessment* para apoiar o processo de avaliação de alunos diversificados. Essa ferramenta foi utilizada na avaliação contínua sumativa de alunos de Engenharia de Telecomunicações, tendo sido aplicados vários mini testes, distribuídos ao longo do semestre, que representavam 100% da avaliação para os alunos que tivessem optado pelo regime de avaliação contínua;
- Holmes (2015) verificou que a utilização de *e-assessment* com a aplicação de pequenos mini testes semanais para avaliação contínua sumativa, para alunos de um curso de Geografia, melhorou o empenho dos alunos.

CAPÍTULO 2. QUESTÕES DE ESCOLHA MÚLTIPLA

Um teste de escolha múltipla convencional consiste numa coleção de QEM. Uma QEM, tipicamente é "uma questão na qual se pretende que o aluno selecione uma só resposta correta a partir de um leque de opções disponíveis" (traduzido de JISC, 2006, p. 74). Neste capítulo, começamos por apresentar as principais características, vantagens e limitações das QEM, prosseguindo com uma abordagem dos formatos que as QEM podem tomar e as suas implicações na avaliação. Em seguida, explicamos alguns conceitos importantes no que diz respeito aos bancos de questões e, por fim, terminamos com uma breve discussão sobre linhas de orientação para o desenvolvimento de QEM.

A utilização das QEM remonta ao início do século XX, muito antes de existir *e-assessment*. Aceita-se que a primeira utilização foi feita por Frederick J. Kelly com o objetivo de reduzir a ambiguidade, e consequente diferenciação, nas avaliações feitas pelos professores aos seus alunos (Watters, 2015). Desde então, a utilização das QEM tem vindo a ganhar popularidade devido à sua objetividade, sendo que a primeira utilização em grande escala foi feita para o recrutamento de pessoal militar na I Guerra Mundial, com os testes Alpha e Beta de E.L. Thorndike (Jordan, 2013; Watters, 2015). No início do século XX, surgiram diversas máquinas para a realização de testes com QEM, facilitando a sua aplicação a imensas pessoas de forma rápida e eficiente (Watters, 2015). Também o aparecimento, na década de 50 do século XX, do leitor ótico de alta velocidade foi um grande impulso para a expansão das QEM (Liu et al., 2011). Durante o século XX, as QEM ganharam grande popularidade em contexto educativo, principalmente nos EUA, à medida que os investigadores foram descobrindo as limitações das questões de resposta aberta e as vantagens das QEM, nomeadamente a sua objetividade e consistência (Jordan, 2013).

Estruturalmente, as QEM são compostas por 3 elementos: (1) um tronco[2] que apresenta o problema e que pode ter a forma de uma frase incompleta ou de uma pergunta; (2) a opção correta ou chave de resposta; e (3) vários "distratores", que são alternativas incorretas, mas igualmente plausíveis para alunos que não dominem completamente as aprendizagens a serem testadas (Burton, Sudweeks, Merrill, & Wood, 1991; Bush, 2015; Clegg & Cashin, 1986). Na Figura 1 apresenta-se um exemplo de uma QEM, no formato tradicional, indicando os seus elementos fundamentais: o tronco e as alternativas, isto é, a resposta correta e as opções distratoras.

Tendo em consideração as várias tipologias de QEM, o tronco pode tomar três formas, a saber: pergunta; frase incompleta; solicitar a melhor resposta. A resposta correta deve sê-lo

[2] Do inglês "Stem".

inquestionavelmente, enquanto que os distratores devem ser plausíveis para aqueles que ainda não têm o conhecimento necessário, mas devem constituir inquestionavelmente opções incorretas para aqueles que já o possuem, tornando-se na parte da questão mais difícil de elaborar (Haladyna, 2004).

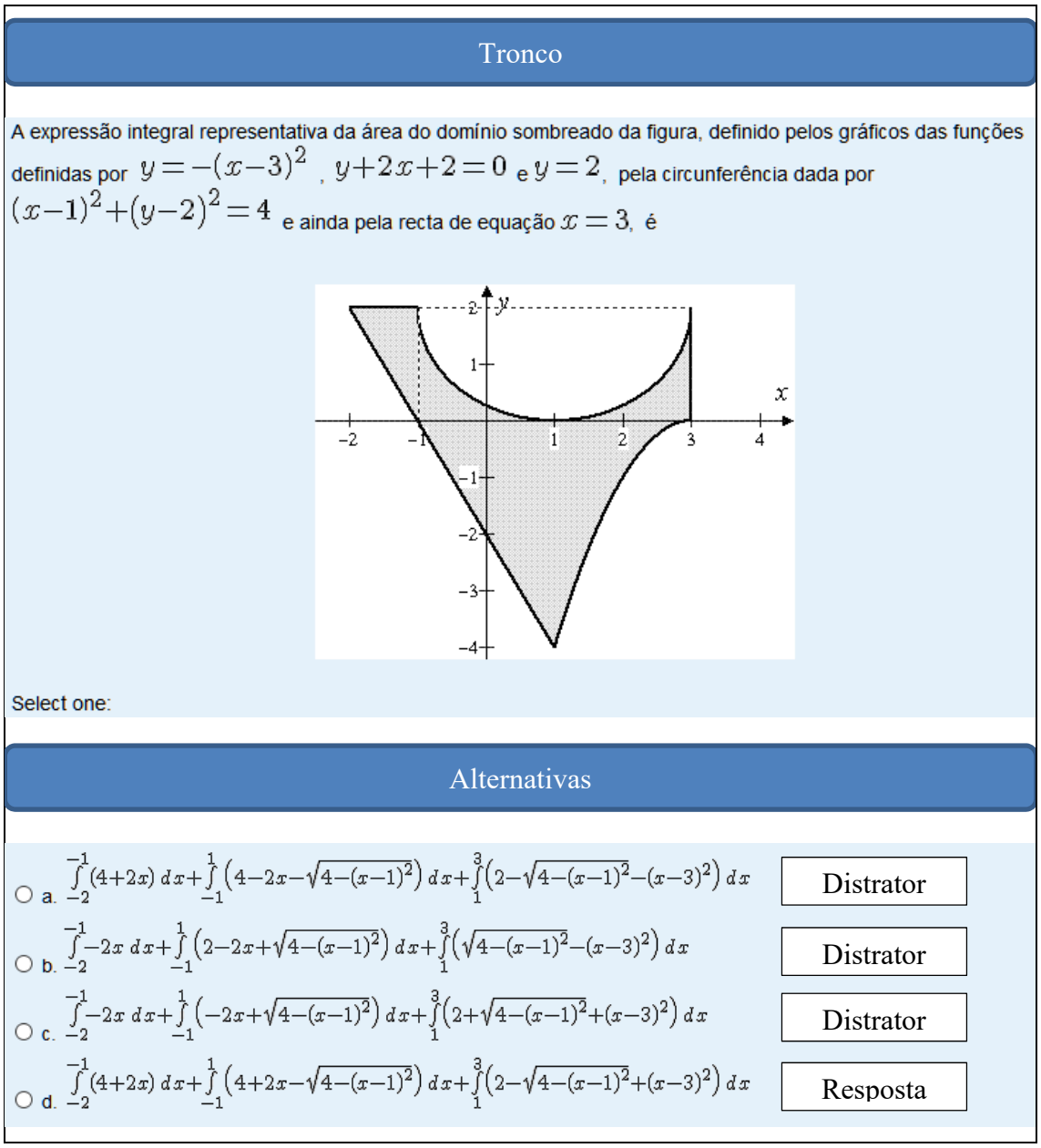


Figura 1: Exemplo de uma QEM.

Os estudos que pretendem comparar a utilização das QEM com as questões que implicam a construção da resposta por parte dos alunos - e que daqui para a frente serão referidas como "questões de resposta aberta"-, não são conclusivos, havendo algum ceticismo quanto à utilização das QEM na avaliação (Bull & Danson, 2001; Haladyna, 2004; Haladyna et al., 2002;

Jordan, 2013; Liu et al., 2011; Rod et al., 2010; Scouller, 1998). A investigação neste domínio foca-se nos resultados obtidos e na sua influência nas estratégias de aprendizagem dos alunos. De seguida apresentamos, por ordem cronológica, alguns desses estudos que consideramos relevantes:

- Bible e colaboradores (2008) apresentaram um estudo feito ao longo de quatro semestres a alunos de Contabilidade no qual, através da utilização de regressão linear múltipla, pretenderam aferir a influência das QEM nas questões de resposta aberta. Verificou-se que os resultados obtidos nas QEM explicavam cerca de dois terços da variabilidade dos resultados das questões de resposta aberta, concluindo os autores que o desempenho dos alunos nas QEM e nas questões de resposta aberta são suficientemente "próximas", permitindo confiar moderadamente nos testes com QEM.
- Ferrão (2010) comparou os resultados obtidos em testes com QEM com os resultados obtidos pelos mesmos alunos em testes de resposta aberta, numa UC de Estatística, concluindo que há uma forte correlação entre ambos, podendo assim os testes com QEM ser utilizados como alternativa aos testes com questões de resposta aberta.
- Lee e colaboradores (2011) estudaram a influência que as respostas obtidas em QEM podem ter em questões de resposta aberta que abordam os mesmos conteúdos, na avaliação da integração do conhecimento em alunos de Ciências. Através da análise das respostas obtidas em 106 QEM e 84 questões de resposta aberta, os autores concluíram que as questões de reposta aberta são melhores que as QEM e que estas não explicam os resultados obtidos nas questões de resposta aberta.
- Mora e colaboradores (2012) realizaram um estudo no qual os mesmos alunos faziam teste de escolha-múltipla e testes tradicionais em papel, abordando os mesmos tópicos. Apesar de reconhecerem a existência de algumas limitações no seu estudo, verificaram que não existem diferenças significativas nos resultados globais obtidos com os dois tipos de testes.
- Heron e Lerpiniere (2013) apresentaram um estudo feito com alunos de um curso de Serviço Social, no qual utilizaram uma metodologia de ensino destinada a fomentar uma estratégia de aprendizagem aprofundada. Concluíram que não houve diferenças na abordagem feita pelos alunos na sua aprendizagem utilizando exames com QEM ou utilizando exames nos quais era utilizada a redação de texto, isto é, a utilização das QEM não reduziu o nível de profundidade na abordagem à aprendizagem feita pelos alunos.

Independentemente de todas estas discrepâncias, as QEM têm vindo a ser cada vez mais utilizadas na avaliação em todos os níveis de ensino e, em particular, no Ensino Superior. Apresenta-se de seguida um resumo das principais vantagens e limitações das QEM.

2.1. Vantagens e limitações das Questões de Escolha Múltipla

Numa das secções anteriores, apresentámos as principais vantagens e limitações do *e-assessment*, sendo algumas delas também atribuídas às QEM. No entanto, há algumas vantagens e limitações que são específicas das QEM e/ou se manifestam nelas com maior intensidade. Conhecer as vantagens e as limitações das QEM ajuda o professor a tomar melhores decisões sobre as situações em que elas devem ou não devem ser utilizadas (Clegg & Cashin, 1986). Na Tabela 4, encontra-se um resumo das vantagens das QEM que se encontram na literatura. Algumas das vantagens focam-se no professor, outras focam-se mais no aluno. Estas últimas encontram-se a sombreado na Tabela 4. Adicionalmente, algumas vantagens, assinaladas a negrito, referem-se especificamente à Matemática.

Tabela 4: Vantagens das Questões de Escolha Múltipla

Vantagens
• Poupança de tempo (por exemplo, na obtenção das classificações) e de recursos (Bible et al., 2008; Brown, 2001; Burton et al., 1991; Camilo & Silva, 2008; Clegg & Cashin, 1986; Douglas et al., 2012; Ferrão, 2010; Green & Mitchell, 2009; Jordan, 2013; Liu et al., 2011; Nicol, 2007; Wild et al., 1997)
• Facilidade na avaliação de um elevado número de alunos, em testes de grande escala (Bible et al., 2008; Brown, 2001; Camilo & Silva, 2008; Clegg & Cashin, 1986; Green & Mitchell, 2009; Haladyna et al., 2002; Heron & Lerpiniere, 2013; Jordan, 2013; Liu et al., 2011; Nicol, 2007; Wild et al., 1997)
• Facilidade no cálculo de análises estatísticas e dos resultados dos testes (Bible et al., 2008; Brown, 2001; Burton et al., 1991; Camilo & Silva, 2008; Douglas et al., 2012; Green & Mitchell, 2009; Guo et al., 2014; Haladyna, 2004)
• Permite obter uma maior e mais rápida abrangência relativamente aos conteúdos do curso, o que permite avaliar um conjunto mais vasto de tópicos e de conhecimentos (Bible et al., 2008; Brown, 2001; Burton et al., 1991; Camilo & Silva, 2008; Clegg & Cashin, 1986; Ferrão, 2010; Green & Mitchell, 2009; Harris et al., 2015; Jordan, 2013; Wild et al., 1997)
• Compatibilidade entre cursos baseados na Web (Bible et al., 2008)
• Maior objetividade e fiabilidade nas classificações (Bible et al., 2008; Brown, 2001; Burton et al., 1991; Camilo & Silva, 2008; Douglas et al., 2012; Ferrão, 2010; Green & Mitchell, 2009; Haladyna, 2004; Jordan, 2013; Wild et al., 1997)
• Existência de bancos de questões para futura utilização (Ferrão, 2010; Guo et al., 2014)
• Facilidade de implementação através de computadores (Ferrão, 2010)
• Mais fáceis de gerir (Brown, 2001; Douglas et al., 2012; Haladyna, 2004; Liu et al., 2011)
• É um método estandardizado (Brown, 2001)
• Grande variedade de formatos (Brown, 2001)
• Existência de equilíbrio entre validade e fiabilidade com a facilidade logística (Harris et al., 2015)

Vantagens
• **Tem potencial para medir a compreensão, a análise, a capacidade de resolução de problemas e a capacidade de cálculo** (Brown, 2001; Burton et al., 1991; Clegg & Cashin, 1986; Kim, Patel, Uchizono, & Beck, 2012; Nicol, 2007)
• **Evita a introdução de notação simbólica por parte dos alunos, no caso específico da Matemática** (Jordan, 2013)
• Maior confiança na obtenção da resposta correta a partir de processos de eliminação das respostas erradas (Bible et al., 2008)
• Permite avaliar os conhecimentos dos alunos *per se* e não as suas capacidades de escrita (Bible et al., 2008; Green & Mitchell, 2009)
• Perceção de que os testes com QEM são mais objetivos e fiáveis (Bible et al., 2008; Brown, 2001; Guo et al., 2014; Liu et al., 2011)
• Úteis para autoavaliação e revisão (Brown, 2001; Clegg & Cashin, 1986; Green & Mitchell, 2009; Nicol, 2007)
• O *feedback* é rápido e impessoal (sem noção de julgamento) (Brown, 2001; Camilo & Silva, 2008; Douglas et al., 2012; Green & Mitchell, 2009; Guo et al., 2014; Jordan, 2013; Nicol, 2007)
• Maior motivação e envolvimento por parte dos alunos (Green & Mitchell, 2009; Jordan, 2013)

Na Tabela 5 encontra-se um resumo de limitações das QEM que se encontram na literatura. Tal como na tabela anterior, as limitações centradas no aluno encontram-se a sombreado e as que se referem especificamente à Matemática estão assinaladas a negrito.

Tabela 5: Limitações das Questões de Escolha Múltipla

Limitações
• Podem não avaliar os mesmos níveis de entendimento que são avaliados pelas questões de resposta aberta (Bible et al., 2008; Burton et al., 1991; Ferrão, 2010; Guo et al., 2014; Jordan, 2013; Lee et al., 2011)
• Possível ambiguidade nas próprias questões (Bible et al., 2008; Clegg & Cashin, 1986)
• Incapacidade para medir de forma adequada determinadas capacidades de níveis cognitivos mais elevados (Bible et al., 2008; Ferrão, 2010; Green & Mitchell, 2009; Lee et al., 2011; Liu et al., 2011; Nicol, 2007; Rod et al., 2010)
• O desenvolvimento de questões devidamente estruturadas é bastante moroso e exige muita formação (Burton et al., 1991; Clegg & Cashin, 1986; Ferrão, 2010; Guo et al., 2014; Jordan, 2013; Liu et al., 2011)
• Podem favorecer a memorização superficial dos conceitos (Heron & Lerpiniere, 2013; Liu et al., 2011; Nicol, 2007)
• Perigo de testar apenas conhecimento trivial (Brown, 2001; Douglas et al., 2012; Green & Mitchell, 2009)

Limitações
• Os alunos podem tentar acertar na resposta de forma aleatória (Burton et al., 1991; Bush, 2015; Clegg & Cashin, 1986; Douglas et al., 2012; Haladyna et al., 2002; Heron & Lerpiniere, 2013; Jordan, 2013; Lee et al., 2011; Liu et al., 2011; Wild et al., 1997)
• **Os alunos podem inverter a resolução e não se estará a avaliar aquilo que é suposto**[3] (Jordan, 2013)
• **Em questões com cálculos o aluno pode chegar a uma solução que não existe nas opções concluindo logo que a sua resposta está incorreta** (Jordan, 2013)
• Não permite que os alunos expliquem as suas respostas, pelo que são limitativas (Liu et al., 2011; Wild et al., 1997)
• Podem penalizar alunos que não têm tendência para tomar riscos (Ávila & Torrubia, 2004; Brown, 2001; Douglas et al., 2012; Jordan, 2013; Triantis & Ventouras, 2012)
• A personalização do *feedback* é limitada (Douglas et al., 2012; Nicol, 2007)

2.2. Formatos das questões de escolha múltipla

Com a utilização de QEM, a tentativa de o aluno tentar acertar na resposta correta, em vez de conduzir um processo de resolução que leve à resposta correta, está sempre presente (Haladyna, 2004). Refere Haladyna (2004) que o aluno ou: i) sabe a resposta correta; ii) tem conhecimento parcial que lhe permite eliminar distratores não plausíveis; iii) tenta simplesmente adivinhar a resposta de forma aleatória na ausência de qualquer conhecimento.

Qual a probabilidade de um aluno poder acertar na resposta correta, sem qualquer tipo de penalização numa resposta errada, a uma QEM na ausência de qualquer conhecimento? Por exemplo, a probabilidade de um aluno acertar na resposta correta a uma QEM de forma aleatória, considerando o seu formato com 4 alternativas, é de 25%. É um valor bastante elevado, mas num teste com duas QEM a probabilidade de acertar em ambas é 6.25%, num teste com três QEM a probabilidade de acertar nas três é aproximadamente igual a 1.56%, num teste com quatro QEM a probabilidade de acertar nas quatro é aproximadamente igual a 0.39%, isto é, aumentando o número de QEM num teste diminui-se significativamente a probabilidade de o aluno acertar em todas as questões, aproximando-se esta probabilidade de zero. Já a probabilidade de um aluno acertar de forma aleatória em pelo menos cinco QEM num teste de 10, isto é, a probabilidade de tirar positiva, é aproximadamente igual a 8%, mas se forem 20 questões a probabilidade desce para cerca de 1%, enquanto a probabilidade de acertar de forma aleatória em pelo menos 14 QEM num teste de 20, isto é, obter 14 ou mais valores (supondo 1 valor para cada questão), é aproximadamente igual a 0.003%. São probabilidades que podemos considerar bastante pequenas, mas que ainda assim serão de considerar. É então necessário

[3] Por exemplo, numa questão para integrar, o aluno pode diferenciar cada uma das opções apresentadas.

utilizar estratégias que levem os alunos a desistir da tentativa de acertar na resposta correta na ausência de qualquer conhecimento.

Uma das estratégias para minimizar este problema é a atribuição de penalizações quando são selecionados distratores. Verifica-se que este procedimento reduz a probabilidade de um aluno obter, por exemplo, uma classificação positiva. No entanto, alguns autores afirmam que a atribuição de uma cotação negativa às opções distratoras pode prejudicar os alunos que têm menos tendência para arriscar, favorecendo, por exemplo, os indivíduos do género masculino (Ávila & Torrubia, 2004; Brown, 2001; Douglas et al., 2012; Jordan, 2013; Triantis & Ventouras, 2012).

Triantis e Ventouras (2012) apresentam uma abordagem interessante para minimizar a tentativa de os alunos tentarem adivinhar a resposta na ausência de qualquer conhecimento, para além da atribuição de penalizações a respostas erradas. Eles conceberam um sistema de verificação dupla: i) todas as QEM do teste são colocadas aos pares, isto é, em cada teste são colocadas duas questões abordando os mesmos tópicos sem que isso possa ser percebido pelos alunos; ii) é atribuído um bónus aos alunos que acertam ambas as questões do par; iii) é atribuída uma penalização se uma das questões do par está errada e a outra está certa. De qualquer das formas, de acordo com Haladyna (2004), a tentativa de os alunos tentarem adivinhar a resposta correta não terá grande influência na classificação final, se forem incluídas questões em número suficiente, portanto, testes mais longos apresentarão menores problemas a este nível.

Uma outra forma de controlar melhor este problema consiste na utilização de formatos diversificados para as questões. Existem vários formatos possíveis para as QEM. No caso de uma questão tradicional com 4 opções de resposta, qualquer opção selecionada de forma aleatória tem três vezes maior probabilidade de estar incorreta do que correta. Assim sendo, atribuir a cotação +3 à opção correta e a cotação -1 a cada uma das opções distratoras resulta num esquema de atribuição de cotações que é neutro para aqueles que tentam acertar na resposta correta de forma aleatória (Bush, 2015). Desta forma, quem faz o teste não tem nada a ganhar ou a perder, em média, isto é, o valor esperado é zero. O esquema de cotações deverá ser sempre desenhado de modo a obter um esquema neutro.

Bush (2015) distingue “adivinhar de forma aleatória” de “adivinhar de forma informada”. Adivinhar de forma aleatória ocorre quando aquele que responde tem a mesma confiança em qualquer uma das opções e, ainda assim, escolhe uma delas como resposta. Adivinhar de forma informada ocorre quando aquele que responde tem confiança diferente em alguma(s) das opções e não é capaz de expressar evidentemente qual é a sua convicção. Em vez disso, tem de escolher arbitrariamente entre duas ou mais opções, para as quais tem uma confiança igual ou aproximadamente igual. A partir destes dois conceitos (adivinhar de forma aleatória vs. adivinhar de forma informada), Bush (2015) apresenta oito formatos diferentes para as questões de escolha múltipla, os quais descrevemos, de forma abreviada, na Tabela 6. Os três formatos

de seleção repetida adaptam-se especialmente bem a uma utilização no contexto do *e-assessment*, enquanto os restantes formatos podem ser utilizados em ambos os contextos, *e-assessment* e avaliação em suporte papel.

Tabela 6: Classificação de formatos de QEM de Bush (2015)

Formato	Descrição/Exemplo[4]
Tradicional	"Selecione a opção que considera ser a mais correta. +3 valores serão atribuídos a uma resposta correta, -1 valor a uma resposta incorreta." *Nota:* Podem ser introduzidas variações a este formato. Por exemplo, solicitar ao aluno que atribua um nível de confiança (baixo, médio ou alto) à resposta por ele dada. A cotação atribuída, quer seja positiva, quer seja negativa, refletirá também este nível de confiança[5].
Seleção de um subconjunto	"Selecione a(s) opção(ões) que considera ser(em) a(s) mais correta(s); pode selecionar até três opções. +3 valores serão atribuídos a uma resposta correta, -1 valor a uma resposta incorreta."
Seleção de um distrator	"Selecione a(s) opção(ões) que considera corresponder(em) a uma (ou mais) resposta(s) errada(s); pode selecionar até três opções. +1 valor será atribuído a cada resposta errada corretamente identificada, -3 valores a uma resposta incorreta." *Nota:* Podemos ver este formato como o inverso do formato seleção de um subconjunto: aqui, devem selecionar-se as opções distratoras, no formato anterior, devem selecionar-se as opções que se acredita serem mais plausíveis.
Ordenação estrita	"Ordene as opções de acordo com a probabilidade que atribui a cada uma delas de estar correta, onde "1" indica a mais provável e "4" indica a menos provável. Serão atribuídos os valores +3, +1, -1 ou -3 dependendo da posição 1ª, 2ª, 3ª ou 4ª (respetivamente) em que colocar a opção correta." *Nota:* +3, +2, +1 ou -6, poderia ser uma cotação alternativa, a qual recompensaria o conhecimento parcial de forma mais generosa.
Seleção repetida	"Em primeiro lugar, selecione a opção que considera ser a mais correta, ou pode optar por não responder. Se a sua primeira seleção estiver incorreta, pode fazer uma segunda seleção ou pode optar por não responder. Se a sua segunda seleção estiver incorreta, pode selecionar uma opção final entre as duas restantes, ou pode optar por não responder." *Nota:* Utilizando-se *e-assessment*, este formato pode tornar-se mais amigável, separando a apresentação da questão em vários passos. Em primeiro lugar: "Selecione a opção que considera mais correta ou não responda." Depois, duas ou três vezes conforme necessário, poderia surgir algo semelhante a: "Incorreto. Selecione outra opção ou não responda."
Seleção repetida de um distrator	"Em primeiro lugar, selecione uma opção que considera estar incorreta, ou pode optar por não responder. Se a sua primeira seleção estiver incorreta, pode fazer uma segunda seleção ou pode optar por não responder. Se a sua segunda seleção estiver incorreta, pode selecionar uma opção final entre as duas restantes, ou pode optar por não responder."

[4] Considerando questões com quatro opções.

[5] Curtis e colaboradores (2013) apresentam um caso interessante de aplicação deste formato de questões na área da Medicina. Distinguem dois tipos de alunos: mal-informados – aqueles que têm uma resposta incorreta, mas têm um nível elevado de confiança que a resposta está correta; e não informados – aqueles que têm uma resposta incorreta mas têm um nível baixo de confiança que a resposta está correta. Consideram que estes dois tipos de alunos exigem diferentes estratégias remediativas de intervenção.

Formato	Descrição/Exemplo[4]
	Nota 1: Utilizando-se *e-assessment* também este formato pode tornar-se mais amigável, separando a apresentação da questão em vários passos. Na resposta à questão, pode-se selecionar um distrator a cada passo, até que já não haja mais distratores, ou até se ter identificado a opção correta como distratora. *Nota 2:* De certa forma, este formato é o inverso do formato seleção repetida, visto que aqui se começa por selecionar a opção que é a menos provável que esteja correta, em vez de ser a mais provável.
Ordenação parcial	"Ordene as opções de acordo com a probabilidade que atribui a cada uma delas de estar correta, onde 1 indica a mais provável e 4 indica a menos provável. Pode atribuir a mesma posição a qualquer das opções, de modo que a ordenação pode ser qualquer uma das seguintes: (1-2-3-4), (1-1-3-4), (1-2-2-4), (1-1-3-3), etc.. Serão atribuídos um dos valores de +3 a -3, dependendo da ordenação feita e da posição em que estiver a resposta correta." *Nota:* A cotação é baseada no formato "Ordenação estrita". Considerando os exemplos de resposta dados, a primeira resposta obteria a cotação de +3, a segunda resposta +2 (média entre +3 e +1), etc. Para uma resposta (1-1-1-1) , a cotação seria 0.
Seleção repetida de um subconjunto	Inicia-se de forma idêntica à do formato seleção de um subconjunto, isto é: "Selecione a(s) opção(ões) que considera ser(em) a(s) mais correta(s); pode selecionar até três opções. +3 valores serão atribuídos a uma resposta correta, -1 valor a uma resposta incorreta." Caso o primeiro conjunto selecionado não inclua a opção correta, poderá ser dada uma segunda oportunidade e, eventualmente, uma terceira.

A atribuição de uma cotação negativa pode desencorajar a tentativa de adivinhar de forma aleatória, mas é esperado que as QEM com formato tradicional incentivem a tentativa de adivinhar de forma informada. O formato "Seleção de um distrator" pode tornar-se mais efetivo que o formato "Seleção de um subconjunto", uma vez que reduz a probabilidade de se tentar adivinhar de forma aleatória. Para o formato "Ordenação estrita", aqueles que respondem à questão poderão ser levados a empreender uma resposta de forma informada. Já os formatos "Ordenação parcial" e "Seleção repetida de um subconjunto" eliminam de forma definitiva qualquer necessidade de adivinhar a resposta, podendo assim ser considerados os mais adequados para utilização em avaliação sumativa. Os formatos "Seleção repetida" e "Seleção repetida de um distrator" poderão ser considerados os formatos mais adequados para utilização em avaliação formativa, devido ao seu *feedback* permanente (Bush, 2015).

Na sequência das diferentes propostas de Bush (2015) para o formato de QEM, e supondo a existência de quatro opções por questão, o número de respostas possível para cada formato é variável. Assim, para o formato "Tradicional" existem 5 respostas possíveis[6]. Para cada um dos formatos "Seleção de um subconjunto" e "Seleção de um distrator", o número de respostas possíveis é de 16 e para o formato "Ordenação estrita" é de 25. Já para os restantes formatos, Bush afirma não ser possível determinar o número de respostas possíveis. Desta forma, verifica-se que a probabilidade de os alunos adivinharem a resposta correta de forma aleatória é menor

[6] Não responder é uma reposta possível.

para formatos mais complexos. No entanto, estes formatos são mais difíceis de perceber por parte daqueles que respondem às questões, e são também mais difíceis de implementar, principalmente em testes que sejam feitos em formato papel. Mesmo no caso do *e-assessment*, os sistemas informáticos que tradicionalmente são utilizados para o implementar, fornecem um número limitado de formatos. Por exemplo, a maioria dos LMS, que são os sistemas com mais forte implantação no Ensino Superior, apresentam apenas o formato tradicional para as QEM. É reconhecido que este continua a ser o formato mais utilizado para *e-assessment*.

Na literatura, podem encontrar-se outros tipos de classificações para os formatos das QEM, mas não com o nível de sofisticação das até aqui apresentadas. Por exemplo, Haladyna e colaboradores (2002, 2004) através da análise de uma vasta gama de trabalhos científicos, identificaram para as QEM seis formatos diferentes, bem como algumas variações possíveis a esses formatos, apresentando vantagens e desvantagens de cada um deles. Estes formatos são apresentados na Tabela 7. Podemos dizer que nesta classificação, os formatos apresentados são variações das QEM de formato "Tradicional", não se tratando efetivamente de novos formatos no sentido daqueles que são apresentados por Bush (2015). Também Burton e colaboradores (1991) elencam uma variedade de formatos para as QEM semelhantes aos apresentados por Haladyna e colaboradores (2002, 2004).

Tabela 7: Classificação de formatos de QEM de Haladyna e colaboradores (2002, 2004)

Formato	Descrição/Exemplo
Escolha Múltipla Convencional	Um tronco, seguido de várias opções, sendo uma delas correta e as restantes incorretas.
Escolha-alternada	Um tronco, seguido de duas opções, oferecendo-se uma comparação entre duas alternativas possíveis.
Verdadeiro-Falso	Uma proposição que é avaliada pelo respondente como sendo falsa ou verdadeira.
Verdadeiro-Falso múltiplo	Um tronco com várias opções. Cada opção é avaliada pelo respondente como sendo falsa ou verdadeira.
Correspondência	Várias opções, seguidas por um grupo de troncos. A cada tronco deve fazer-se corresponder uma opção, podendo haver mais troncos do que opções.
Escolha Múltipla Complexa	Um tronco seguido de opções que estão reagrupadas em conjuntos para que os respondentes selecionem o conjunto correto.
Conjunto de questões dependentes do contexto	Um "estímulo" seguido por uma ou mais questões de escolha múltipla convencionais, que com ele se relacionam.

Por fim, referimos o trabalho de Liu e colaboradores (2011), as quais apresentam um outro formato de QEM que não está incluído em nenhuma das classificações anteriormente descritas. Este formato, que as autoras designam como "Escolha Múltipla com Explicação", consiste numa

QEM com duas partes: a primeira parte apresenta quatro opções, entre as quais o aluno deve escolher uma; a segunda parte apresenta seis possíveis explicações para a escolha que foi feita anteriormente, sendo que o aluno deve escolher uma como sendo a que melhor explica a sua opção anterior. Este formato tem a vantagem de permitir obter do aluno uma justificação e raciocínio, sem perder a objetividade das QEM tradicionais.

Independentemente das vantagens dos vários formatos de QEM existentes, não é recomendada a utilização de formatos complexos – ainda que sejam mais efetivos no que se reporta à minimização dos efeitos de respostas dadas de forma aleatória – dado que além de poderem confundir os alunos, a sua elaboração consome demasiado tempo por parte dos docentes (Brown, 2001).

2.3. Banco de questões

No nosso entender, uma das limitações da utilização das QEM, que não encontrámos explicitamente referida na literatura, tem a ver com a possibilidade de os alunos copiarem com mais facilidade neste caso, do que copiam quando as questões são de resposta aberta. Para obviar esta limitação, nos testes que utilizam QEM em formato papel é habitual elaborar várias versões do mesmo teste, introduzindo-lhe ligeiras alterações, mas de modo a manter a viabilidade e a fiabilidade da avaliação. No caso do *e-assessment,* um banco de questões devidamente concebido e implementado representa aqui um papel fundamental, podendo mesmo chegar-se ao limite de obter uma versão diferente para cada aluno, gerada de forma aleatória pelo sistema informático (Azevedo, 2015). Quando se pretende utilizar o *e-assessment,* em geral, e as QEM, em particular, na avaliação sumativa é primordial a construção de um banco de questões (Yorke, 2001). Os bancos de questões podem contribuir para assegurar a validade e a fiabilidade do processo de avaliação, poupando recursos, tempo e dinheiro (Bull & Danson, 2001; McAlpine, 2002b).

Bancos de questões são coleções de questões e podem ser vistos como repositórios especializados ou como bases de dados de questões, nos quais estas podem ser armazenadas de acordo com categorias de assuntos/temas, sendo cada uma delas identificada de forma única e armazenada de modo a permitir a criação automática ou manual de testes, em papel e/ou no ecrã, de forma aleatória se necessário, de modo a que sejam satisfeitos determinados critérios. Cada questão tem descritores associados que podem definir um certo número de características, tais como nível académico, tópico, dificuldade e competência ou conhecimento abordado por ela. Atualmente, quase todos os bancos de questões são eletrónicos (Bull & Danson, 2001; Green & Mitchell, 2009; JISC, 2006; McAlpine, 2002b).

Um aspeto importante no desenvolvimento de QEM a incluir num banco de questões é a garantia de que as questões elaboradas têm qualidade. Harris e colaboradores (2015) apresentam uma abordagem interessante e bem-sucedida, na qual os alunos escrevem as questões a incluir num banco de questões para avaliação formativa na área da Medicina. Trata-se de um processo de cinco passos, envolvendo alunos, docentes e especialistas, para a criação e avaliação das questões e sua inclusão no banco de questões. Em Azevedo (2015), apresentamos a implementação de um banco de questões para avaliação contínua sumativa contendo QEM. Para isso, foi utilizado um processo de revisão sistemático, o qual foi fundamental para a obtenção de QEM com qualidade e para o sucesso na implementação do banco de questões e da avaliação contínua sumativa. Haladyna (2004) defende que, para validação das questões de um banco de dados, é fundamental: i) seguir um conjunto de procedimentos durante o seu desenvolvimento, nomeadamente seguir um conjunto de linhas de orientação; ii) realizar a análise estatística das respostas dadas às questões. Neste seguimento, consideramos que para garantir a qualidade das questões do banco (de questões) são fundamentais três cuidados, que serão abordados com mais detalhe nas secções seguintes, a saber: i) seguir linhas de orientação aquando da escrita das questões (secção 2.4); ii) analisar os testes e as questões implementadas, utilizando técnicas adequadas, tais como a Teoria de Resposta ao Item ou a Teoria Clássica de Análise de Testes (capítulo 3); e iii) classificar as questões desenvolvidas de acordo com taxonomias adequadas (capítulo 4).

2.4. Linhas de orientação para a escrita de Questões de Escolha Múltipla

Diversas linhas orientadoras para a escrita de QEM podem ser encontradas na literatura, tendo em vista assegurar a sua qualidade. Por exemplo, Clegg e Cashin (1986) apresentam quatro aspetos que consideram ser fundamentais antes de começar a escrever questões de escolha múltipla: i) necessidade de um grande domínio dos conteúdos a serem testados, pois caso contrário pode não se estar alertado para as falácias e confusões mais comuns; ii) desenvolver e utilizar um conjunto de objetivos educacionais, bem como os níveis de aprendizagem que se desejam testar; iii) conhecer os alunos e adaptar a complexidade e dificuldade dos testes de acordo com as suas características; e iv) dominar a comunicação escrita, sendo capaz de comunicar com precisão e simplicidade e de utilizar linguagem que os alunos possam entender.

Clegg e Cashin (1986) apresentam uma lista de 34 recomendações para a construção de QEM; Burton e colaboradores (1991) apresentam um conjunto de 16 linhas de orientação, com exemplos bastante ilustrativos; Camilo e Silva (2008) dividem as orientações em dois pontos: regras que devem ser seguidas na escrita das questões e erros/falhas técnicas que se devem evitar. Já Haladyna e colaboradores (2002), a partir de uma análise sistemática exaustiva de

trabalhos científicos que abordam este tópico, apresentam um total de 31 linhas de orientação que são divididas em 5 grupos, os quais dizem respeito a cuidados a considerar nos seguintes aspetos: conteúdo, formatação, estilo, escrita do enunciado da questão e escrita das opções da questão. Estas linhas de orientação apresentam-se na Tabela 8.

Tabela 8: Linhas de orientação para a escrita de QEM (traduzido de Haladyna et al., 2002, p. 312)

Grupo	Linhas de Orientação
CUIDADOS COM O CONTEÚDO	1. Cada questão deve refletir conteúdo específico e um único comportamento mental concreto, tal como preconizado nas especificações dos testes. 2. Fundamentar cada questão em termos de conteúdos de aprendizagem importantes; evitar conteúdo trivial. 3. Utilizar materiais inovadores para testar aprendizagens de nível mais elevado. Reescrever a linguagem utilizada no livro de apoio ou a linguagem utilizada durante as aulas, quando incluídas nas questões de um teste, de modo a evitar testes apenas de memorização. 4. Manter o conteúdo de cada questão independente do conteúdo de outras questões do teste. 5. Evitar conteúdos demasiado específicos ou demasiado genéricos ao escrever as questões. 6. Evitar questões baseadas em opiniões. 7. Evitar questões com artimanhas. 8. Manter o vocabulário simples, tendo em conta o grupo de alunos que está a ser testado.
CUIDADOS COM A FORMATAÇÃO	9. Utilizar todos os formatos[7], exceto o formato "Escolha Múltipla complexa" que deve ser evitado. 10. Formatar a questão verticalmente e não horizontalmente.
CUIDADOS COM O ESTILO	11. Editar e rever as questões. 12. Usar corretamente a gramática, a pontuação, as letras maiúsculas e a ortografia. 13. Minimizar a quantidade de leitura necessária em cada questão.
CUIDADOS NA ESCRITA DO ENUNCIADO DA QUESTÃO	14. Certificar-se que as instruções no enunciado são muito claras. 15. Incluir a ideia central no enunciado ao invés de nas opções. 16. Evitar palavreado excessivo. 17. Escrever o enunciado na forma afirmativa, evitando negações tais como NÃO ou EXCETO. Se forem utilizadas negações, usar as palavras com cautela e garantir sempre que a palavra aparece em maiúsculas e em negrito.
CUIDADOS NA ESCRITA DAS OPÇÕES DA QUESTÃO	18. Desenvolver tantas opções eficazes quantas seja possível, mas a investigação sugere que três é adequado. 19. Certificar-se que apenas uma dessas opções é a resposta correta. 20. Variar a localização da resposta correta de acordo com o número de opções. 21. Colocar as opções por ordem, lógica ou numérica. 22. Garantir opções independentes; as opções não devem ter elementos comuns. 23. Garantir opções homogéneas, quer em termos de conteúdo quer em termos de estrutura gramatical. 24. Manter o tamanho das opções aproximadamente igual.

[7] "Todos os formatos" refere-se aos formatos apresentados na Tabela 7.

Grupo	Linhas de Orientação
	25. Utilizar cuidadosamente "Nenhum dos anteriores". 26. Evitar utilizar “Todos os anteriores”. 27. Escrever as opções na forma afirmativa; evitar negações tais como NÃO. 28. Evitar dar dicas para a resposta correta, tais como: • Determinantes específicos incluindo sempre, nunca, completamente e absolutamente; • Associações de palavras com sons idênticos, escolhas idênticas ou parecidas com termos utilizados no enunciado; • Incoerências gramaticais que deem pistas ao aluno sobre a resposta correta. • Resposta correta evidente; • Pares ou tripletos de opções que irão indicar ao aluno a resposta correta; • Opções ostensivamente absurdas ou ridículas. 29. Garantir que todos os distratores são plausíveis. 30. Usar erros típicos dos alunos para escrever os distratores. 31. Utilizar humor, se ele é compatível com o professor e com o ambiente de aprendizagem.

Não é dada a mesma importância a todas as linhas de orientação, sendo que algumas delas se apresentam como mais relevantes na revisão feita por Haladyna e colaboradores (2002).

As mais relevantes são:

- “Incluir a ideia central no enunciado ao invés de nas opções”, que é favoravelmente indicada em 100% das fontes utilizadas no estudo;
- “Evitar dar dicas para a resposta correta” e “Garantir que todos os distratores são plausíveis”, que são favoravelmente indicadas em 96% das fontes utilizadas no estudo;
- “Utilizar materiais inovadores para testar aprendizagens de nível mais elevado. Reescrever a linguagem utilizada no livro de apoio ou a linguagem utilizada durante as aulas, quando incluídas nas questões de um teste, de modo a evitar testes apenas de memorização” e “Manter o tamanho das opções aproximadamente igual”, que são favoravelmente indicadas em 85% das fontes utilizadas no estudo;
- “Certificar-se que as instruções no enunciado são muito claras”, que são favoravelmente indicadas em 82% das fontes utilizadas no estudo.

As menos relevantes são:

- “Utilizar humor, se ele é compatível com o professor e com o ambiente de aprendizagem”, que apenas é citada em 15% das referências utilizadas no estudo e, ainda assim, de forma desfavorável;
- “Evitar conteúdos demasiado específicos ou demasiado genéricos ao escrever as questões”, que apenas é citada em 15% das referências utilizadas no estudo;

- "Garantir opções independentes; as opções não devem ter elementos comuns", que apenas é citada em 30% das referências utilizadas no estudo.

Há ainda algumas linhas que podem ser consideradas como gerando alguma controvérsia, visto haver fontes que são desfavoráveis a estas linhas de orientação, ao contrário das anteriores:

- "Utilizar cuidadosamente "Nenhum dos anteriores"", que é citada favoravelmente em 44% das referências utilizadas no estudo e desfavoravelmente em 48% das mesmas;
- "Formatar a questão verticalmente e não horizontalmente", que é citada favoravelmente em 37% das referências utilizadas no estudo e desfavoravelmente em 11% das mesmas;
- "Escrever o enunciado na forma afirmativa, evitando negações tais como NÃO ou EXCETO. Se forem utilizadas negações, usar as palavras com cautela e garantir sempre que a palavra aparece em maiúsculas e em negrito", que é citada favoravelmente em 63% das referências utilizadas no estudo e desfavoravelmente em 18% das mesmas;
- "Evitar utilizar "Todos os anteriores"", que é citada favoravelmente em 70% das referências utilizadas no estudo e desfavoravelmente em 22% das mesmas;
- "Desenvolver tantas opções eficazes quantas seja possível, mas a investigação sugere que três é adequado", que é citada favoravelmente em 70% das referências utilizadas no estudo e desfavoravelmente em 4% das mesmas.

Haladyna e colaboradores (2002) salientam que este não é um trabalho encerrado e que as linhas de orientação para elaboração de QEM evoluem constantemente. Como consequência, novos trabalhos de investigação devem ser levados a cabo para se conseguir um maior entendimento sobre cada uma das linhas de orientação por eles apresentadas. Neste seguimento, Haladyna (2004) apresenta uma versão atualizada da proposta inicialmente formulada, que contém 4 grupos e 26 linhas de orientação.

CAPÍTULO 3. ANÁLISE DE TESTES E DE QUESTÕES

Nas instituições de Ensino Superior, uma das principais funções dos testes é medir aquilo que os alunos conseguiram alcançar. Um teste é, então, "um instrumento de medida com o qual se pretende descrever numericamente o grau ou quantidade de aprendizagem sob condições uniformes, padronizadas" (traduzido de Haladyna, 2004, p. 4). Assim sendo, é importante avaliar a sua qualidade, de modo a saber até que ponto podemos confiar neles para realizarem essa medição. Os testes contêm uma ou mais questões, sendo que cada questão pode ser vista como "a unidade básica de observação de qualquer teste" (traduzido de Haladyna, 2004, p. 3). A análise das questões incluídas nos testes é uma forma de avaliar a sua qualidade, olhando para as suas partes constituintes. Pode também ser vista como uma forma de obter evidências da validade das questões (Haladyna, 2004).

Há duas grandes teorias no que diz respeito à análise de questões em testes de avaliação: a Teoria Clássica de Testes (TCT) e a Teoria de Resposta ao Item (TRI). Para a TCT a unidade de análise é o teste, enquanto para a TRI a unidade de análise é a questão (item) (Baker, 2001; Hambleton, Swaminathan, & Rogers, 1991; McAlpine, 2002b). Estas formas de análise visam principalmente garantir a qualidade da avaliação assegurando que as questões têm um nível de dificuldade apropriado e que discriminam de forma adequada os alunos que estão a ser avaliados, distinguindo entre os melhores alunos e os piores alunos (McAlpine, 2002c). Ambas as teorias rivalizam sobre qual delas é preferível à outra, sendo este aspeto ainda um fator de constante debate entre os defensores de uma e de outra teoria (Haladyna, 2004). Assim sendo, vamos neste capítulo, começar por referir os aspetos fundamentais da TCT e da TRI, terminando com algumas considerações adicionais sobre a análise de testes e questões.

Dado que este trabalho aborda a utilização de QEM, iremos apenas descrever modelos que foram desenvolvidos para dados dicotómicos, isto é, com apenas dois valores possíveis. É este o caso das QEM, nas quais os dados obtidos com as respostas dos examinandos podem apenas tomar dois valores, a saber, resposta correta ou resposta incorreta, que poderão ser representadas por 1 e 0, respetivamente.

3.1. Teoria clássica dos testes (TCT)

A TCT remonta ao início do século XX e teve origem na Psicologia, sendo muito usada em Inglaterra (McAlpine, 2002a). A TCT concentra-se em duas grandes áreas, a saber, a dificuldade das questões e a discriminação das questões. A dificuldade de uma questão relaciona-se com a quantidade de sujeitos que conseguem responder acertadamente a essa questão. De uma forma

simplificada, podemos dizer que quanto mais difícil for a questão, menor será a proporção de indivíduos que respondem corretamente a essa questão. O principal indicador da dificuldade de uma questão é o Índice de Dificuldade[8]. A discriminação de uma questão tem a ver com a capacidade que uma questão tem para distinguir os sujeitos "melhores" dos "piores"[9]. Em termos simples, podemos dizer que quanto mais alta for a discriminação de uma questão, maior será o número de indivíduos do grupo dos melhores que responderá a essa questão de forma acertada, e menor será o número dos indivíduos do grupo dos piores que responderá acertadamente a essa questão. O principal indicador de discriminação de uma questão é o Índice de Discriminação.

3.1.1. Índice de Dificuldade

O Índice de Dificuldade de uma questão, que podemos representar como *IDif*, é habitualmente apresentado como a proporção de sujeitos que acertam nessa questão (Equação 1).

$$IDif = \frac{n.º\ de\ acertos}{n.º\ de\ respostas}$$

(Equação 1: Índice de Dificuldade)

De acordo com vários autores (Camilo & Silva, 2008; McAlpine, 2002a, 2002b, 2002c), é aconselhável que: o Índice de Dificuldade seja próximo de 0.5[10]; num teste contendo várias questões, o Índice de Dificuldade seja variável; situando-se entre 0.15 e 0.85 e que fora desse intervalo as questões sejam rejeitadas, a não ser em condições especiais.

Uma das limitações do Índice de Dificuldade prende-se com o facto de ser dependente da amostra, isto é, as mesmas questões poderão ter índices de dificuldade diferentes no caso de sujeitos diferentes responderem a essas mesmas questões (Haladyna, 2004; McAlpine, 2002b).

3.1.2. Índice de Discriminação

O Índice de Discriminação de uma questão, que podemos representar como *IDisc*, pretende medir até que ponto a questão distingue os sujeitos: um aluno com melhor aprendizagem tende a responder corretamente e um aluno com pior aprendizagem tende a responder de forma errada (Haladyna, 2004). Há vários métodos para determinar o Índice de Discriminação. Um dos mais comuns é o coeficiente de correlação de Pearson entre as classificações obtidas em cada questão e a classificação total obtida no teste (Equação 2). Assume-se unidimensionalidade, isto é, todas as questões medem uma determinada área de conteúdos ou competências.

[8] Alguns autores utilizam índice de facilidade, dado que na verdade estamos a medir o grau de facilidade: quanto maior o índice de facilidade, mais fácil será a questão, visto que há uma maior proporção de acertos.
[9] Os indivíduos "melhores" são aqueles que têm melhor nota no teste e os "piores" são os que têm pior nota no teste no qual a questão está incluída.
[10] Utilizamos nesta tese o ponto (.) como separador decimal. É este o procedimento normal nas nossas UC e, além disso, facilita a utilização das várias aplicações informáticas utlizadas nesta tese.

$$IDisc = \frac{Cov(X,Y)}{\sqrt{Var(X)}.\sqrt{Var(Y)}}$$

onde

X é a variável das classificações obtidas nas respostas à questão;

Y é a variável das classificações totais obtidas no teste.

(Equação 2: Índice de Discriminação)

Quanto à discriminação, pode-se afirmar que (Camilo & Silva, 2008; Lee et al., 2011; McAlpine, 2002a, 2002b, 2002c):

- dado tratar-se de um coeficiente de correlação, varia entre -1 e 1, sendo que 1 significa uma correlação perfeita entre as classificações obtidas nessa questão e as classificações obtidas no teste, isto é, quanto mais alta a classificação na questão, mais alta será a classificação no teste, e -1 significa uma correlação perfeita inversa entre as classificações nessa questão e as classificações no teste, isto é, quanto mais alta a classificação nessa questão, mais baixa a classificação no teste;
- no geral deve ser positiva (a não ser que não haja unidimensionalidade), dado que não se espera que os indivíduos com melhor desempenho tenham menor probabilidade de acertar numa questão do que os que têm pior desempenho;
- um bom poder de discriminação significa que ID >0.4;
- um baixo poder de discriminação significa que ID<0.2; no entanto, um Índice de Discriminação baixo pode significar apenas que a pergunta testa conhecimentos básicos e nesse caso a questão deverá/poderá ser mantida no banco de questões; uma questão com um Índice de Discriminação zero, não discrimina;
- as questões com níveis de dificuldade extremos têm mais tendência para discriminações baixas.

3.2. Teoria da resposta ao item (TRI)

A TRI foi originalmente desenvolvida na década de 1940, tendo obtido grande sucesso nas décadas de 1960 e 1970 nos EUA e, desde então, tem vindo a passar por grandes desenvolvimentos. Atualmente é muito utilizada por associações americanas de certificação, nomeadamente a "Educational Testing Service" (ETS) (McAlpine, 2002a).

Há várias situações, entre as quais se encontra a avaliação, em que se pretende medir uma variável de interesse, por exemplo a competência para resolver equações, que se pode descrever mas que não se pode medir diretamente, como se mede a altura ou o peso de uma pessoa. Diz-se que essa variável representa uma característica não observável ou latente. Na

TRI o objetivo é determinar o quanto dessa característica o examinando possui e utiliza-se o termo geral "capacidade", normalmente representado por θ. Para cada nível de capacidade, θ, há uma certa probabilidade, $P(\theta)$, de o examinando dar a resposta correta. Esta probabilidade é mais baixa para examinandos com capacidade mas baixa e é mais alta para examinandos com capacidade mais elevada.

De acordo com Hambleton e colaboradores (1991), a TRI baseia-se em dois postulados básicos:

- O desempenho de um examinando num teste pode ser previsto, ou explicado, através de um conjunto de fatores chamados características latentes.
- A relação entre o desempenho do examinando e o conjunto de características de uma questão, subjacentes ao seu desempenho, pode ser descrita através de uma função monótona crescente, chamada função característica da questão ou curva característica do item (CCI).

Uma CCI é o gráfico de $P(\theta)$, que é uma curva sigmoide logística, isto é, relaciona a probabilidade de dar uma resposta correta numa questão, com a capacidade medida pelo teste e com as características da questão. Às características da questão chamam-se parâmetros. O modelo matemático padrão para a CCI é a forma cumulativa da função logística, definindo-se uma família de curvas com a mesma forma, sendo que todos os modelos TRI contêm um ou mais parâmetros que descrevem a questão, e um ou mais parâmetros que descrevem o examinando. A principal diferença que permite distinguir entre os diversos modelos TRI em utilizações comuns está no número e no tipo de parâmetros da questão que afetam o desempenho do examinando. Há três modelos básicos na TRI, distinguindo-se entre si através do número de parâmetros utilizados para descrever a questão:

- modelo logístico de 1-parâmetro;
- modelo logístico de 2-parâmetros;
- modelo logístico de 3-parâmetros.

Estes modelos são apropriados apenas para dados dicotómicos, isto é, para situações em que as respostas às questões apenas podem tomar dois valores (o indivíduo acertou, representado normalmente por 1, ou o indivíduo não acertou, representado normalmente por 0), tal como acontece com os testes com QEM. No entanto, há adaptações aos modelos que permitem lidar com outros tipos de dados (Hambleton et al., 1991). Um determinado modelo TRI pode ser, ou não, apropriado para um certo conjunto de dados, sendo necessário avaliar o ajustamento do modelo aos dados, examinando quão bem o modelo explica os resultados obtidos. Apresentam-se de seguida as características de cada um destes modelos.

3.2.1. Modelo logístico de 1-parâmetro

De acordo com a TRI, no modelo logístico de 1-parâmetro considera-se que apenas a dificuldade influencia o desempenho do examinando e que as questões discriminam todas da mesma forma

(Baker, 2001; Hambleton et al., 1991). Assim sendo, cada questão é um membro de uma família de curvas dada pela (Equação 3:

$$\boldsymbol{P_i(\theta) = \frac{e^{\theta - b_i}}{1 + e^{\theta - b_i}} = \frac{1}{1 + e^{-(\theta - b_i)}}, i = 1, 2, \ldots, n}$$

onde

$\mathrm{P}_i(\theta)$ é a probabilidade de que um examinando com capacidade θ, escolhido aleatoriamente, responda corretamente à questão i;

b_i é a dificuldade da questão i;

n é o número de questões;

e é o número de nepper.

(Equação 3: CCI de 1-parâmetro)

Na Figura 2 apresenta-se um exemplo de uma CCI para uma questão, considerando o modelo logístico de 1-parâmetro. Verifica-se que $\mathrm{P}(\theta)$ é uma curva em forma de S, com valores que variam entre 0 e 1. O valor de θ teoricamente varia de $-\infty$ a $+\infty$, mas tipicamente varia de -3 a 3. O valor de b_i é o ponto na escala da capacidade (eixo dos xx) onde a probabilidade de obter uma resposta correta é 0.5. Quanto maior for o valor do parâmetro b_i, maior é a dificuldade de questão. Caso existam várias CCI no mesmo referencial, as curvas mais à direita apresentam valores de b_i mais elevado. Os valores de b_i, tipicamente, variam entre de -2 a 2 (Hambleton et al., 1991) ou de -3 a 3 (Baker, 2001), mas teoricamente podem variar de $-\infty$ a $+\infty$. Uma assíntota horizontal ao gráfico de $P(\theta)$ é $y = 0$ quando $\theta \to 0,\ P(\theta) \to 0,$ o que significa que um examinando com uma capacidade muito baixa tem uma probabilidade nula de acertar na questão.

O modelo Rasch é bastante comum na literatura (Aziz, Salleh, Khatimin, & Zaharim, 2013; Baker, 2001; Hambleton et al., 1991; Lee et al., 2011; Liu et al., 2011; McAlpine, 2002a, 2002b). No entanto, tal como referem Hambleton e colaboradores (1991), embora a sua forma seja diferente desta que é apresentada para o modelo logístico de 1-parâmetro, é matematicamente equivalente a este. Já Baker (2001), não faz sequer distinção entre os dois modelos, apresentando em paralelo Rasch e o modelo logístico de 1-parâmetro.

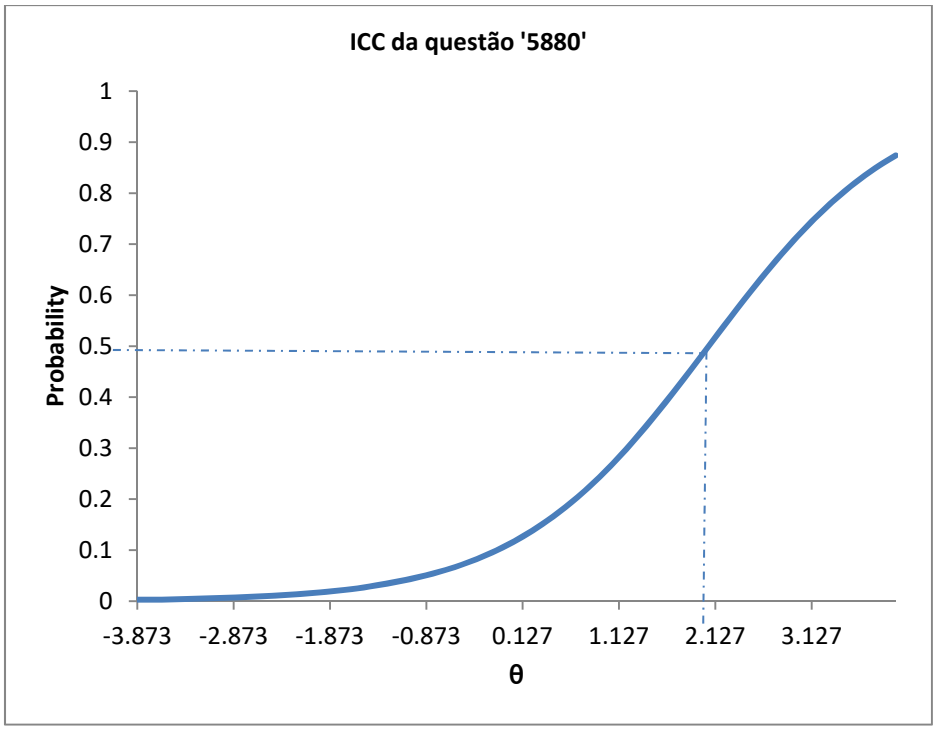


Figura 2: CCI para modelo logístico de 1-parâmetro.

3.2.2. Modelo logístico de 2-parâmetros

De acordo com a TRI, no modelo logístico de 2-parâmetros considera-se que quer a dificuldade quer a discriminação influenciam o desempenho do examinando (Baker, 2001; Hambleton et al., 1991). Assim sendo, cada questão é um membro de uma família de curvas dada pela (Equação 4:

$$\boldsymbol{P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}, i = 1, 2, \ldots, n}$$

onde

$P_i(\theta)$ é a probabilidade de que um examinando com capacidade θ, escolhido aleatoriamente, responda corretamente à questão i;

b_i é a dificuldade da questão i;

D é um fator de escala para tornar a função logística o mais próximo possível da normal. Demonstrou-se que para $D = 1.7$ os valores de $P_i(\theta)$ normal difere do logístico em menos de 0.01 para todos os valores de θ;

a_i é a discriminação da questão i;

n é o número de questões;

e é o número de Nepper.

(Equação 4: CCI de 2-parâmetros)

Na Figura 3 apresenta-se um exemplo de uma CCI para uma questão, considerando o modelo logístico de 2-parâmetros. Considera-se que o modelo logístico de 2-parâmetros é uma generalização do modelo logístico de 1-parâmetro, portanto os aspetos relativos à capacidade θ, à probabilidade $P_i(\theta)$, ao parâmetro b_i e à CCI referidos para o modelo logístico de 1-parâmetro mantêm-se para o modelo logístico de 2-parâmetros. Quanto ao valor de a_i, este é proporcional ao declive da CCI no ponto b_i da escala da capacidade (eixo dos xx). As questões com discriminações mais elevadas, isto é, nas quais as CCI apresentam declives mais acentuados, são mais úteis a separar os examinandos em diferentes níveis de capacidade do que as questões com discriminações menos elevadas. Teoricamente a discriminação varia de $-\infty$ a $+\infty$, mas tipicamente varia entre 0 e 2 (Hambleton et al., 1991) ou -3 a 3 (Baker, 2001). Questões com discriminações negativas devem ser rejeitadas, porque há algo de muito errado com uma questão cuja probabilidade de o examinando responder corretamente diminui à medida que a capacidade do examinando aumenta. A dificuldade e a discriminação não medem a capacidade do examinando, apenas descrevendo a forma da CCI. Podem definir-se níveis para as escalas quer da dificuldade quer da discriminação. Por exemplo, numa escala de 5 níveis, a escala de dificuldade poderia ser "Muito fácil", "Fácil", "Médio", "Difícil" e "Muito difícil" e a escala da discriminação poderia ser "Nenhuma", "Baixa", "Moderada", "Alta" e "Perfeita" (Baker, 2001).

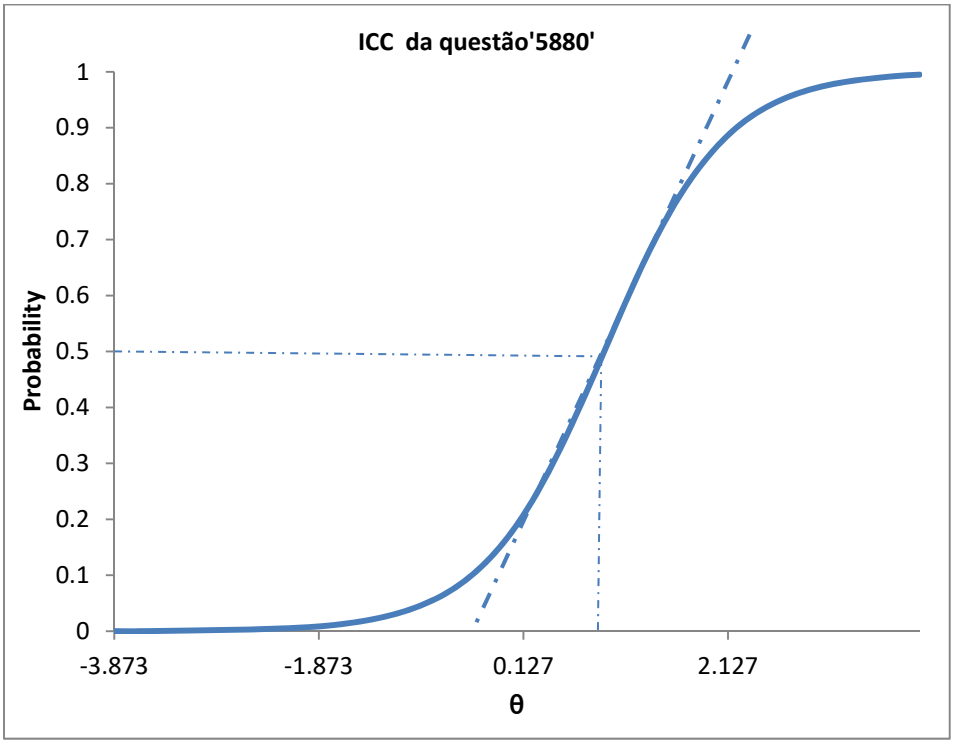


Figura 3: CCI para modelo logístico de 2-parâmetros.

3.2.3. Modelo logístico de 3-parâmetros

De acordo com a TRI (Baker, 2001; Hambleton et al., 1991) no modelo logístico de 3-parâmetros considera-se que além da dificuldade e da discriminação, também um outro parâmetro, a que habitualmente se chama acerto casual[11], influencia também o desempenho do examinando. Assim sendo, cada questão é um membro de uma família de curvas dada pela (Equação 5:

$$\boldsymbol{P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}, i = 1, 2, \dots, n}$$

onde

$P_i(\theta)$ é a probabilidade de que um examinando com capacidade θ, escolhido aleatoriamente, responda corretamente à questão i;

b_i é a dificuldade da questão i;

D é um fator de escala para tornar a função logística o mais próximo possível da normal. Demonstrou-se que para $D = 1.7$ os valores de $P_i(\theta)$ normal difere do logístico em menos de 0.01 para todos os valores de θ. No entanto, alguns autores, por exemplo Baker (2001), consideram $D = 1$;

a_i é a discriminação da questão i;

c_i é o nível de acerto casual;

n é o número de questões;

e é o número de Nepper.

(Equação 5: CCI de 3-parâmetros)

Na Figura 4 apresenta-se um exemplo de uma CCI para uma questão, considerando o modelo logístico de 3-parâmetros. Pode-se considerar que o modelo logístico de 3-parâmetros é uma generalização do modelo logístico de 2-parâmetros, portanto os aspetos relativos à capacidade θ, à probabilidade $P_i(\theta)$, ao parâmetro b_i, ao parâmetro a_i e à CCI referidos para o modelo logístico de 2-parâmetros mantêm-se para o modelo logístico de 3-parâmetros. No entanto, Baker (2001) refere que a definição do parâmetro dificuldade, b_i, se altera para o ponto na escala de θ, onde $P_i(\theta) = \frac{1+c_i}{2}$. Quanto ao parâmetro c_i, fornece a possibilidade de qualquer aluno, inclusive com capacidade baixa, poder responder acertadamente à questão. Podemos verificar que a assimptota inferior da CCI já não se aproxima de zero, mas sim de um outro

[11] Traduziu-se o parâmetro c (guessing), por acerto casual, por ser mais usual. No entanto, existem outras designações tais como pseudo-escolha, pseudo-casualiudade e adivinhação.

valor, que corresponde ao parâmetro c_i. Teoricamente $0 \leq c_i \leq 1$. Habitualmente este parâmetro deverá assumir valores menores que o valor que resultaria se o examinando escolhesse a resposta de forma aleatória (Hambleton et al., 1991). Já Baker (2001) considera que valores de $c_i > 0.35$ não são aceitáveis.

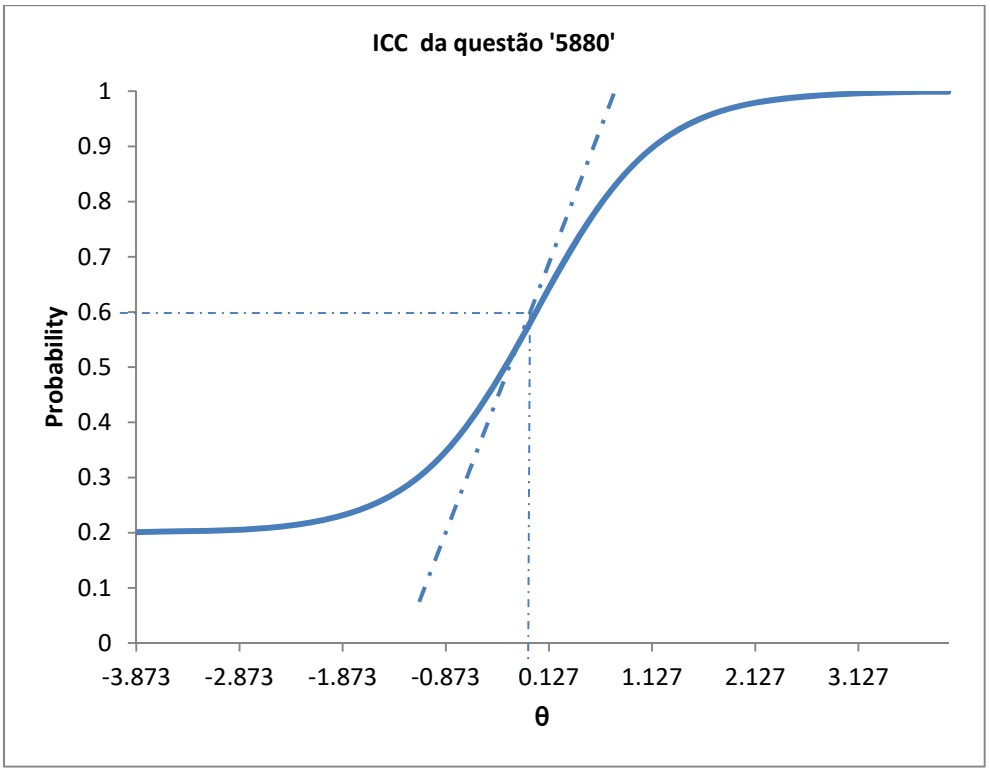


Figura 4: CCI para modelo logístico de 3-parâmetros.

3.2.4. Condições, propriedades dos parâmetros e ajustamento do modelo

Para poder aplicar qualquer um dos modelos anteriormente apresentados, é necessário que se verifiquem duas condições (Hambleton et al., 1991): unidimensionalidade e independência local.

A unidimensionalidade diz-nos que apenas uma capacidade do examinando pode ser medida por um determinado conjunto de questões num teste. É reconhecido na literatura que não é garantido que a unidimensionalidade se possa obter, sendo no entanto necessária a presença de uma componente dominante que influencia o desempenho do examinando no teste. Haladyna (2004) refere que o índice 20 de Kuder-Richardson (Equação 6) pode ser utilizado para estudar a unidimensionalidade, embora o considere um método pouco convencional. Também aponta a análise fatorial confirmatória como sendo um método mais seguro.

Quanto à independência local, significa que quando os parâmetros que influenciam o teste se mantêm constantes, as respostas dos examinandos a quaisquer pares de questões são estatisticamente independentes, ou seja, os parâmetros especificados pelo modelo são os únicos fatores que influenciam a resposta dos examinandos às questões.

Para estimar cada um dos parâmetros efetua-se um processo semelhante ao que é feito para um modelo de regressão, utilizando-se, no entanto, o método da máxima verosimilhança em vez do método dos mínimos quadrados (Baker, 2001; Hambleton et al., 1991). Hambleton e colaboradores (1991) apresentam ainda outros exemplos de métodos que podem ser utilizados, como por exemplo, estimativa bayesiana ou heurística.

Depois de um destes modelos ter sido ajustado aos dados, há algumas propriedades, consideradas desejáveis, que se obtêm. Uma das propriedades é o facto de os parâmetros da questão e a capacidade serem invariantes, isto é, as estimativas da capacidade e dos parâmetros da questão não são dependentes do teste, e serão as mesmas se forem obtidas a partir de diferentes conjuntos de dados, exceto para erros de medição. A propriedade da invariância implica que os parâmetros que caracterizam uma questão não dependem da distribuição da capacidade dos examinandos e significa também que a capacidade que caracteriza os examinandos não depende do conjunto de questões. Outra propriedade concerne no facto de serem fornecidas estimativas dos erros padrão para cada estimativa da capacidade, em vez de uma só estimativa de erro igual para todos os examinandos.

Para julgar o ajustamento do modelo aos dados de teste, Hambleton e colaboradores (1991) propõem uma abordagem empírica defendendo que se devem procurar três tipos de evidências:

- verificar a validade das condições (unidimensionalidade e independência local) nos dados de teste;
- verificar em que medida são obtidas as propriedades do modelo (invariância e estimativas dos erros da capacidade);
- verificar a precisão das previsões dos modelos.

No entanto, outros autores (Baker, 2001; Hall, Jung & Pilant, 2012) referem que o ajustamento do modelo pode ser medido pelo índice de ajustamento do $\mathcal{X}^2$.

3.3. Considerações adicionais sobre a análise de testes e questões

3.3.1. Limitações de cada uma das teorias de análise

Há várias limitações que são apontadas à TCT, principalmente pelos defensores da TRI. Uma das principais limitações apontadas prende-se com o facto de a TCT ser orientada para os testes e não para as questões e, portanto, não se poderem separar as características do examinando, das características do teste (Hambleton et al., 1991). Assim sendo, afirmam os defensores da TRI, que com a TCT não se podem comparar devidamente examinandos que responderam a testes diferentes, dado que as características dos testes são distintas. Também se aponta como limitação o facto de os parâmetros calculados dependerem da amostra utilizada, por exemplo, o Índice de Dificuldade para a mesma questão pode ser maior ou menor consoante os sujeitos incluídos na amostra tenham mais ou menos capacidade (Haladyna, 2004). Outra limitação tem a ver com o facto de não ser plausível que os erros de medição sejam iguais para todos os examinandos (Hambleton et al., 1991). Pelo contrário, na TRI consideram-se estimativas de erros distintas para as diferentes capacidades estimadas. No entanto, a análise das questões utilizando TCT acaba por ser mais intuitiva.

Há também algumas críticas apontadas à TRI, que estão relacionadas com o tamanho e a heterogeneidade das amostras utilizadas, sendo que caso as amostras sejam pequenas e não sejam heterogéneas, no que diz respeito aos examinandos, os valores dos parâmetros calculados não podem ser considerados boas estimativas (Haladyna, 2004; Zickar & Broadfoot, 2009). Acresce que o *software* disponível para aplicação da TRI apresenta ainda bastantes limitações (Zickar & Broadfoot, 2009).

Em qualquer dos casos, há autores que defendem que para amostras suficientemente grandes, os valores dos parâmetros podem ser considerados boas estimativas, sendo o tamanho da amostra mais crítico no caso do cálculo da discriminação do que no caso do cálculo da dificuldade (Haladyna, 2004). Também a heterogeneidade e a representatividade da amostra têm um papel relevante, na medida em que a discriminação pode ser enviesada por uma amostra demasiado homogénea.

3.3.2. Análise da Fiabilidade ou Consistência Interna

Determinar a consistência interna consiste em saber "até que ponto as diferentes partes de um teste ou procedimentos de avaliação têm as mesmas características, capacidades ou qualidades. As medidas de fiabilidade são frequentemente baseadas na consistência interna" (traduzido de JISC, 2006, p. 61).

Uma medida utilizada como estimador de fiabilidade, ou seja, como estimador da consistência interna de um dado teste é o índice 20 de Kuder-Richardson (KR20), cuja fórmula se encontra na (Equação 6.

$$KR20 = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} p_i(1-p_i)}{s^2}\right]$$

onde

K é o número de questões no teste;

p_i é a proporção de respostas corretas;

s é o desvio padrão das respostas

(Equação 6: Fórmula do índice 20 de Kuder-Richardson)

Quanto à correlação KR20 (Ferrão, 2010):

- varia entre 0 e 1;
- quanto mais próximo estiver de 1, maior será a consistência interna dos teste;
- considera-se que KR20>0.8 representa uma consistência razoável.

Outra medida utilizada como estimador de fiabilidade, ou seja, como estimador da consistência interna de um dado teste é o coeficiente α de Cronbach[12], cuja fórmula se encontra na (Equação 7.

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

onde

K é o número de questões no teste;

p_i é a proporção de respostas corretas;

s é o desvio padrão das respostas

(Equação 7: Índice α de Cronbach)

Quanto ao α de Cronbach, Maroco e Garcia-Marques (2006) afirmam que valores superiores a 0.8 representam uma fiabilidade moderada a elevada.

[12] O α de Cronbach está relacionado com o primeiro componente principal da análise fatorial.

CAPÍTULO 4. TAXONOMIAS DE APRENDIZAGEM

Uma limitação das QEM prende-se com o facto, identificado por muitos autores, de que estas questões podem não avaliar níveis cognitivos mais elevados[13] (Bible et al., 2008; Lee et al., 2011; Liu et al., 2011; Rod et al., 2010), apesar de haver autores que defendem que com um esforço adicional na elaboração das questões se podem avaliar níveis cognitivos superiores (Clegg & Cashin, 1986; Curtis, Lind, Boscardin, & Dellinges, 2013; Kim et al., 2012; Nicol, 2007; Yonker, 2011). De qualquer das formas, para desenhar e implementar a avaliação é necessário ter conhecimento das exigências cognitivas e, nesse sentido, as chamadas taxonomias de aprendizagem têm um papel importante na identificação das exigências, apesar de lhes serem reconhecidas algumas limitações (Brown, 2001; Darlington, 2014; Haladyna, 2004; Smith et al., 1996).

A Taxonomia de Bloom, cujo nome deriva de Benjamin Bloom, também designada como Taxonomia dos objetivos cognitivos (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956), é a mais divulgada. No entanto, podem encontrar-se na literatura outras taxonomias. John Biggs é conhecido por desenvolver a Taxonomia SOLO (Structure of Observed Learning Outcomes) que nos lembra a importância de prestar atenção aos resultados observáveis da aprendizagem (Jonh Biggs & Collis, 1982). Estas duas taxonomias irão ser abordadas com mais algum detalhe em seguida.

4.1. Taxonomia de Bloom

Tal como já foi referido, a Taxonomia de Bloom é uma das mais populares para a classificação de objetivos de aprendizagem. Inicialmente estaria prevista uma estrutura de três domínios: cognitivo, afetivo e psicomotor. No entanto, apenas o primeiro domínio foi definido por Bloom (Haladyna, 2004; Kim et al., 2012; Munzenmaier & Rubin, 2013). Esta taxonomia, no que diz respeito ao domínio cognitivo, pressupõe uma hierarquia de seis níveis de aprendizagem, como se apresenta na Figura 5, sendo que a cada um dos níveis estão associadas palavras-chave, apresentando-se alguns exemplos na referida figura. Os níveis são os seguintes (Bloom et al., 1956; Gelade & Fursenko, 2007; Imrie, 1995; Kim et al., 2012; Munzenmaier & Rubin, 2013):

- **Conhecimento** – corresponde ao nível mais baixo da hierarquia; consiste em relembrar informação apropriada, previamente aprendida, estando assim associado a situações que valorizam a memorização, evocação e reconhecimento de informação.

[13] Considerando os níveis cognitivos da Taxonomia de Bloom e outras semelhantes a ela, que serão abordadas de seguida.

- **Compreensão** – consiste em extrair significado de materiais com informação e explicar ideias, traduzindo-se no entendimento de uma mensagem comunicada.
- **Aplicação** – é a utilização de informação previamente aprendida em situações novas e concretas para resolver problemas que têm uma ou várias respostas, estando assim associada à capacidade de abstração.
- **Análise** – implica desintegrar os materiais com informação nos seus componentes, examinando a relação entre esses elementos e a forma como se organizam, para desenvolver conclusões divergentes através da identificação de motivos ou causas, fazendo inferências e/ou encontrando evidências para apoiar generalizações.
- **Síntese** – é a aplicação do conhecimento e competências previamente adquiridos na produção de algo novo e original, integrando todos os elementos num conjunto coerente.
- **Avaliação** – consiste em julgar ou produzir opiniões pessoais, com um determinado objetivo pelo que não existem repostas certas ou erradas quando é requerida.

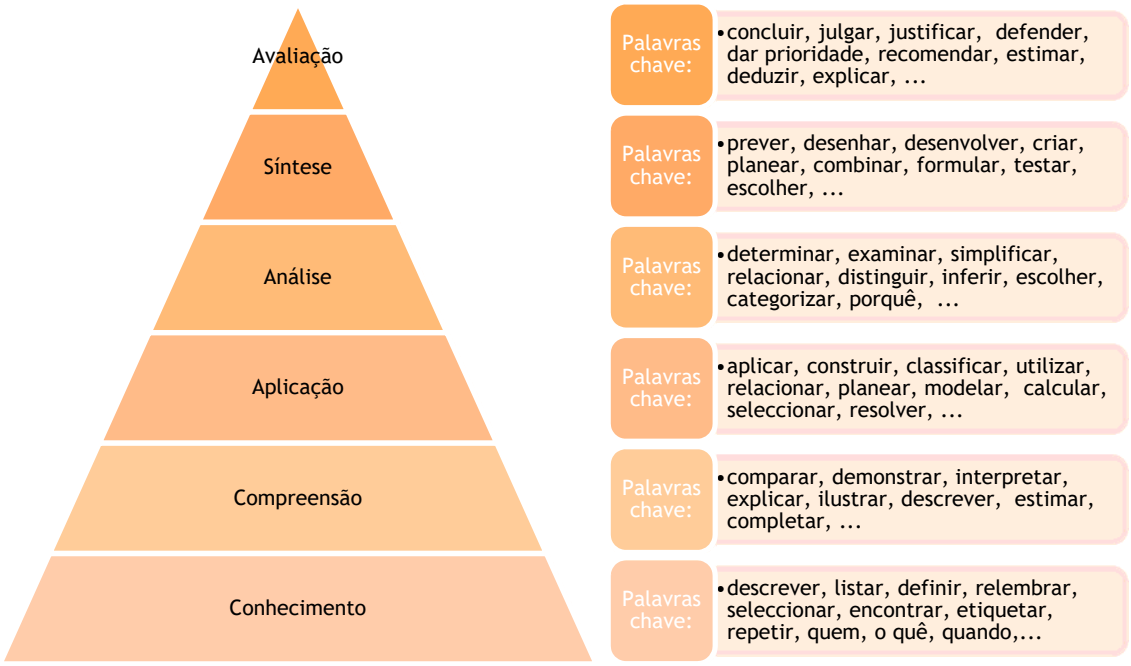


Figura 5: Taxonomia de Bloom (Adaptado de Bloom et al., 1956; Gelade & Fursenko, 2007; Imrie, 1995; Kim et al., 2012; Munzenmaier & Rubin, 2013).

Ao longo dos anos, a Taxonomia de Bloom tem vindo a sofrer várias adaptações através da apresentação de outras taxonomias, como por exemplo, a Taxonomia RECAP (Burrow et al., 2005; Imrie, 1995), a Taxonomia MATH (Ball et al., 1998; Smith et al., 1996; Smith & Wood, 2000) e, durante os anos noventa, uma versão revista da Taxonomia de Bloom por outros autores (Anderson et al., 2000; Munzenmaier & Rubin, 2013), entre outras (Bennie, 2013; Haladyna, 2004; Huntley, Engelbrecht, & Harding, 2009). Enquanto as duas primeiras taxonomias (RECAP e MATH) propõem modificações ajustando novos níveis - que acabam por ser equivalentes aos

que são apresentados na original Taxonomia de Bloom -, a Taxonomia de Bloom Revista poderá ser considerada a mais relevante, dado que propõe alterações a 3 níveis (terminologia, de estrutura e ênfase) e não apenas alterações nos níveis e respetivas nomenclaturas, considerados inicialmente por Bloom.

4.2. Taxonomia SOLO

Outra taxonomia bastante divulgada é a SOLO, do inglês, Structure of the Observed Learning Outcome[14] (John Biggs & Tang, 2011; Jonh Biggs & Collis, 1982). Esta taxonomia, baseada na teoria do desenvolvimento cognitivo, define uma estrutura que descreve a evolução da aprendizagem, sendo um meio de classificar a evolução dos resultados de aprendizagem em termos da sua complexidade, desde o entendimento superficial até ao entendimento aprofundado. Foram identificadas cinco fases, bem como verbos ou termos que podem ser utilizados para formar os resultados de aprendizagem em cada uma dessas cinco fases, que são as seguintes (Jonh Biggs & Collis, 1982; Frankland, 2007a; Leung, 2000; Wong, 2007):

- Pré-estrutural – neste nível, a aprendizagem é considerada irrelevante ou não apropriada para a tarefa, o aluno não entendeu o ponto crucial e como tal é incompetente na execução da tarefa;
- Uni-estrutural – neste nível, apenas um dos aspetos considerados relevantes para a tarefa é considerado e utilizado; verbos/termos: identificar, efetuar um procedimento simples;
- Multi-estrutural – neste nível, vários aspetos da tarefa são adquiridos, mas tratados de forma separada, são vistos como não estando ligados; verbos/termos: enumerar, descrever, listar, combinar, executar algoritmos;
- Relacional – neste nível, os componentes quantitativos tornam-se integrados como um todo, o que normalmente significa um entendimento adequado do tópico; verbos/termos: comparar/contrastar, explicar causas, analisar, relacionar, aplicar;
- Abstração alargada – neste nível, o todo previamente integrado pode ser conceptualizado a um nível de abstração mais elevado e generalizado para um novo tópico ou área; verbos/termos: teorizar, generalizar, propor hipóteses, refletir.

As três primeiras etapas são normalmente identificadas como quantitativas e as duas últimas como qualitativas. A Taxonomia SOLO tem vindo a manter-se estável ao longo dos anos.

[14] Que podemos traduzir como Estrutura do Resultado de Aprendizagem Observado.

4.3. Considerações sobre as Taxonomias de aprendizagem

Uma das principais vantagens do uso de taxonomias em contexto de aprendizagem tem a ver com a garantia de obtenção de qualidade, na medida em que poderão servir como evidência documental da qualidade que se pretende e como um enquadramento teórico que assegura a ligação entre a avaliação e a qualidade das aprendizagens (Haladyna, 2004; Imrie, 1995). Independentemente disso, alguns autores identificam algumas limitações a este tipo de taxonomias, maioritariamente à Taxonomia de Bloom referindo-se, por exemplo, que lhe falta consistência interna, que não foi validada e que por vezes é demasiado simplista (Burrow et al., 2005; Haladyna, 2004; Munzenmaier & Rubin, 2013). Também lhe são apontadas algumas limitações em determinadas áreas de aplicação, nomeadamente no contexto da Matemática (Darlington, 2014; Smith et al., 1996).

Depois de analisarmos várias taxonomias, e tendo em conta as necessidades de avaliação que deram origem ao estudo realizado nesta tese, identificamo-nos com a visão de Haladyna (2004). Este autor classifica três tipos de aprendizagem, interrelacionados e complementares, aos quais se pode associar uma hierarquia, no sentido de que cada nível depende dos anteriores, conforme apresentado de seguida:

- **Conhecimento** que é "o corpo de verdades acumuladas ao longo do tempo" (traduzido de Haladyna, 2004, p. 29), como, por exemplo, identificar números primos;
- **Competência** que "envolvem o desempenho de um ato físico ou mental" (traduzido de Haladyna, 2004, p. 34), como, por exemplo, calcular uma derivada;
- **Capacidade cognitiva** que se refere a "capacidades mentais complexas que podem ser desenvolvidas ao longo do tempo e com prática" (traduzido de Haladyna, 2004, p. 35), como, por exemplo, resolver um problema matemático.

Neste contexto, Haladyna (2004) ainda refere que:

I. O conhecimento revela-se através da colocação de questões, sendo a avaliação através do uso de QEM bastante adequada neste caso.
II. As competências devem ser executadas e observadas para verificar que foram adquiridas, podendo, no entanto, construir-se QEM adequadas, de modo a avaliar competências de forma apropriada.
III. A demonstração de uma capacidade cognitiva requere o uso dos conhecimentos e das capacidades numa combinação única, de modo a conseguir alcançar um resultado complexo, não sendo as QEM adequadas, neste caso.

PARTE II. ESTUDO EMPÍRICO

CAPÍTULO 5. METODOLOGIA DE INVESTIGAÇÃO

Atendendo à natureza do problema apresentado na introdução, optámos por implementar na investigação desenvolvida a Investigação–Ação (IA). Como o próprio nome indica, a IA caracteriza-se pela existência de duas vertentes: a investigação e a ação. É dada particular ênfase à adoção por parte do investigador de um papel de intervenção e de colaboração ativa com os restantes participantes no estudo, provocando mudanças que têm como objetivo atingir melhorias (Given, 2008; Hughes, 2008; Sousa & Baptista, 2011; Yin, 2011). Abrimos este capítulo com um preâmbulo, de forma a especificar os objetivos do estudo, antes de passarmos para a descrição da metodologia de investigação propriamente dita. Em seguida, abordamos a opção metodológica seguida - a Investigação-Ação -, com uma caraterização geral deste tipo de investigação, uma breve apresentação histórica, uma reflexão sobre a sua importância na Educação, em geral, e na Matemática, em particular, seguindo-se a apresentação dos ciclos que tipicamente caracterizam a IA. Serão depois apresentados o contexto e os participantes do estudo, assim como o procedimento adotado em termos de desenho da investigação, especificando cada um dos 3 ciclos de IA desenvolvidos. Os instrumentos de recolha de dados utilizados são enumerados e descritos em seguida. Começamos pelo banco de questões e explicamos como foram definidas as categorias para as questões, como foram criadas as questões e os testes, como foi efetuado o processo de revisão, caracterizando o banco de questões por ciclo, como foram aplicados os testes opcionais do 1º ciclo de IA e como foram avaliadas as questões usando a Teoria Clássica dos Testes e a Teoria da Resposta ao Item. De seguida, apresentam-se os questionários aos docentes no 1.º ciclo de IA, a entrevista aos docentes no 3.º ciclo de IA e, na última subsecção deste capítulo, o questionário aos alunos no 3.º ciclo de IA.

Objetivos do Estudo

O problema que deu início a esta investigação teve como origem a implementação do Processo de Bolonha. Este apontava não só para a necessidade de realizar avaliação contínua ao longo do(s) semestre(s), mas também para a necessidade de englobar metodologias diversificadas (Boticki & Milasinovic, 2008; Llamas-Nistal et al., 2013; Mora et al., 2012; Rod et al., 2010). A implementação do Processo de Bolonha levou a uma redução da carga horária destinada à lecionação, devido à reestruturação dos cursos e, além disso, houve também um aumento no número de alunos por turma. Assim sendo, o objetivo geral do estudo é refletir sobre o processo de avaliação da aprendizagem dos alunos em UC de Matemática, utilizando *e-assessment* com testes contendo QEM. Como objetivos mais específicos pretende-se:

- perceber como o *e-assessment* pode influenciar o processo de ensino-aprendizagem por parte dos alunos;

- perceber como o *e-assessment* pode influenciar o processo de ensino-aprendizagem por parte dos docentes;
- definir boas práticas para o desenvolvimento de QEM na área da Matemática;
- descobrir formas adequadas de análise das QEM de modo a fomentar uma avaliação tão justa quanto possível para os alunos.

No seguimento destes objetivos pretendemos dar resposta a questões tais como:

- Como evolui o rendimento académico dos estudantes nas UC em que é implementada a estratégia de *e-assessment*?
- Quais as vantagens e limitações da utilização de testes de QEM, do ponto de vista dos docentes e dos estudantes?
- Será que a implementação do *e-assessment* provoca alterações no processo de aprendizagem dos estudantes?
- Será que a implementação do *e-assessment* provoca alterações nas práticas educativas dos docentes?
- Como assegurar a qualidade no desenvolvimento de QEM?
- Como assegurar a validade e fiabilidade no processo de avaliação com *e-assessment*?

5.1. Opção Metodológica: a Investigação-Ação

5.1.1. Características Gerais da Investigação-Ação

A IA refere-se a um processo de investigação que é disciplinado e conduzido por indivíduos que executam alguma ação para seu uso durante a sua implementação. O processo é também caracterizado pelas atividades desses indivíduos, que aprendem através da realização dessas ações. A IA envolve a identificação de um problema por um grupo de indivíduos que dedicam os seus melhores esforços para resolver esse problema, medindo o seu sucesso e, se os resultados não forem satisfatórios, repetem o processo – dando origem aos chamados ciclos de IA (Burns, 2007; Coghlan & Brydon-Miller, 2014; Dick, Stringer, & Huxham, 2009; Given, 2008; Herr & Anderson, 2005; Reason & Bradbury, 2008). De acordo com Capobianco e Ní Ríordáin (2015), IA define-se como uma contribuição que visa a resolução de problemas de indivíduos e, além disso, os ajuda na consecução dos seus objetivos. Como resultado, o compromisso é dual, apoiando no estudo dos sistemas, bem como na colaboração com os seus membros, alterando-o para aquilo que é desejado por todos os intervenientes. Há assim a necessidade de colaboração entre os intervenientes e o investigador, o que ajuda a enfatizar a importância da co-aprendizagem como uma das vertentes essenciais da IA. O facto de a IA ser em grande parte relevante para os seus participantes, poderá motivá-los a um maior envolvimento ao longo do processo. Em termos gerais, a IA aumenta a capacidade dos investigadores para o

desenvolvimento da sua investigação e para o desenvolvimento de abordagens sistemáticas, à medida que levam a cabo as suas práticas. Desta forma, na sua vasta maioria, essas práticas destinam-se a trazer mudanças positivas para os indivíduos e a sua comunidade (Mostofo & Zambo, 2015).

A IA pode ser vista como "um processo participativo, preocupado com o desenvolvimento de conhecimento prático na busca de atingir objetivos humanos que sejam úteis. Procura juntar ação e reflexão, teoria e prática, na participação com outros, na procura de soluções para aspetos de grande preocupação das pessoas e, de forma mais geral, a prosperidade das pessoas como indivíduos e das suas comunidades" (traduzido de Reason & Bradbury, 2008, p. 4). Por outro lado, a IA "é uma metodologia de investigação flexível especialmente adequada à investigação e ao apoio à mudança. Ela integra a investigação social com a ação exploratória para promover o desenvolvimento. Na forma clássica, a IA envolve ciclos fluidos e sobrepostos de investigação, planeamento de ações, execução de novas práticas e avaliação dos resultados, incorporando em todas as fases a recolha e a análise de dados e a geração de conhecimento. Os resultados da Investigação–Ação envolvem assim, aspetos práticos e teóricos: "conhecimento que gera tem um impacto direto e permanente na mudança da prática para os participantes e numa audiência mais vasta através das publicações resultantes" (traduzido de Given, 2008, p. 4). A IA pode utilizar um conjunto diversificado de métodos para recolha e análise de dados, quer qualitativos quer quantitativos, tais como questionários, entrevistas, análise de textos, conversas informais ou revisão de documentos. Esta diversidade de métodos é uma das grandes riquezas da IA sendo também uma das suas mais-valias (Burns, 2007; Given, 2008; Herr & Anderson, 2005; Ivankova, 2015).

Desde a sua criação que a IA se tem vindo a desenvolver, existindo diferentes tradições e abordagens e, dependendo delas e do objetivo de estudo, temos variações quanto ao nome. Estas abordagens são habitualmente designadas como uma família de métodos (Heller, 2004; Herr & Anderson, 2005; Reason & Bradbury, 2008). Para além da IA a que podemos chamar mais tradicional, mais centrada no desenvolvimento e na aprendizagem organizacional, uma das tradições comuns e bastante divulgada é a Investigação-Ação Participada (IAP, em inglês "Participatory Action Research" - PAR) e uma extensão a ela a Investigação–Ação Crítica. Algum relevo é também dado à chamada Ciência Ação (do inglês, Action Science).

A IAP tem as suas origens na segunda metade do século XX, segundo as linhas definidas pelo trabalho de Kurt Lewin nas décadas de 40 e 50 desse século. As abordagens mais contemporâneas têm vindo a ser influenciadas por diversas tradições intelectuais, tais como o Marxismo, o Feminismo e o Post-Positivismo. A IAP é uma metodologia que tem levantado alguma controvérsia, devido ao facto de criticar e desafiar a relação entre investigador e investigado, proposta pelas metodologias mais tradicionais, segundo as quais o investigador pode colaborar com indivíduos e grupos durante a investigação, mantendo ainda assim a sua integridade como especialista (Given, 2008, p. 601).

A IA Crítica é vista como uma extensão da IA ou dos processos da PAR. Os processos de IA Crítica invertem o poder hierárquico tradicional do investigador sobre o investigado, trabalhando estes em conjunto no sentido de encontrar novas formas de ver a situação e para desenvolver soluções, capacitando tanto o investigador como o investigado (Given, 2008, p. 139).

A Ciência Ação caracteriza-se pela compreensão das pessoas como investigadores das próprias práticas, envolvendo-as na investigação do seu próprio comportamento de modo a tentarem alcançar os seus objetivos e a testar teorias tácitas de ação. As pessoas colocam-se assim perante um processo reflexão crítica, alargando o seu leque de escolhas relativamente ao seu comportamento e relações" (Coghlan & Brydon-Miller, 2014, p. 15).

Em suma, em qualquer uma destas tradições e abordagens, há, no geral, um denominador comum que consiste na integração da Teoria com a Ação, num processo de reflexão, com o qual se pretende uma intervenção numa organização ou comunidade, de modo a resolver um problema e a produzir mudança. Este processo passa pela identificação de um problema, pela proposta de resolução, pela implementação e pela avaliação, num ciclo que se pode repetir de modo a resolver o problema inicial ou modificá-lo através de melhorias. Podemos ainda dizer que se trata de um processo cíclico ou em espiral. Por meio destes ciclos e respetivos resultados finais temos criação de novo conhecimento, podendo contribuir para o desenvolvimento de novas teorias (Burns, 2007; Given, 2008; Heller, 2004; Herr & Anderson, 2005; Reason & Bradbury, 2008).

5.1.1.1. Breve Perspetiva Histórica

Apesar de não existir unanimidade sobre quem foi o fundador da IA, Kurt Lewin, após a publicação do artigo "Action Research and Minority Problems" (Lewin, 1946), é amplamente reconhecido como sendo o seu pai pela maioria dos investigadores, sendo ele o responsável pela criação de todo o processo (Burns, 2007; Coghlan & Brydon-Miller, 2014; Given, 2008; Heller, 2004; Herr & Anderson, 2005; Kitchen & Stevens, 2008). No entanto, há investigadores que afirmam que a IA foi utilizada pela primeira vez por Jonh Collier, nos seus esforços para promover a melhoria dos relacionamentos entre comunidades raciais durante a II Guerra Mundial, e que este teve um papel fulcral no desenvolvimento da IA (Burns, 2007; Neilsen, 2006; Reese, 2015; Salleh, 2006). Também John Dewey é considerado um percursor da IA, apresentando os fundamentos teóricos da IA, alicerçada na experiência do investigador de modo a criar novo conhecimento (Helskog, 2014; Herr & Anderson, 2005).

Nos seus primórdios, a IA estava maioritariamente preocupada com a resolução de problemas sociais, tentando explicar o comportamento humano e introduzindo melhorias. Lewin não acreditava que fosse possível fazer uma generalização dos comportamentos humanos a todos os contextos (Given, 2008), o que está alinhado com o facto de o objetivo da IA ser a produção de conhecimento mais ligado à realidade das pessoas que vivem os problemas e, portanto, ser mais relevante e viável a resolução desses mesmos problemas (Coghlan & Brydon-Miller, 2014). A IA surgiu como uma mudança radical em relação à ciência que tradicionalmente era

desenvolvida nas universidades, pelo que não admira que, apesar do sucesso inicial nas décadas de 1940 a 1960, tanto nos EUA como na Europa, tenha sido desconsiderada mais tarde. A principal crítica apontada à IA era a sua incapacidade para produzir generalizações, como era apanágio das metodologias de investigação mais qualitativas. No entanto, muitos investigadores desenvolveram trabalhos importantes nesta área que levaram ao seu sucesso e à aceitação que usufrui nos nossos dias. Como exemplo apontamos Paulo Freire pelo seu trabalho ligado à IA Participada, William Torbet, Chris Argyris e Donald Schon pelo seu trabalho ligado à Ciência Ação e outros tais como John Elliott, Eric Trist, Wilfred Carr, Stephen Kemmis, Davydd Greenwood, John Dewey (Brydon-Miller, Greenwood, & Maguire, 2003; Burns, 2007; Coghlan & Brydon-Miller, 2014; Given, 2008; Herr & Anderson, 2005).

Atualmente pode-se pensar na IA como uma família de abordagens, as quais são diversificadas e diferentes em alguns aspetos, que se afirmam como desenvolvendo um tipo de investigação diferente e ao mesmo tempo uma investigação rica e diversificada, com aplicações nas mais diferentes áreas e realidades (Heller, 2004; Reason & Bradbury, 2008). A IA teve algumas dificuldades em afirmar-se como ciência, devido ao seu caráter iminentemente prático e ao facto de ser aplicada em casos muito concretos, não se podendo generalizar. No entanto, tem vindo a ganhar reconhecimento, devido aos trabalhos de qualidade desenvolvidos por diversos investigadores na área, sendo hoje amplamente reconhecida como uma metodologia importante na geração de conhecimento, havendo já exemplos de modelos que foram objeto de generalização (Elliott, 2007; Helskog, 2014; Herr & Anderson, 2005).

5.1.1.2. A Investigação-Ação na Educação

A IA é amplamente utilizada na área da Educação (Coghlan & Brydon-Miller, 2014; Given, 2008; Herr & Anderson, 2005; Kitchen & Stevens, 2008) e é conhecida por poder ser aplicada diretamente na sala de aula, fornecendo aos educadores uma perspetiva crítica e mais detalhada sobre o trabalho por eles desenvolvido, levando-os a obter melhores resultados, oferecendo evidências de que o seu trabalho está a fazer uma diferença real na vida dos seus alunos e a produzir melhorias do processo em termos de qualidade (Brydon-Miller et al., 2003; Capobianco & Ní Ríordáin, 2015; Moreno, 2015; Mostofo & Zambo, 2015; Sommer, 2009). Assim sendo, a IA pode ser vista como uma estratégia importante dos professores, especialmente aqueles que desejam desenvolver os seus métodos de trabalho, de modo a ajudar os alunos na sua aprendizagem (Kitchen & Stevens, 2008). O facto de a IA estar bem adaptada para ser aplicada na área da Educação, na medida em que permite produzir melhorias no sistema, é uma motivação para que os professores a utilizem (Kitchen & Stevens, 2008; Reese, 2015). A IA preenche a lacuna que existe entre a prática e a investigação e ainda facilita o desenvolvimento profissional dos educadores, encorajando-os a assegurar uma análise atenta da dinâmica da sala de aula, a garantir as ações e interações dos alunos, a desafiar e a validar práticas atualmente em uso e a aceitar maiores riscos nos esforços desenvolvidos para melhorar todo o processo (Capobianco & Ní Ríordáin, 2015; Mostofo & Zambo, 2015; Reese, 2015; Salleh, 2006).

Assim sendo, podemos afirmar que a utilização da IA na Educação é bastante vantajosa quer para os professores, quer para os alunos.

As raízes da IA na Educação provêm do trabalho de Jonh Dewey e da importância que ele deu à experiência humana na produção de conhecimento e também dos trabalhos de Schon sobre a noção de prática reflexiva e de aprendizagem profissional. Na década de 1950, apesar de toda a oposição existente à IA, foi relevante o trabalho desenvolvido por Corey, o qual acreditava que o professor poderia obter resultados da sua investigação mais úteis do que o poderiam fazer pessoas estranhas ao meio. O movimento ressurgiu em finais da década de 1960 e princípios da década de 1970 em Inglaterra, com o chamado movimento "O professor como investigador", associado a Lawrence Stenhouse e a John Elliott e Clem Adelman. Animados por este movimento, um grupo de australianos liderados por Stephen Kemmis debruçou-se, nas décadas de 1980 e 1990, sobre a IA como metodologia de investigação, levando a desenvolvimentos importantes. Também, mais tarde nos EUA, a IA na Educação se desenvolveu de forma dinâmica (Burns, 2007; Herr & Anderson, 2005).

No caso do Ensino Superior, nesta era de mudança, provocada nomeadamente pela evolução da sociedade no geral e na Europa, em particular, pela implementação do Processo de Bolonha, a IA pode representar uma ferramenta fundamental. A sua importância prende-se com o facto de o seu objetivo principal ter a ver com a implementação de mudanças que visam o alcance de melhorias, tornando-se assim capaz de confrontar as organizações de Ensino Superior com os aspetos que estas mudanças levantam e apresentando o potencial necessário para trazer contributos significativos para levar a cabo uma mudança positiva dos processos para todos os intervenientes (Reason & Bradbury, 2008).

5.1.1.3. A Investigação-Ação na Matemática

A IA tem vindo a ser aplicada também no contexto específico da Matemática. Vejamos alguns exemplos encontrados na literatura:

- Capobianco e Ní Ríordáin (2015) apresentam um estudo efetuado com professores de Matemática, nos EUA e no Reino Unido, com o qual concluem que a utilização da IA ajuda a reconhecer, aceitar e abordar de forma positiva e produtiva as incertezas que surgem quando se tornam professores e investigadores.
- Já Moreno (2015) apresenta um estudo realizado nos EUA, em aulas de Matemática para adultos pertencentes a minorias étnicas, que tiveram problemas na sua escolarização no seu tempo de crianças, os quais descobriram que os seus problemas eram comuns aos de outros e que derivavam da desconexão entre os conteúdos lecionados e o seu mundo real.
- Mostofo e Zambo (2015), também nos EUA, apresentam a utilização da metodologia IA para a formação inicial de professores, sendo que os participantes nesta investigação aprenderam com a sua prática e que a sua eficácia melhorou bastante através do processo utilizado.

- Um outro estudo de Clarke e Fournillier (2012) aborda a utilização da IA na formação profissional de professores de Matemática nos EUA, a qual os ajudou a explorar o desenvolvimento das suas capacidades como professores–investigadores na sala de aula.
- Um estudo conduzido no departamento de Matemática, no Politécnico de Singapura, permitiu concluir que a IA é útil no ensino–aprendizagem, tendo sido identificados alguns fatores únicos que influenciam o processo (Khiat, Chia, Tan-Yeoh, & Kok-Mak, 2011).
- Larkin, Jamieson-Proctor e Finger (2012), na Austrália, realizaram um estudo para analisar a utilização das TIC no ensino e na aprendizagem da Matemática. Tal como eles próprios afirmam, "esta investigação–ação forneceu uma análise para ilustrar como e quando a utilização das TIC foi efetiva e quando foi problemática e os dados nesta análise foram usados para realizar alterações concretas à tecnologia utilizada e à abordagem pedagógica no uso da tecnologia" (traduzido de Larkin et al., 2012, p. 223)

Em Portugal, parece-nos que João Pedro da Ponte e Maria de Lurdes Serrazina são os percussores da utilização da IA na área da Matemática, nomeadamente na formação inicial de professores (Ponte, 2002; Serrazina & Oliveira, 2002). Apesar de tudo, João Pedro da Ponte utiliza a designação investigação da própria prática, não havendo uniformidade nas designações utilizadas nos estudos associados a estes e de outros investigadores. Serrazina e Oliveira (2002, p. 286), afirmam que "Muitas vezes o termo professor como investigador aparece associado ao de investigação–ação. Nesta, as motivações enraízam-se no envolvimento dos profissionais na definição de problemas a resolver e na identificação de soluções viáveis". A criação do Grupo de Investigação em Matemática (GTI), da Associação Portuguesa de Matemática (APM), tem vindo a desenvolver um papel relevante e podem já encontrar-se diversas teses e dissertações na área que utilizam a IA com sucesso (Ponte, 2008).

5.1.2. Ciclos da Investigação–Ação

Na sua forma clássica, a IA consiste de vários ciclos de investigação, incorporando em todos esses ciclos a recolha e análise de dados e a geração de conhecimento (Given, 2008). Cada ciclo divide-se em várias fases. Há diversas variantes, mas o modelo mais conhecido consiste de 4 fases, que são as seguintes (Figura 6):

- Planificação – corresponde à fase inicial, obtendo-se como resultado o conjunto das ações a serem tomadas pelo investigador ou as alterações pretendidas. Além disso, nesta fase definem-se os limites de tempo durante os quais as alterações se devem tornar efetivas.
- Ação – corresponde à fase da implementação, durante a qual os planos definidos são executados, de forma deliberada, controlada e criticamente informada. Quaisquer novas descobertas no trabalho dos investigadores chegam nesta fase, podendo assim ser incorporadas no projeto atual e podendo também ser utilizadas no futuro.

- Observação – corresponde à fase na qual os dados são recolhidos, os resultados monitorizados e guardados de forma sistemática, de modo a que o investigador possa avaliar os efeitos das ações efetuadas.
- Reflexão – a reflexão constante sobre o trabalho realizado é um aspeto fundamental da IA. O ciclo deve acabar com uma reflexão sobre aquilo que aconteceu, testando-se a eficácia das alterações e também o que se aprendeu. O investigador examina também que barreiras poderão ter dificultado o processo e como é que se pode melhorar a implementação das mudanças no futuro.

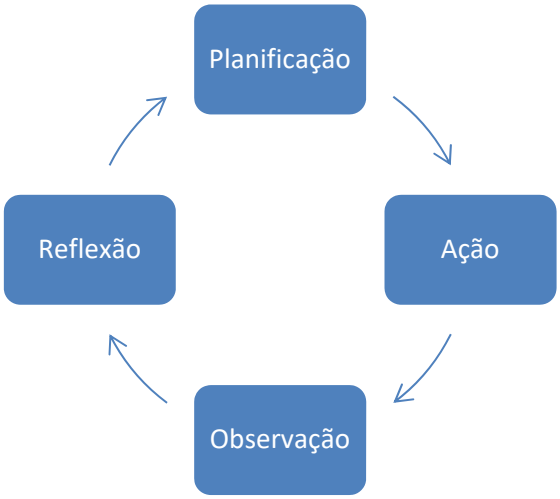


Figura 6: Fases do Ciclos de IA.

5.2. Contexto da Investigação e Participantes

A instituição onde foi desenvolvida toda a investigação descrita nesta tese é uma escola do Instituto Politécnico do Porto (IPP), a saber, o Instituto Superior de Contabilidade e Administração do Porto (ISCAP). O ISCAP é anterior ao IPP, tendo sido criado em 1985. Tem origem numa antiga e reconhecida escola do Porto criada em 1886, o Instituto Industrial e Comercial do Porto. A designação atual de ISCAP foi oficializada em 1975. Somente em 1988 foi integrada no IPP, o qual se insere no Ensino Superior Politécnico.

O foco principal do ISCAP é o ensino de Cursos de Ensino Superior dos 1.º e 2.º ciclos. Além destes cursos, ministra cursos de pós-graduação e formações mais específicas nas áreas da Contabilidade, Gestão, Tradução e Sistemas de Informação.

O objetivo principal do ISCAP e subjacente ao IPP é o seguinte (realce do autor):

> "O ensino politécnico, orientado por uma constante perspetiva de investigação aplicada e de desenvolvimento, dirigido à compreensão e solução de problemas concretos, visa proporcionar uma sólida formação cultural e técnica de nível superior, desenvolver a

> capacidade de inovação e de análise crítica e ministrar conhecimentos científicos de índole teórica e prática e as suas aplicações **com vista ao exercício de atividades profissionais**."(*Lei n.º 49/2005, de 30 de agosto. Segunda alteração à lei de bases do sistema educativo e primeira alteração à lei de bases do financiamento do ensino superior*, 2005, p. 5122)

De momento, o ISCAP ministra 6 cursos do 1.º ciclo e 8 do 2.º ciclo. Os cursos do 1.º ciclo além de serem ministrados em regime diurno, são também ministrados em regime noturno, enquanto que os cursos do 2.º ciclo são todos ministrados em regime noturno. As licenciaturas que fazem parte do 1.º ciclo são as seguintes: Contabilidade e Administração, Comércio Internacional, Marketing, Assessoria e Tradução, Comunicação Empresarial e Gestão de Atividades Turísticas. Os mestrados que fazem parte do 2.º ciclo são os seguintes: Assessoria e Tradução, Empreendedorismo e Internacionalização, Auditoria, Contabilidade e Finanças, Finanças Empresariais, Marketing Digital, Tradução e Interpretação Especializadas e Logística.

Durante os anos em que decorreu esta investigação, de 2008 a 2014, o número de alunos inscritos no ISCAP variou entre 3394 e 3863, tendo vindo a manifestar-se uma evolução crescente, apesar de em alguns dos anos haver um ligeira diminuição, conforme apresentado na Tabela 9.

Tabela 9: Número total de alunos inscritos no ISCAP, por ano letivo

Anos	2008	2009	2010	2011	2012	2013	2014
n.º de alunos	3394	3457	3743	3836	3825	3863	3842

Os cursos que estiveram envolvidos neste estudo foram o curso de Licenciatura em Contabilidade e Administração (LCA) e o curso de Licenciatura em Comércio Internacional (LCI).

O número de alunos que estiveram envolvidos na investigação, entre 2008 e 2014, variou entre 1152 e 959, com poucas variações de ano para ano. Os valores exatos e por semestre encontram-se na Tabela 10.

Tabela 10: Número total de alunos envolvidos na investigação por semestre e por ano letivo

Anos	2008	2009	2010	2011	2012	2013	2014
1.º Semestre	558	721	686	637	608	593	489
2.º Semestre	594	696	610	575	578	569	470
Total	**1152**	**1417**	**1296**	**1212**	**1186**	**1162**	**959**

As Unidades Curriculares (UC) que foram objeto da investigação foram Matemática I da LCI e Matemática da LCA, no 1.º semestre, e no 2.º semestre foram Matemática II da LCI e Matemática

Aplicada da LCA, todas do 1.º ano dos ciclos de estudos. Apesar de terem nomes diferentes, as UC do mesmo semestre apresentam os mesmos conteúdos programáticos, a mesma avaliação e o mesmo funcionamento.

As aulas nestas UC, tanto diurnas como noturnas, foram sempre teórico-práticas, nunca havendo distinção formal entre aulas teóricas e práticas. A carga horária em todas as UC em estudo foi de 90 minutos por aula, duas vezes por semana.

Os conteúdos programáticos do 1.º Semestre, apesar de um ou outro pequeno ajuste ao longo dos anos, consistem essencialmente de Funções Reais de Variáveis Reais e Álgebra Linear. Em termos genéricos, os conteúdos abordam os seguintes tópicos:

- **Funções Reais de Variável Real** (Definição, Domínio, Operações com Funções, Funções Especiais, Limites, Continuidade, Cálculo Diferencial, Derivação da Função Implícita, Diferenciais, Aplicação do Cálculo Diferencial ao Estudo de Funções, Aplicação do Cálculo Diferencial em Ciências Empresariais);
- **Funções Reais de Várias Variáveis Reais** (Definição, Domínio, Limites, Continuidade, Derivadas Parciais, Aplicações em Ciências Empresariais);
- **Álgebra Linear** (Matrizes, Determinantes).

No Anexo D encontra-se em mais detalhe o programa destas UC.

Os conteúdos programáticos do 2.º Semestre, apesar de um ou outro pequeno ajuste ao longo dos anos, consistem essencialmente de Integrais, Cálculo Combinatório e Séries. Em termos genéricos, os conteúdos abordam os seguintes tópicos:

- **Cálculo Integral** (Integral Indefinido, Integral Definido, Integral Múltiplo, Aplicação do Cálculo Integral na Resolução de Problemas de Economia);
- **Análise Combinatória** (Introdução, Métodos de Contagem, Triângulo de Pascal. Binómio de Newton);
- **Séries Numéricas Reais** (Sucessões, Séries e Convergência).

No Anexo E apresenta-se em mais detalhe o programa destas UC.

Os objetivos gerais comuns a estas UC eram:

- Proporcionar aos alunos um conjunto de métodos matemáticos básicos indispensáveis ao sucesso em Ciências Empresariais.
- Proporcionar aos alunos uma aprendizagem de conteúdos matemáticos básicos e a partir destes desenvolver a capacidade de raciocinar, aprender e aplicar conteúdos mais elaborados, nas várias unidades curriculares.
- Facilitar a compreensão e aprendizagem dos alunos, através de uma abordagem intuitiva, ilustrando os diversos tópicos com um ou mais exemplos de aplicação relevantes.
- Incentivar os alunos a usar ferramentas computacionais para resolver alguns dos exercícios propostos.

Em relação à avaliação existem dois regimes, a saber, o regime de avaliação contínua e o regime de avaliação final. Os alunos, salvo duas situações particulares que são identificadas no capítulo seguinte, puderam sempre optar pelo regime de avaliação contínua ou pelo regime de avaliação por Exame Final. Se optassem pelo regime de avaliação contínua, caso reprovassem poderiam fazer somente um Exame em época de Recurso, caso optassem somente pelo regime de avaliação final poderiam fazer dois exames à UC, um na Época Normal e outro na Época de Recurso.

O regime de avaliação final, em ambas as épocas, consiste de um exame com QEM em formato papel contendo 20 questões. No regime de avaliação contínua são também apresentadas aos alunos 20 questões, sendo estas questões distribuídas por dois ou mais testes ao longo do semestre. Em todos os anos do estudo foram sempre considerados, para além dos testes, outros parâmetros de avaliação para os alunos que escolhiam a avaliação contínua. Estes parâmetros apresentavam variações conforme o ciclo de IA, sendo os detalhes apresentados no próximo capítulo. Em termos globais, estes parâmetros envolveram trabalhos de casa, assiduidade e participação.

O número de docentes foi sensivelmente constante ao longo de todo o estudo, sendo de 6/7 em cada semestre letivo. A
Tabela 11 apresenta a caraterização dos vários docentes que com regularidade lecionaram as UC aqui estudadas.

Tabela 11: Breve caraterização dos docentes que lecionaram as UC no decurso da Investigação

Professor	Grau Académico no início do estudo	Anos de Ensino Superior	Anos na(s) unidade(s) curricular(es)
1	PhD	> 20	> 20
2	Lic	> 30	> 30
3	Lic	> 30	> 30
4	PhD	> 20	> 20
5	PhD	> 20	> 10
6	PhD	> 20	> 20
7	MSc	> 20	> 20
8	MSc	> 20	> 20
9	MSc	> 20	> 20
10	MSc	> 20	> 20
11	MSc	> 10	> 10

Como se pode comprovar, a maioria dos docentes apresentava mais de 20 anos de docência no Ensino Superior e lecionavam estas UC, ou semelhantes em conteúdos, há mais de 20 anos, pelo que se trata de docentes com vasta experiência no Ensino Superior.

5.3. Desenho da Investigação

Foram implementados 3 Ciclos de IA, tendo em cada ciclo sido seguidas todas as fases que caracterizam a IA, nomeadamente, Planificação, Ação, Observação e Reflexão. Em seguida descrevem-se sucintamente os 3 ciclos de IA considerados.

5.3.1. 1.º Ciclo de IA – Implementando uma estratégia de avaliação contínua com *e-assessment*

O 1.º ciclo de IA decorreu ao longo dos anos letivos de 2008/09, 2009/10 e 2010/11.

Com a implementação do Processo de Bolonha no ISCAP foi necessário colocar em prática um regime de avaliação contínua, ou mais apropriadamente, uma avaliação distribuída. A palavra "distribuída" é aqui usada no sentido de que há vários momentos de avaliação sumativa e/ou formativa distribuídos ao longo do semestre. Daqui em diante, o termo avaliação contínua será usado neste sentido de avaliação distribuída, tendo em conta que também há momentos para avaliação formativa, sendo alguns deles tanto formativos como sumativos.

Para os professores de Matemática, envolvidos neste estudo, a concretização da avaliação contínua após as alterações decorrentes da implementação do processo de Bolonha constituía uma tarefa difícil, uma vez que duas situações antagónicas se verificaram. Em primeiro lugar, a duração das UC foi substancialmente reduzida, mantendo a necessidade de ensinar quase os mesmos tópicos a fim de fornecer atempadamente os fundamentos matemáticos necessários para outras UC. Em segundo lugar, o número de alunos por turma aumentou devido a limitações orçamentais, o que conduziu à existência de turmas numerosas. Estamos a falar de cerca de 800 alunos, distribuídos por turmas de cerca de 40 ou mais alunos. O tempo necessário para os professores fazerem avaliação é proporcional ao número de alunos. Assim, a existência de turmas numerosas desencoraja os professores a implementarem processos de avaliação contínua.

Para resolver este problema, foi estudada uma estratégia de *e-assessment*. Esta incluiu uma componente que consistiu no uso de testes QEM, pelas vantagens já referidas anteriormente (Secção 2.1. do Capítulo 2). Uma vez que o LMS de código aberto *Moodle*[15] já estava disponível na instituição, ele surgiu como a escolha natural para implementar os testes QEM. O *Moodle* tem a vantagem de ser capaz de gerar testes aleatoriamente através da seleção de um número fixo de perguntas, existentes num banco de questões divididas em categorias, definidas pelos professores. É assim apresentado a cada aluno um teste diferente, evitando a necessidade de desenvolver vários testes distintos. Uma vez que é mais fácil os alunos copiarem em testes QEM, apresentar versões diferentes de testes QEM para alunos adjacentes é importante.

No sentido de implementar toda uma estrutura de avaliação adaptada às novas condições e treinar os alunos para o uso da plataforma *Moodle*, começou-se por desenvolver um conjunto

[15] https://moodle.org/

de trabalhos de casa que consistiam em testes implementados através do *Moodle*. Esta componente de *e-assessment* foi tanto de natureza formativa como de natureza sumativa. A natureza sumativa esteve presente porque os testes tiveram um peso nas notas finais obtidas pelos alunos. A natureza formativa esteve presente porque os testes foram respondidos 7 a 10 dias antes dos testes formais para ajudar os alunos a verificarem e a terem consciência dos seus conhecimentos, possibilitando a autorregulação da aprendizagem. Os testes formais foram a outra componente de avaliação, a qual incluiu 3 testes QEM respondidos pelos alunos em formato papel, ao mesmo tempo para todos os alunos, numa data previamente agendada pela escola. É importante dizer que estes testes também podem ser considerados como um tipo de *e-assessment*, já que as notas foram obtidas e armazenadas em suporte eletrónico. O suporte eletrónico consistiu num arquivo MS Excel™ com fórmulas adequadas. As respostas dos alunos foram convertidas para formato eletrónico, as notas foram geradas e armazenadas automaticamente neste formato e foram calculadas as estatísticas relativamente aos testes. É importante referir que a apresentação de testes diferentes a cada aluno é também um problema nestes testes QEM em formato escrito. Pelo menos 8 versões diferentes foram necessárias para cada teste.

A implementação foi feita gradualmente de forma a testar cuidadosamente o sistema. A primeira etapa consistiu no desenvolvimento de um banco de QEM. Este banco de QEM foi cuidadosamente planeado e implementado de modo que os testes fossem gerados aleatoriamente pelo *Moodle*, permitindo que cada aluno tivesse um teste diferente mantendo, mesmo assim, uniformidade. A construção do banco QEM é discutida na Secção 8.1 e os resultados da sua implementação são apresentados na Secção 9.2.

De seguida, decidiu-se implementar 3 testes como trabalhos de casa, durante períodos específicos previamente definidos e comunicados aos alunos. Inicialmente, era pretendido fazerem-se os testes durante as aulas, mas não havia condições técnicas para tal. Devido a estas restrições, estes testes QEM implementados com o *Moodle* tiveram inicialmente um peso de 10%, sendo que os 3 testes QEM em formato papel tiveram os restantes 90%. A decisão, para os pesos, tomada pelos docentes, foi considerada boa já que esta foi a primeira vez que o banco de QEM foi utilizado e algumas situações incontroláveis poderiam acontecer. Além disso, os testes foram opcionais para os alunos e estes responderam fora do ambiente da sala de aula. Como afirmado anteriormente, o seu objetivo era principalmente servir como avaliação formativa. No entanto, foi também considerado que a atribuição de um peso, ainda que pequeno, para a avaliação sumativa poderia servir como incentivo para os alunos realizarem os testes. Na seção 9.3 apresentam-se as taxas de resposta ao longo dos anos dos alunos a estes testes opcionais. Este formato foi mantido durante 3 anos académicos, mas com algumas mudanças pequenas no seu formato e no peso.

Os docentes responderam a um questionário com o qual se pretendia efetuar uma análise sobre as linhas de orientação para o desenvolvimento de questões de escolha múltipla, apresentadas

por Haladyna e colaboradores (2002). A descrição do questionário é feita na secção 8.2 e os resultados são apresentados na secção 9.4.

Após a conclusão deste ciclo de IA, foi feita uma reflexão, em reunião de professores realizada para o efeito, algumas conclusões foram retiradas e foram apresentadas sugestões de alterações para serem implementadas em anos posteriores:

- os docentes envolvidos consideraram que após a experiência destes 3 anos se poderia e deveria implementar um processo de *e-assessement* para avaliação contínua sumativa com o apoio do *Moodle*, substituindo os testes em formato papel pelos realizados no *Moodle*;
- verificou-se que alguns alunos referiam que as questões que tinham realizado em casa eram todas elas complexas e outros referiram mesmo que as suas questões eram muito mais difíceis que as de outros colegas;
- foi sugerido que, atendendo ao ponto anterior, se deveria fazer uma análise mais aprofundada das questões para averiguar a sua equidade e qualidade;
- foi sugerido que se deveriam estudar melhor as várias opções do *Moodle* de modo a evitar fraudes na realização dos testes.

5.3.2. 2.º Ciclo de IA – Implementação de uma estratégia de e-assessement para avaliação contínua sumativa

O 2.º ciclo de IA decorreu nos anos letivos 2011/12 e 2012/13 para ambas as UC de cada semestre, e acrescentou-se, no ano letivo 2013/14, somente as UC do 1.º semestre.

A maior mudança introduzida neste 2.º ciclo foi a realização de testes sumativos gerados aleatoriamente pelo *Moodle,* para avaliação formal, realizados na escola, mas fora da respetiva sala de aula. Recorde-se que durante o 1.º ciclo de IA os testes eram gerados também no *Moodle*, mas realizados como trabalhos de casa. Neste 2.º ciclo de IA, os testes *Moodle* vieram substituir os testes em formato papel que se realizavam no 1.º ciclo de IA.

Este ciclo consistiu, então, na utilização das QEM para avaliação contínua sumativa presencial, através de testes realizados em período letivo, fora das salas de aula habituais das turmas. Estas salas consistiam em 4 grandes salas especiais destinadas à realização de avaliações, as quais não eram normalmente destinadas à lecionação. Para a implementação desta avaliação, levantaram-se dois tipos de questões:

- Tecnológicas – falta de equipamentos para os alunos realizarem os testes, nível de segurança da rede da escola e da plataforma *Moodle*, capacidade dos servidores para responder a um nível de serviço elevado;
- Científico-pedagógicas – definição/ajustamentos das categorias das QEM, número de questões por categoria, nível de dificuldade de cada questão, uniformidade dos testes gerados aleatoriamente pelo *Moodle*.

É relevante referir aqui que, muitos docentes não se sentiam motivados a realizar testes sumativos nos próprios horários de aulas, com as condições normais de avaliação. Algumas razões foram apontadas:

- número elevado de alunos por turma;
- não disponibilidade de salas com capacidade suficiente durante o tempo letivo normal;
- necessidade de um grande número de diferentes versões de um mesmo teste para a mesma turma e, mais ainda, para turmas que realizam teste em horários diferentes;
- não existência de espaço suficiente entre alunos para que estejam concentrados no seu próprio teste e não no teste do colega ao lado;
- dificuldade de vigilância do teste por um único docente;
- número elevado de correções de testes para um docente com muitas turmas.

Evidentemente, a maioria dos problemas apontados surgiram devido ao número elevado de alunos por turma.

Para resolver estes problemas e como tivemos sempre como objetivo a realização de testes na sala de aula, os testes gerados aleatoriamente pelo *Moodle* surgiram naturalmente como uma solução para alguns dos problemas apontados. Para facilitar a correção e de modo a que todo o processo de avaliação fosse mais rápido, uma vez que não existiam no ISCAP computadores em número suficiente, foi necessário encontrar uma solução, a qual passou pela realização destes testes nos computadores pessoais dos alunos. Levantaram-se aqui uma série de problemas de segurança, e não só, que em termos genéricos denominamos de Problemas Tecnológicos. Vejamos com mais detalhe os problemas e soluções encontradas relativamente a estes Problemas Tecnológicas.

Em relação ao uso dos computadores na avaliação, inicialmente verificámos ser impossível usar os computadores da escola, que existem em determinadas salas, porque além de estas salas serem em número reduzido, estavam quase sempre ocupadas com aulas da Área Científica de Informática, bem como de outras áreas científicas que também usam estas salas para a manipulação de aplicativos específicos de apoio à lecionação. Ainda foi considerada a possibilidade de equipar um conjunto de salas com computadores da escola, mas não foi possível conseguir verbas para tal.

A solução encontrava-se então na utilização dos computadores portáteis dos próprios alunos. Verificámos, conforme a "Resolução do Conselho de Ministros n.º 137/2007" que o Plano Tecnológico da Educação teve o seu início em 2007 e com ele o programa e.escola[16] (entre outros programas). Este programa englobava inicialmente somente os alunos do 2.º e 3.º ciclos

[16] O Programa e.escola visava promover o acesso à Sociedade da Informação e fomentar a info-inclusão, através da disponibilização de computadores portáteis e ligações à internet de banda larga, em condições vantajosas. Os alunos mais carenciados, através da Ação Social, dependendo dos escalões em que se encontravam, não pagavam um valor inicial pela aquisição de um computador, mas pagavam somente 5 euros por mês durante 3 anos para o acesso à Internet. O endereço web http://eescola.pt/ relacionado com o programa ainda está disponível com as informações à altura do projeto.
Este programa terminou em 2011, mas foi extinto formalmente somente em 2015.

do ensino básico. Em 2008, a "Resolução do Conselho de Ministros n.º 51/2008" determinava a continuidade do Programa e.escola e a sua extensão aos alunos do Ensino Secundário. Mais tarde este programa foi estendido ao Ensino Superior. Ora, estes alunos em 2011, ano do começo do 2.º ciclo de IA, já todos teriam um portátil pessoal. Assim sendo, tínhamos um grande problema resolvido: poderíamos usar os portáteis pessoais dos alunos para fazerem a avaliação *online* via *Moodle*. Esta solução permitia o uso do horário das aulas para realização dos testes.

Com esta solução, levantaram-se aqui uma série de problemas de segurança, como por exemplo a necessidade de assegurar que os alunos não tivessem acesso à Internet e consequentemente a informação não autorizada e a outras informações no computador pessoal. Estes problemas não seriam fáceis de resolver.

Quanto à questão de contornar o problema de os alunos tentarem aceder a informação não autorizada, a solução consistiu na definição de um tempo limitado adequado para a realização do teste. O tempo limite que se decidiu colocar para a realização de cada teste, depois de devidamente testado e calculado em experiências piloto com alunos de 3 turmas, não permitia ao aluno ter tempo para poder estar constantemente a procurar e a ler informação relevante sobre os conteúdos que eram avaliados nos testes. Permitimos, no entanto, que o aluno levasse consigo um formulário manuscrito, por ele elaborado, com as fórmulas que considerasse serem necessárias. Definimos ainda uma outra forma para contornar estes tipos de fraude, que passou pela colocação do professor sempre no fundo da sala de aula, nas costas dos alunos, de modo a que se pudesse ter um maior controlo de todos eles. Claro que os alunos também sabiam que qualquer tentativa de fraude teria como punição serem excluídos da avaliação contínua. Tentou-se ainda usar uma opção de segurança do *Moodle* para que não fosse permitido abrir outra janela do *browser*, mas tivemos alguns problemas com o uso dessa opção que envolvia "JavaScript". Para reforçar a segurança na realização de testes e ajudar à resolução de alguns problemas detetados anteriormente, implementaram-se ainda outras soluções, que a seguir se apresentam.

Quanto à questão de não permitir que os alunos acedessem à Internet, uma das soluções implementadas foi uma rede *wireless* específica para a avaliação contínua através do *Moodle*. Esta solução garantia, em primeiro lugar, uma rede rápida, sem problemas de acesso e sem falhas na ligação; em segundo lugar, uma maior segurança de modo a não se ter a intromissão de agentes externos a aceder a esta rede exclusiva para a avaliação contínua. Assim, neste ciclo de IA foi solicitada à Presidência da escola a implementação de uma rede interna, gerida pela escola e não pelos Serviços Centrais do IPP, via *Routers* específicos de acesso aos servidores, os quais conteriam um *Moodle* especialmente estruturado para realizar somente avaliações *online* – o *Moodle* de Avaliação. Esta rede assim criada não permitia o acesso à Internet ou a qualquer outro sítio a não ser ao *Moodle* de Avaliação.

Para a implementação desta rede informática, foram realizadas as seguintes etapas implementadas pelo Gabinete Técnico de Informática do ISCAP e pelo atual GAIE (Gabinete de Apoio à Inovação em Educação)[17]:

- configuração do Servidor de modo a possibilitar a sua identificação na Intranet e não permitir qualquer tipo de acesso por parte dos alunos a não ser ao *software* específico para avaliação;
- instalação e configuração de uma versão do *Moodle* unicamente dedicada à avaliação; os alunos apenas podiam aceder ao teste quando estava visível e nada mais podiam fazer neste *Moodle*;
- importação das informações de cada aluno, diretamente da Secretaria Online[18], de modo a que o aluno pudesse aceder ao *Moodle* com as suas credenciais, da mesma forma que acedia a qualquer outro serviço da escola;
- definição do acesso a esta plataforma de modo a não permitir qualquer alteração do perfil de aluno;
- instalação de *Routers* nas salas de aula onde se iriam realizar os testes;
- atribuição de um endereço IP[19] fixo a cada um dos *Routers*;
- instalação nas salas de aula, selecionadas para a implementação desta avaliação, de extensões elétricas por mesa/secretária dos alunos para que estes pudessem ligar os portáteis.

Foram equipadas 4 salas com capacidade máxima para aproximadamente 140 alunos a realizar a avaliação *online* simultaneamente.

Foram ainda considerados outros aspetos relacionados com a segurança, nomeadamente:

- inserção dos IPs dos *Routers* num campo específico aquando da criação do teste no *Moodle*, conforme apresentado na Figura 7 (campo "Exigir endereço de rede");
- colocação de uma senha por teste/turma no *Moodle*, conforme apresentado na Figura 7 (campo: "Exigir senha");
- definição de períodos específicos para abertura e encerramento por teste/turma, conforme apresentado na Figura 8.

[17] O GAIE, cuja sigla anterior era PAOL, tem por missão promover a conciliação das necessidades pedagógicas inerentes ao ensino superior com a eficiente introdução e utilização da tecnologia ao serviço da educação, sendo responsável pela manutenção do *Moodle*.
[18] A Secretaria Online é um serviço que permite que docentes e alunos possam aceder a um vasto conjunto de informações relacionadas com as aulas em si (horário, mapa das aulas, notas, plano do curso, inscrições), bem como a toda a área de secretariado (pedido de certidões/documentos, inscrição em exames, pagamento de propinas, reclamações de notas).
[19] Protocolo de Internet, em Inglês, *Internet Protocol* (IP) é, em termos simples, a atribuição de um conjunto de números, num determinado formato, para identificação de um computador ou impressora ou outro equipamento numa rede local ou pública de modo a poderem comunicar entre si.

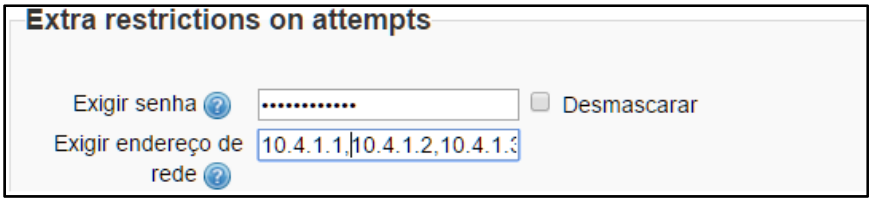


Figura 7: Exemplo de dois campos no *Moodle* para restrição de acessos não autorizados.

Figura 8: Exemplo de campos para controlo de tentativas de acesso fora de horas das aulas, com indicação do tempo limite para terminar o teste.

Sem dúvida que o facto de o aluno ter de estar ligado a um dos *Routers* específicos na sala, para realizar o teste, bem como a implementação dos outros controlos descritos acima, permitiu que a fraude tendesse para zero.

Acrescentem-se alguns dos cuidados, mais específicos, na configuração do *Moodle* que se devem ter em conta para evitar outro tipo de tentativas de fraude:

- impossibilitar a alteração do nome e do n.º de utilizador;
- impossibilitar a consulta do perfil de qualquer outro utilizador da plataforma;
- impossibilitar o acesso à lista de utilizadores autenticados;
- impossibilitar a consulta de listagens de colegas inscritos em páginas de disciplinas;
- desativar quaisquer mecanismos de comunicação interna no *Moodle*, tais como: *chat*, sistema de mensagens, fóruns e *blogs* pessoais.

Na prática, é importante, conforme a versão do *Moodle*, desativar todas as opções que não sejam necessárias à realização do teste e não permitir fazer qualquer tipo de outras alterações que o *Moodle* ou perfil de entrada das credenciais permitam. Desta forma minimiza-se a possibilidade de os alunos cometerem fraudes.

Durantes os anos que compuseram este ciclo de IA, começaram a surgir no mercado os *Tablets*. Apesar de em 2011 ninguém ter pedido para usar os *Tablets* nos testes, em 2012 surgiram vários alunos a solicitar o seu uso. Contudo, sabíamos que seria mais difícil controlar os acessos à Internet por parte destes alunos, pois este tipo de equipamentos, na maioria dos casos, permitia o acesso a dados via cartão *SIM*. Assim, sabendo já em 2013 da existência de muitos alunos com este equipamento, decidiu-se averiguar da possibilidade de bloquear o sinal de telemóvel. Contudo, verificada a legislação nacional, não foi permitido o seu bloqueio. Apesar destes

constrangimentos e da possibilidade de poder existir alguma fraude, foi-se permitindo o seu uso, em especial nos últimos anos, mas apenas desde que, numa mesma turma, o número de alunos com *Tablets* fosse razoável de modo a que o docente tenha possibilidade de um controlo maior destes mesmos alunos. Apesar de muitos alunos solicitarem o uso do telemóvel ou *smartphone*, em especial nos últimos anos, o seu uso nunca foi permitido nos testes. Ainda que tenham existido poucos alunos com *Tablets*, decidiu-se avançar um pouco mais no controlo das fraudes, visto que este número tendia a aumentar. Verificou-se que começava a haver na literatura alguns artigos a relatarem problemas de fraude no acesso ao *Moodle* e, após a análise de alguns deles, centrámo-nos no trabalho desenvolvido por Matos, Torrão e Vieira (2012), já que apresentava alguns problemas que iam de encontro às nossas preocupações e, simultaneamente, apresentava uma solução para a maioria desses problemas. A solução passava por instalar um suplemento ao *Moodle* desenvolvido pelos autores. Apesar de não necessitarmos de todas as opções da aplicação referida, pensou-se em fazer algo semelhante para usarmos na escola. Assim, foi desenvolvido um suplemento para o *Moodle* por intermédio do GAIE com o nome "Unique login". Este suplemento permitia:

- evitar que dois ou mais estudantes se autenticassem no *Moodle* com as mesmas credenciais de acesso;
- a visualização rápida do IP por parte do docente e, também, saber a localização do *Router* a que o aluno se tinha ligado;
- o encerramento automático de sessão no *Moodle*, após X minutos de inatividade (se o utilizador não clicasse em nada);
- acesso por parte do Docente a um painel que permita visualizar todos os utilizadores ativos, com indicação acerca da data/hora de acesso ao *Moodle* e de um botão que permita ao professor encerrar a sessão do aluno, caso fosse necessário.

Apesar de o *Moodle* ter, por defeito, algumas das informações que o "Unique login" fornece, com este suplemento foi mais fácil a consulta e controlo dos acessos dos alunos. Este suplemento veio permitir um controlo mais efetivo e minimizar várias possibilidades de fraude (por exemplo, foi possível identificar estudantes a cometerem fraudes).

Com todas as alterações realizadas, a saber, equipamentos nas salas, implementação da rede informática unicamente para avaliação contínua e configurações gerais do Servidor e do *Moodle*, foram criadas as condições tecnológicas necessárias para garantir a realização dos testes sumativos para a avaliação contínua. Realça-se um procedimento importante que foi implementado e que era obrigatório realizar-se todos os anos e para todos os alunos que escolhessem o regime de avaliação contínua, nomeadamente a realização de um teste de simulação. Este teste aplicava-se aos alunos que pela primeira vez eram submetidos a este tipo de avaliação e devia-se realizar com o devido tempo de antecedência em relação à marcação do 1.º teste sumativo. O seu principal objetivo era dar ao aluno capacidade para realizar o 1.º teste sem qualquer tipo de dificuldade. Este teste de simulação, de modo a não comprometer o número de aulas dedicadas à lecionação, era realizado fora do horário normal de aulas. Para

que esta simulação fosse o mais possível perfeita, os alunos eram devidamente avisados do horário do teste e duração do mesmo. Era, também, deixado no *Moodle* das UC um documento que tinham de imprimir e levar para o dia do teste de simulação. Para ajudar a resolver problemas técnicos relacionados com a rede informática tinha-se sempre o apoio de um técnico de Informática da escola que era o responsável pela manutenção da rede. Contava-se ainda com o apoio de um elemento do GAIE.

A folha que os alunos levavam impressa para o teste de simulação e que se aconselhava fosse previamente analisada, continha várias instruções, entre as quais, a forma de acesso à rede de avaliação e *Moodle* de Avaliação. No Anexo F encontram-se os detalhes das instruções fornecidas aos alunos.

No final do teste de simulação os docentes faziam um levantamento dos problemas encontrados para, em reunião, se produzir um relatório sobre este teste. O relatório continha os problemas encontrados e resolvidos, e em especial os problemas que não se conseguiam resolver e para os quais urgia encontrar uma solução. Este relatório era depois enviado aos departamentos competentes da escola, para que pudessem resolver os problemas a tempo de realizar o teste sumativo. O exemplo de um destes relatórios encontra-se no Anexo G.

Os problemas no acesso à rede e ao *Moodle* nos últimos anos foram praticamente inexistentes, mas apresenta-se no Anexo G o primeiro relatório realizado no final do 1.º teste de simulação e primeiro ano deste ciclo de IA. Como se pode notar, à altura existiam ainda muitos problemas para serem resolvidos antes do 1.º teste sumativo. O responsável pela gestão da rede informática e o responsável pela gestão do *Moodle* foram informados dos horários de realização dos testes de avaliação sumativa. Apesar de atualmente continuarmos a contar com o apoio destas duas áreas da escola, os docentes já conseguem resolver a maioria dos problemas que vão surgindo, dado que aprenderam a lidar com eles.

O segundo tipo de problemas para serem resolvidos eram de natureza científico–pedagógico e tinham a ver essencialmente com a qualidade das QEM do banco de questões. Com vista a apresentar uma solução para os problemas encontrados no 1.º ciclo de IA, relacionados com a qualidade das questões e com a uniformização dos testes apresentados pelo *Moodle* a cada aluno, foram analisadas as QEM que compunham o banco de questões do *Moodle* utilizando a Teoria Clássica de Testes (TCC) e a Teoria de Resposta ao Item (TRI). A descrição dos instrumentos utilizados encontra-se na secção 8.3 e os resultados encontram-se na secção 9.4.

Os resultados destas análises permitiram definir um conjunto de mudanças nas questões, tentando assim obter testes mais uniformes e mais justos. Atendendo aos resultados verificou-se ser necessário eliminar algumas questões e manter outras, criando níveis de dificuldade.

Refira-se que alguns dos problemas e soluções implementadas neste 2.º ciclo de IA e aqui descritos, não se colocaram todos no mesmo ano. No entanto, as condições mínimas para o

arranque dos testes foram garantidas no primeiro ano deste ciclo antes da realização do 1.º teste sumativo, que se realizou em meados de outubro.

Após a conclusão deste ciclo de IA, foi feita uma reflexão, em reunião de professores realizada para o efeito, e algumas conclusões foram retiradas e apresentadas sugestões de alterações, a serem implementadas em anos posteriores:

- houve uma evolução positiva nas classificações dos alunos (como se pode comprovar na análise apresentada na secção 9.1);
- seria importante que os alunos realizem os testes na sala onde decorrem normalmente as aulas, de modo a evitar grandes turbulências;
- houve necessidade de apetrechar um conjunto de salas de aula onde normalmente decorria a lecionação com as condições para que se pudessem realizar os testes sumativos de avaliação contínua e atribuir estas salas aos docentes no horário que lecionavam as UC que eram objeto desta avaliação;
- decidiu-se criar um teste, a que apelidamos de "Repescagem", para os alunos cuja classificação final seja inferior a 10 valores ;
- necessidade de se encontrarem formas de minimizar as fraudes por parte dos alunos na realização dos testes.

5.3.3. 3.º Ciclo de IA – Análise de mudança nas práticas educativas

O terceiro ciclo de IA corresponde aos anos letivos 2013/14 e 2014/15 para as UC que eram lecionadas no 2.º semestre e corresponde somente ao ano letivo 2014/15 para as UC do 1.º semestre. Conseguiu-se que no 2.º semestre de 2013/14 as salas que pretendíamos já estivessem devidamente equipadas para a realização dos testes. Este ciclo consistiu na execução de melhoramentos no processo de avaliação e na realização dos testes durante as aulas e nas salas habituais das turmas. Dois aspetos essenciais caracterizaram o 3.º ciclo de IA: i) realização dos testes sumativos de avaliação contínua na sala de aula normal de cada turma; ii) realização de Testes de "Repescagem".

Como foi descrito na reflexão do ciclo anterior, era importante que os alunos fossem avaliados na própria sala de aula, não tendo necessidade de se deslocarem para outras instalações, que era o que acontecia no ciclo anterior. Assim, foi proposto à escola que se apetrechassem mais algumas salas de aula, que eram usadas regularmente para lecionação, com condições iguais às 4 salas que em anos anteriores tinham sido usadas na nossa avaliação. Assim, foram equipadas mais 14 salas com este tipo de equipamento. Foi pedido aos responsáveis pelos horários que na atribuição de salas de aula aos docentes que lecionavam as UC da Área Científica de Matemática com esta forma de avaliação, lhes fossem atribuídas determinadas salas (as que continham as condições necessárias para a realização da avaliação), sendo essa indicação dada pelo Coordenador da Área Científica de Matemática. Na primeira reunião de docentes neste novo ciclo de IA decidiu-se aprofundar as condições necessárias à realização do Teste de "Repescagem" que tinha sido sugerido no final do ciclo de IA anterior. Assim, foram definidas

as condições para que os alunos pudessem realizar este teste nomeadamente, os alunos teriam de ter realizado todos os testes e não ter conseguido classificação final positiva. Apenas estes poderiam fazer um Teste de "Repescagem" na última semana do semestre. Esse teste substituiria um dos testes de avaliação contínua, sendo a classificação final recalculada. A escolha deste teste era feita no próprio dia do teste, não tendo os alunos necessidade de indicar com antecedência qual o teste a realizar. Observe-se que os alunos que faltassem a algum dos testes não poderiam fazer o Teste de "Repescagem".

Para controlar os alunos que poderiam ou não fazer o teste e para que as classificações fossem automaticamente recalculadas, foi definido um conjunto de procedimentos implementados em MS Excel™ para este efeito. As informações do *Moodle* eram exportadas em formato MS Excel™ e, depois de eliminada alguma informação desnecessária, essas informações eram copiadas para determinadas folhas de cálculo. Era automaticamente verificado se o aluno tinha efetuado somente um teste ou não, e ainda se estava nas condições assinaladas (em caso afirmativo, a classificação final era recalculada automaticamente).

Em termos de avaliação geral, como se poderá verificar na secção 9.1, houve claramente melhores médias e percentagens de classificações positivas dos alunos. Observou-se ainda uma menor desistência da avaliação contínua por parte dos alunos.

Para avaliação deste 3.º ciclo, foram ainda efetuadas entrevistas a docentes e um questionário aos alunos. Foi feita a análise dos dados recolhidos, de modo a aferir as mudanças nas práticas educativas resultantes da implementação deste processo de avaliação através da realização de testes com QEM implementados na plataforma *Moodle*.

Após a conclusão deste ciclo de IA, foi feita uma reflexão, em reunião de docentes realizada para o efeito, e algumas conclusões foram retiradas e apresentadas sugestões de alterações, a serem implementadas em anos posteriores:

- houve um aumento considerável de alunos que assistiam às aulas até final do semestre, em particular no 2.º;
- houve uma melhoria considerável das classificações;
- haverá ainda necessidade de reavaliar as questões através da TRI;
- dever-se-á melhorar o controlo à fraude no que diz respeito ao uso dos telemóveis;
- dever-se-ão resolver alguns problemas com a versão do *Moodle*, no que concerne ao uso do TeX, instalando uma nova versão.

Na Tabela 12 encontra-se um resumo dos 3 ciclos de IA desenvolvidos nesta investigação, considerando-se as diversas fases de cada um deles.

Tabela 12: Resumo dos 3 ciclos de IA

	1.º Ciclo	2.º Ciclo	3.º Ciclo
Planificação	Foi diagnosticado, em reuniões de área científica, o problema de existir um número demasiado elevado de alunos, o que dificultava a realização de avaliação contínua de forma justa e eficaz; Foi definida uma estratégia de avaliação contínua, utilizando QEM implementadas no *Moodle*, quer como avaliação formativa quer sumativa; Foram definidas regras para a elaboração das QEM e dos testes gerados aleatoriamente pelo *Moodle*; Foi definido um processo de revisão para a elaboração das QEM.	Foi planeada a extensão da utilização das QEM para avaliação sumativa presencial;	Foi planeada a implementação da avaliação sumativa presencial durante o horário normal das turmas;
Ação	Construção de um banco de QEM; Implementação de testes com QEM, pondidos pelos alunos como trabalho de casa. Utilização de um questionário para a análise, por parte dos intervenientes, das linhas de orientação existentes para elaboração de QEM.	Análise das QEM do banco de questões utilizando a TRI e a TCT. Criação das condições tecnológicas necessárias para a realização dos testes presenciais para avaliação sumativa. Ajustamentos e introdução de melhorias no banco de QEM.	Criação das condições tecnológicas e organizacionais necessárias para a realização dos testes presenciais para avaliação sumativa na sala de aula durante o horário normal das turmas; Implementação do teste de "Repescagem" Realização de um questionário aos alunos; Realização de entrevistas aos docentes.
Observação	Análise dos dados obtidos na fase anterior;	Análise dos dados obtidos na fase anterior;	Análise dos dados obtidos na fase anterior;
Reflexão	Análise crítica do processo desenvolvido, por parte do autor da tese e de todos os docentes envolvidos.	Análise crítica do processo desenvolvido, por parte do autor da tese e de todos os docentes envolvidos.	Análise crítica do processo desenvolvido, por parte do autor da tese e de todos os docentes envolvidos.

	1.º Ciclo	2.º Ciclo	3.º Ciclo
	Foi feita a deteção da necessidade de uma análise mais aprofundada das questões desenvolvidas. Possibilidade de implementar um processo de *e-assessement* para avaliação contínua sumativa com o *Moodle*.	Necessidade de apetrechar mais salas com as condições necessárias para a realização dos testes no *Moodle*. Necessidade de realizar um Teste de "Repescagem". Necessidade de encontrar formas para minimizar as fraudes por parte dos alunos.	Verificou-se uma melhoria considerável das classificações. Necessidade de reavaliar as questões com a TRI. Necessidade de melhorar ainda mais o controlo à fraude. Necessidade de resolver problemas com as novas versões do *Moodle*.

5.4. Instrumentos de Recolha de Dados

Nesta secção descrevem-se os instrumentos que foram utilizados durante esta investigação. Na subsecção 5.5.1 começamos por descrever o banco de questões, explanando como foram definidas as categorias para as questões, como foram criados as questões e os testes, como foi efetuado o processo de revisão, caracterizando o banco de questões por ciclo, explicando como foram aplicados os testes opcionais do 1º ciclo de IA e como foram avaliadas as questões usando a Teoria Clássica dos Testes e a Teoria da Resposta ao Item. Os questionários aos docentes no 1.º ciclo de IA apresentam-se na subsecção 5.5.2, a entrevista aos docentes no 3.º ciclo de IA na subsecção 5.5.3 e o questionário aos alunos no 3.º ciclo de IA na subsecção 5.5.4

5.4.1. Banco de questões

Podemos afirmar que a tarefa mais importante da implementação deste processo de *e-assessment* foi a construção do banco de QEM. A evolução do número de questões do banco de questões encontra-se na secção 9.2. Foram considerados três aspetos: definição de categorias para as questões, construção das questões e testes, e processo de revisão. Estes aspetos são explicados de seguida.

5.4.1.1. Definindo categorias para as questões

O *Moodle* pode gerar testes aleatoriamente selecionando um número fixo de questões a partir de categorias ou subcategorias pré-definidas, as quais contêm um determinado número de questões, havendo assim um teste diferente para cada aluno. Isto coloca duas questões importantes:

- como garantir que os testes avaliam os mesmos tópicos para todos os alunos?
- como garantir que os testes têm o mesmo grau de dificuldade para todos os alunos?

Foi definido que a garantia de que os mesmos tópicos são avaliados para todos os alunos poderia ser alcançada com a definição de categorias ou subcategorias nas quais classificar cada uma das perguntas elaboradas, sendo que cada categoria corresponderia a um resultado de aprendizagem. Os resultados de aprendizagem foram cuidadosamente definidos pelo grupo de professores com base nos resultados de aprendizagem que foram definidos para cada uma das UC. Estes foram definidos no início de cada semestre pelo grupo de professores com base nas necessidades dos alunos, mas são essencialmente uniformes ao longo dos anos. Detetou-se que se os testes incluíssem mais de uma pergunta por categoria, o *Moodle* poderia selecionar a mesma pergunta pelo menos duas vezes, o que de alguma forma é comum em testes gerados aleatoriamente pelo *Moodle*. Assim, para evitar este problema, os testes gerados aleatoriamente pelo *Moodle* e apresentados a cada estudante incluem somente uma questão por categoria, de modo a evitar que a mesma questão apareça mais que uma vez no teste. Para obviar este problema de saída de uma pergunta por categoria, foram criadas em algumas delas subcategorias.

Quanto à garantia de que os testes são uniformes em dificuldade para todos os alunos, foi decidido que os professores devem desenvolver perguntas classificando-as desde dificuldade baixa até dificuldade média. As perguntas devem também ser uniformes em formato: por exemplo, não é aceitável ter uma questão com três opções e outra com sete, uma vez que é mais difícil para os alunos analisar esta última.

5.4.1.2. Criando as questões e os testes

Como mencionado anteriormente, é importante definir um formato a ser seguido pelos professores na conceção das questões. Foi definido que todas as questões teriam 4 opções: 1 correta e 3 distratores. Uma penalização de 33% foi introduzida para os distratores, de maneira a tentar evitar que os alunos respondam de forma aleatória. Também foi decidido que a primeira opção deve ser a correta para facilitar possíveis revisões posteriores. Este não é um problema para os alunos, uma vez que o *Moodle* baralha as várias opções antes de apresentar a questão aos alunos nos testes por ele gerados.

Ao gerar os testes no *Moodle* foi prestada atenção especial aos seguintes aspetos:

- gerar um teste diferente para cada turma, definindo a duração, data, hora a que o teste fica disponível e hora a que o teste deixa de ficar disponível;
- foi definido um tempo limite durante o qual o teste está disponível para o aluno responder às diversas questões;
- apenas foi permitida uma tentativa para o aluno resolver o teste;
- os testes foram gerados aleatoriamente pelo *Moodle* através da seleção de uma pergunta de cada uma das categorias predefinidas (cada categoria corresponde a um resultado de aprendizagem predeterminado);
- as opções em cada umas QEM foram misturadas aleatoriamente para cada teste gerado;
- retirar qualquer seleção nas “Opções de Revisão”.

O *Moodle* permite definir todas estas opções e muitas outras na configuração do recurso "Teste" conforme é apresentado na Figura 9.

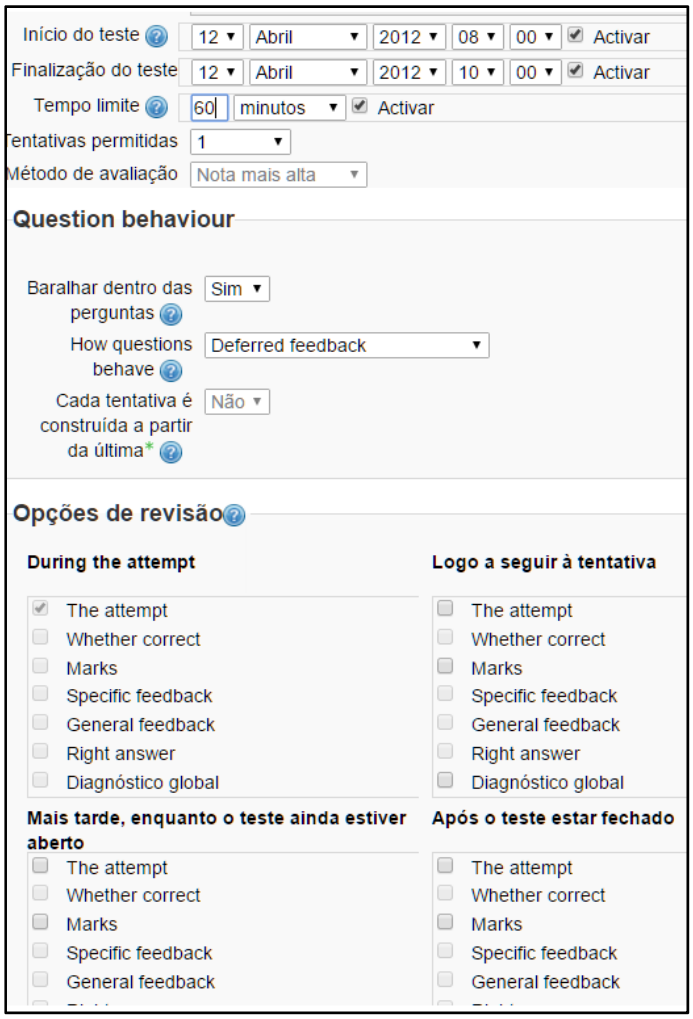


Figura 9: Algumas opções no recurso "Teste" no *Moodle*.

5.4.1.3. O processo de revisão das questões

Produzir questões sem erros é crucial para desenvolver confiança no processo de avaliação em qualquer situação, mas é mais difícil de garantir ao construir QEM do que construir questões de resposta aberta. Um processo de revisão cuidadosa foi desenhado para que os erros pudessem ser minimizados. O processo consistiu nos seguintes passos:

1. foram designados grupos de dois professores com a responsabilidade de preparar um número apropriado de questões para cada uma das categorias definidas;
2. o coordenador do curso avaliou as questões e sugeriu mudanças;
3. a mesma equipa concretizou as alterações e preparou uma resolução detalhada das questões;
4. uma segunda equipa de dois professores analisou as questões e resoluções em detalhe propondo alterações considerando, por exemplo, o tempo necessário para resolvê-las,

o nível de dificuldade em consonância com todas as perguntas da mesma categoria e os erros encontrados;
5. a primeira equipa realizou as alterações;
6. o coordenador do curso analisou a versão final das questões e propôs alterações, nesta fase foram mínimas;
7. a última versão das perguntas foi verificada por todo o grupo e o acordo final foi dado.

É relevante dizer que um processo de revisão semelhante foi seguido para os testes QEM em formato papel (referidos na secção 7.1), incluindo a necessidade de gerar várias versões. Mais tarde, depois dos estudantes responderem aos testes QEM em formato papel, a resolução (não apenas as respostas) de uma das versões foi disponibilizada aos alunos.

Atualmente o processo é mais simplificado. Tendo em conta a experiência anteriormente adquirida, foi possível eliminar as etapas 3, 4 e 6.

Este processo de revisão revelou ser eficaz uma vez que até agora não foram encontrados erros graves nos testes.

5.4.1.4. O Banco de Questões por Ciclos

A implementação do banco de questões pode ser considerada um sucesso, devido ao árduo trabalho de todos os docentes durante todos os anos de implementação do projeto. Na Tabela 13 e Tabela 14 apresenta-se a evolução do número de questões e as respetivas categorias, existentes no banco de questões ao longo dos três ciclos de IA para os cursos do 1.º e do 2.º semestres letivos, respetivamente.

Neste momento, para os cursos do 1.º semestre letivo, existem 17 categorias principais no banco de questões, sendo que 7 dessas categorias estão subdivididas em subcategorias. Entre categorias principais e subcategorias, consideraram-se então 33 categorias como sendo aquelas que são utilizadas para, em cada ano letivo, selecionar 20 delas, diferentes de ano para ano, de modo a que o *Moodle* possa gerar aleatoriamente os testes a apresentar aos alunos. Estas categorias foram definidas atendendo aos objetivos de aprendizagem da UC. Durante o 1.º ciclo de IA foram desenvolvidas 742, correspondendo a uma média de 23 questões por categoria. Durante o 2.º ciclo de IA foram desenvolvidas 730, correspondendo a uma média de 23 questões por categoria. Durante o 3.º ciclo de IA foram desenvolvidas 86, correspondendo a uma média de 3 questões por categoria. Neste momento, o banco de questões contém um total de 1558 questões o que corresponde a uma média de 45 questões por categoria.

Neste momento, para os cursos do 2.º semestre letivo, existem 21 categorias principais no banco de questões, sendo que apenas uma dessas categorias estava subdividida em subcategorias. Entre categorias principais e subcategorias, consideraram-se então 26 categorias como sendo aquelas que são utilizadas para, em cada ano letivo, selecionar 20 delas, diferentes de ano para ano, de modo a que o *Moodle* possa gerar aleatoriamente os testes a apresentar aos alunos. Estas categorias foram definidas atendendo aos objetivos de aprendizagem da UC.

Durante o 1.º ciclo de IA foram desenvolvidas 756, correspondendo a uma média de 29 questões por categoria. Durante o 2.º ciclo de IA foram desenvolvidas 561, correspondendo a uma média de 22 questões por categoria. Durante o 3.º ciclo de IA foram desenvolvidas 23, correspondendo a uma média de 2 questões por categoria. Neste momento, o banco de questões contém um total de 1340 questões o que corresponde a uma média de 51 questões por categoria.

Podemos afirmar que o número de QEM incluídas no banco de questões é bastante elevado, o que permite obter testes diferentes para cada aluno. Cada um destes testes é equivalente a uma versão diferente dos testes em papel. Acresce que o número de erros encontrados é residual ao longo de todos os anos de implementação do projeto. Contudo, esses erros nunca inviabilizaram a realização dos testes, pelo que se podem considerar como pequenas de gralhas. Estas gralhas foram sendo corrigidas pelos professores à medida que foram sendo encontradas.

Tabela 13: N.º de questões elaboradas em cada categoria do banco de questões para os cursos do 1.º semestre letivo

categoria	Número de Questões			
	1.º Ciclo de IA	2.º Ciclo de IA	3.º Ciclo de IA	Total por categoria
(Funções Reais de Variável Real)				
Domínios	30	27	3	60
Funções Tipo I				
Função definida por ramos	9	10	3	22
Função Polinomial	8	6	0	14
Função Racional	7	5	0	12
Operações com funções	20	20	2	42
Funções Tipo II				
Função Exponencial	5	6	1	12
Função Inversa	28	26	4	58
Função Logarítmica	10	9	1	20
Limites				
Sem regra L'Hospital	18	14	3	35
Com regra L'Hospital	29	30	2	61
Continuidade				
Sem regra L'Hospital	30	28	2	60
Com regra L'Hospital	1			1
Cálculo Diferencial (reta tg/normal)	44	41	4	89
Diferencial e Aproximação Linear	24	22	3	49
Aplicação do Cálculo Diferencial ao estudo de Funções				
Monotonia e Extremos	32	30	1	63
Concavidades e Pontos de Inflexão	32	40	1	73
Assímptotas	22	24	1	47
Formas Indeterminadas – Exponencial-Potência (0^0;inf^0;1^inf)	6	4		10
Função Derivada		3		3
F. R. V. R. 2				
Domínios de funções reais de duas variáveis (SubCat.)	31	35	4	70
Derivadas Parciais (SubCat.)	46	39	4	89
Extremos de funções reais de duas variáveis	25	22	10	57

categoria	Número de Questões 1.º Ciclo de IA	2.º Ciclo de IA	3.º Ciclo de IA	Total por categoria
Álgebra Linear				
Operações com matrizes I (explicitadas)	38	34	4	76
Operações com matrizes II (não explicitadas)	39	35	5	79
Sistemas de Gauss-Jordan - Discussão	36	33	4	73
Sistemas de Gauss-Jordan - Resolução	15	12	1	28
Matriz Inversa				
Cálculo da Inversa	20	17	2	39
Determinantes (Matriz Regular/Singular)	21	18	2	41
Equações Matriciais	13	30	1	44
Equações Matriciais Mais Simples	13	10	2	25
Determinantes - Tipo I - Propriedades (explicitando a matriz)	30	29	4	63
Determinantes - Tipo II - Propriedades (não explicitando a matriz)	33	46	6	85
Sistemas - Resolução com Determinantes	27	25	6	58
Total	**742**	**730**	**86**	**1558**

Tabela 14: N.º de questões elaboradas em cada categoria do banco de questões para os cursos do 2.º semestre letivo

categoria	Número de Questões 1.º Ciclo de IA	2.º Ciclo de IA	3.º Ciclo de IA	Total por categoria
Integrais Indefinidos - Imediatos sem valor inicial	44	36		80
Integrais Indefinidos - Imediatos com valor inicial	32	28		60
Integrais Indefinidos - por Partes	33	24	3	60
Integrais Indefinidos - Substituição	13	14	5	32
Integrais Indefinidos - Racionais	28	20		48
Integrais Definidos I				
Int. Def. Imediatos	5	3		8
Int. Def. de Funções definidas por ramos	14	8		22
Int. Def. Propriedades	11	8		19
Int. Def. Valor Médio	32	20		52
Integrais Definidos - 2.º TFC	28	24		52
Integral Definido - Partes	21	14		35
Integral Definido - Substituição	34	36	1	71
Integral Definido - Áreas	49	38	1	88
Integral Definido - Impróprios	33	24	1	58
Integral Múltiplo - I. P.	34	14	1	49
Integral Definido - Regiões	28	16	2	46
Análise Combinatória				
Análise Combinatória - Permutações	36	29	1	66
Análise Combinatória - Combinações	27	24		51
Análise Combinatória - Triângulo de Pascal	20	17	3	40
N01 - Triângulo de Pascal **	7	5		12
Séries Numéricas Reais				
Séries Numéricas Reais - Definição/Geométrica/Telescópica	47	40		87
N01 - Series Def/Geom/Telesc**	1	3		4

Séries Numéricas Reais – Propriedades/CNC/Resto/Integral	44	30	2	76
Séries Numéricas Reais – Teste de Comparação no limite	46	28	1	75
Séries Numéricas Reais – Teste de D'Alembert	43	26	1	70
Séries Numéricas Reais – Teste de Cauchy	46	32	1	79
Total	**756**	**561**	**23**	**1340**

5.4.1.5. Teste opcionais do 1.º ciclo de IA

Tal como já foi referido, na secção 5.3.1, foram implementados testes com QEM como trabalho de casa, que eram de resolução opcional para os alunos. Apesar de os testes serem opcionais, obtiveram-se boas taxas de resposta. Na Tabela 15 apresenta-se o número de alunos que responderam aos três testes em cada um dos três semestres/anos letivos relativos à implementação do 1.º Ciclo de IA. Foi decidido que durante o 1.º semestre do ano letivo 2010/11 apenas se realizariam dois testes em vez de três por considerarmos mais adequado para o funcionamento da UC neste semestre.

Tabela 15: Número de alunos por teste opcional (1.º ciclo de IA)

	2008/09		2009/10		2010/11	
	1S (nº/%)	2S (nº/%)	1S (nº/%)	2S (nº/%)	1S (nº/%)	2S (nº/%)
N.º de Alunos Avaliados	**558**	**594**	**721**	**696**	**686**	**610**
Teste 1	559/100%	536/90%	624/87%	588/84%	573/84%	546/90%
Teste 2	468/84%	478/80%	598/83%	524/75%	506/74%	466/76%
Teste 3	287/51%	388/65%	554/77%	466/67%	-	366/60%

O número de alunos decresce do primeiro para o terceiro teste, dado haver muitos alunos a desistirem da avaliação contínua, tendo mesmo alguns desistido da inscrição na unidade curricular. Este aspeto é mais evidente durante o 1.º semestre de 2008, primeiro ano de implementação do projeto. Consideramos que isto se deve ao facto de, quer os alunos quer os docentes, não estarem a par das implicações burocráticas relacionadas com a realização da avaliação contínua no início do semestre, dado tratar-se do primeiro ano da sua implementação. No 2.º semestre de 2010/11 não foi possível implementar a avaliação contínua devido a problemas técnicos e organizacionais. Ainda assim, foi decidido que seria dada opção aos alunos em continuar a realizar estes trabalhos de casa e a considerar a sua classificação em exame final, com um peso opcional máximo de 20%. Esta decisão prendeu-se com o facto de se considerar ser importante que os alunos continuassem a fazer estes trabalhos para testarem o sistema e se ambientarem a esta tecnologia.

5.4.1.6. Teoria Clássica dos Testes e Teoria de Resposta ao Item

Pretendemos utilizar a Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI), de forma a analisar a qualidade das QEM para se poder ajustar o banco de questões desenvolvido. Estes instrumentos introduzidos no 0tentam, em termos globais, fazer uma avaliação da qualidade das questões através da utilização de métricas. Pretende-se que os testes que os alunos realizam sejam o mais possível uniformes.

Visto que os testes são gerados de forma aleatória pelo *Moodle*, não podemos aplicar as teorias diretamente nas respostas obtidas em cada teste, como é habitual encontrar na literatura e como se pode encontrar no *Moodle*. Isto prende-se com o facto de as questões não serem as mesmas para todos os alunos que fazem o teste, apesar de pertencerem todas à mesma categoria do banco de questões.

As questões foram avaliadas por semestre. O número de questões e categorias/subcategorias por semestre existentes no banco de questões e que foram objeto de análise, encontra-se resumido na Tabela 16.

Tabela 16: Número de questões e categorias avaliadas através das TCT e TRI

UC	Questões	Categorias e subcategorias
1.º Semestre	1472	33
2.º Semestre	1317	26

Para que fosse possível realizar a análise pretendida tivemos de extrair os dados necessários da base de dados do *Moodle* de Avaliação, a qual continha os testes e respetivas respostas de todos os alunos. Foi solicitado o acesso a esta base de dados à Presidência do ISCAP, tendo sido concedida ao autor da tese a autorização necessária. O responsável pela gestão técnica do *Moodle* forneceu ao autor da tese a base de dados com os dados necessários em formato MySQL. Para a realização deste trabalho foram ainda utilizadas as seguintes soluções tecnológicas: driver ODBC, MS Excel™ e VBA.

O MySQL é um dos mais populares sistemas de gestão de base dados relacionais, de código aberto. É muito versátil, sendo suportado por muitas plataformas atuais, é compatível com muitos drivers de ligação, especialmente o ODBC, e também é utilizado pelo *Moodle* de onde foram extraídos os dados para análise. Foi utilizada a versão MySQL server 5.5[20].

A tecnologia ODBC permite ligar sistemas, como por exemplo o MS Excel™, a uma base de dados externa, como por exemplo o MySQL, de modo a obter os dados aí contidos. Neste trabalho utilizamos o driver MySQL ODBC 5.1[21]. Este driver forneceu-nos o acesso à base de dados MySQL

[20] Fez-se o *download* do site http://dev.mysql.com/downloads/installer/ e a instalação foi efetuada seguindo o *wizard* do binário.

[21] Procedeu-se também ao *download* e instalação do *Connector* ODBC 5.1 a partir do site http://dev.mysql.com/downloads/connector/odbc/5.1.html. Após a instalação do *Connector* ODBC procedeu-se à configuração do DSN no sistema operativo Windows (painel de controlo → Ferramentas administrativas → ODBC), conforme a imagem no anexo H. Após estas configurações procedeu-se, com os

a partir do VBA no MS Excel™, permitindo assim trabalhar a grande quantidade de dados do *Moodle*.

O VBA for MS Excel™, é uma linguagem de programação de alto nível que permite aos utilizadores do MS Excel™ tirarem um maior partido desta ferramenta, potenciando a automatização de tarefas repetitivas e personalizadas que eram necessárias no trabalho efetuado. Para a execução deste trabalho foi utilizada a versão MS Excel™ 2013.

Utilizando então estas soluções tecnológicas, os dados necessários foram extraídos e organizados em folhas de cálculo MS Excel™. Esta organização correspondeu às necessidades de análise inerentes a cada uma das teorias TCT e TRI e também às restrições relacionadas com os dados existentes.

Como o MS Excel™ não tem integradas, por defeito, as funcionalidades necessárias para aplicar a TRI e como os cálculos da TRI são complexos e levariam imenso tempo a programar em VBA, optou-se pela procura de suplementos de qualidade e de código aberto já desenvolvidos para o MS Excel™. Estão disponíveis algumas aplicações específicas para este fim, a maioria comerciais, umas com mais usabilidade do que outras[22].

Encontram-se alguns trabalhos na literatura, de entre os quais salientamos o trabalho desenvolvido por Valois, Houssemand, Germain e Abdous (2011), no qual apresentam o suplemento e*irt* no âmbito do projeto "*libirt*"[23]. Outros autores (Langlois, Lapointe, Valois, & de Leeuw, 2014; Poitras, Guay, & Ratelle, 2012) usaram também este suplemento para MS Excel™, o qual também existe para o R.

O suplemento permite aplicar os Modelos logísticos da TRI com 1, 2 ou 3 parâmetros, além de apresentar várias opções associadas, em especial a escolha do método de estimação. Este suplemento já existe desde o ano de 2006 estando de momento na versão 1.3, a qual foi utlizada para o desenvolvimento deste trabalho.

5.4.2. Questionário aos docentes no 1.º ciclo de IA

No final do 1.º ciclo de IA considerámos importante conhecer a opinião dos docentes, que elaboraram as QEM incluídas no banco de questões, sobre cada uma das linhas de orientação indicadas por Haladyna e colaboradores (2002), já apresentadas na secção 2.4. Para isso foi desenvolvido um questionário, que se encontra no anexo A.

Considerando os propósitos apresentados, pretendeu-se atingir os seguintes objetivos:

devidos comandos, à importação dos dados que o *Moodle* continha e com o VBA foi-se preenchendo as folhas de Excel com as informações para análise.

[22] No mercado existem duas empresas principais a trabalhar nestas ferramentas, a Xcalibre (http://www.assess.com/product/xcalibre-4/) e a SSI (http://www.ssicentral.com/), a qual apresenta vários produtos.

[23] O Projeto pode encontrar-se em https://sourceforge.net/projects/libirt/. Há um outro sítio associado no endereço http://psychometricon.net/libirt/.
Os seus autores são Stephane Germain, Pierre Valois e Belkacem Abdous. No geral, o programa tem um conjunto de funções escritas em C para estimar os parâmetros das questões e capacidades a partir das respostas obtidas em testes e questionários. Permite ajustar modelos logísticos para 1, 2 e 3 parâmetros.

- conhecer as opiniões dos docentes em relação às linhas de orientação na elaboração de QEM segundo Haladyna e colaboradores (2002);
- comparar os resultados obtidos com os apresentados por Haladyna e colaboradores (2002).

O questionário era constituído por 7 questões, 5 das quais contêm um total de 30 afirmações nas quais se utiliza uma escala de Likert de "1 – Discordo Totalmente" a "5 – Concordo Totalmente". Estas questões foram de resposta obrigatória. No final o questionário apresenta ainda duas perguntas de resposta aberta, ambas de reposta opcional. Uma delas para acrescentar alguma linha de orientação que não estivesse contemplada nas perguntas anteriores e fosse útil para a construção da QEM. A outra servia para serem apresentados comentários adicionais, caso os respondentes assim o entendessem.

O questionário *online* foi elaborado com a aplicação livre *LimeSurvey*[24], instalada no servidor do GAIE e acessível através do endereço: http://paol.iscap.ipp.pt/iscapsurvey/index.php.

Foi enviado *email* a todos os potenciais respondentes, que eram todos os docentes que já tinham participado na elaboração das QEM ao longo do 1.º ciclo de IA. O questionário foi respondido por 12 docentes.

Os resultados deste questionário apresentam-se na secção 6.3.

5.4.3. Entrevista aos docentes no 3.º ciclo de IA

Chegados a esta fase da investigação considerou-se necessário recolher de forma mais sistemática a opinião dos docentes sobre o processo de *e-assessment* com QEM, para avaliação contínua. Para isso, foi efetuada uma entrevista semi-estruturada de modo a permitir uma melhor organização dos tópicos a abordar e, ainda assim, dar liberdade aos entrevistados para expressarem livremente as suas ideias. Os objetivos da entrevista foram os seguintes:

- refletir sobre o processo de *e-assessment* implementado;
- verificar a existência de mudanças nas práticas educativas por parte dos docentes;
- aferir quais as vantagens e desvantagens para o docente deste tipo de avaliação;
- verificar a existência de mudança no processo de aprendizagem, por parte dos alunos.

Quanto à sua estrutura, a entrevista consistiu em primeiro lugar na caracterização do entrevistado, quanto ao sexo, idade, área disciplinar e tempo de serviço no ISCAP. Depois abordaram-se as seis dimensões definidas para análise, as quais coincidem com as questões apresentadas no guião de entrevista, que se encontra no Anexo C. As dimensões são então as seguintes:

- opinião sobre a forma de *e-assessment* implementada;
- principais dificuldades encontradas na implementação;
- em que medida houve ou não mudanças nas práticas pedagógicas do docente;

[24] https://www.limesurvey.org/

- vantagens desta forma de avaliação para o docente;
- desvantagens desta forma de avaliação para o docente;
- perceção por parte dos docente de alterações nas práticas dos alunos no seu processo de aprendizagem.

A questão 7 destina-se a aferir se os docentes identificam mais alguma dimensão para além das consideradas pelo autor da tese.

Os docentes a entrevistar foram contactados telefonicamente ou por email, conforme necessário, com a finalidade de marcar um horário conveniente quer para o docente (entrevistado), quer para o autor da tese (entrevistador).

Durante a sua realização, depois de obtido consentimento dos entrevistados, as entrevistas foram gravadas. As entrevistas foram posteriormente transcritas na sua totalidade, de modo a permitir efetuar a sua análise. Nesta utilizámos técnicas de Análise de Conteúdo seguindo duas etapas. A primeira etapa consistiu na leitura atenta de todas as entrevistas, no sentido de captar o sentido geral dos discursos. A segunda etapa consistiu no agrupamento das respostas por cada uma das dimensões inicialmente definidas e já referidas acima. Os resultados de todo o trabalho de análise encontram-se na seção 6.5.

5.4.4. Questionário aos alunos no 3.º ciclo de IA

Neste 3.º ciclo de IA, era importante saber a opinião dos alunos sobre o processo de *e-assessment* com QEM para avaliação contínua.

Visto que não se pretendia condicionar previamente as respostas dos alunos através de um questionário contendo várias afirmações sobre as quais se deveria apresentar o nível de concordância/discordância através de uma escala de Likert, este não foi considerado o método mais adequado. Foi considerado que o ideal seria utilizar entrevistas. No entanto, a utilização de entrevistas no formato habitual, como foi feito para os docentes, revelou-se impraticável devido ao elevado número de alunos e à sua pouca disponibilidade para este tipo de abordagens. Assim sendo, foi elaborado um questionário com perguntas abertas e fechadas que permitiu a sua realização em tempo útil. Os objetivos deste questionário foram os seguintes:

- conhecer a opinião dos alunos sobre o processo de *e-assessment* implementado;
- Identificar quais as vantagens e/ou desvantagens dos QEM do ponto de vista dos alunos;
- verificar a existência de mudanças nas práticas dos alunos no seu processo de aprendizagem.

O questionário apresenta dois grupos. O grupo I contém 7 questões, as quais visavam a caraterização dos inquiridos. A caraterização foi feita a nível de sexo, idade, regime de frequência, situação (trabalhador estudante ou não) e primeiro ano letivo de inscrição na UC.

O Grupo II consistiu de 8 perguntas com as quais foram introduzidas as 8 dimensões definidas:

- os testes QEM são justos (questão 2.1);
- é melhor o formato em papel ou o uso de novas tecnologias (questão 2.2);

- os testes QEM influenciam as práticas educativas (questão 2.3)
- o regime de avaliação influência a presença nas aulas (questão 2.4);
- o número de testes é importante na escolha do regime de avaliação (questão 2.5);
- opinião sobre o Teste de "Repescagem" (2.6);
- quais as vantagens das QEM para os alunos (questão 2.7);
- quais as desvantagens dos QEM para os alunos (questão 2.8).

Na Tabela 17 resume-se a correspondência entre os objetivos do questionário e as questões aí incluídas.

Tabela 17: Correspondência entre os objetivos e as questões incluídas no questionário aos alunos (3.º ciclo de IA)

Objetivos	Questões
Conhecer a opinião dos alunos sobre o processo de *e-assessment* implementado.	2.1, 2.2, 2.6
Identificar quais as vantagens e/ou desvantagens dos QEM do ponto de vista dos	2.7; 2.8
Verificar a existência de mudanças por parte dos alunos nas suas práticas educativas.	2.3; 2.4; 2.5

Cada uma destas questões, ligava a questões de resposta aberta, as quais dependiam do valor da resposta dada anteriormente. Estas questões são consideradas sub-questões das 8 questões principais. Por exemplo, a questão "2.1 – Considera que os testes … são justos?" tem duas sub-questões: i) a questão "2.1.1 – Porque não os considera justo?" que surge no questionário caso o aluno responda "Não" à questão 2.1, ii) a questão "2.1.2 – Porque os considera justo?" que surge no questionário caso o aluno responda "Sim" à questão 2.1. Todas estas questões eram de resposta obrigatória. O questionário termina solicitando comentários adicionais, com vista a verificar a existência de dimensões para além das que foram consideradas pelo autor da tese. O questionário encontra-se no anexo B.

O questionário foi implementado utilizando o *LimeSurvey*, conforme descrito na secção 8.3.

O questionário foi testado com um grupo de teste, a fim de ser validado. Os alunos responderam ao questionário durante a última semana de aulas do 1.º semestre. Dado que os alunos são praticamente os mesmos no 1.º e 2.º semestres, o questionário foi respondido apenas no 1.º semestre.

Foram utilizadas técnicas de análise de conteúdo. Dado que as respostas já estavam agrupadas de acordo com as dimensões a analisar, procedeu-se à sua leitura cuidadosa no sentido de captar o sentido geral dos discursos. A aplicação MaxQDA, na sua versão 12, foi utilizada para apoiar nesta análise de conteúdo, devido ao elevado volume de dados qualitativos disponíveis para análise.

CAPÍTULO 6. APRESENTAÇÃO E ANÁLISE DOS DADOS

Tal como já foi exposto nos capítulos anteriores, foram elaborados inquéritos e entrevistas aos docentes, inquéritos aos alunos e ainda recolhidos dados envolvendo os resultados das classificações das avaliações realizadas pelos alunos. Além disso, foram ainda recolhidos dados do *Moodle* relativos ao banco de questões, tendo sido feita uma análise da qualidade dessas questões.

A análise da evolução das classificações referentes ao 1.º Semestre é apresentada na secção 6.1 e na secção 6.2 é apresentada a análise da evolução das classificações referentes ao 2.º semestre. Em ambos os casos, a análise relativamente à evolução da média e à evolução de proporções de positivas, é feita quer por anos quer por ciclos de IA. A análise das respostas ao questionário aos docentes no 1.º ciclo de IA é feita na secção 6.3. A análise da qualidade dos testes e das questões é feita em 6.4, considerando-se a análise usando TCT em 6.4.1 e a análise usando TRI em 6.4.2. A análise das respostas às entrevistas aos docentes no 3.º ciclo de IA efetua-se na secção 6.5, fazendo-se a caracterização dos entrevistados em 6.5.1 e a análise das dimensões consideradas na entrevista em 6.5.2. A análise das respostas ao questionário aos alunos no 3.º ciclo de IA apresenta-se na secção 6.6, fazendo-se caracterização dos alunos que responderam ao questionário na secção 6.6.1 a análise das dimensões consideradas no questionário na secção 6.6.2.

6.1. Análise da Evolução das Classificações Referentes ao 1.º Semestre

Consideramos útil analisar, antes de mais, a evolução das classificações finais dos alunos ao longo dos anos e ciclos em que decorreu a IA descrita nesta tese. Vamos por isso focar-nos no período entre 2008 e 2014.

Todos os dados referentes às classificações dos alunos que se apresentam nesta tese foram recolhidos da base de dados da Secretaria do ISCAP, e com a devida autorização da Presidência a escola.

Depois de recolhidos, os dados foram posteriormente tratados, pois continham algumas informações codificadas que foi necessário corrigir, por exemplo, alunos com classificação "88" eram alunos que entretanto tinham desistido. Estes alunos foram retirados da base de dados. Uma outra situação que foi corrigida prendia-se com o facto de muitos alunos apresentarem mais do que uma classificação num mesmo ano letivo, porque a base de dados continha as classificações dos vários exames que o aluno tinha realizado nesse semestre (avaliação contínua, época de recuso, etc). Foram eliminadas as repetições deixando somente a

classificação maior, dado ser essa a classificação que será atribuída ao aluno. Existiam ainda outros pequenos ajustes que foram efetuados, mas de muito menor importância.

Para analisar e interpretar os dados fizemos uso da Estatística Descritiva e da Inferência Estatística, recorrendo ao *MS Excel*™ como principal ferramenta de trabalho. No âmbito da Estatística Descritiva foi efetuada a construção de tabelas e gráficos e foram realizados cálculos de algumas medidas de localização e de dispersão, os quais, essencialmente resumem e descrevem os dados. No âmbito da Inferência Estatística, entre outras ferramentas, fizemos uso de vários testes de hipóteses, em particular com recurso à Análise de Variância, o que permitiu tirar conclusões sobre os dados.

Como já foi referido anteriormente, o *MS Excel*™ foi a principal ferramenta utlizada para trabalhar os dados. Contudo, o seu suplemento "Análise de Dados" não tem a maioria dos testes estatísticos necessários ou estes não são suficientemente completos para permitir a sua aplicação neste contexto. Assim sendo, decidiu-se instalar o suplemento de distribuição gratuita *Real Statistics*[25] (Zaiontz, 2015). Portanto, a análise de dados para esta tese foi realizada utilizando o suplemento *Real Statistics Resource Pack software* (Release 4.3) Copyright (2013 – 2015). Este suplemento não contém, o Método de *Marascuilo*, o teste de Qui-Quadrado para proporções de mais do que duas variáveis nem o teste de *Bartlett*. No que se segue, nas utilizações destes, fizemos os cálculos com as fórmulas adequadas do *MS Excel*™.

6.1.1. Análise da evolução da média das classificações

Tal como já foi referido, o 1.º semestre compreende as UC Matemática e Matemática I dos cursos de licenciatura "Contabilidade e Administração" e "Comércio Internacional", respetivamente. Foram recolhidas 4292 classificações correspondentes a 7 anos letivos.

Na Tabela 18 apresenta-se a análise descritiva das classificações dos alunos durante o 1.º semestre entre os anos letivos de 2008 e 2014. Destaca-se na tabela, com cores diferentes, a informação dos anos letivos que compõem cada ciclo de IA.

Tabela 18: Análise descritiva de alguns parâmetros estatísticos referente às classificações dos alunos durante o 1.º semestre entre 2008 e 2014

	1.º Ciclo			2.º Ciclo			3.º Ciclo
	2008	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
Contagem	558	721	686	637	608	593	489
Média	7.2	7.0	6.1	6.7	7.4	9.3	9.7

[25] Este suplemento (http://www.real-statistics.com/) foi desenvolvido e é atualizado pelo Dr. Charles Zaiontz. A documentação associada ao suplemento apresenta todas as fórmulas estatísticas programadas bem como as definições, propriedades e algumas demonstrações. Apresenta ainda muitos exemplos concretos e particularidades a considerar nas análises.

Erro-padrão	0.21	0.17	0.17	0.18	0.19	0.19	0.22
Mediana	7	7	5	7	8	11	11
Moda	10	11	1	10	10	11	11
Variância da amostra	25.0	20.1	20.7	20.5	21.2	22.2	22.7
Desvio padrão	5.00	4.48	4.55	4.53	4.61	4.72	4.77
Mínimo	0	0	0	0	0	0	0
Máximo	20	18	19	20	20	20	20
Coeficiente de Variação de Pearson (CVP) em %	69%	64%	75%	67%	62%	51%	49%

Na Tabela 18 podemos observar que apesar de haver uma ligeira queda no valor da média das classificações até 2010, posteriormente houve uma recuperação. Os dois últimos anos destacam-se na melhoria deste parâmetro, melhoria essa que é reforçada pela observação dos valores da mediana, que também são mais elevados. A moda é idêntica em todos os anos (10 ou 11), com exceção do ano 2010 no qual, surpreendentemente, é 1. No entanto, a frequência de 1 é 77, de 10 é 75 e de 11 é 76.

Em relação à variabilidade/dispersão das classificações, verificamos que os valores do desvio padrão em cada ano estão muito próximos uns dos outros. No entanto, observamos que, para os valores do Coeficiente de Variação, apesar de nenhum dos anos ser considerado homogéneo[26], os dois últimos anos destacam-se por apresentar menor heterogeneidade nas classificações. Verifica-se que nos cinco primeiros anos do estudo, a média Coeficiente Variação de Pearson é 68%, isto é, em média as classificações têm um desvio de 68% em relação à média. Nos dois últimos anos, a média dos Coeficiente Variação de Pearson é 50%, isto é, em média as classificações têm um desvio de 50% em relação à média. Assim sendo, podemos afirmar que as classificações começam a aproximar-se da homogeneidade.

Estas informações são corroboradas com o gráfico da Figura 10.

[26] O Coeficiente de Variação de Pearson é calculado dividindo o valor do desvio padrão pela respetiva média. Consideram-se heterogéneas as variáveis para as quais o coeficientes de variação é superior a 30%.

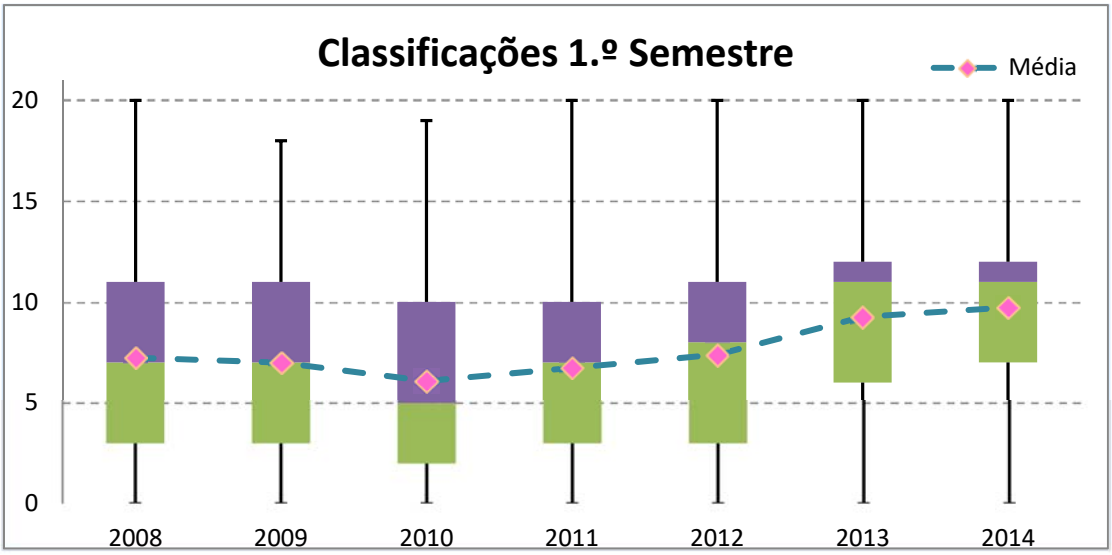


Figura 10: Diagrama de extremos e quartis das classificações entre os anos 2008 e 2014 do 1.º Semestre.

Poderemos ainda assinalar que em 2009 e 2010 não há alunos com a classificação máxima. Esta situação poderá ser explicada com o facto de os alunos, entre 2009 e 2011, para terem uma classificação superior a 17, terem de fazer uma prova para defesa de nota. A maioria dos alunos optou por não fazer a defesa de nota e, dos que a fizeram, nenhum deles conseguiu acertar em todas as questões.

Para melhor comparar os diferentes anos e compreender melhor alguns dos seus resultados, acrescentam-se ainda as seguintes informações:

- o número de testes em avaliação contínua foi sempre três, à exceção de 2010 e 2012 que foi dois.
- em 2012 e anos posteriores, foi acrescentado um novo parâmetro à avaliação contínua com a classificação máxima de um valor, a acrescentar à classificação final como bónus, dependendo da assiduidade e participação dos alunos nas aulas;
- nos dois últimos anos, os alunos que obtivessem classificação final negativa à avaliação contínua, tinham a possibilidade de fazer um teste suplementar, que chamamos Teste de "Repescagem" o que está explicado na secção 7.3.

O ano de 2011 corresponde a um ano de mudança nas classificações dos alunos, em termos globais. Neste ano e em anos posteriores, a evolução positiva nas classificações é notória. Claramente, a mediana a partir de 2010 começou a subir, atingindo valores positivos em 2013 e 2014. Visivelmente, em 2013 e 2014 o intervalo interquartis Q3-Q1 é menor, logo podemos afirmar que as classificações estão mais concentradas em torno da média.

Em termos gerais, parece haver uma evolução positiva nas classificações dos alunos ao longo dos anos, que se poderá observar no gráfico da Figura 10 e ainda na Tabela 18. Contudo, é

conveniente verificar se as diferenças descritas são ou não estatisticamente significativas. Para isso, recorremos a alguns testes estatísticos que vamos apresentar de seguida. Iremos testar as diferenças entre as médias das classificações e a diferença entre as proporções de positivas nas classificações.

Para testar se as diferenças entre as médias nos diferentes anos letivos são estatisticamente significativas, vamos comparar este parâmetro entre as diferentes amostras (mais do que duas) usando a Análise de Variância, vulgo ANOVA. No nosso caso utilizamos ANOVA a um fator. Esta análise permite que vários grupos (anos letivos) sejam comparados no que diz respeito às medidas de localização, nomeadamente no que diz respeito à média.

Antes de mais, temos de verificar as condições de aplicabilidade deste teste, que são as seguintes:

- independência mútua;
- normalidade da distribuição;
- homogeneidade da variância (σ^2 constante).

Atendendo às classificações dos alunos que são objeto de estudo, quer por semestre quer por ano, em relação à independência, ela é assegurada porque qualquer que seja o valor particular que uma amostra toma, ela não influencia a distribuição de outra.

Visto que as suas dimensões são grandes (neste caso maiores que 50) e são independentes, pelo Teorema do Limite Central, podemos considerar que cada amostra segue uma distribuição Normal (Guimarães & Cabral, 2007).

Quanto à terceira condição, a homogeneidade da variância, é por norma mais difícil de provar, porque em muitos casos é necessário recorrer a testes estatísticos para a sua verificação. No entanto, no nosso caso, não é necessário recorrer a estes testes estatísticos para verificar a homogeneidade, pois "Na prática, a homogeneidade da variância só se torna importante quando as dimensões das amostras (grupos ou células) forem muito diferentes, isto é, quando a maior amostra tiver uma dimensão pelo menos dupla da dimensão da menor amostra. Quando as amostras não são fortemente desequilibradas, o efeito da heterogeneidade da variância, mesmo se acentuada, é pouco significativo" (Guimarães & Cabral, 2007, p. 332).

Pretendemos então, testar se as médias das classificações nos diferentes anos diferem de forma estatisticamente significativa entre si. Isto é, pretendemos testar as seguintes hipóteses:

H_0: A classificação é, em média, **idêntica** em todos os anos letivos.
H_1: A classificação é, em média, **diferente em pelo menos um** par de anos letivos

ou, em linguagem Matemática,

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$
$H_1 : \mu_i \neq \mu_j$ para algum $i \neq j$ $\left(i, j \in \{1,..,7\}\right)$.

Aplicando ANOVA a um fator, obtiveram-se os resultados apresentados na Tabela 19.

Tabela 19: Resultados da aplicação do teste ANOVA – às classificações dos alunos nos 7 anos letivos, no 1.º semestre

DESCRIPTION					Alpha	0.05		
Groups	Count	Sum	Mean	Variance	SS	Std Err	Lower	Upper
2008	558	4041	7.2419	25.0491	13952.34	0.1969	6.8551	7.6288
2009	721	5042	6.9931	20.0597	14442.97	0.1732	6.6529	7.3332
2010	686	4173	6.0831	20.6632	14154.26	0.1776	5.7344	6.4318
2011	637	4288	6.7316	20.5206	13051.1	0.1843	6.3696	7.0935
2012	608	4492	7.3882	21.2132	12876.39	0.1887	7.0176	7.7587
2013	593	5494	9.2648	22.2423	13167.43	0.1910	8.8896	9.6399
2014	489	4759	9.7321	22.7170	11085.91	0.2104	9.3188	10.1454
	4292							
ANOVA								
Sources	SS	df	MS	F	P value	F crit	RMSSE	Omega Sq
Between Groups	6264.319	6	1044.053	48.2449	1.4E-57	2.1007	0.2898	0.0620
Within Groups	92730.4	4285	21.6407					
Total	98994.72	4291	23.0703					

A partir da leitura da tabela, podemos verificar que o valor da estatística do teste F é de aproximadamente **48.2449** e considerando um intervalo de confiança para a média de 95% (α=0.05), este valor é consideravelmente superior ao seu valor crítico que é aproximadamente **2.1007**. Temos um valor p aproximadamente igual a 1.4×10^{-57} (praticamente nulo) que é inferior ao valor alfa. Como $F_{(4291,6)} = 48.24 (p<0.001)$, podemos assim rejeitar a hipótese nula e concluir que existem diferenças estatisticamente significativas entre pelo menos um par de anos em relação à média das classificações.

Consideramos importante referir que em muitas análises estatísticas, em especial as que utilizam ANOVA, o valor da estatística do teste, o valor crítico e nível de significância não são por vezes suficientes para se poderem tirar conclusões sobre o significado e a importância prática dos resultados. Em especial, amostras muito grandes podem originar resultados estatisticamente significativos, mesmo que as diferenças observadas entre grupos sejam pequenas. Assim sendo, é importante avaliar a significância prática, isto é, o tamanho ou

magnitude do efeito baseado em Estatística Descritiva, que não dependam do tamanho da amostra, que vai servir de complemento aos testes estatísticos usados habitualmente. Para este efeito, existem várias medidas para calcular o tamanho do efeito dos resultados encontrados. No Anexo I encontram-se mais alguns detalhes sobre este assunto.

Na nossa análise, pode-se referir que a magnitude da diferença entre as médias é elevada mas muito próxima de ser moderada, quer calculada pela medida RMSSE (0.2898), quer pela medida Omega Sq (0.0620).

É ainda necessário saber qual ou quais os pares de anos para os quais existem diferenças estatisticamente significativas em relação à média das classificações. Existem diversos métodos que permitem abordar este aspeto, tendo-se optado pelo método de *Tukey HSD (Honest Significant Difference)/Tukey-Kramer* para amostras pouco desequilibradas. Consideram-se aqui as condições para usar o método de *Tukey* sugeridas em (Guimarães & Cabral, 2007, p. 310), isto é, "as dimensões das amostras relativas aos diferentes grupos são moderadamente diferentes". No nosso caso as amostras consideram-se pouco desequilibradas, ou seja, moderadamente diferentes, porque as suas dimensões, segundo os mesmos autores, possuem a propriedade de que a dimensão da maior das amostras é inferior a duas vezes a dimensão da menor.

Na Tabela 20 apresenta-se o resultado da aplicação método de *Tukey* para dois dos anos (2010 e 2014), utilizando os chamados Contrastes Ortogonais[27].

Tabela 20: Resultados da aplicação do Método de *Tukey* usando Contrastes Ortogonais, para os anos 2010 e 2014 no 1.º semestre

TUKEY'S HSD / TUKEY-KRAMER					Alpha	0.05
Groups	*c*	*mean*	*n*	*ss*	*c^2/n*	*c*mean*
2008		7.2419	558	13952.34	0	0
2009		6.9931	721	14442.97	0	0
2010	1	6.0831	686	14154.26	0.001458	6.0831
2011		6.7316	637	13051.1	0	0
2012		7.3882	608	12876.39	0	0
2013		9.2648	593	13167.43	0	0
2014	-1	9.7321	489	11085.91	0.002045	-9.7321
			4292	92730.4	0.003503	-3.6490

Q TEST

std err	*q-stat*	*df*	*q-crit*	*lower*	*upper*	*sig*	*x-crit*	*Cohen d*	*effect r*
0.19468	-18.7436	4285	4.17	-4.46083	-2.8372	yes	0.8118	0.7844	0.2753

[27] Contrastes Ortogonais podem ser usados para testar a diferença entre as médias de vários grupos, testando a média de um deles contra a média de um outro, desde que a soma dos coeficientes usados nos grupos em estudo seja igual a 0 (zero).

A partir da Tabela 20, verifica-se que existe uma diferença significativa entre as médias dos anos 2010 e 2014 (sig= yes). Considerando que a medida *Cohen d* é igual 0.7844, considera-se que a diferença entre as médias é elevada (ver Anexo I).

De forma análoga utilizamos o Método de *Tukey HSD/Tukey-Kramer* para todos os possíveis pares de anos, recorrendo aos Contrastes. Na Tabela 21 apresenta-se um resumo dos resultados obtidos.

Tabela 21: Resultados da aplicação do Método de *Tukey* em relação à diferença ou não entre as médias das classificações aplicados a todos os pares de anos letivos, no 1.º semestre

	S: Sim	**N**: Não	(*Cohen d*)			Alpha 0.05	
TUKEY HSD	2008	2009	2010	2011	2012	2013	2014
2008		N (0.05) p≈0.964	S (0.24) p≈0.000	N (0.11) p≈0.486	N (0.03) p≈0.998	S (0.43) p≈0.000	S (0.53) p≈0.000
2009			S (0.2) p≈0.005	N (0.06) p≈0.946	N (0.08) p≈0.719	S (0.49) p≈0.000	S (0.59) p≈0.000
2010				N (0.14) p≈0.148	S (0.28) p≈0.000	S (0.68) p≈0.000	S (0.78) p≈0.000
2011					N (0.14) p≈0.163	S (0.54) p≈0.000	S (0.65) p≈0.000
2012						S (0.40) p≈0.000	S (0.50) p≈0.000
2013							N (0.10) p≈0.653
2014							

Realçam-se as diferenças estatisticamente significativas entre as médias de cada ano e os últimos dois anos letivos. Destaca-se ainda que os valores *Cohen d* indicam que essas diferenças são de nível moderado a elevado. Há ainda diferenças estatisticamente significativas entre as médias dos pares de anos 2008/2010, 2009/2010 e 2010/2012, sendo que, atendendo ao valor da medida *Cohen d*, estas diferenças são moderadas, mas muito próximas de serem consideradas pequenas. Quanto aos restantes pares de anos as diferenças entre as médias não são estatisticamente significativas.

6.1.2. Análise da evolução da proporção de classificações positivas

A Figura 11 apresenta um gráfico que ilustra a evolução da percentagem das classificações positivas e negativas dos alunos no 1.º semestre, ao longo dos anos 2008 a 2014. Verifica-se que o número de positivas apresenta uma clara tendência crescente a partir de 2010.

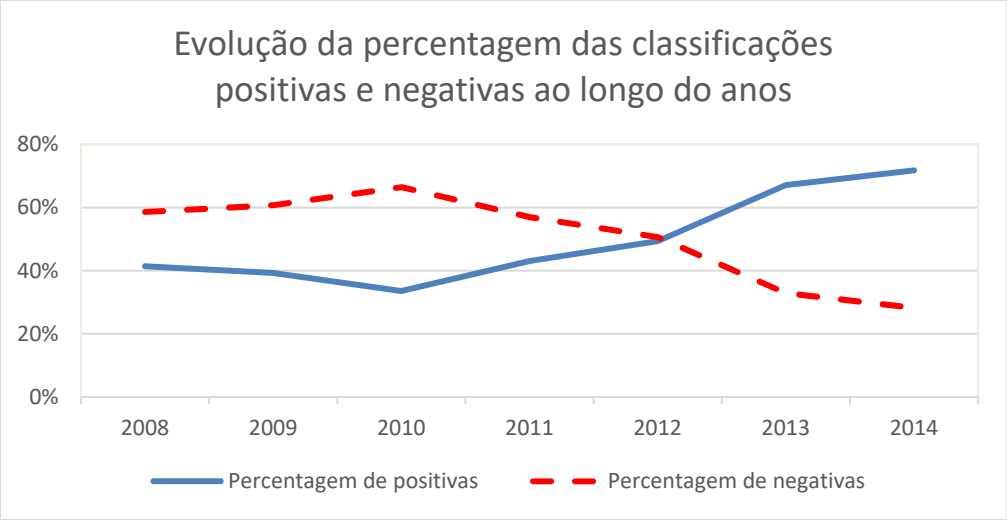


Figura 11: Evolução da percentagem das classificações positivas e negativas no 1.º semestre.

Analisemos agora as proporções de classificações positivas e sua evolução ao longo dos anos 2008 a 2014 no 1.º semestre. Na Tabela 22 apresenta-se a evolução do número de classificações positivas e de negativas ao longo dos anos, bem como a proporção de classificações positivas. Verifica-se que a proporção de positivas apresenta uma tendência claramente crescente.

Tabela 22: Número de positivas e negativas por ano letivo e proporção de classificações positivas, no 1.º semestre

	2008	2009	2010	2011	2012	2013	2014
Negativas	327	438	456	363	308	195	138
Positivas	231	283	230	274	300	398	351
Total	**558**	**721**	**686**	**637**	**608**	**593**	**489**
Proporção Posit. (p)	0.414	0.393	0.335	0.430	0.493	0.671	0.718

Pretendemos agora testar se as proporções de positivas nos diferentes anos diferem entre si de forma estatisticamente significativa. Para isso aplicamos o teste do Qui-Quadrado a uma tabela de contingência envolvendo as proporções de positivas e negativas das classificações dos alunos, cujos valores se encontram na Tabela 22.

Formulamos assim, as seguintes hipóteses:

H_0: A proporção de positivas é idêntica em todos os anos letivos.
H_1: A proporção de positivas é diferente em pelo menos um par de anos letivos.

Para testar estas hipóteses, não podemos utilizar o suplemento "*Real Statistics Resource Pack*" porque não contém, na versão atual, o teste de Qui-Quadrado para proporções de mais do que duas variáveis, por isso todos os cálculos foram realizados em *MS Excel*™, usando as fórmulas adequadas. Assim, considerando um valor de significância de **0.05**, e 6 graus de liberdade, obtivemos os seguintes valores aproximados:

Estatística do Teste: **293.686**
Valor Crítico: **12.592**
o valor p: **0.000**

Como $\chi^2_{(0.95,6)} = 12.592 \ (p < 0.001)$, podemos assim rejeitar a hipótese nula e concluir que existem diferenças estatisticamente significativas entre pelo menos um par de anos em relação às proporções de positivas.

Contudo, este procedimento não nos diz quais os pares de anos para os quais existem essas diferenças. Assim, iremos averiguar entre que anos existem estas diferenças utilizando o Método de *Marascuilo,* o qual efetua a comparação das proporções entre todos os pares de anos.

No cálculo dos valores críticos manteve-se 0.05 como nível de significância. Os resultados encontram-se na Tabela 23.

Tabela 23: Resultados da aplicação do Método de *Marascuilo* para existência de diferenças entre as proporções de classificações positivas entre os diferentes pares de anos letivos, no 1.º semestre

	S: Sim	N: Não					
Marascuilo	2008	2009	2010	2011	2012	2013	2014
2008		N	N	N	N	S	S
2009			N	N	S	S	S
2010				N	S	S	S
2011					N	S	S
2012						S	S
2013							N
2014							

Tal como acontece para as médias das classificações, realçam-se as diferenças estatisticamente significativas entre as proporções de positivas de cada ano e os últimos dois anos letivos. Há ainda diferenças estatisticamente significativas entre as proporções dos pares de anos 2009/2012 e 2010/2012. Quanto aos restantes pares de anos, a diferença entre as proporções de positivas não é estatisticamente significativa. Destaca-se que para o par de anos 2009/2012, a diferença entre as proporções de positivas é estatisticamente significativa, mas o mesmo não acontece com a média das classificações. Já para os pares de anos 2008/2010 e 2009/2010, há diferenças estatisticamente significativas para as médias das classificações, mas não para a proporção de positivas.

Os resultados do Método de *Marascuilo* vêm confirmar os resultados anteriores e assim realçar as conclusões já descritas anteriormente. Em termos globais e em função dos testes estatísticos aplicados às classificações dos alunos durante o 1.º semestre, podemos confirmar a subida continuada da média e da proporção de positivas das classificações dos alunos depois de 2010 e em especial o aumento em 2013. Também se confirma que nos dois últimos anos, 2013 e 2014, os resultados estão a estabilizar. A Tabela 24 resume os valores testados, a saber a média das classificações e a proporção de positivas aos longos dos anos em estudo, no 1.º semestre.

Tabela 24: Média e percentagem de positivas das classificações dos alunos por ciclos de estudo do 1.º Semestre

	1.º Ciclo			2.º Ciclo			3.º Ciclo
	2008	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
Negativas	327	438	456	363	308	195	138
Positivas	231	283	230	274	300	398	351
Total	558	721	686	637	608	593	489
Média	**7.2**	**7.0**	**6.1**	**6.7**	**7.4**	**9.3**	**9.7**
Percentagem Positivas	**41%**	**39%**	**34%**	**43%**	**49%**	**67%**	**72%**

6.1.3. Análise da evolução das Classificações por ciclos de IA

Importa agora verificar a evolução das classificações dos alunos por ciclos de IA. Isto é, averiguar se existem diferenças estatisticamente significativas entre as médias das classificações entre os diferentes ciclos de IA e analisar ainda se existem diferenças para as proporções de positivas. Recordamos que no 1.º semestre o 1.º ciclo de IA envolve os anos de 2008 a 2010, o 2.º ciclo entre os anos 2011 e 2013 e o 3.º ciclo corresponde somente ao ano de 2014.

6.1.3.1. Análise da evolução da média das classificações

Em primeiro lugar, apresenta-se a Tabela 25 com uma Análise Descritiva sumária das classificações dos alunos em cada um dos três ciclos de IA. Podemos constatar, em termos genéricos, que parece haver diferenças entre os ciclos, no que concerne à média das classificações durante o 1.º semestre de aulas.

Podemos observar que o valor da média das classificações apresenta uma tendência crescente acentuada. Verifica-se ainda uma melhoria acentuada na mediana, que passa de um valor negativa no 1.º ciclo para um valor positivo no 2.º e ainda aumentando ligeiramente no 3.º ciclo. A moda é positiva e idêntica nos dois primeiros ciclos (10 valores) e aumenta ligeiramente para 11 valores no 3.º ciclo.

Em relação à variabilidade/dispersão das classificações, verificamos que os valores do desvio padrão em cada ciclo de IA estão muito próximos uns dos outros. No entanto, observamos que, para os valores do Coeficiente de Variação de Pearson, apesar de nenhum dos ciclos de IA ser considerado homogéneo, o último ciclo de IA destaca-se por apresentar muito menor heterogeneidade nas classificações, sendo a diferença em relação aos dois ciclos anteriores grande.

Tabela 25: Análise Descritiva sumária das classificações dos alunos pelos respetivos ciclos de IA, no 1.º semestre

	1.º Ciclo	2.º Ciclo	3.º Ciclo
Contagem	1965	1838	489
Média	6.7	7.8	9.7
Erro-padrão	0.11	0.11	0.22
Mediana	6	10	11
Moda	10	10	11
Variância da amostra	21.91	22.43	22.72
Desvio padrão	4.68	4.74	4.77
Mínimo	0	0	0
Máximo	20	20	20
Coeficiente Variação de Pearson em %	69%	61%	49%

É agora conveniente verificar se as diferenças descritas são ou não estatisticamente significativas. Para isso aplicamos o teste ANOVA a um fator.

O teste ANOVA pressupõe que três condições, já referidas anteriormente, sejam satisfeitas para podermos aplicar o teste aos dados.

Atendendo às classificações dos alunos que são objeto de estudo, e como os Ciclos são compostos por grupos de classificações de anos letivos distintos, em relação à independência, esta é assegurada porque qualquer que seja o valor particular que uma amostra (Ciclo) toma, ela não influencia a distribuição de outra.

Visto que as suas dimensões são grandes (neste caso maiores que 50) e são independentes, pelo Teorema do Limite Central, podemos considerar que cada Ciclo segue uma distribuição Normal (Guimarães & Cabral, 2007).

A condição que terá de se averiguar é se existe homogeneidade das variâncias. Usamos o teste de *Bartlett* para testar a homogeneidade das variâncias.

Para testar a homogeneidade das Variâncias entre os Ciclos, formulamos as seguintes hipóteses:

H_0: Há homogeneidade das variâncias nos ciclos de IA.
H_1: Não há homogeneidade das variâncias nos ciclos de IA.

ou em linguagem Matemática,

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$
$$H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ para algum } i \neq j \left(i, j \in \{1,2,3\}\right)$$

Usando as fórmulas adequadas ao teste para amostras com dimensões diferentes, apresentam-se, na Tabela 26, os resultados obtidos.

Tabela 26: Resumo dos valores obtidos com aplicação do teste de Bartlett para os três ciclos de IA, no 1.º semestre

Teste de Bartlett para igualdade das Variâncias com Alpha 0.05			
Ciclos de IA	Contagem	Variância	Desv. Padrão
1.º Ciclo	1965	21.91	4.68
2.º Ciclo	1838	22.43	4.74
3.º Ciclo	489	22.72	4.77
ET	0.390		
Valor Crítico	5.991		
p–valor	0.8227		

Vejamos na Tabela 26, que a Estatística do Teste (0.390) é inferior ao valor Crítico (5.991) e o valor p (0.8227) é superior ao nível de significância (0.05). Podemos então concluir que não há evidência estatística para rejeitar a Hipótese Nula. Assim, podemos concluir que existe homocedasticidade, isto é, há homogeneidade das variâncias entre os diferentes Ciclos e pode-se aplicar ANOVA.

De forma análoga ao que foi feito anteriormente, testemos se as médias das classificações entre os diferentes Ciclos diferem entre si de forma estatisticamente significativa.

Formulámos assim, as seguintes hipóteses estatísticas:

H_0: A classificação é, em média, idêntica em todos os ciclos de IA.
H_1: A classificação é, em média, diferente em pelo menos um par de ciclos de IA.

Aplicando ANOVA a um fator, obtiveram-se os resultados apresentados na Tabela 27.

Tabela 27: Resultados da aplicação do teste ANOVA às classificações dos alunos entre os ciclos de IA no 1.º semestre

ANOVA: Single Factor								
DESCRIPTION					Alpha	0.05		
Groups	Count	Sum	Mean	Variance	SS	Std Err	Lower	Upper
1.º Ciclo	1965	13256	6.7461	21.91053	43032.28	0.1063	6.5375	6.9546
2.º Ciclo	1838	14274	7.7661	22.42537	41195.4	0.1100	7.5504	7.9817
3.º Ciclo	489	4759	9.7321	22.71702	11085.91	0.2132	9.3132	10.1510

ANOVA								
Sources	SS	df	MS	F	P value	F crit	RMSSE	Omega Sq
Between Groups	3681.127	2	1840.5635	82.8232	0	2.9978	0.3220	0.0367
Within Groups	95313.59	4289	22.2228					
Total	98994.72	4291	23.0703					

A partir da leitura da tabela, podemos verificar que o valor da estatística do teste F é de aproximadamente **82.8232** e, considerando um intervalo de confiança para a média de 95% (α=0.05), este valor é consideravelmente superior ao seu valor crítico que é aproximadamente **2.9978,** temos um valor p aproximadamente igual a 0 (praticamente nulo), que é inferior ao valor alfa. Como $F_{(4291,6)} = 82.82 (p<0.001)$ podemos assim rejeitar a hipótese nula e concluir que existem diferenças estatisticamente significativas entre pelo menos um par de ciclos de IA em relação à média das classificações.

Na nossa análise, pode-se referir que a magnitude da diferença entre as médias, calculada por RMSSE (0.3220) é considerada grande, mas a calculada por Omega Sq (0.0367) é considerada moderada.

É ainda necessário saber qual ou quais os pares de anos para os quais existem diferenças estatisticamente significativas em relação à média das classificações

Na comparação entre ciclos, não podemos aplicar o mesmo teste anterior (Tukey) usado para as classificações por ano letivo. Como existem diferenças acentuadas no número de dados entre os ciclos de IA, o método adequado para os testar é o *Hochberg's GT2*. A sua escolha, em primeiro lugar, é devida ao grande desequilíbrio entre o número de dados, como aponta Stoline (1981) e Field (2013) e ainda segudo Larry Toothaker (citado por Cramer e Howitt 2004) refere que para aplicar o teste, para além da grande diferença na dimensão das amostras, exige-se a igualdade das variâncias entre os grupos em estudo, o que já foi testado anteriormente e se verifica. O poder deste teste está bem demonstrado pelos seus autores Benjamini e Hochberg (1995).

Para aplicar o Método de *Hochberg's GT2*, não podemos utilizar o suplemento "*Real Statistics Resource Pack*" porque o suplemento não contém este método implementado na versão atual. A sua implementação em *MS Excel*™ não foi possível e por isso usou-se o *IBM SPSS Statistics* versão 22. Os resultados deste teste encontram-se na Tabela 28.

Tabela 28 Resultados estatísticos do teste de *Hochberg GT2* às classificações por ciclos de IA no 1.º semestre

Comparações múltiplas						
Hochberg GT2						
		Diferença			Intervalo de Confiança 95%	
(I) Ciclos	(J) Ciclos	média (I-J)	Erro Padrão	Sig.	Limite inferior	Limite superior
1	2	-1.01999*	.15297	.000	-1.3853	-.6547
	3	-2.98605*	.23823	.000	-3.5550	-2.4171
2	1	1.01999*	.15297	.000	.6547	1.3853
	3	-1.96606*	.23987	.000	-2.5389	-1.3932
3	1	2.98605*	.23823	.000	2.4171	3.5550
	2	1.96606*	.23987	.000	1.3932	2.5389
*. A diferença média é significativa no nível 0.05.						

Como se observa na Tabela 28, o teste de *Hochberg GT2* mostra que existem diferenças estatisticamente significativas em relação à média entre todos os pares de ciclos de IA durante o 1.º semestre, porque todos os valores p (Sig.) são inferiores a 0.001. Na Tabela 29 apresenta-se o resumo dos resultados do teste de *Hochberg GT2*.

Tabela 29: Resultados da aplicação do teste de *Hochberg GT2* em relação à diferença, ou não, entre as médias das classificações entre os pares de Ciclos, no 1.º semestre

	S: Sim	**N**: Não	*(Cohen d)*
Hoct GT"	**1.º Ciclo**	**2.º Ciclo**	**3.º Ciclo**
1.º Ciclo		**S (0.21) p≈0.000**	**S (0.63) p≈0.000**
2.º Ciclo			**S (0.41) p≈0.000**
3.º Ciclo			

Realçam-se as diferenças estatisticamente significativas entre as médias de todos os pares de Ciclos. Destaca-se ainda que os valores *Cohen d* indicam que essas diferenças são de nível moderado a elevado. Realça-se o nível elevado na magnitude do efeito entre o 1.º ciclo e o 3.º ciclo de IA.

Atendendo às evidências estatísticas, bem como ao nível dos efeitos, podemos afirmar que a evolução das médias das classificações entre os ciclos foi bastante positiva.

6.1.3.2. Análise da evolução da proporção de positivas

Analisemos agora as proporções de classificações positivas e sua evolução ao longo dos Ciclos de IA. Na Tabela 30 apresenta-se a evolução do número de positivas e de negativas ao longo dos ciclos de IA, bem como a proporção de positivas. Verifica-se que esta proporção apresenta uma tendência claramente crescente.

Tabela 30: Número de positivas e negativas por ciclo de IA e proporção de classificações positivas

	1.º Ciclo	**2.º Ciclo**	**3.º Ciclo**
Negativas	1221	866	138
Positivas	744	972	351
Total	**1965**	**1838**	**489**
Proporção Positivas (p)	0.379	0.529	0.718

Pretendemos testar se as proporções de positivas nos diferentes Ciclos diferem de forma estatisticamente significativa entre si. Para isso aplicamos o teste do Qui-Quadrado a uma tabela de contingência envolvendo as proporções de positivas e negativas das classificações dos alunos, cujos valores se encontram na Tabela 30.

Formulamos assim, as seguintes hipóteses estatísticas:

H_0: A proporção de positivas é idêntica em todos os ciclos de IA.
H_1: A proporção de positivas é diferente em pelo menos um par de ciclos de IA.

Assim, considerando um nível de significância de **0.05**, e 2 graus de liberdade, obtivemos os seguintes valores aproximados:

Estatística do Teste: **209.15**
Valor Crítico: **5.991**
o valor p: **0.000**

Como $\chi^2_{(0.95,2)} = 5.991$ $(p<0.001)$ podemos assim rejeitar a hipótese nula e concluir que existem diferenças estatisticamente significativas entre pelo menos um par de ciclos de IA em relação às proporções de positivas.

Contudo, este procedimento não nos diz quais os pares de Ciclos para os quais existem essas diferenças. Assim sendo, iremos averiguar entre que Ciclos existem estas diferenças utilizando o Método de *Marascuilo*, o qual permite efetuar a comparação das proporções entre todos os pares de Ciclos.

A versão atual do suplemento "*Real Statistics Resource Pack*" também não contém o Método de *Marascuilo*, tendo os cálculos sido realizados com as fórmulas adequadas do *MS Excel*™. No cálculo dos valores críticos manteve-se 0.05 de nível de significância. Os resultados encontram-se na Tabela 31.

Tabela 31: Resultados da aplicação do Método de *Marascuilo* para existência de diferenças entre as proporções de positivas nos ciclos de IA, no 1.º semestre

	S: Sim	N: Não	
Marascuilo	**1.º Ciclo**	**2.º Ciclo**	**3.º Ciclo**
1.º Ciclo		S	S
2.º Ciclo			S
3.º Ciclo			

Tal como acontece para as médias dos ciclos de IA, realçam-se as diferenças estatisticamente significativas entre as proporções de positivas entre todos os ciclos de IA.

Os resultados da aplicação do Método de *Marascuilo* vêm confirmar os resultados anteriores e assim realçar as conclusões já descritas anteriormente. Em termos globais e em função dos testes estatísticos aplicados às classificações dos alunos por ciclos de IA neste 1.º semestre, podemos confirmar a subida continuada e acentuada da média e proporção de positivas das

classificações dos alunos. A Tabela 32 resume os valores testados: a média das classificações e a proporção de positivas aos longos dos ciclos de IA neste semestre.

Tabela 32: Média e percentagem de positivas das classificações dos alunos ao longo do 1.º Semestre por ciclos de IA

	1.º Ciclo	2.º Ciclo	3.º Ciclo
Negativas	1221	866	138
Positivas	744	972	351
Total	**1965**	**1838**	**489**
Média	6.7	7.8	9.7
Proporção de Positivas	**0.38**	**0.53**	**0.72**
Percentagem de Positivas	**38%**	**53%**	**72%**

6.2. Análise da evolução das classificações referentes ao 2.º Semestre

Os procedimentos iniciais para este 2.º semestre foram iguais aos do 1.º, tal como consta no início da secção 6.1.

Recordemos que o 2.º semestre compreende as UC Matemática Aplicada e Matemática II dos cursos de Licenciatura "Contabilidade e Administração" e "Comércio Internacional", respetivamente.

Vejamos de seguida se os bons resultados, em termos de evolução positiva das classificações durante o 1.º Semestre, acontecem também no 2.º Semestre. Tratando-se de UC distintas do 1.º Semestre, atendendo aos seus conteúdos são, de uma forma geral, consideradas pelos alunos e professores mais exigentes.

6.2.1. Análise da evolução da média das classificações

Relativamente ao 2.º semestre, foram recolhidas 4092 classificações correspondentes a 7 anos letivos.

Na Tabela 33 apresenta-se a análise descritiva das classificações dos alunos durante o 2.º semestre entre os anos letivos 2008 e 2014. Destaca-se na tabela, com cores diferentes, a informação dos anos letivos que compõem cada ciclo de IA.

Tabela 33: Análise descritiva de alguns parâmetros estatísticos referente às classificações dos alunos durante o 2.º semestre entre 2008 e 2014

	1.º Ciclo			2.º Ciclo		3.º Ciclo	
	2008	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
Contagem	594	696	610	575	578	569	470
Média	6.4	5.9	5.9	6.8	7.3	8.6	8.9
Erro-padrão	0.19	0.18	0.17	0.19	0.20	0.19	0.24
Mediana	6	5	5	7	10	10	10
Moda	11	0	10	11	10	11	11
Variância da amostra	20.9	22.3	16.8	21.5	23.4	20.8	27.4
Desvio padrão	4.57	4.73	4.10	4.64	4.83	4.56	5.23
Mínimo	0	0	0	0	0	0	0
Máximo	20	18	18	20	20	20	20
Coeficiente Variação de Pearson (CVP) em %	71%	80%	69%	68%	66%	53%	59%

Na Tabela 33 podemos observar que apesar de em 2009 haver uma queda no valor da média das classificações e mantendo-se o mesmo valor em 2010, posteriormente houve uma recuperação. Os dois últimos anos destacam-se na melhoria deste parâmetro, melhoria essa que é reforçada pela observação dos valores da mediana, que também são mais elevados, embora o valor da mediana tenha dado um grande salto em 2012. A moda é idêntica em todos os anos (10 ou 11), com exceção do ano 2009 no qual, surpreendentemente, a moda é 0. No entanto, a frequência de 0 é 109, de 10 é 88 e de 11 é 71.

Em relação à variabilidade/dispersão das classificações, verificamos que os valores do desvio padrão em cada ano estão muito próximos uns dos outros, com exceção de um ligeiro aumento em 2014. No entanto, observamos que, para os valores do Coeficiente de Variação, apesar de nenhum dos anos ser considerado homogéneo, os dois últimos anos destacam-se por apresentar maior homogeneidade nas classificações apesar de uma ligeira subida em 2014. Verifica-se que nos cinco primeiros anos do estudo, a média Coeficiente Variação de Pearson é 71%, isto é, em média as classificações têm um desvio de 68% em relação à média. Nos dois últimos anos, a média dos Coeficiente Variação de Pearson é 56%, isto é, em média as classificações têm um desvio de 50% em relação à média. Assim sendo, podemos afirmar que as classificações tendem a ser menos heterogéneas.

Estas informações são corroboradas com o gráfico da Figura 12.

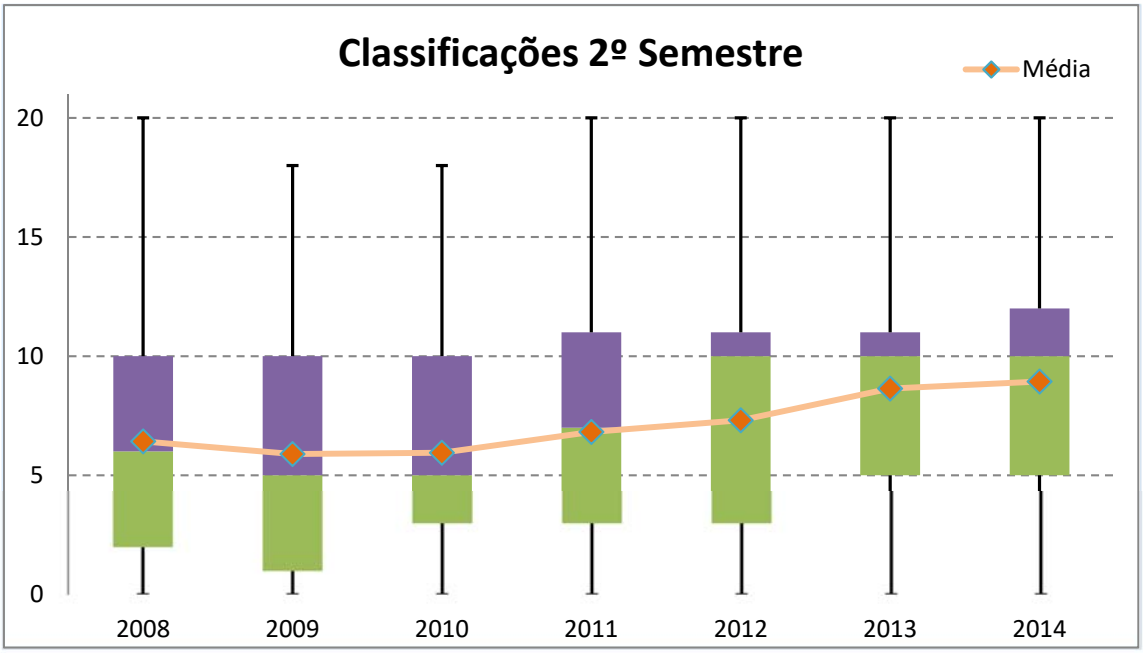


Figura 12: Diagrama de extremos e quartis das classificações entre os anos 2008 e 2014 do 2.º Semestre.

Poderemos assinalar que em 2009 não há alunos com a classificação máxima. Esta situação poderá ser explicada com o facto de nesse ano os alunos, para terem uma classificação superior a 17, terem de fazer uma prova para defesa de nota. A maioria dos alunos optou por não fazer a defesa de nota e, dos que a fizeram, nenhum deles conseguiu acertar em todas as questões. No ano de 2010, os alunos foram avaliados somente por Exame Final e pelos trabalhos realizados no *Moodle*. Contudo, este exame tinha dois grupos de perguntas, das quais o segundo grupo tinha 3 perguntas com um grau de dificuldade maior, pelo que poderá também estar aqui um dos fatores para não haver alunos com classificação máxima. Estes procedimentos deixaram de acontecer nos anos seguintes.

Para melhor comparar os diferentes anos e compreender melhor alguns dos seus resultados, acrescentam-se ainda as seguintes informações:

- o número de testes em avaliação contínua foram sempre três à exceção de 2010 que não houve avaliação contínua, 2012 que foram dois e em 2013 quatro;
- a implementação de um parâmetro à avaliação contínua bem como o Teste de "Repescagem", os moldes de implementação foram iguais aos do 1.º semestre.

O ano de 2011 corresponde a um ano de mudança nas classificações dos alunos, em termos globais. Neste ano e em anos posteriores, a evolução positiva nas classificações é notória. Claramente, a mediana a partir de 2011 começou a subir, atingindo valores positivos em 2012 e anos seguintes. Visivelmente, em 2013 (aqui mais notório) e 2014 o intervalo interquartis Q3-Q1 é menor. Logo, podemos afirmar que as classificações estão mais concentradas em torno da média. Parece contudo, ter havido um ligeiro agravamento no último ano (2014).

Em termos gerais, parece haver uma evolução positiva nas classificações dos alunos ao longo dos anos, que se poderá observar no gráfico da Figura 12 e ainda na Tabela 33. Contudo, é conveniente verificar se as diferenças descritas são ou não estatisticamente significativas. Para isso, recorremos a alguns testes estatísticos que vamos apresentar de seguida. Iremos testar as diferenças entre as médias das classificações e a diferença entre as proporções de positivas nas classificações.

Para testar se as diferenças entre as médias nos diferentes anos letivos são estatisticamente significativos, vamos comparar este parâmetro entre as diferentes amostras (mais do que duas) usando a Análise de Variância, vulgo ANOVA a um fator, de modo análogo ao que foi realizado para o 1.º semestre.

As condições de aplicabilidade deste teste, já foram expostas no 1.º semestre e verificam-se aqui, de forma idêntica.

Pretendemos testar se as médias das classificações nos diferentes anos diferem entre si de forma estatisticamente significativa. Isto é, pretendemos testar as seguintes hipóteses:

H_0: A classificação é, em média, idêntica nos diferentes anos letivos.
H_1: A classificação é, em média, diferente em pelo menos um par de anos letivos

Aplicando ANOVA a um fator, obtiveram-se os resultados, apresentados na Tabela 34.

Tabela 34: Resultados da aplicação do teste ANOVA às classificações dos alunos nos 7 anos letivos, no 2.º semestre

ANOVA: Single Factor								
DESCRIPTION					Alpha	0.05		
Groups	Count	Sum	Mean	Variance	SS	Std Err	Lower	Upper
2008	594	3816	6.4242	20.9192	12405.09	0.1912	6.0488	6.7997
2009	696	4101	5.8922	22.3438	15528.92	0.1766	5.5455	6.2390
2010	610	3628	5.9475	16.8051	10234.32	0.1886	5.5771	6.3180
2011	575	3921	6.8191	21.5491	12369.19	0.1943	6.4375	7.2008
2012	578	4225	7.3097	23.3684	13483.57	0.1938	6.9291	7.6903
2013	569	4914	8.6362	20.7988	11813.69	0.1953	8.2526	9.0198
2014	470	4199	8.9340	27.3752	12838.96	0.2149	8.5117	9.3563
ANOVA	4092							
Sources	SS	df	MS	F	P value	F crit	RMSSE	Omega Sq
Between Groups	5076.009	6	846.0015	38.9734	2.61E-46	2.1008	0.2640	0.0527
Within Groups	88673.74	4085	21.7072					
Total	93749.74	4091	22.9161					

A partir da leitura da tabela, podemos verificar que o valor da estatística do teste F é de aproximadamente **38.9734** e, considerando um intervalo de confiança para a média de 95% (α=0.05), este valor é consideravelmente superior ao seu valor crítico que é aproximadamente **2.1008** e temos um valor p aproximadamente igual a 2.6×10^{-46} (praticamente nulo) que é inferior ao valor alfa. Como $F_{(4091,6)} = 38.97 (p < 0.001)$ podemos assim rejeitar a hipótese nula e concluir que existem diferenças estatisticamente significativas entre pelo menos um par de anos em relação à média das classificações.

Na nossa análise, pode-se referir que a magnitude da diferença entre as médias é moderada, quer calculada pela medida RMSSE (0.2640), quer pela medida Omega Sq (0.0527). Apesar de que o tamanho do efeito pela medida RMSSE já estar dentro do intervalo Elevado, mas está longe do valor extremo desse efeito.

É ainda necessário saber qual ou quais os pares de anos para os quais existem diferenças estatisticamente significativas em relação à média das classificações. Como as amostras neste 2.º semestre verificam as condições para que se aplique o método de *Tukey HSD (Honest Significant Difference)/Tukey-Kramer*, é este o escolhido.

Na Tabela 35 apresenta-se o resultado da aplicação do método de *Tukey* para dois dos anos (2008 e 2013), utilizando os chamados Contrastes Ortogonais.

Tabela 35: Resultados da aplicação do Método de *Tukey* usando Contrastes Ortogonais para os anos 2008 e 2013 no 2.º semestre

TUKEY'S HSD / TUKEY-KRAMER					Alpha	0.05
Groups	c	mean	n	ss	c^2/n	c*mean
2008	1	6.4242	594	12405.09	0.0017	6.4242
2009		5.8922	696	15528.92	0	0
2010		5.9475	610	10234.32	0	0
2011		6.8191	575	12369.19	0	0
2012		7.3097	578	13483.57	0	0
2013	-1	8.6362	569	11813.69	0.0018	-8.6362
2014		8.9340	470	12838.96	0	0
			4092	88673.74	0.0034	-2.2120

Q TEST

std err	q-stat	df	q-crit	lower	upper	sig	x-crit	*Cohen d*	effect r
0.1933	-11.4459	4085	4.17	-3.0178	-1.4061	yes	0.8059	0.4748	0.1763

A partir da Tabela 35, verifica-se que existe uma diferença significativa entre as médias dos anos 2008 e 2013 (sig= yes). Considerando que a medida *Cohen d* é igual 0.4748 considera-se que a diferença entre as médias é moderada.

De forma análoga utilizamos o Método de *Tukey HSD/Tukey-Kramer*, para todos os possíveis pares de anos, recorrendo aos Contrastes. Na Tabela 36 apresenta-se um resumo dos resultados obtidos.

Tabela 36: Resultados da aplicação do Método de *Tukey* em relação à diferença ou não entre as médias das classificações aplicados os pares de anos letivos, no 2.º semestre

	S: Sim	**N**: Não	(*Cohen d*)			Alpha 0.05	
TUKEY HSD	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
2008		N (0.11) p≈0.387	N (0.10) p≈0.565	N (0.08) p≈0.775	S (0.19) p≈0.020	S (0.47) p≈0.000	S (0.54) p≈0.000
2009			N (0.01) p≈1	S (0.20) p≈0.008	S (0.30) p≈0.000	S (0.59) p≈0.000	S (0.65) p≈0.000
2010				S (0.19) p≈0.022	S (0.29) p≈0.000	S (0.58) p≈0.000	S (0.64) p≈0.000
2011					N (0.11) p≈0.557	S (0.39) p≈0.000	S (0.45) p≈0.000
2012						S (0.28) p≈0.000	S (0.35) p≈0.000
2013							N (0.06) p≈0.948
2014							

Realçam-se as diferenças estatisticamente significativas entre as médias de cada ano e os últimos dois anos letivos. Destaca-se ainda que os valores *Cohen d* indicam que essas diferenças são de nível moderado a elevado. Há ainda diferenças estatisticamente significativas entre as médias dos pares de anos 2008/2012, 2009/2011, 2009/2012, 2010/2011 e 2010/2012, sendo que, atendendo ao valor da medida *Cohen d*, estas diferenças são moderadas e duas delas são pequenas. Quanto aos restantes pares de anos, as diferenças entre as médias não são estatisticamente significativas. Verifica-se ainda que há diferenças estatisticamente significativas entre a média das classificações observada em 2013 e todos os anos anteriores, o que confirma o valor superior da média registada este ano.

6.2.2. Análise da evolução da proporção de classificações positivas

A Figura 13 apresenta um gráfico que ilustra a evolução da percentagem das classificações positivas e negativas dos alunos no 2.º semestre, ao longo dos anos 2008 a 2014. Verifica-se que

desde 2009, a percentagem de positivas apresenta uma tendência crescente e a percentagem de negativas uma tendência decrescente. No entanto, os dois últimos anos mostram alguma estagnação.

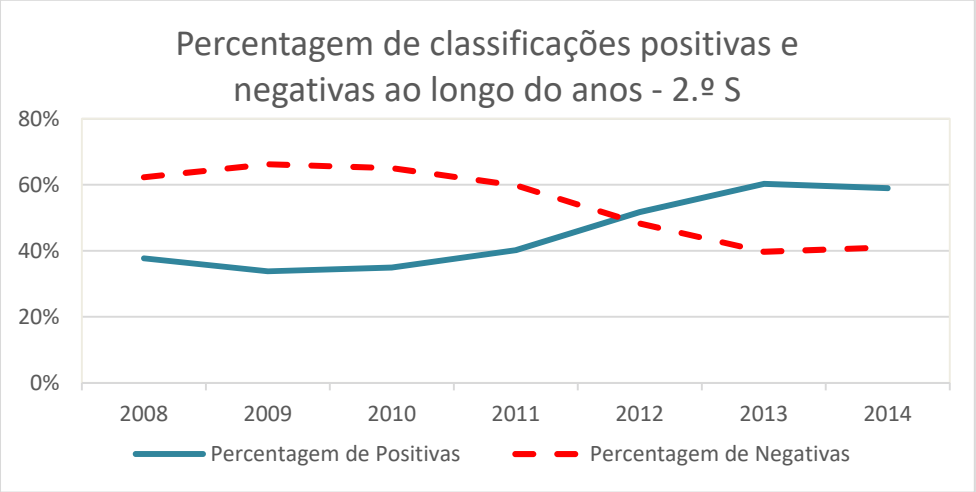


Figura 13: Evolução da percentagem das classificações positivas e negativas no 2.º semestre.

Analisemos agora as proporções de classificações positivas e sua evolução ao longo dos anos 2008 a 2014 no 2.º semestre. Na Tabela 37 apresenta-se a evolução do número de classificações de positivas e de negativas ao longo dos anos, bem como a proporção de classificações positivas. Verifica-se que a proporção de positivas apresenta uma tendência claramente crescente entre 2009 e 2013, apresentando um ligeira redução em 2014.

Tabela 37: Número de positivas e negativas por ano letivo e proporção de classificações positivas, no 2.º semestre

	2008	2009	2010	2011	2012	2013	2014
Negativas	311	364	319	301	303	298	246
Positivas	283	332	291	274	275	271	224
Total	594	696	610	575	578	569	470
Proporção Posit. (p)	0.377	0.338	0.349	0.402	0.517	0.603	0.589

Pretendemos testar se as proporções de positivas nos diferentes anos diferem entre si de forma estatisticamente significativa, para isso aplicamos o teste do Qui-Quadrado a uma tabela de contingência envolvendo as proporções de positivas e negativas das classificações dos alunos, cujos valores se encontram na Tabela 37.

Formulamos assim, as seguintes hipóteses:

H_0: A proporção de positivas é idêntica em todos os anos letivos.
H_1: A proporção de positivas é diferente em pelo menos um par de anos letivos.

Para testar estas hipóteses, não podemos utilizar o suplemento "*Real Statistics Resource Pack*" porque não contém, na versão atual, o teste de Qui-Quadrado para proporções de mais do que duas variáveis, por isso todos os cálculos foram realizados em *MS Excel*™, usando as fórmulas adequadas. Assim, considerando um valor de significância de **0.05**, e 6 graus de liberdade, obtivemos os seguintes valores aproximados:

Estatística do Teste: **194.194**
Valor Crítico: **12.592**
o valor p: **0.000**

Como $\chi^2_{(0.95,6)} = 12.592 \ (p < 0.001)$, podemos assim rejeitar a hipótese nula e concluir que existem diferenças estatisticamente significativas entre pelo menos um par de anos em relação às proporções de positivas

Contudo, este procedimento não nos diz quais os pares de anos para os quais existem essas diferenças. Assim, iremos averiguar entre que anos existem estas diferenças utilizando o Método de *Marascuilo,* o qual efetua a comparação das proporções entre todos os pares de anos.

No cálculo dos valores críticos manteve-se 0.05 com nível de significância. Os resultados encontram-se na Tabela 38.

Tabela 38: Resultados da aplicação do Método de *Marascuilo* para existência de diferenças entre as proporções de classificações positivas entre os diferentes pares de anos letivos, no 2.º semestre

S: Sim N: Não

Marascuilo	2008	2009	2010	2011	2012	2013	2014
2008		N	N	N	S	S	S
2009			N	N	S	S	S
2010				N	S	S	S
2011					S	S	S
2012						N	N
2013							N
2014							

Notam-se algumas diferenças de resultados em relação às médias das classificações. Enquanto nas médias se realçava as diferenças estatisticamente significativas entre as médias de cada ano e os últimos dois anos letivos, relativamente às proporções de positivas somente há diferenças estatisticamente significativas entre as proporções de positivas de cada ano e os últimos três anos letivos. Quanto aos restantes pares de anos, a diferença entre as proporções de positivas não é estatisticamente significativa. Destaca-se que para o par de anos 2011/2012 a diferença entre as proporções de positivas é estatisticamente significativa, mas o mesmo não acontece para a média das classificações. Já para os pares de anos 2009/2011, 2010/2011, 2012/2013 e 2012/2014, há diferenças estatisticamente significativas para as médias das classificações, mas não para a proporção de positivas.

Os resultados do Método de *Marascuilo* vêm confirmar parte dos resultados anteriores e assim realçar as conclusões já descritas anteriormente. Em termos globais e em função dos testes estatísticos aplicados às classificações dos alunos durante o 2.º semestre, podemos confirmar a subida continuada da média e da proporção de positivas das classificações dos alunos depois de 2011, e em especial o aumento em 2013. Também se confirma que nos dois últimos anos, 2013 e 2014, os resultados não são iguais em ambos os testes estatísticos, o que parece confirmar haver algum problema em 2014. A Tabela 39 resume os valores testados, a saber a média das classificações e de proporção de positivas aos longos dos anos em estudo, no 2.º semestre.

Tabela 39: Média e percentagem de positivas das classificações dos alunos ao longo dos anos e por ciclos de estudo no 2.º Semestre

	1.º Ciclo			2.º Ciclo		3.º Ciclo	
	2008	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
Negativas	370	461	397	344	279	226	193
Positivas	224	235	213	231	299	343	277
Total	594	696	610	575	578	569	470
Média	**6.4**	**5.9**	**5.9**	**6.8**	**7.3**	**8.6**	**8.9**
Percentagem Posit.	**38%**	**34%**	**35%**	**40%**	**52%**	**60%**	**59%**

6.2.3. Análise da evolução das Classificações por Ciclos de IA

Importa agora verificar a evolução das classificações dos alunos por ciclos de IA. Isto é, averiguar se existem diferenças estatisticamente significativas entre as médias das classificações entre os diferentes ciclos de IA e analisar ainda se existem diferenças para as proporções de positivas. Recordamos que, neste 2.º semestre, o 1.º ciclo de IA envolve os anos de 2008 a 2010, o 2.º ciclo envolve os anos 2011 e 2012 e o 3.º ciclo corresponde aos anos 2013 e 2014.

6.2.3.1. Análise da evolução da média das classificações

Em primeiro lugar, apresenta-se a Tabela 40 com uma Análise Descritiva sumária das classificações dos alunos em cada um dos três ciclos de IA. Podemos constatar, em termos genéricos, que parece haver diferenças entre os Ciclos, no que concerne à média das classificações durante o 2.º semestre de aulas.

Na Tabela 40 podemos observar que o valor da média das classificações apresenta uma tendência crescente. Verifica-se ainda uma melhoria acentuada na mediana que passa de um valor negativo no 2.º ciclo para um valor positivo no 3.º ciclo. A moda é positiva e idêntica nos dois primeiros ciclos (10 valores) e aumenta ligeiramente para 11 valores no 3.º ciclo.

Em relação à variabilidade/dispersão das classificações, verificamos que os valores do desvio padrão em cada Ciclo estão muito próximos uns dos outros. No entanto, observamos que, para os valores do Coeficiente de Variação de Pearson, apesar de nenhum dos Ciclos ser considerado homogéneo, o último Ciclo destaca-se por apresentar muito menor heterogeneidade nas classificações.

Tabela 40: Análise Descritiva sumária das classificações dos alunos pelos respetivos ciclos de IA, no 2.º semestre

	1.º Ciclo	2.º Ciclo	3.º Ciclo
Contagem	1900	1153	1039
Média	6.1	7.1	8.8
Erro-padrão	0.10	0.14	0.15
Mediana	5	7	10
Moda	10	10	11
Variância da amostra	20.15	22.50	23.77
Desvio padrão	4.49	4.74	4.88
Mínimo	0	0	0
Máximo	20	20	20
Coeficiente Variação de Pearson (CVP) em %	74%	67%	56%

É conveniente verificar se as diferenças descritas são ou não estatisticamente significativas. Para isso aplicamos ANOVA a um fator.

Atendendo às dimensões de cada Ciclo, as três condições para aplicar ANOVA são satisfeitas.

De forma análoga ao que foi feito anteriormente, testemos se as médias das classificações entre os diferentes Ciclos diferem entre si de forma estatisticamente significativa.

Formulamos assim, as seguintes hipóteses estatísticas:

H_0: A classificação é, em média, idêntica em todos os ciclos de IA.
H_1: A classificação é, em média, diferente em pelo menos um par de ciclos de IA.

Aplicando ANOVA a um fator, obtiveram-se os seguintes resultados, apresentados na Tabela 41.

Tabela 41: Resultados da aplicação do teste ANOVA aplicado às classificações dos alunos entre os ciclos de IA no 2.º semestre

ANOVA: Single Factor

DESCRIPTION					Alpha	0.05		
Groups	Count	Sum	Mean	Variance	SS	Std Err	Lower	Upper
1.º Ciclo	1900	11545	6.0763	20.1548	38273.9342	0.1070	5.8666	6.2861
2.º Ciclo	1153	8146	7.0650	22.5018	25922.1214	0.1373	6.7957	7.3344
3.º Ciclo	1039	9113	8.7709	23.7721	24675.4822	0.1446	8.4871	9.0547

ANOVA

Sources	SS	df	MS	F	P value	F crit	RMSSE	Omega Sq
Between Groups	4878.206	2	2439.103	112.2237	0	2.9979	0.2924	0.0516
Within Groups	88871.54	4089	21.7343					
Total	93749.74	4091	22.9161					

A partir da leitura da tabela, podemos verificar que o valor da estatística do teste F é de aproximadamente **112.2237** e, considerando um intervalo de confiança para a média de 95% (α=0.05), este valor é consideravelmente superior ao seu valor crítico que é aproximadamente **2.9979,** temos um valor p aproximadamente igual a 0 (praticamente nulo), que é inferior ao valor alfa. Como $F_{(4091,6)} = 112.22 (p < 0.001)$ podemos assim rejeitar a hipótese nula e concluir que existem diferenças estatisticamente significativas entre pelo menos um par de ciclos de IA em relação à média das classificações.

Na nossa análise, pode-se referir que a magnitude da diferença entre as médias, calculada por RMSSE (0.2924) é considerada grande, mas o valor da medida Omega Sq (0.0516) é considerada moderada.

É ainda necessário saber qual ou quais os pares de anos para os quais existem diferenças estatisticamente significativas em relação à média das classificações

Atendendo às dimensões de cada ciclo, podemos aplicar o Método de *Tukey HSD / Tukey-Kramer*. Na Tabela 42 representa-se o resultado de aplicação do método de *Tukey* para dois dos ciclos (1.º e 3.º), utilizando os chamados Contrastes Ortogonais.

Tabela 42: Resultados de aplicação do Método de *Tukey*, usando contrastes ortogonais para o 1.º ciclo e o 3.º ciclo no 2.º semestre

TUKEY'S HSD / TUKEY-KRAMER					Alpha	0.05
Groups	c	mean	n	ss	c^2/n	c*mean
1.º Ciclo	1	6.0763	1900	38273.93	0.0005	6.0763
2.º Ciclo		7.0650	1153	25922.12	0	0

3.º Ciclo	-1	8.7709	1039	24675.48	0.0010	-8.7709
			4092	88871.54	0.0015	-2.6946

Q TEST

std err	q-stat	df	q-crit	lower	upper	sig	x-crit	*Cohen d*	effect r
0.1272	-21.1848	4089	3.314	-3.1161	-2.2731	yes	0.4215	0.5780	0.3145

Podemos verificar na Tabela 42, que existe uma diferença estatisticamente significativa entre as médias dos 1.º e 3.º Ciclos de IA (sig = yes). Considerando que a medida *Cohen d* é igual 0.5780 considera-se que a diferença entre as médias é elevada, mas apenas um pouco acima do efeito moderado.

De forma análoga utilizamos o Método de *Tukey HSD/Tukey-Kramer*, para todos os possíveis pares de Ciclos, recorrendo aos Contrastes. Na Tabela 43 apresenta-se um resumo dos resultados obtidos.

Tabela 43: Resultados de aplicação do Método de *Tukey* em relação à diferença entre as médias das classificações aplicados a todos os pares de Ciclos, no 2.º semestre

	S: Sim	**N**: Não	*(Cohen d)*
TUKEY'S HSD	**1.º Ciclo**	**2.º Ciclo**	**3.º Ciclo**
1.º Ciclo		**S (0.21) p≈0.000**	**S (0.58) p≈0.000**
2.º Ciclo			**S (0.37) p≈0.000**
3.º Ciclo			

Realçam-se as diferenças estatisticamente significativas entre as médias de todos os Ciclos. Destaca-se ainda que os valores *Cohen d* indicam que essas diferenças são de nível moderado a elevado. Realça-se o nível elevado na magnitude do efeito entre o 1.º ciclo e o 3.º ciclo de IA.

Atendendo às evidências estatísticas, bem como ao nível dos efeitos, a evolução entre os ciclos foi bastante positiva ao nível das médias das classificações.

6.2.3.2. Análise da evolução da proporção de positivas

Analisemos agora as proporções de classificações positivas e sua evolução ao longo dos ciclos de IA. Na Tabela 44 apresenta-se a evolução do número de positivas e de negativas ao longo dos ciclos de IA, bem como a proporção de positivas. Verifica-se que esta proporção de positivas apresenta uma tendência claramente crescente.

Tabela 44: Número de positivas e negativas por ciclo de IA e proporção de classificações positivas

	1.º Ciclo	2.º Ciclo	3.º Ciclo
Negativas	1228	623	419
Positivas	672	530	620
Total	1900	1153	1039
Proporção Posit. (p)	0.354	0.460	0.597

Pretendemos testar se as proporções das classificações positivas entre os diferentes Ciclos de IA diferem entre si de forma estatisticamente significativa, para isso aplicamos o teste do Qui-Quadrado a uma tabela de contingência envolvendo as proporções de positivas e negativas das classificações dos alunos, cujos valores se encontram na Tabela 44.

Formulamos assim, as seguintes hipóteses:

H_0: A proporção de positivas é idêntica em todos os ciclos de IA.
H_1: A proporção de positivas é diferente em pelo menos um par de ciclos de IA.

Assim, considerando um nível de significância de **0.05**, e 2 graus de liberdade, obtivemos os seguintes valores aproximados:

Estatística do Teste: **161.98**
Valor Crítico: **5.991**
o valor p: **0.000**

Como $\chi^2_{(0.95,2)} = 5.991\ (p<0.001)$ podemos assim rejeitar a hipótese nula e concluir que existem diferenças estatisticamente significativas entre pelo menos um par de ciclos de IA em relação às proporções de positivas.

Contudo, este procedimento não nos diz quais os pares de Ciclos para os quais existem essas diferenças. Assim, iremos averiguar entre que Ciclos existem estas diferenças utilizando o Método de *Marascuilo,* o qual permite efetuar a comparação das proporções entre todos os pares de Ciclos.

No cálculo dos valores críticos manteve-se 0.05 de nível de significância. Os resultados encontram-se na Tabela 45.

Tabela 45: Resultados de aplicação do Método de *Marascuilo* para existência ou não de diferenças entre as proporções de positivas entre os diferentes ciclos de IA, no 2.º semestre

	S: Sim	N: Não	
Marascuilo	1.º Ciclo	2.º Ciclo	3.º Ciclo
1.º Ciclo		S	S
2.º Ciclo			S
3.º Ciclo			

Tal como acontece para as médias dos ciclos de IA, realçam-se as diferenças estatisticamente significativas entre as proporções de positivas entre todos os ciclos de IA.

Os resultados de aplicação do Método de *Marascuilo* vêm confirmar os resultados anteriores e assim realçar as conclusões já descritas anteriormente. Em termos globais e em função dos testes estatísticos aplicados às classificações dos alunos por ciclos de IA neste 2.º semestre, podemos confirmar a subida continuada e acentuada da média e da proporção de positivas das classificações dos alunos. A Tabela 46 resume os valores testados: a média das classificações e a proporção de positivas aos longos dos ciclos de IA neste semestre.

Tabela 46: Média e percentagem de positivas das classificações dos alunos ao longo do 2.º Semestre por ciclos de IA

	1.º Ciclo	2.º Ciclo	3.º Ciclo
Negativas	1228	623	419
Positivas	672	530	620
Total	1900	1153	1039
Média	6.1	7.1	8.8
Proporção de Positivas	0.35	0.46	0.60
Percentagem de Positivas	35%	46%	60%

Dos resultados apresentados, quer por ano letivos quer por ciclos de IA, nota-se uma evolução positiva nas classificações dos alunos. As intervenções realizadas em cada ciclo de IA e explicadas ao longo da tese confirmam que alguns dos objetivos a que se propunham foram atingidos.

6.2.4. Síntese da evolução das classificações nos dois semestres

Em jeito de síntese, foi possível verificar que houve uma evolução positiva das avaliações dos alunos nas amostras estudadas. Nota-se alguma dificuldade nos anos iniciais do estudo, mas que foram evoluindo positivamente ao longo dos anos até 2014, apesar de parecer haver alguma estabilidade nos dois últimos anos. Contudo, verificamos que os resultados foram globalmente

melhores no 1.º semestre. As análises estatísticas aplicadas e valores apresentados anteriormente confirmam estas conclusões.

Se consideramos a análise por ciclos tanto no 1.º semestre como no 2.º, os resultados das classificações foram muito animadores e consideravelmente muito positivos. As análises estatísticas aqui aplicadas revelaram valores estatisticamente bastante significativos entre os ciclos de IA em ambos semestres.

6.3. Análise das respostas ao questionário aos docentes no 1.º ciclo de IA

Responderam ao questionário 11 docentes, o que corresponde à totalidade dos docentes envolvidos no desenvolvimento das QEM, neste 1.º ciclo de IA. Todas as respostas foram consideradas válidas. Nas Figura 14, Figura 15, Figura 16, Figura 17 e Figura 18, apresenta-se a distribuição das respostas obtidas ao questionário, tomando cada uma das categorias que agrupam as diferentes linhas de orientação. No caso das questões de resposta aberta, apenas um dos docentes introduziu um comentário, a saber, "Penso que o número de opções adequado deverá ser 4". Apresentam-se na Tabela 47 as medidas estatísticas da média, desvio padrão e moda das respostas dos docentes por linha de orientação.

Tabela 47: Média, desvio padrão e moda nos itens do questionário aos docentes no 1º ciclo de IA

Linhas de Orientação	Média	Desvio padrão	Moda
CUIDADOS COM O CONTEÚDO			
Cada questão deve refletir conteúdo específico e um único comportamento mental concreto, tal como preconizado nas especificações dos testes.	4.2	0.94	5
Fundamentar cada questão em termos de conteúdos de aprendizagem importantes; evitar conteúdo trivial.	4.0	0.85	4
Utilizar materiais inovadores para testar aprendizagens de nível mais elevado. Reescrever a linguagem utilizada no livro de apoio ou a linguagem utilizada durante as aulas, quando incluídas nas questões de um teste, de modo a evitar testes apenas de memorização.	4.4	0.77	5
Manter o conteúdo de cada questão independente do conteúdo de outras questões do teste.	3.8	1.19	5

Linhas de Orientação	Média	Desvio padrão	Moda
Evitar conteúdos demasiado específicos ou demasiado genéricos ao escrever as questões.	3.3	0.86	3
Evitar questões baseadas em opiniões.	4.6	0.64	5
Evitar questões com artimanhas.	3.6	1.07	4 e 5
Manter o vocabulário simples, tendo em conta o grupo de alunos que está a ser testado.	4.0	0.74	4
CUIDADOS COM A FORMATAÇÃO			
Formatar a questão verticalmente e não horizontalmente.	3.5	0.89	3
CUIDADOS COM O ESTILO			
Editar e rever as questões.	4.6	0.77	5
Usar corretamente a gramática, a pontuação, as letras maiúsculas e a ortografia.	5.0	0.00	5
Minimizar a quantidade de leitura necessária em cada questão.	3.9	0.90	4
ENUNCIADO DA QUESTÃO			
Certificar-se que as instruções no enunciado são muito claras.	4.9	0.29	5
Incluir a ideia central no enunciado ao invés de nas opções.	4.1	0.67	4
Evitar palavreado excessivo.	4.2	0.94	5
Escrever o enunciado na forma afirmativa, evitando negações tais como NÃO ou EXCETO. Se forem utilizadas negações, usar as palavras com cautela e garantir sempre que a palavra aparece em maiúsculas e em negrito.	3.3	1.21	3 e 4
OPÇÕES DA QUESTÃO			
Desenvolver tantas opções eficazes quantas seja possível, mas a investigação sugere que três é adequado.	2.8	1.11	2
Certificar-se que apenas uma dessas opções é a resposta correta.	4.8	0.39	5
Variar a localização da resposta correta de acordo com o número de opções.	4.6	0.64	5
Colocar as opções por ordem, lógica ou numérica.	2.3	1.42	1 e 2

Linhas de Orientação	Média	Desvio padrão	Moda
Garantir opções independentes; as opções não devem ter elementos comuns.	3.2	1.19	3
Garantir opções homogéneas, quer em termos de conteúdo quer em termos de estrutura gramatical.	4.0	0.95	4 e 5
Manter o tamanho das opções aproximadamente igual.	3.6	1.07	4
Utilizar cuidadosamente "Nenhum dos anteriores".	3.9	1.38	5
Evitar utilizar “Todos os anteriores”.	4.1	1.16	5
Escrever as opções na forma afirmativa; evitar negações tais como NÃO.	3.4	1.23	4
Evitar dar dicas para a resposta correta, tais como: a) Determinantes específicos incluindo sempre, nunca, completamente e absolutamente; b) Associações de palavras com sons idênticos, escolhas idênticas ou parecidas com termos utilizados no enunciado; c) Incoerências gramaticais que deem pistas ao aluno sobre a resposta correta. d) Resposta correta evidente; e) Pares ou tripletos de opções que irão indicar ao aluno a resposta correta; f) Opções ostensivamente absurdas ou ridículas.	3.8	1.11	5
Garantir que todos os distratores são plausíveis.	4.0	0.85	3 e 5
Usar erros típicos dos alunos para escrever os distratores.	3.8	1.34	5
Utilizar humor, se ele é compatível com o professor e com o ambiente de aprendizagem.	2.8	1.19	3

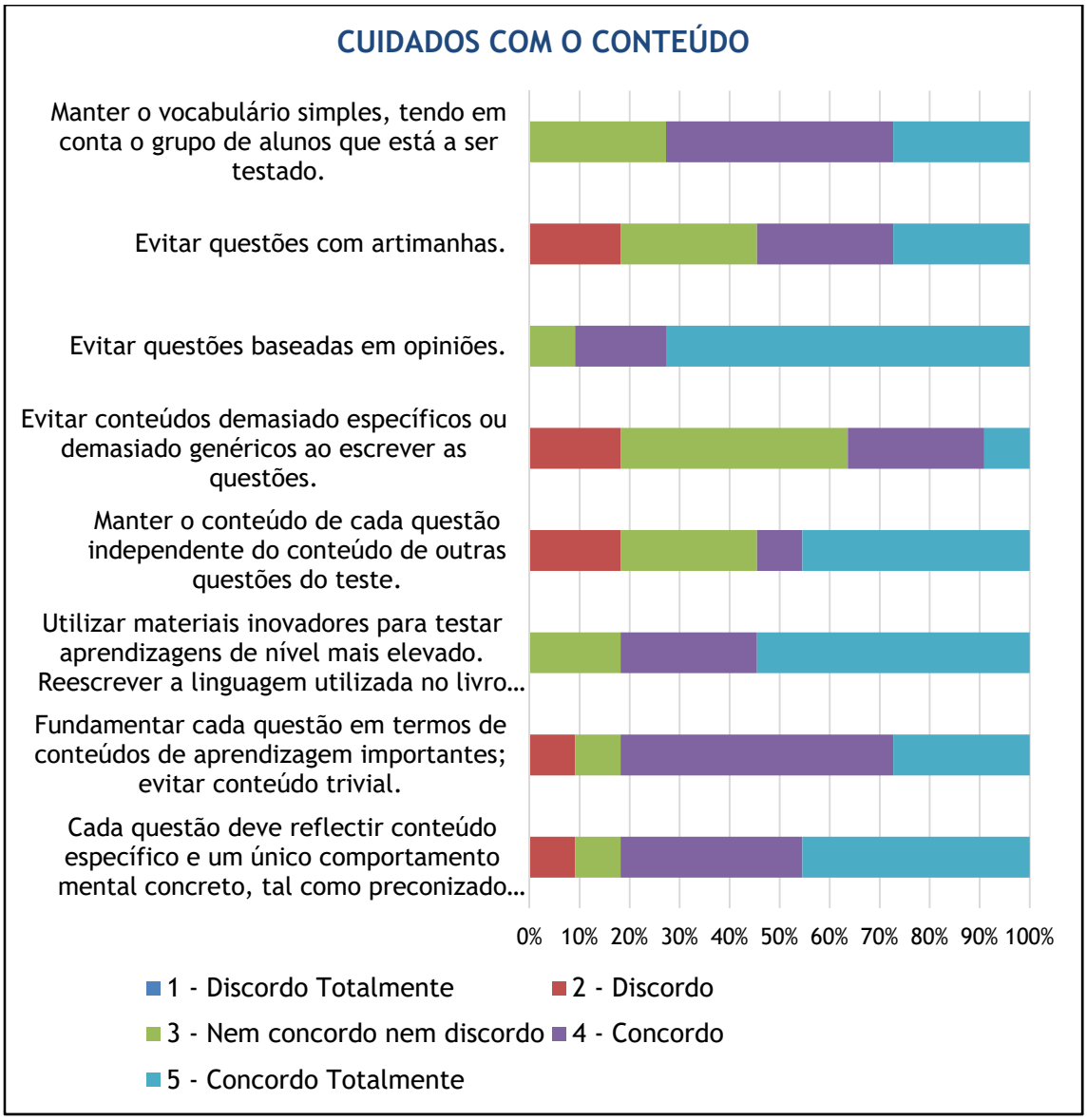


Figura 14: Frequência das repostas dos docentes quanto aos "Cuidados com o Conteúdo".

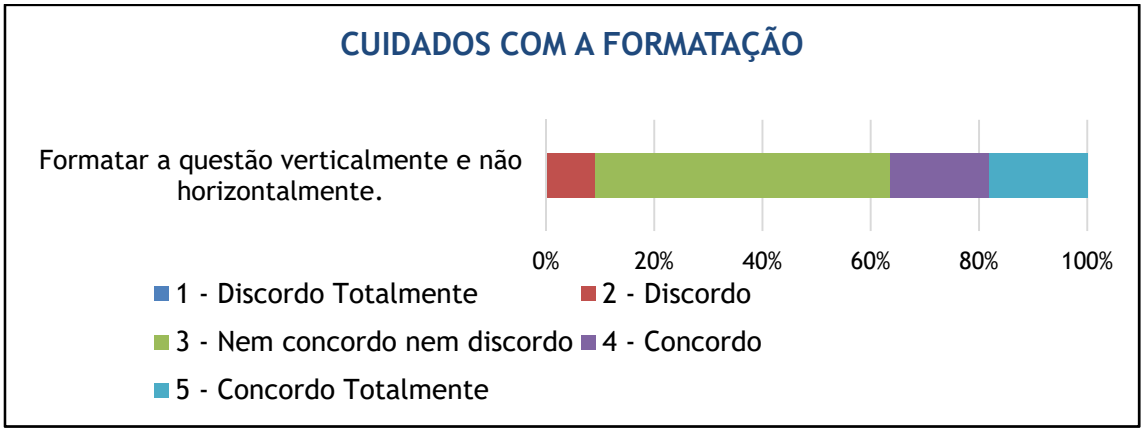


Figura 15: Frequência das repostas dos docentes quanto aos "Cuidados com a Formatação".

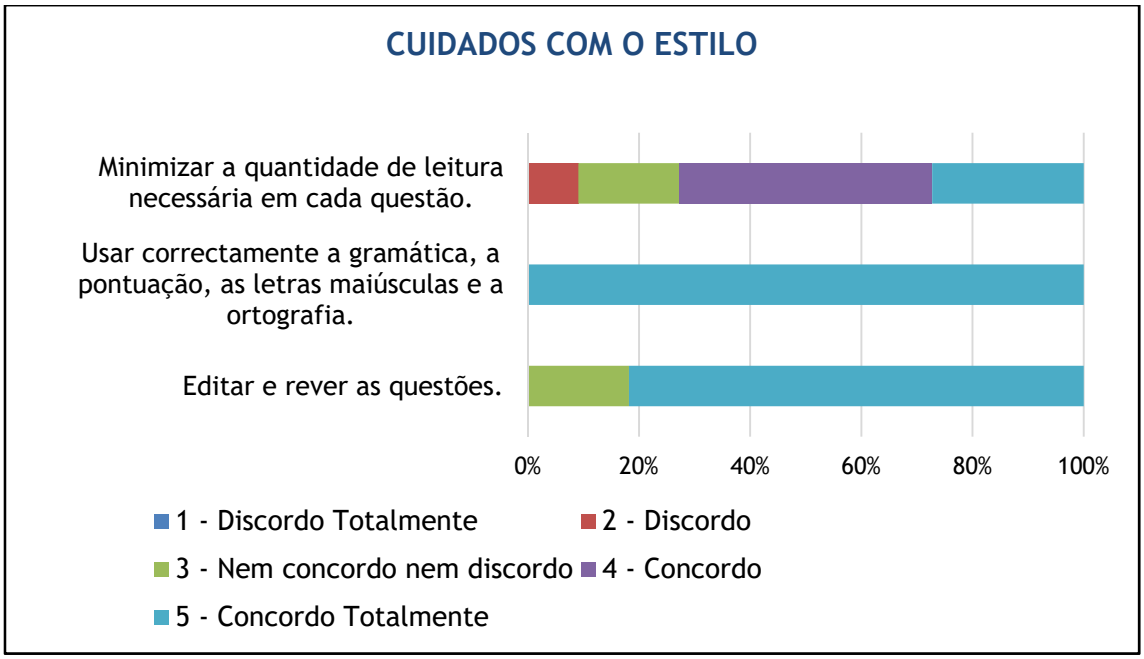


Figura 16: Frequência das repostas dos docentes quanto aos "Cuidados com o Estilo".

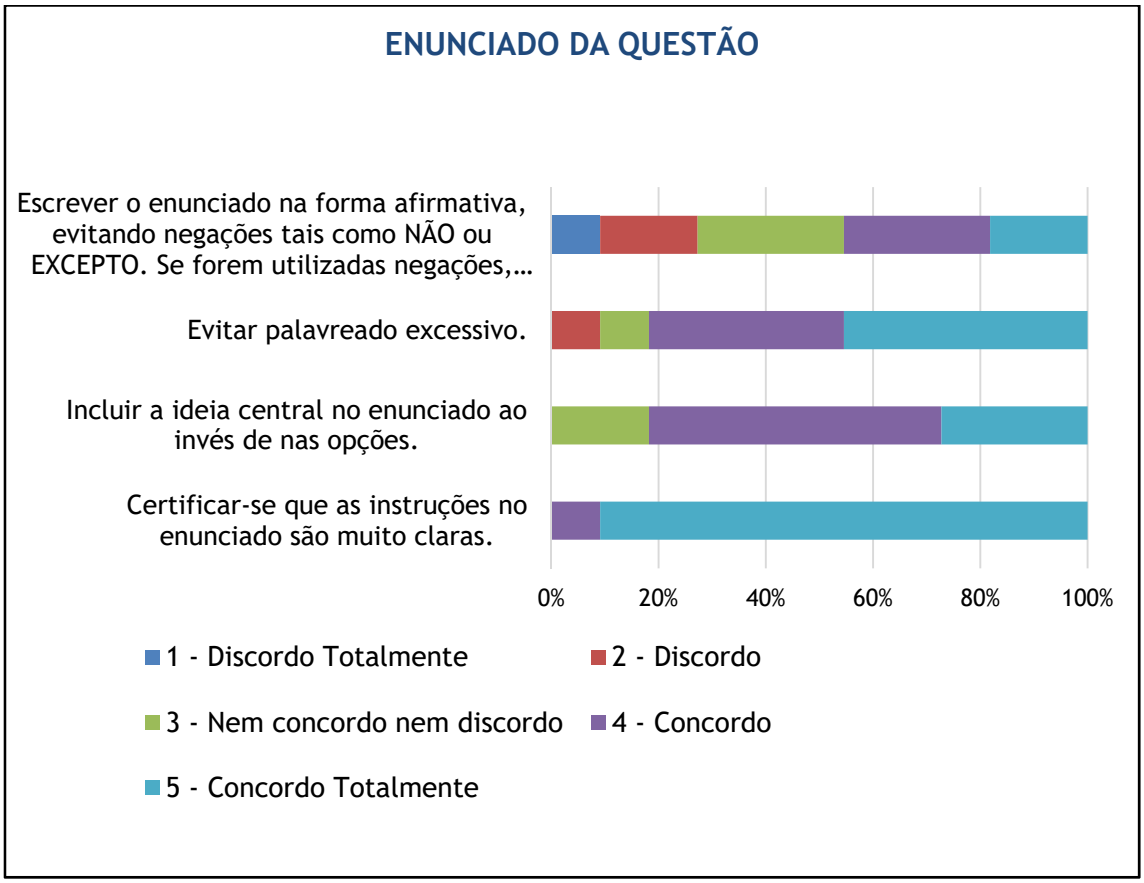


Figura 17: Frequência das repostas dos docentes quanto ao "Enunciado da Questão".

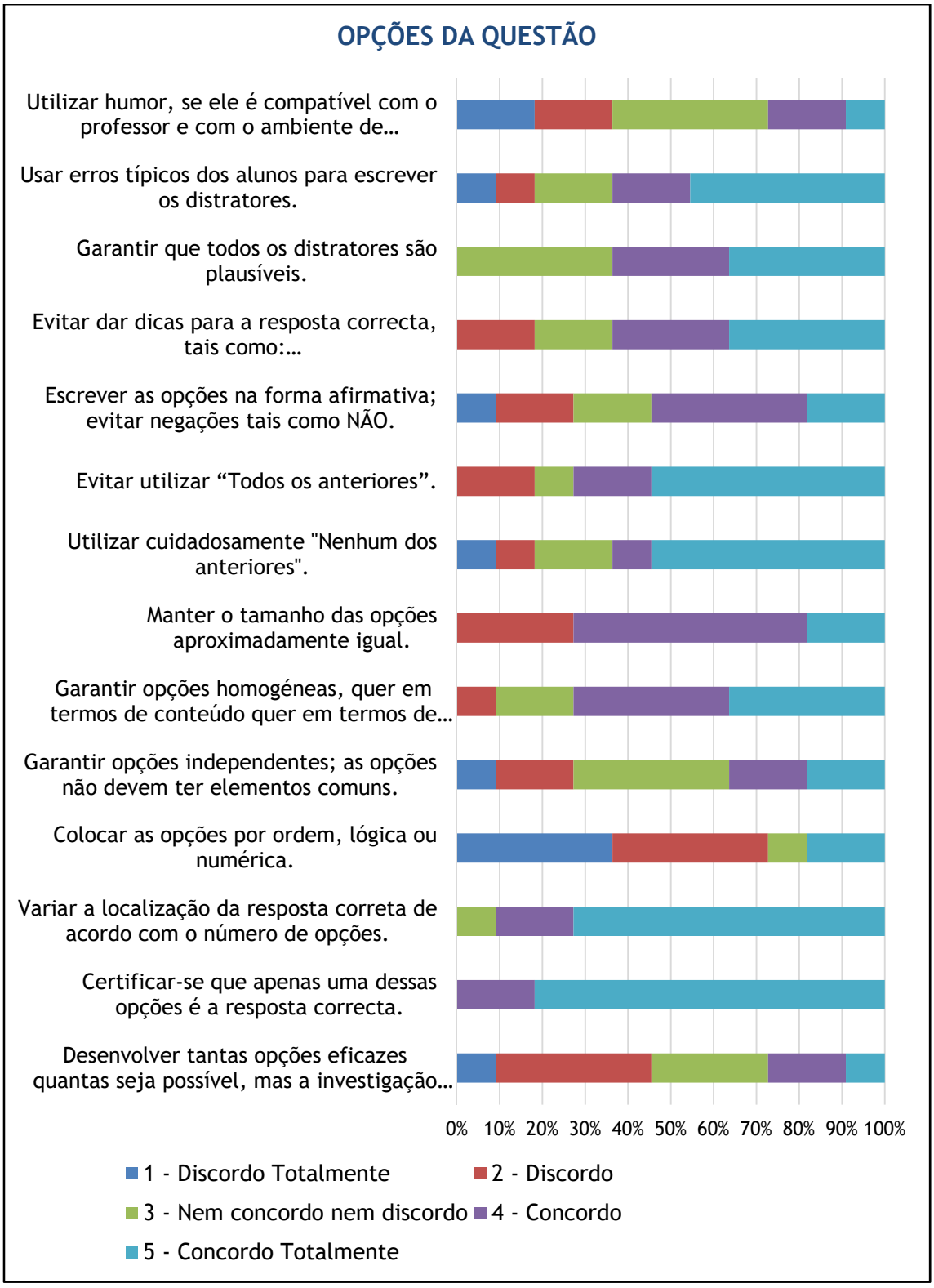


Figura 18: Frequência das repostas dos docentes quanto aos às “Opções da Questão”.

Verificamos que, no geral, os docentes concordam com a maioria das linhas de orientação, dado que apenas 3 delas apresentam um valor para a média inferior ao valor intermédio na escala de Likert, que é 3. Estas são as 3 linhas de orientação menos relevantes para os docentes que responderam ao questionário:

- "**Colocar as opções por ordem, lógica ou numérica**" – esta linha de orientação apresenta 2.3 como valor médio. Apesar de a maioria dos respondentes (n=4) terem respondido 2 e 1, ou seja, discordam ou discordam totalmente desta linha de orientação, o desvio padrão é bastante elevado (1.42) e o intervalo de valores das respostas vai de 1 a 5 (um docente respondeu 3 – "Nem concordo nem discordo" e dois responderam 5 - "Concordo totalmente").
- "**Desenvolver tantas opções eficazes quantas seja possível, mas a investigação sugere que 3 é adequado**" – esta linha de orientação apresenta 2.8 como valor médio. A maioria dos docentes (n=4) respondeu 2, ou seja, discordam com esta linha de orientação. De qualquer forma, o desvio padrão é elevado (1.11), sendo que a gama de valores das respostas vai de 1 a 5 (um docente respondeu "Discordo Totalmente", três responderam "Nem concordo nem discordo", dois responderam "Concordo" e um respondeu "Concordo Totalmente").
- "**Utilizar humor, se ele é compatível com o professor e com o ambiente de aprendizagem**" – esta linha de orientação apresenta 2.8 como valor médio. No entanto, o maior número de docentes (n=4) responderam 3 - "Nem concordo nem discordo". O desvio padrão é também elevado (1.19), sendo que a gama de valores das respostas vai de 1 a 5 (dois docentes responderam "Discordo Totalmente", dois responderam "Discordo", dois "Concordo" e um respondeu "Concordo Totalmente").

Em seguida apresentam-se as linhas de orientação que os docentes apontaram como mais relevantes. Talvez de forma surpreendente, dado os docentes lecionam no domínio da Matemática, a linha de orientação que consideraram mais relevante foi:

"**Usar corretamente a gramática, a pontuação, as letras maiúsculas e a ortografia**" – todos os respondentes responderam 5 ("Concordo Totalmente").

Por ordem decrescente, em termos de relevância, as restantes 4 questões foram: i) "**Certificar-se que as instruções no enunciado são muito claras**" – esta linha de orientação apresenta 4.9 como valor médio (0.29) e a quase totalidade dos docentes (n=10) responderam 5 ("Concordo Totalmente"), sendo que apenas um respondeu 4 ("Concordo") pelo que o desvio padrão é baixo (0.29); ii) "**Certificar-se que apenas umas dessas opções é a resposta correta**" – esta linha de orientação apresenta 4.8 como valor médio, a grande maioria dos docentes (n=10) respondeu 5 e dois responderam 4, pelo que o desvio padrão é pequeno (0.64); iii) "**Evitar questões baseadas em opiniões**" e "**Variar a localização da resposta correta de acordo com o n.º de opções**" – estas linhas de orientação apresentam 4.6 como valor médio. Podemos considerar que o desvio padrão é relativamente baixo (0.64), sendo que a gama de valores das respostas vai de 3 a 5: um docente respondeu 3 ("Nem concordo nem discordo"), dois responderam 4 ("Concordo") e oito responderam 5 ("Concordo Totalmente").

Para além destas, consideramos também importante salientar aquelas linhas de orientação com as quais nenhum dos respondentes discordou, isto é, não obtiveram as respostas 1 ("Discordo Totalmente") ou 2 ("Discordo"). Foram elas as seguintes: i) **"Editar e rever as questões"** – esta linha de orientação apresenta 4.6 como valor médio e 0.77 como desvio padrão, nove docentes responderam 5 ("Concordo Totalmente") e dois responderam 3 ("Nem concordo nem discordo"); ii) **"Utilizar materiais inovadores para testar aprendizagem de nível mais elevado…"** – esta linha de orientação apresenta 4.4 como valor médio e 0.77 como desvio padrão, seis docentes responderam 5 ("Concordo Totalmente"), três responderam 4 ("Concordo") e dois responderam 3 ("Nem concordo nem discordo"); iii) **"Incluir a ideia central no enunciado ao invés de nas opções"** – esta linha de orientação apresenta 4.1 como valor médio e 0.67 como desvio padrão, três docentes responderam 5 ("Concordo Totalmente"), seis responderam 4 ("Concordo") e dois responderam 3 ("Nem concordo nem discordo"); iv) **"Manter o vocabulário simples, tendo o grupo de alunos que está a ser testado"** – esta linha de orientação apresenta 4 como valor médio e 0.74 como desvio padrão, três docentes responderam 5 ("Concordo Totalmente"), cinco responderam 4 ("Concordo") e três responderam 3 ("Nem concordo nem discordo"); v) **"Garantir que todos os distratores são plausíveis"** – esta linha apresenta 4 como valor médio e 0.85 como desvio padrão, quatro docentes responderam 5 ("Concordo Totalmente"), três responderam 4 ("Concordo") e quatro responderam 3 ("Nem concordo nem discordo").

Para todas as restantes linhas de orientação, há pelo menos um docente que respondeu 2 ("Discordo") ou 1 ("Discordo totalmente").

Tal como no estudo de Haladyna e colaboradores (2002) há linhas de orientação que suscitam menos concordância, isto é, geram mais controvérsia. No caso deste estudo, podemos aceitar que nesta situação se encontram aquelas linhas de orientação que apresentam maior desvio padrão, a saber:

- **"Colocar as opções por ordem lógica ou numérica"** – desvio padrão 1.42
- **"Utilizar cuidadosamente 'nenhum dos anteriores'"** – desvio padrão 1.38
- **"Usar erros típicos dos alunos para escrever os distratores"** – desvio padrão 1.34
- **"Escrever as opções na forma afirmativa; evitar negações tais como NÃO"** – desvio padrão 1.23
- **"Escrever o enunciado na forma afirmativa, evitando negações tais como NÃO ou EXCETO"** – desvio padrão 1.21

Síntese da opinião dos docentes sobre as linhas orientadoras para a elaboração de QEM

A análise das respostas ao questionário aos docentes no 1.º ciclo de IA permite-nos concluir que há uma grande concordância com a maior parte das linhas de orientação, sendo residual o

número de linhas de orientação para as quais a média é inferior ao valor intermédio da escala de Likert.

Salienta-se em particular a preocupação de todos os docentes, com uma boa utilização da língua portuguesa e nesse seguimento, podemos verificar também que todos os professores se preocupam com o facto de ser necessário que as questões elaboradas sejam claras no sentido de permitir uma fácil compreensão por parte dos discentes.

6.4. Análise da qualidade dos testes e questões

Devido ao facto de os testes serem gerados aleatoriamente pelo *Moodle*, as questões não foram todas propostas aos alunos o mesmo número de vezes, havendo questões que foram apresentadas mais vezes do que outras. Dado que os instrumentos de análise utilizados (TCT e TRI) têm origem na Estatística, é importante começar por determinar o número de vezes que cada questão saiu nos testes. A este valor chamaremos, daqui em diante, número de respostas. Na Tabela 48 apresentam-se as frequências do número de respostas obtidas nas várias questões, quer no 1.º semestre, quer no 2.º semestre.

Tabela 48: Frequências do número de respostas obtidas às questões

Número de respostas	Frequência	
	1.º semestre	2.º semestre
$[0, 20[$	379	534
$[20, 50[$	1001	723
$[50, 100[$	90	60
$[100, +\infty[$	2	0
	1472	**1317**

No 1.º semestre verificamos que do total de 1472 questões existentes no banco de questões, 379 têm menos de 20 respostas, 1001 têm entre 20 e 50 (exclusive) respostas, 90 têm entre 50 e 100 (exclusive) respostas e 2 têm 100 ou mais respostas.

Quanto ao 2.º semestre, verificamos que do total de 1317 questões existentes no banco de questões, 534 têm menos de 20 respostas, 723 têm entre 20 e 50 (exclusive) respostas e 60 têm 50 ou mais respostas.

Consideramos que o número de respostas obtidas a cada questão é já considerável. No entanto, as 379 questões do 1.º semestre e as 534 questões do 2.º semestre que obtiveram menos de 20 respostas poderá ser considerado reduzido[28], não possibilitando que os resultados à análise da qualidade dessas questões seja o mais adequado. Assim sendo, a análise apenas foi realizada às questões que têm mais de 20 respostas, 1093 no 1.º semestre e 783 no 2.º semestre.

6.4.1. Análise das questões com a Teoria Clássica dos Testes

Iniciou-se a análise com a TCT. Neste caso, os dados e cálculos necessários foram organizados da seguinte forma, para cada um dos semestres:

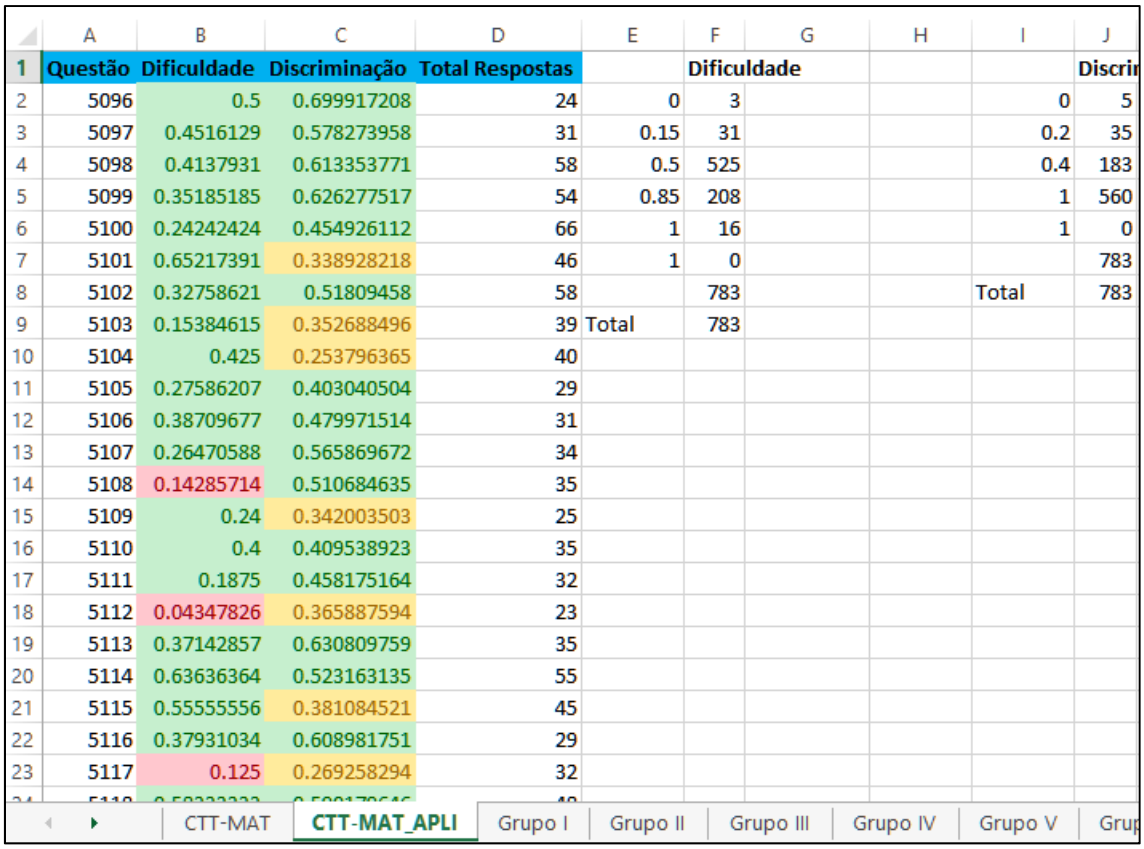
- todas as respostas dadas por todos os alunos a cada uma das questões foram colocadas numa folha de cálculo de um livro MS Excel™, uma folha para cada questão;
- foram calculados os totais de respostas existentes para cada questão, utilizando fórmulas convenientes do MS Excel™;
- foram calculados os Índices de Dificuldade (Equação 1) e o Índice de Discriminação (Equação 2), utilizando fórmulas e funções convenientes do MS Excel™ (ver Figura 19)
- foi feito um resumo dos valores obtidos para todas as questões numa única folha de cálculo MS Excel™ (ver Figura 20) de modo a poder efetuar a sua análise.

2						4756					
3	TURNO ou TURMA	teste	Nº aluno	classificação no teste		A expressão da função inversa de [f(x)=-4+3e^{-4x}] é : [y=\frac{1}{4} \ln \left(\frac{x+4}{3} \right)]; [y=-\frac{1}{4} \ln \left(\frac{x-4}{3} \right)]; [y=\frac{1}{4} \ln \left(\frac{x-4}{3} \right)]; [y=-\frac{1}{4} \ln \left(\frac{x+4}{3} \right)]					
4	ID Categoria					198					
5	Nome Categoria					Função Inversa			Frequência	Dificuldade	Discriminação
6	Categoria Pai				Data	196		1	17	0.515151515	0.612899266
7	C10D	1º teste		4	18-11-2011	1		-0.33333	9		
8	C10D	1º teste		0.66667	18-11-2011	1		0	7		
9	C11N1	1º teste		-0.66667	15-11-2011	-0.3333333		Total	33		
10	C12N	1º teste		0	15-11-2011	0					
11	C13D2	1º teste		-2	16-11-2011	-0.3333333					
12	C13D2	1º teste		2.33333	16-11-2011	1					
13	C12D1	1º teste		0.66667	16-11-2011	1					
14	C12D2	1º teste		1.33333	16-11-2011	0					
15	C16D	1º teste		2.33333	17-11-2011	1					
16	C16D	1º teste		3.66667	17-11-2011	1					
17	C14D2	1º teste		-0.33333	17-11-2011	0					
18	C13N2	1º teste		2	15-11-2011	1					
19	C17D	1º teste		4.66667	18-11-2011	1					
20	C18D	1º teste		4	18-11-2011	1					
21	Q11D	1º teste		-0.66667	23-11-2011	0					
22	Dia 20 - Turno 20h	1º teste		3	20-11-2012	-0.3333333					
23	Dia 20 - Turno 11h	1º teste		6	20-11-2012	1					
24	Dia 20 - Turno 11h	1º teste		2	20-11-2012	0					
25	Dia 20 - Turno 11h	1º teste		2	20-11-2012	-0.3333333					
26	Dia 21 - Turno 8h3	1º teste		0.66667	21-11-2012	-0.3333333					
27	Dia 21 - Turno 8h3	1º teste		5.33333	21-11-2012	1					
28	Dia 11 Nov -Turno	1º teste		0.66667	11-11-2013	-0.3333333					
29	Dia 12 Nov -Turno	1º teste		2	12-11-2013	-0.3333333					

Resumo | 3815 | 4866 | 4756 | 4707 | 4607 | 4551 | 3797 | 4882 | 4784 | 4684 | 4644

Figura 19: Extrato da análise TCT de uma questão.

[28] Empiricamente em Estatística considera-se que uma amostra de tamanho inferior a 20 é pequena. Não sendo consensual este valor, pois ele depende de inúmeros fatores, há autores que referem 30 como sendo o mais aconselhado devido ao Teorema do Limite Central. Mesmo assim, Guimarães e Cabral (2007, p. 175) afirmam que dependendo da distribuição original, esta dimensão até pode estar entre 10 e 50. Aliás, sem referir o que se entende por pequeno, Zickar e Broadfoot (2009, p. 51) referem que quando há limitações de dados, uma entre outras razões, a TCT é preferível à TRI quado se tem um pequeno tamanho da amostra.



	A	B	C	D	E	F	G	H	I	J
1	Questão	Dificuldade	Discriminação	Total Respostas		Dificuldade				Discri
2	5096	0.5	0.699917208	24	0	3			0	5
3	5097	0.4516129	0.578273958	31	0.15	31			0.2	35
4	5098	0.4137931	0.613353771	58	0.5	525			0.4	183
5	5099	0.35185185	0.626277517	54	0.85	208			1	560
6	5100	0.24242424	0.454926112	66	1	16			1	0
7	5101	0.65217391	0.338928218	46	1	0				783
8	5102	0.32758621	0.51809458	58		783			Total	783
9	5103	0.15384615	0.352688496	39	Total	783				
10	5104	0.425	0.253796365	40						
11	5105	0.27586207	0.403040504	29						
12	5106	0.38709677	0.479971514	31						
13	5107	0.26470588	0.565869672	34						
14	5108	0.14285714	0.510684635	35						
15	5109	0.24	0.342003503	25						
16	5110	0.4	0.409538923	35						
17	5111	0.1875	0.458175164	32						
18	5112	0.04347826	0.365887594	23						
19	5113	0.37142857	0.630809759	35						
20	5114	0.63636364	0.523163135	55						
21	5115	0.55555556	0.381084521	45						
22	5116	0.37931034	0.608981751	29						
23	5117	0.125	0.269258294	32						

Figura 20: Extrato da folha de cálculo com o resumo da análise TCT de várias questões.

Para a análise das questões com a TCT, começamos por calcular o seu Índice de Dificuldade.

Na Tabela 49 encontra-se a distribuição das frequências dos índices de dificuldade das várias questões quer no 1.º semestre, quer no 2.º semestre.

Tabela 49: Frequência dos índices de Dificuldade das questões

Índice de Dificuldade	Frequência	
	1.º Semestre	2.º Semestre
$[0, 0.15[$	50	34
$[0.15, 0.5[$	642	525
$[0.5, 0.85]$	389	210
$]0.85, 1]$	12	14

Relativamente ao 1.º semestre, verificamos que 50 das questões apresentam para o Índice de Dificuldade valores abaixo dos recomendados (< 0.15) e que 12 delas apresentam valores acima do recomendado (>0.85). Estas questões deverão ser alvo de uma análise mais aprofundada. Questões com Índice de Dificuldade igual a 0 (zero), o que significa que todos os alunos erraram

a questão, ou com Índice de Dificuldade igual a 1 (um), o que significa que todos os alunos acertaram na questão, não são de considerar na análise. Verificamos que no 1.º semestre não existem questões nesta situação.

As restantes questões apresentam valores para o Índice de Dificuldade dentro dos valores recomendados (≥ 0.15 e ≤ 0.85), sendo que considerámos dois intervalos de dificuldade, um com questões com Índice de Dificuldade 0.5 ou maior, com 389 questões, e outro com Índice de Dificuldade abaixo dos 0.5, com 642 questões.

Relativamente ao 2.º semestre, verificámos que 34 das questões apresentam para o Índice de Dificuldade valores abaixo dos valores recomendados (< 0.15) e que 14 delas apresentam valores acima do recomendado (≥0.85). Estas questões deverão ser alvo de uma análise mais aprofundada. Verificamos que no 2.º semestre existem 3 questões com Índice de Dificuldade igual a 0 (zero). Não existe qualquer questão com Índice de Dificuldade igual a 1 (um). As restantes questões apresentam valores para o Índice de Dificuldade dentro dos valores recomendados (≥ 0.15 e ≤ 0.85), sendo que considerámos também dois intervalos de dificuldade, um com questões com Índice de Dificuldade de 0.5 ou maior, com 208 questões, e outro com Índice de Dificuldade abaixo dos 0.5, com 525 questões.

Passemos agora à análise dos Índices de Discriminação. Na Tabela 50 encontra-se a distribuição de frequências dos Índices de Descriminação das várias questões quer no 1.º semestre, quer no 2.º semestre.

Tabela 50: Frequência dos Índices de Discriminação das questões

Índice de Discriminação	Frequência	
	1.º Semestre	2.º Semestre
$[-1, 0[$	13	5
$[0, 0.2[$	64	35
$[0.2, 0.4]$	237	183
$]0.4, 1]$	779	560

Antes de mais, verificámos que no 1.º semestre há 13 questões e no 2.º semestre há 5 questões com discriminação negativa. Tal como já foi referido na secção 3.1, é no mínimo estranho, dado que um valor negativo para a discriminação significa que os "melhores alunos" têm menos probabilidade de acertar na resposta correta e os "piores alunos" têm maior probabilidade de acertar na resposta correta. Estas questões deveriam ser retiradas do banco de questões para serem analisadas cuidadosamente.

Em relação ao 1.º semestre, verificámos que 779 questões apresentam, para o Índice de Discriminação, valores que estão dentro dos recomendados (> 0.4), isto é, têm um bom poder de discriminação. Também verificámos que 237 das questões apresentam um poder de discriminação razoável (≥ 0.2 e ≤ 0.4). As restantes 64 questões apresentam um baixo poder de discriminação.

Em relação ao 2.º semestre, verificámos que 560 questões apresentam, para o Índice de Discriminação, valores que estão dentro dos recomendados, isto é, apresentam um bom poder de discriminação. Também verificámos que 183 das questões apresentam um poder de discriminação razoável (entre 0.2 e 0.4). As restantes 35 questões apresentam um baixo poder de discriminação.

Vamos agora fazer uma análise considerando as duas variáveis, Índice de Dificuldade e Índice de Discriminação em conjunto, de modo a obter uma informação mais abalizada sobre a qualidade das questões. Na Figura 21 apresenta-se o Diagrama de Dispersão que ilustra a correlação entre o Índice de Dificuldade e o Índice de Discriminação das questões do banco de questões do 1.º semestre. Na Figura 22 apresenta-se o Diagrama de Dispersão que apresenta a correlação entre o Índice de Dificuldade e Índice de Discriminação das questões do banco de questões do 2.º semestre. Em ambos os casos verifica-se que a nuvem de pontos nos diagramas é bastante dispersa, pelo que a correlação entre os dois Índices é pequena. O cálculo dos Coeficientes de Correlação confirma este facto: 0.186 no 1.º semestre e 0.018 no 2.º semestre. Verifica-se em ambos os casos que a maioria das questões se situa dentro dos limites aconselhados quer para o Índice de Dificuldade, quer para o Índice de Discriminação.

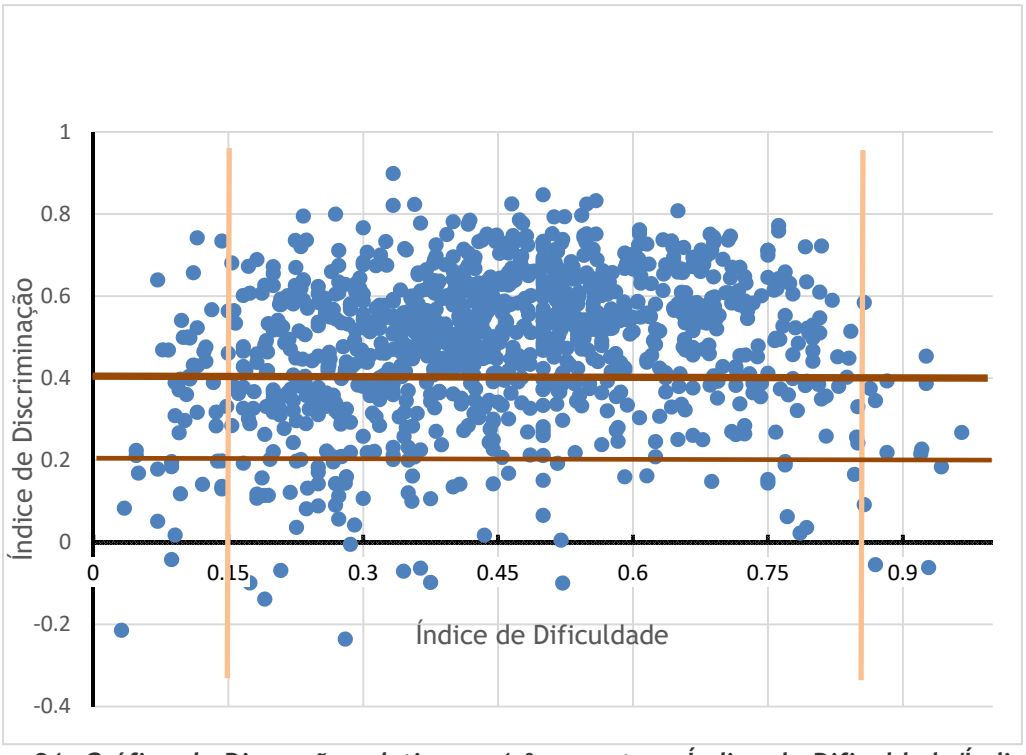


Figura 21: Gráfico de Dispersão relativo ao 1.º semestre - Índice de Dificuldade/Índice de Discriminação.

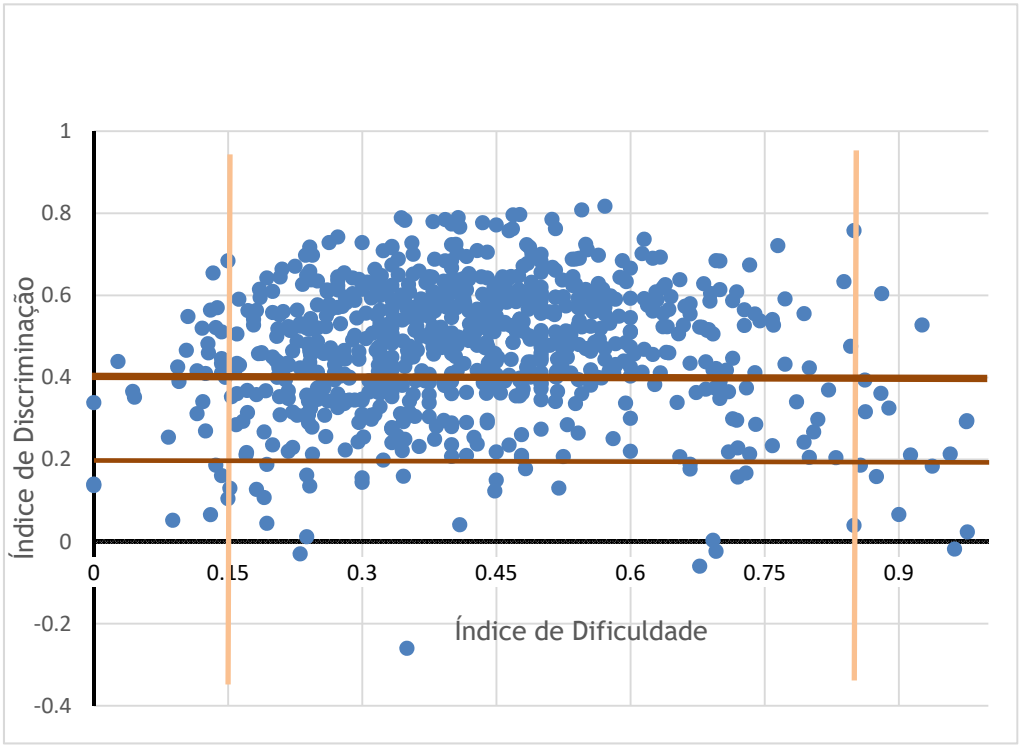


Figura 22: Gráfico de Dispersão relativo ao 2.º semestre - Índice de Dificuldade/Índice de Discriminação.

Após a análise houve alguns grupos de questões que foram retirados do banco de questões para futura análise, criando-se uma categoria específica para esse fim. Outros grupos foram mantidos no banco de questões e classificados em grupos com caraterísticas específicas. De seguida apresentam-se esses grupos.

Grupo I – Questões com número de respostas inferior a 20

Encontram-se neste grupo 379 questões do 1.º semestre e 534 questões do 2.º semestre. Estas questões não foram analisadas devido ao reduzido número de respostas. No entanto, estas questões serão mantidas no banco de questões para serem utilizadas em testes futuros de forma controlada de modo a obter um número de respostas que permita, posteriormente, fazer uma análise de qualidade das questões.

Grupo II – Questões com Índices de Discriminação negativos

Encontram-se neste grupo 13 questões do 1.º semestre e 5 questões do 2.º semestre. Devido ao facto de que, tal como já foi referido, valores negativos para a discriminação indicarem que os "melhores alunos" têm menor probabilidade de acertar na resposta correta e os "piores alunos" têm maior probabilidade de acertar na resposta correta, decidiu-se retirar estas questões do banco de questões de modo a serem alvo de uma análise criteriosa por parte dos docentes envolvidos no processo.

Grupo III – Questões com Índice de Dificuldade menor que 0.15 ou maior que 0.85 e com Índice de Discriminação menor ou igual que 0.4

Encontram-se neste grupo 38 questões do 1.º semestre e 29 questões do 2.º semestre. Estas questões apresentam problemas graves de qualidade, dado que nem o Índice de Dificuldade nem o Índice de Discriminação apresentam valores dentro daqueles que são recomendados. Decidiu-se retirar estas questões do banco de questões, de modo a serem alvo de uma análise criteriosa por parte dos docentes envolvidos no processo.

Grupo IV – Questões com Índice de Dificuldade menor que 0.15 e com Índice de Discriminação maior que 0.4

Encontram-se neste grupo 18 questões do 1.º semestre e 18 questões do 2.º semestre. Neste grupo encontram-se questões muito difíceis (menos de 15% dos alunos acertaram na resposta correta), mas são questões que apresentam um bom Índice de Discriminação. Assim sendo, decidiu-se mantê-las no banco de questões para serem utilizadas em situações de teste nas quais seja importante a inclusão de questões com níveis de dificuldade muito elevados.

Grupo V – Questões com Índice de Dificuldade maior que 0.85 e com Índice de Discriminação maior a 0.4

Encontram-se neste grupo 2 questões do 1.º semestre e 3 questões do 2.º semestre. Neste grupo encontram-se questões muito fáceis (mais de 85% dos alunos acertaram na resposta correta), mas são questões que apresentam um bom Índice de Discriminação. Assim sendo, decidiu-se mantê-las no banco de questões para serem utlizadas em situações de teste nas quais seja importante a inclusão de questões com níveis de dificuldade muito reduzidos.

Grupo VI – Questões com Índice de Dificuldade maior ou igual a 0.15 ou menor ou igual que 0.85 e com Índice de Discriminação menor que 0.2

Encontram-se neste grupo 49 questões do 1.º semestre e 24 questões do 2.º semestre. Neste grupo, apesar de os Índices de Dificuldade das questões se encontrarem dentro dos valores aconselhados, os Índices de Discriminação são muito baixos, pelo que as questões praticamente não fornecem a discriminação necessária. Assim sendo, decidiu-se retirar estas questões do banco de questões, de modo a serem alvo de uma análise criteriosa por parte dos docentes envolvidos no processo.

Grupo VII – Questões com Índice de Dificuldade maior ou igual a 0.15 e menor ou igual que 0.85 e com Índice de Discriminação maior ou igual a 0.2

Encontram-se neste grupo 973 questões do 1.º semestre e 707 questões do 2.º semestre. Neste grupo encontram-se as questões que apresentam valores adequados quer para os Índices de Dificuldade, quer para os Índices de Discriminação. Podemos assim, considerar que este grupo de questões representa o núcleo fundamental do nosso banco de questões. Decidimos considerar dois subgrupos:

1. **Índice de Dificuldade menor que 0.5** – consideramos que este subgrupo contém questões de nível básico. Este subgrupo contém 560 questões do 1.º semestre e 484 questões do 2.º semestre.
2. **Índice de Dificuldade maior ou igual que 0.5** – consideramos que este subgrupo contém questões de nível médio/avançado. Este subgrupo contém 411 questões do 1.º semestre e 222 questões do 2.º semestre.

6.4.2. Análise das questões com a Teoria da Resposta ao Item (TRI)

De seguida foi efetuada a análise com TRI. Neste caso, os dados e cálculos foram analisados da seguinte forma, para cada um dos semestres:

- todas as respostas dadas por todos os alunos a todas as questões foram colocadas numa única folha de cálculo;
- ajustou-se o modelo logístico com 2 parâmetros utilizando o suplemento do MS Excel™ já referido. Na Figura 23 mostra-se um dos passos do suplemento.

Inicialmente tinha-se optado pela análise das questões por categoria, mas não foi possível ajustar o modelo para todas as categorias, devido ao reduzido número de questões e/ou respostas existentes em algumas categorias. Assim sendo, ajustou-se o modelo utilizando todas as questões e todas as respostas em simultâneo. Um outro aspeto que consideramos relevante está relacionado com a escala de valores das respostas dadas pelos alunos às questões. Para a análise TRI, utilizou-se uma escala dicotómica, isto é, com dois valores, a saber, " 1 – acertou na resposta correta" e "0 – não acertou na resposta correta". Na realidade os dados apresentam 3 valores: a saber, os dois precedentes e " $-\frac{1}{3}$ selecionou um dos distratores". No entanto, dado que o " $-\frac{1}{3}$ " funciona apenas como uma penalização para desincentivar os alunos a tentarem acertar na resposta de forma aleatória, pode-se claramente considerar uma escala dicotómica " 1 – acertou na resposta correta" e "0 – não acertou na resposta correta".

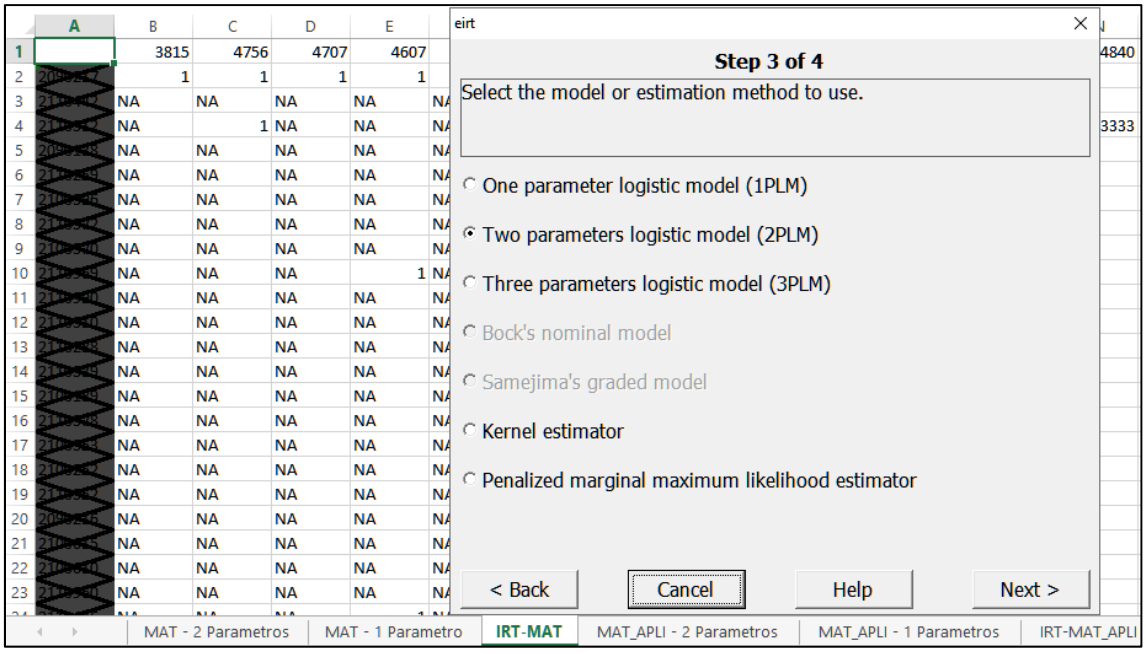


Figura 23: Um dos passos do assistente do suplemento do MS Excel™ "eirt".

A análise das questões com a TRI levantou sérias dificuldades ao autor da tese. Inicialmente foi planeado analisar as questões por categorias. Ajustou-se o modelo logístico com 2-parâmetros por se considerar o mais adequado: após a análise com a TCT verificou-se que as questões apresentavam discriminações bastante diferenciadas, pelo que considerar o modelo logístico

com apenas 1-parâmetro, o qual considera que todas as questões apresentam a mesma discriminação, seria bastante redutor. Para fazer o ajustamento do modelo desta forma, criou-se uma folha de cálculo para cada categoria com todas as questões e com todas as respostas existentes para essa categoria, eliminando todas as questões que continham menos de 20 respostas, tal como foi feito na análise com a TCT. No entanto, os resultados foram muito pouco animadores:

- em muitas das categorias houve questões para as quais o modelo não convergiu, isto é, não se conseguiu ajustar o modelo;
- mesmo para as questões para as quais o modelo convergiu, os Índices de Dificuldade são muito elevados, acima de 8, o que ultrapassa em muito a variação típica que é de -3 a +3;
- o índice alfa (α) de Cronbach é menor que 0.013 para todas as categorias, sendo 0,000 para a maioria.

Aquando da revisão de literatura, verificámos que uma das principais limitações do TRI se prende com o grande volume de dados necessários para que o modelo seja ajustado de forma conveniente. No caso deste projeto, o volume de dados em termos de número de questões é bastante elevado, mas o mesmo não acontece em relação ao número de respostas existentes para cada uma das questões. Decidiu-se então analisar todas as questões em simultâneo, colocando todas as questões e todas as respostas numa só folha de cálculo.

Obtiveram-se valores bastante animadores para o alfa de Cronbach: 0.953 no 1.º semestre e 0.943 no 2.º semestre. No entanto, os resultados restantes, relativos aos Índice de Dificuldade, foram, podemos afirmá-lo, pouco animadores.

- no caso do 1.º semestre, o modelo convergiu apenas para duas questões. Mesmo para essas duas questões, o Índice de Dificuldade foi 15.676 para uma das questões, foi 16.602 para outra, o que representa valores muito acima dos valores típicos (-3 a +3);
- no caso do 2.º semestre, o modelo convergiu apenas para três questões e tal como no caso do 1.º semestre, os valores dos Índices de Dificuldades ultrapassam em muito os valores típicos (-3 a +3).

Na tentativa de conseguir valores que nos permitissem retirar algumas conclusões, decidimos ajustar, apesar de redutor, o modelo logístico com 1–parâmetro. No entanto, em ambos os semestres, as questões apresentam Índices de Dificuldade superiores a 3.

Estes valores levariam à conclusão de que todas as questões apresentam níveis de dificuldade demasiado elevados, pelo que todas elas deveriam ser revistas. No entanto, dado que o volume de dados é reduzido, não podemos garantir a validade dos resultados, pelo que considerámos válidos os resultados obtidos com a TCT.

6.4.3. Síntese de resultados sobre a qualidade dos testes e questões

No que diz respeito à análise com a TCT, foram calculados os índices de dificuldade e de discriminação para todas as questões que tinham um número de respostas considerado suficiente. Verificou-se que a maioria das questões apresentam valores para esses índices que permitem concluir que têm qualidade e que portanto podem ser utilizadas no âmbito de uma avaliação sumativa. Foram definidos grupos de perguntas com características semelhantes, de modo a permitir a obtenção de testes mais homogéneos.

A realização da análise com a TRI gerou valores que não permitiram retirar qualquer conclusão relevante. Este problema prende-se com o facto de não haver um número suficiente de respostas para cada questão, o que é exigido para realizar uma análise com a TRI de modo a permitir que os modelos convirjam.

6.5. Análise das respostas às entrevistas aos docentes no 3.º ciclo de IA

Para analisar os dados das entrevistas efetuadas aos docentes no 3.º ciclo de IA, começamos por caracterizar os docentes entrevistados e depois analisaremos as suas respostas no que diz respeito a cada uma das dimensões definidas.

6.5.1. Caracterização dos docentes entrevistados

Foram entrevistados 6 docentes. Quatro deles são do sexo feminino e dois do sexo masculino. A média de idades é 55 anos, variando entre os 44 e os 71 anos. Quanto ao tempo de serviço no ISCAP, a média é de 22 anos. Podemos assim concluir que os docentes entrevistados são bastante experientes.

No sentido de salvaguardar o anonimato dos participantes, identificámos cada uma das entrevistas com "E" seguido de um número de ordem (E1, E2, …). Vejamos a caracterização individual de cada um dos entrevistados:

- **Entrevistado 1 (E1)** – mulher, com 47 anos de idade e 15 anos de tempo de serviço no ISCAP;
- **Entrevistado 2 (E2)** – mulher, com 44 anos de idade e 15 anos de tempo de serviço no ISCAP;
- **Entrevistado 3 (E3)** – mulher, com 66 anos de idade e 30 anos de tempo de serviço no ISCAP;
- **Entrevistado 4 (E4)** – homem, com 71 anos de idade e 22 anos de tempo de serviço no ISCAP;

- **Entrevistado 5 (E5)** – mulher, com 50 anos de idade e 23 anos de tempo de serviço no ISCAP;
- **Entrevistado 6 (E6)** – homem, com 56 anos de idade e 27 anos de tempo de serviço no ISCAP.

6.5.2. Análise das dimensões consideradas na entrevista

Apresentamos de seguida a análise do conteúdo das entrevistas relativamente a cada uma das dimensões definidas. Relembramos que cada questão, apresentada no guião da entrevista, corresponde a uma dimensão a estudar.

Opinião sobre a forma de *e–assessment* implementada

Todos docentes apresentam uma opinião muito positiva sobre a forma de avaliação implementada, sendo que todos os docentes têm a opinião de que este é um bom sistema de avaliação. Apresentamos as afirmações de dois dos entrevistados que resumem de forma clara a opinião de todos:

> "*Foi muito apelativa para os alunos e acho que lhes aumentou o interesse e que teve bons resultados.*" (E3)

> "*Eu acho que é uma avaliação que valeu a pena desenvolver. Gosto da avaliação. Acho que é uma avaliação muito adequada quando se pretende fazer diversas provas durante o semestre aos alunos e o número de alunos é bastante elevado.*" (E6)

De qualquer das formas houve dois docentes entrevistados que afirmaram que estavam reticentes no início do processo. Esses dois docentes afirmaram que:

> "*No início estava muito reticente quando nós começamos a utilizar este sistema, em especial no que diz respeito à escolha múltipla.*" (E1)

> "*No início não estava muito recetivo porque pensei que não fosse tão funcional como foi.*" (E2)

Dois dos docentes entrevistados salientaram que houve uma evolução positiva com o tempo, que resultou de um processo de aprendizagem. A afirmação de um deles resume bem essa opinião:

> "*Com o passar dos anos, acho que nós fomos aperfeiçoando, aprendendo a construir as questões de escolha múltipla e acho que melhorou bastante.*" (E1)

Foi ainda referido, por dois docentes, que houve a necessidade de uma maior aprendizagem em relação à forma como devem ser desenvolvidas as questões e que essa aprendizagem foi feita. A afirmação de um dos docentes entrevistados reflete essa opinião:

> "*Nós lemos e refletimos e estudamos um bocadinho como melhorar a elaboração, em especial, a escolha múltipla de modo a que realmente houvesse ou pudéssemos obter melhores resultados.*" (E1)

Principais dificuldades encontradas na implementação

O desconhecimento inicial sobre como elaborar boas questões foi apontado por três docentes como uma das dificuldades encontradas na implementação deste processo de *e-assessment*. A seguinte afirmação reflete isso mesmo:

> "*Tivemos que estudar a melhor maneira de colocar as questões porque a maneira de elaborar é diferente de um teste normal. E no início houve uma dificuldade a elaborar as questões de maneira a que fossem objetivas e não avaliassem mais do que um objetivo em cada questão.*" (E5)

Também apontados por três docentes como uma grande dificuldade, foram os problemas relacionados com a tecnologia, nomeadamente com o funcionamento do *Moodle*, dos computadores para os alunos realizarem os testes ou os servidores onde se alojava o *Moodle*. Disse um dos docentes entrevistados:

> "*Primeiro, a utilização dos portáteis. No início alguns não tinham. Depois isso acho que se aliviou. Também de início, por vezes as falhas de sistema que bloqueavam bastante. Às vezes também a adaptação dos professores aos métodos eletrónicos/informáticos.*" (E3)

Outra das dificuldades apontadas pelos docentes entrevistados foi a introdução das fórmulas matemáticas complexas nos testes *Moodle*, a qual foi referida por dois docentes. Apesar de ser uma dificuldade relacionada com a utilização do *Moodle*, é importante referi-la, dado que está diretamente relacionada com a Matemática. Um docente referiu:

> "*Quando ainda não dominávamos bem o TeXaide*[29] *e então aquilo aparecia tudo cheio de pontos de interrogação e bastava termos um espaço que aparecia lá um ponto…*

[29] O *TeXaide* era uma versão gratuita e especial do *MathType*© (http://www.dessci.com/) que foi utilizada para a escrita das fórmulas matemáticas nas QEM do banco de questões em substituição do TeX. A maioria dos docentes não dominava o TeX e em 2008 conseguiu-se este software que "convertia" as fórmulas matemáticas existentes no Word para TeX, obtendo-se um conjunto de carateres que eram copiados para intercalar com o texto das questões e das opções que eram inseridas no *Moodle*.
O *Moodle* dispunha de um suplemento que convertia o TeX assim gerado em símbolos matemáticos adequados, desde que fossem devidamente assinalados no início e no final com os símbolos $$ ou, em alternativa, com os símbolos \[no início e com o símbolo \] no final.

> *Primeiro que nós conseguíssemos corrigir um erro era complicado e demorávamos imenso tempo.*" (E1)

Dois docentes apontaram como dificuldade o colocar a equipa a funcionar devido à resistência inicial dos intervenientes. Disse um dos docentes entrevistados:

> "*De início havia uma certa resistência porque era um método novo.*" (E3)

Por fim, foi apontado por um docente que inicialmente houve falta de apoio por parte dos órgãos de gestão da escola. Verificamos pela afirmação desse docente que o processo poderia ter começado a ser implementado antes de 2008:

> "*Essa foi a grande dificuldade que pôs ou poderia por em causa o arranque desta avaliação. Não quiseram, os Órgãos de Gestão da época, em 2006, que a avaliação avançasse.*" (E6)

Em que medida houve ou não mudanças nas práticas pedagógicas do docente

Dois docentes referiram que não houve quaisquer mudanças nas suas práticas pedagógicas, afirmando um deles que teve de haver um processo de adaptação maior por parte dos alunos. Quando questionados, esses docentes afirmaram o seguinte:

> "*Basicamente não. Quer dizer, as aulas continuaram a ser dadas na mesma….*" (E3)

> "*Teve de haver uma adaptação mais da parte dos alunos do que nossa.*" (E4)

No entanto, três docentes reconheceram que houve mudanças, mas que elas não foram consequência da forma de avaliação. Um deles afirmou mesmo que a mudança na forma de avaliação foi ela sim resultado do processo de mudança global que foi implementada nas UC. Vejamos as afirmações dos docentes em relação a este aspeto:

> "*Não acho que as minhas práticas pedagógicas tenham mudado devido à escolha deste método de avaliação. Vamos mudando em função de outras coisas: a nível de conhecimento que os alunos trazem.*" (E1)

> "*Pouco mudou. Não foi por causa do sistema de avaliação que houve mudanças.*" (E5)

> "*Mas o tipo de avaliação foi mais uma das consequências das mudanças todas que foram feitas do que o contrário.*" (E6)

Apenas um docente admitiu ter feito algumas mudanças, mas afirmou que apenas mudou a forma como abordava a resolução dos exercícios nas aulas práticas:

> "*Eu acho que é diferente de dizer "resolva este exercício" e pronto… porque eles têm que saber analisar as respostas. E portanto, eu acho que as práticas orientam-se noutro sentido. Mas em termos teóricos eu penso que não mudou nada…*" (E2)

Vantagens desta forma de avaliação para o docente

A principal vantagem para os docentes e que foi apontada por todos os entrevistados tem a ver com a obtenção automática das classificações dos alunos, o que representa uma grande poupança de tempo:

> "*A vantagem na correção é evidente, é um ganho de muitas horas.*" (E4)

Outra vantagem, apontada por dois docentes, foi o facto de a avaliação ser mais objetiva. Afirmaram os docentes:

> "*Avaliação muito mais objetiva, mais seguida e a correção muito mais fácil*" (E3)

> "*Os critérios são muito objetivos e portanto não há disparidade de correção, mesmo elaborando grelhas de correção detalhadas nos testes normais que saiu agora a grelha especificamente os critérios… Há sempre disparidade de correção entre docentes. Aqui portanto as questões são objetivas, ou está certo ou está errado.*" (E5)

Outra das vantagens, apontada por três docentes, é que a existência do banco de questões permite a criação mais prática e mais rápida de testes, por exemplo para os chamados exames de estatuto, que podem ser pedidos pelos alunos a qualquer altura. Disseram eles a esse respeito:

> "*Quando nós construímos os nossos testes também se torna muito simples, porque é só irmos à Base de Dados e escolhermos quero esta categoria ou aquela subcategoria e portanto rapidamente também construímos o teste. Por isso é tudo mais rápido. Se tu fosses construir um teste agora de raiz, tinhas que perder mais tempo.*" (E1)

> "*Em qualquer momento pode-se recorrer a um teste, a uma prova.*" (E4)

> "*Se há necessidade de fazer um teste para o dia seguinte ou para a hora seguinte é fácil tendo o banco de questões.*" (E5)

Um dos docentes afirmou que este processo permite uma mais fácil integração de novos docentes na UC.

Esse mesmo docente apontou como grande vantagem o facto de ter possibilidade de realizar um trabalho mais criativo:

> "*Na questão do tempo que é preciso gastar com a disciplina é mais com aspetos criativos e menos com aspetos 'menores' como aqueles de corrigir testes.*" (E6)

Desvantagens desta forma de avaliação para o docente

A principal desvantagem, a qual foi apresentada por quatro docentes é o facto de ser necessário um grande esforço inicial e muito tempo para desenvolver o banco de questões, mas que vai melhorando ao longo dos anos. A esse respeito apresentamos as afirmações de dois docentes, que espelham bastante bem a opinião de todos:

> "*Colocar as questões, depois as opções e criar a base de dados em si, o dividir nas categorias e subcategorias… Tudo isso, nós fomos construindo ao longo dos anos, demorou imenso tempo. Foram muitas as horas ali investidas.*" (E1)

> "*Claro que dão bastante trabalhão a construir as questões, não é… É mais o trabalho da construção das questões. Mas também com alguma prática… inicialmente investe-se bastante tempo nisso e a gente escreve e não gosta e depois não está como deveria estar… Com a prática essa desvantagem vai-se diluindo.*" (E5)

Outra desvantagem apontada por dois docentes, desta vez quando comparando a elaboração de questões de resposta aberta com as de escolha múltipla, é o facto de ser maior a dificuldade na elaboração de questões de qualidade no caso das questões de escolha múltipla. Disseram esses docentes:

> "*Se tiveres de fazer uma pergunta aberta é muito rápido, não é. E quando estás a formular as respostas tens de ter muita atenção… é uma atenção muito mais redobrada…*" (E2)

> "*No início houve uma dificuldade a elaborar as questões de maneira a que fossem objetivas e não avaliassem mais do que um objetivo em cada questão.*" (E5)

Outra desvantagem, apresentada por dois entrevistados, tem a ver com o facto de o professor não conseguir avaliar a criatividade e o raciocínio dos alunos. A afirmação seguinte espelha bem a opinião desses docentes:

> "*Não permitiu nos testes ver o raciocínio dos alunos. Portanto, bastava ter um erro no caminho e às vezes até podiam chegar ao resultado certo com raciocínios errados ou ao contrário.*" (E3)

Perceção por parte dos docentes de alterações nas práticas educativas dos alunos

Os docentes percecionaram algumas alterações nas práticas dos alunos. A maior parte dos docentes entrevistados, cinco deles, referiram que houve maior assiduidade às aulas por parte dos alunos. As seguintes afirmações representam a opinião de todos os docentes:

> "*O que eu acho em que aspetos eles mudaram é que foram muito mais "seduzidos" pela avaliação contínua nestes termos… dá-me impressão que se fosse num regime muito mais aberto, nós tínhamos mais faltas do que aqueles que havia. Se tivéssemos a avaliação contínua tradicional eles faltariam mais.*" (E2)

> "*E eu acho que com o nosso sistema de avaliação conseguimos que os alunos viessem mais às aulas. E ao virem mais às aulas, portanto eles acabam por aprender mais.*" (E6)

Três docentes salientaram o facto de ter havido uma evolução na forma como os alunos se comportavam ao responder às questões do exame, verificando-se que inicialmente os alunos tentavam responder de forma aleatória e depois tomavam consciência das penalizações em caso de respostas erradas, e iam começando a ser mais cuidadosos com a seleção das respostas. Afirmou um docente que:

> "*Nota-se que por exemplo, eles vêm habituados do secundário a ter escolha múltipla. No caso da Matemática, eles têm escolha múltipla no exame nacional só que não desconta. E isso, eu acho que no início, quando os alunos chegam ao primeiro ano eles tentavam preencher a grelha toda não tendo bem a noção, isto no primeiro teste, do que iria descontar. Mas depois eles vão aprendendo e vão selecionando.*" (E1)

Outra mudança de comportamento identificada por três docentes é o facto de os alunos estudarem de forma mais regular. Apresentamos duas afirmações que resumem convenientemente este aspeto:

> "*Estudam mais assiduamente, não digo diariamente porque acho que eles deviam trabalhar ainda mais mas pronto, num momento próximo do teste sempre vão trabalhando mais.*" (E5)

> "*Havendo vários momentos de avaliação, os alunos também vão ter de estudar mais alguma coisa e isso também é bom.*" (E6)

Um docente apontou como aspeto negativo o facto de os alunos se limitarem a mecanizar os processos sem se preocuparem com o raciocínio. Disse esse docente:

> "*Mecanizou mais os alunos. Os alunos tornaram-se mais mecânicos. Eu notei um desinteresse, pode-se dizer, no raciocínio. (…) Muito mais o atingir o resultado.*" (E3)

Outros aspetos referidos na entrevista

Para além dos aspetos relativos às dimensões em estudo, durante as entrevistas houve alguns aspetos referidos pelos docentes entrevistados que consideramos relevantes. Um aspeto apontado por quatro docentes foi o facto de o número elevado de alunos nas turmas dificultar a realização de avaliação contínua e que realmente esta forma de avaliação possibilitou que se conseguisse implementar avaliação contínua. A afirmação seguinte resume este aspeto:

> "(esta forma de avaliação) *Foi uma boa opção, dado o grande número de alunos que nós temos. Como queremos fazer avaliação continua, a única maneira de conseguir implementar o sistema de avaliação contínua era optar por um sistema deste género.*" (E5)

Três docentes referiram que a possibilidade de realizarem avaliação contínua foi boa para os alunos e que eles melhoraram as classificações. Salientamos duas das afirmações:

> "*Eu penso que, atendendo aos resultados, tem corrido melhor agora. Pelo que eu me lembro das nossas estatísticas, os resultados têm sido bastante melhores. Bem, também temos a vantagem de fazer alguns testes, mais de três pelo menos… três ou quatro… já tivemos quatro. E a matéria repartida talvez facilite um bocadinho.*" (E2)

> "*A avaliação continua para eles é muito melhor tanto que os resultados com a avaliação contínua em termos de aprovações melhoraram muito.*" (E5)

Dois docentes referiram as vantagens da realização do Teste de "Repescagem", dizendo que é bom porque permite que os alunos não desistam logo no 1º teste, caso a nota seja fraca. Vejamos as afirmações desses docentes:

> "*Foi muito bom porque, de facto, quando eles faziam o primeiro teste e lhes corria mal, eles ficavam desmotivados… E agora eles sabem que há hipótese de repescar o primeiro ou um qualquer e faz com que se mantenham nas aulas. E eu acho isso bastante importante.*" (E2)

> "*Um outro aspeto que não tem a ver diretamente com esta avaliação mas pelo facto de haver os testes de repescagem eles… eu acho que isso foi um aspeto muito importante, foi que eles assim desistem menos a meio do semestre. E portanto… porque eles têm ainda uma outra hipótese no caso de alguma coisa correr mal, de recuperarem. E esse aspeto foi fundamental.*" (E6)

Um docente referiu a sua preocupação com a fraude por parte dos alunos, tendo mesmo sido confrontado com esse problema por parte de alguns alunos. Afirmou esse docente:

> "*Eu acho que as minhas maiores preocupações são mesmo nesse sentido. De não copiarem. E não nos dizerem 'Ah! Eu sei deste aqui, Ah! Passou mas eu sabia'…*" (E2)

Outro dos docentes entrevistados afirmou que houve necessidade de fazer uma série de mudanças:

> "*O objetivo foi… o primeiro passo foi tomar consciência que as coisas não estavam bem, porque havia um número excessivo de reprovações, insucesso escolar e, tendo em conta isso, procurou-se diagnosticar os problemas, onde é que estavam os problemas, o que é que era preciso mudar e sem a pretensão de diagnosticar tudo de uma só vez, mas porque houve alguns problema que foram sendo acertados ao longo destes anos, como se deve lembrar… mas havia coisas que eram claramente necessárias fazer… procurarmos que todos tivéssemos ou déssemos a matéria mais ou menos da mesma maneira, sem limitar a liberdade de cada um. Mas haver uma orientação bem definida sobre a matéria e aquilo que era importante nós fazermos. E para isso todos nós estivemos envolvidos de alguma maneira na elaboração dos apontamentos das aulas. Portanto, houve ai, no meu ponto de vista, quando as pessoas são envolvidas nesse processo, logo uma mudança. E as pessoas também tiveram de procurar agir sempre por consensos, gerar consensos na equipa… havia um programa que tínhamos de cumprir e a partir do programa começamos a criar ferramentas e que discutimos muitas vezes. Lembra-se, com certeza, das inúmeras reuniões que tivemos para fazer isto. E mais. Como é que os próprios apontamentos foram elaborados. O trabalho foi distribuído, depois passou por mim para… digamos que eu talvez fosse a pessoa que depois dava alguma unidade às coisas e fazia umas revisões… mas todas as pessoas estavam envolvidas nesse processo. E dessa forma foi possível fazer uma mudança consensual das coisas e que as pessoas aderissem e participassem. Depois também o próprio sistema de avaliação que foi também, e também me recordo que no início nem toda a gente estava de acordo. E acabou por toda a gente, julgo eu, aderir ao processo e perceber, ou pelo menos que depois passado algum tempo toda a gente começou a acreditar que aquele era o caminho. De acordo com os objetivos que tínhamos e de acordo com as circunstâncias era o melhor caminho a seguir.*" (E6)

Esse mesmo docente realçou a importância do trabalho em equipa com motivação:

> "*Eu acho que foi engraçado por uma equipa relativamente grande a funcionar, a puxar todos para o mesmo lado e todos a colaborarem com vontade. O que demonstra mais uma vez, que sobretudo as pessoas são capazes de fazer e que haja o mínimo de motivação e que acreditem naquilo que estão a fazer.*" (E6)

6.5.3. Síntese da opinião dos docentes sobre o processo de e-assessment implementado

Podemos concluir que os inquiridos, apesar de todas as resistências iniciais e das dificuldades quer ao nível das tecnologias quer ao nível da elaboração das questões, manifestaram uma opinião favorável e veem esta estratégia de avaliação de forma muito positiva. No entanto,

esta implicou esforço e empenho da sua parte, no sentido de reformularem todo um processo a que não estavam habituados. Contudo, começaram a verificar que a mudança introduzida se mostrava atrativa. A resistência inicial foi provocada por um certo "medo" pelo desconhecido. Houve uma necessidade de refazer um certo percurso que se encontrava enraizado e abraçar uma nova metodologia que depois de experimentada ir-se-ia tornar fundamental dada a sua objetividade e economia de tempo devido à obtenção automática das classificações e todo um processo de automatizações que tornavam certas tarefas burocráticas muito mais leves. Contudo, há algumas desvantagens, como sejam um grande esforço inicial que foi necessário para se ter um banco de questões e que a construção de novas QEM consome algum tempo para ficarem devidamente construídas. Informam ainda que de ponto de vista do aluno notaram grandes melhorias, em especial a forma como estudam e um aumento de assiduidade.

6.6. Análise das respostas ao questionário aos alunos no 3º ciclo de IA

Para analisar os dados do questionário, efetuadas aos alunos no 3.º ciclo de IA, começamos por caracterizar a amostra de alunos e depois analisaremos as suas respostas no que diz respeito a cada uma das dimensões definidas e indicadores identificados durante a análise de conteúdo.

6.6.1. Caracterização dos alunos que responderam ao questionário

O total de estudantes que acederam responder ao questionário foi de 427. No entanto, apenas 386 completaram efetivamente ao questionário. Destes, 15 não apresentaram respostas válidas a nenhuma das questões, pelo que as suas respostas não foram consideradas válidas e foram, portanto, eliminadas do conjunto de documentos a analisar. Assim sendo, foram considerados válidos 371 questionários, correspondentes ao mesmo número de alunos. São estes os alunos que iremos caracterizar. Identificaremos cada um dos alunos com "id" seguido de um número de ordem (id1, id2, id3, …).

Entre os 371 estudantes, 203 (55%) são do sexo feminino; 261 (70%) frequentavam a UC em regime diurno e os restantes 110 (30%) em regime noturno; 76 alunos (20%) eram trabalhadores estudantes. A maioria dos estudantes (297, 80%) frequentavam a UC Matemática, logo eram alunos da Licenciatura em Contabilidade e Administração, e 74 (20%) frequentavam a UC Matemática I, da Licenciatura em Comércio Internacional, sendo que do total dos estudantes, 107 (29%) estavam a repetir a frequência da UC. Destes, 28 alunos (8%) frequentavam a UC pela segunda vez, 28 (8%) alunos frequentavam a UC pela terceira vez, 19 (5%) alunos frequentavam a UC pela quarta vez e 32 (9%) alunos frequentavam a UC pela quinta vez. A idade dos alunos variou entre os 17 anos (1 aluno) e os 56 anos (1 aluno), sendo que a maioria tinha 18 anos (147 alunos). A média dos alunos é cerca de 21 anos, apresentando um desvio padrão de 5.5. Na Figura 24 encontra-se a distribuição das idades dos alunos que responderam ao questionário.

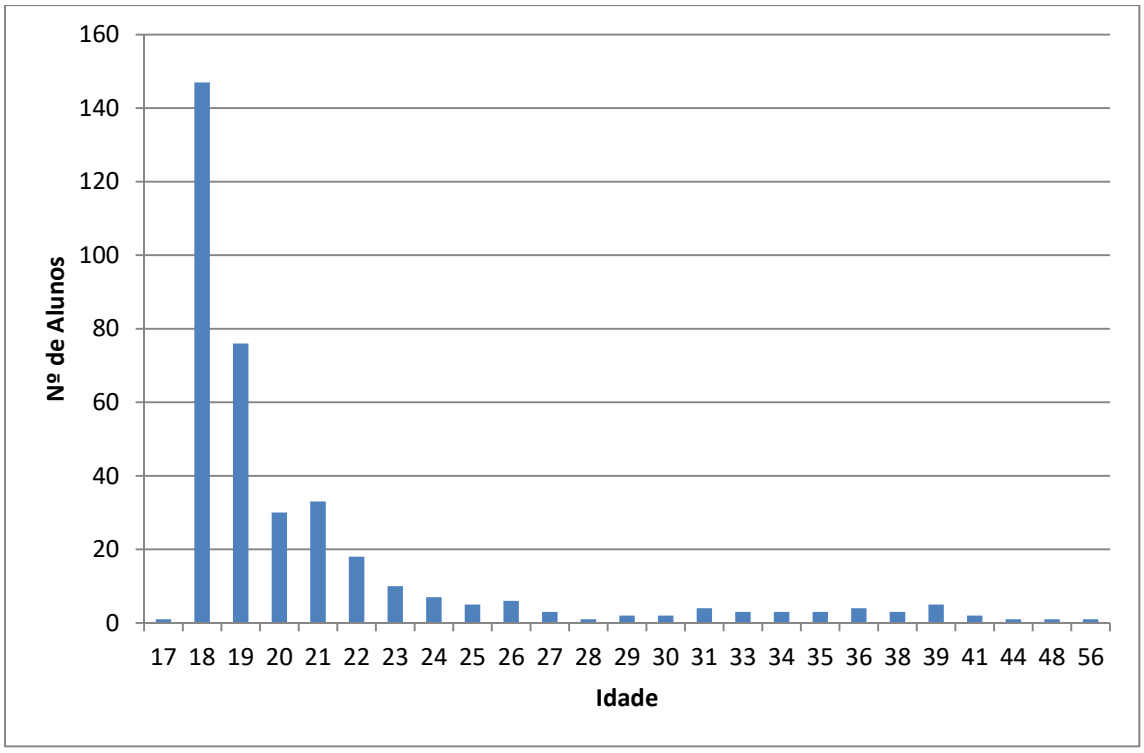


Figura 24: Distribuição das idades dos alunos que responderam ao questionário.

6.6.2. Análise das dimensões consideradas no questionário

Na Tabela 51 apresenta-se um resumo da percentagem de respostas obtidas ao questionário considerando as 9 perguntas aí incluídas, respetivas dimensões e alguns indicadores.

Tabela 51: Perguntas e resumo em percentagem de respostas obtidas ao questionário indicando as dimensões e alguns indicadores.

PERGUNTAS, (DIMENSÕES) e alguns indicadores		%
Considera que os testes de escolha múltipla realizados no *Moodle*, na Unidade Curricular são justos?		
Sim (Os testes QEM são justos?)	*	55%
Não (Os testes QEM são justos?)	*	45%
Considera que se estes testes (EM) fossem realizados em papel em vez de serem realizados no *Moodle*, seriam:		
Melhor em Papel (É melhor o formato em papel ou o uso de novas tecnologias?)	*	23%
Pior em papel (É melhor o formato em papel ou o uso de novas tecnologias?)	*	16%
Igual		61%
O facto de os testes serem de escolha múltipla alterou de alguma forma o modo como estudou?		
Sim (Alterou a forma como estudou, em que aspetos?)	*	12%
Não		88%
A sua presença regular nas aulas depende do regime Avaliação (Contínua ou Final) escolhido?		
Sim (O regime de avaliação influencia a presença nas aulas)	*	34%
Não (O regime de avaliação influencia a presença nas aulas)	*	66%
O facto de existirem 3 testes foi importante para que escolhesse o Regime de Avaliação Contínua?		

PERGUNTAS, (DIMENSÕES) e alguns indicadores		%
Sim		88%
Não (3 testes não foi importante na opção do Reg. Avaliação. Porquê?)	*	12%
Qual a sua opinião sobre a existência de um teste de Repescagem?		
Positiva (Opinião sobre o teste de "Repescagem")	*	74%
Outras Opiniões	*	26%
Na sua opinião quais são as vantagens, para os alunos, dos testes de escolha múltipla?		
(Quais as vantagens das QEM para os alunos)	*	
Não há	**	19%
Opções de resposta ajudam a encontrar solução	**	41%
Outras	*	40%
Na sua opinião quais são as desvantagens, para o aluno, dos testes de escolha múltipla?		
(Quais as desvantagens dos testes com QEM para os alunos)	*	
Não há	**	19%
Raciocínio não ser considerado	**	39%
Outras	*	42%
Comentários adicionais		
(Comentários adicionais)	*	9%

* (Com vários indicadores)

** (Um dos Indicadores)

Apresentamos de seguida a análise do conteúdo com base nas perguntas realizadas através do questionário, relativamente a cada uma das dimensões definidas e respetivos indicadores. No Anexo J encontra-se o resumo de todos os indicadores (códigos) identificados para as várias dimensões consideradas. No Anexo K encontra-se o resumo de todos os indicadores identificados para as várias dimensões consideradas, cruzando-os com as variáveis que foram utilizadas para a caracterização dos alunos que responderam ao questionário.

Os testes QEM são justos?

Verificamos que 203 (55%) alunos consideraram que os testes são justos e 168 (45%) consideraram que os testes não são justos. Considerando as variáveis utilizadas para a caracterização dos alunos que responderam ao questionário, estas percentagens mantêm-se idênticas para todas as variáveis, exceto para a variável sexo. Neste caso, mais de 50% dos alunos do sexo feminino consideram que os testes com QEM não são justos, enquanto apenas cerca de 35% dos alunos do sexo masculino considera que os testes com QEM não são justos, conforme se verifica através do gráfico da Figura 25.

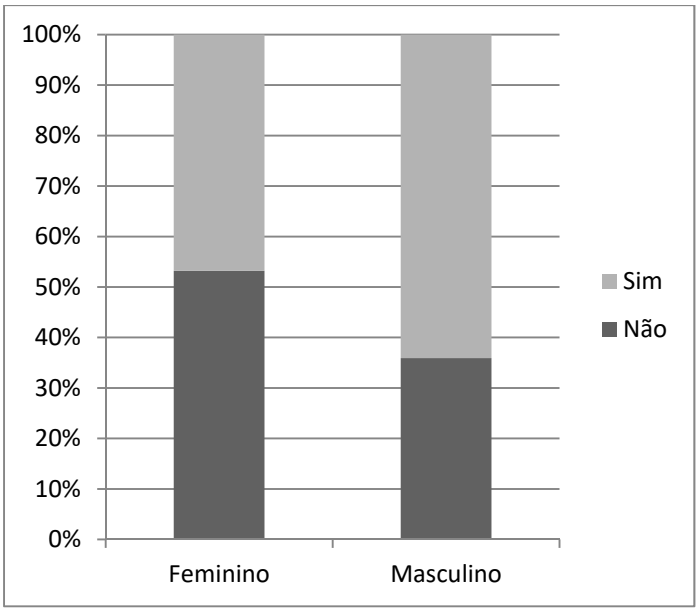


Figura 25: Respostas dos alunos à pergunta "Os testes QEM são justos?", em função do género.

Quanto aos indicadores dos motivos pelos quais os alunos consideraram os testes justos, foram identificados os seguintes:

- Testes uniformes

40 alunos indicaram o facto de os testes serem uniformes como motivo para considerarem os testes QEM justos. Quanto aos testes, disseram os alunos que "*todos têm o mesmo número de perguntas e o mesmo tempo disponível, e o grau de dificuldade provavelmente é o mesmo para todos também*" (id12). Acrescentaram ainda que se trata de "*testes escolhidos aleatoriamente pelo sistema com um grau de dificuldade igual*" (id60), isto é "*os testes são feitos a modo de proporcionar as mesmas dificuldades a todos os alunos*" (id231), "*pois são diferentes de aluno para aluno, mas o grau de dificuldade é o mesmo*" (id271). Destaco ainda a afirmação de um aluno que referiu que os testes são justos "*porque há aleatoriedade na escolha das perguntas, e, teoricamente, todos os testes terão perguntas mais fáceis e outras mais difíceis, o que fará, com que, no geral, a dificuldade global do teste seja semelhante para todos*" (id390).

- Testes mais simples/acessíveis

34 alunos consideraram que os testes com QEM são justos porque se tornam mais simples e acessíveis. Disseram em os alunos que "*escolha múltipla torna mais fácil*" (id32), que "*os testes até são acessíveis*" (id227) e mesmo que "*são bastante acessíveis*" (id 125). Também consideraram que "*se fossem realizados em papel era muito mais difícil*" (id253). Em relação a este indicador, saliente-se que apesar de haver mais alunos do sexo masculino a considerarem os testes justos (108 alunos do sexo masculino e 95 alunos do sexo feminino), o número de

alunos do sexo feminino a referirem este indicador é igual ao número de alunos do sexo masculino (17 alunos).

- Avaliam verdadeiramente os conhecimentos

20 alunos consideraram os testes com QEM justos porque avaliam verdadeiramente os seus conhecimentos. Afirmam os alunos que os testes são justos "*porque testam justamente a capacidade de cada um*" (id48) e, além disso, "*porque mostram exatamente aquilo que nós fizemos e o nosso grau de conhecimento*" (id149)

- Mais difícil cometer fraudes

20 alunos consideraram os testes com QEM justos porque é mais difícil cometer fraudes. Disseram os alunos que os testes "*são todos diferentes e não há maneira de copiar ou tentar adivinhar*" (id206). Por outro lado, "*o facto das respostas erradas ser descontadas no final é algo que, de certa maneira, evita que existam tantos alunos a fazerem a "lotaria"*" (id136). Na verdade, "*ao fim de tudo somos obrigados a resolver os exercícios e não meter à sorte*" (id302) "*porque uma vez que desconta mesmo que se meta à sorte não compensa*" (id286). Em relação a este indicador, saliente-se que apesar de haver mais alunos do sexo masculino a considerarem os testes justos (108 alunos do sexo masculino e 95 alunos do sexo feminino), houve mais alunos do sexo feminino a referirem este indicador (8 alunos do sexo masculino e 12 alunos do sexo feminino).

- Avaliam os alunos como qualquer outro teste

19 alunos consideraram os testes com QEM justos porque avaliam os alunos como qualquer outro teste. Afirmaram os alunos que "*todos os testes em papel ou computador, com perguntas escolha múltipla ou não, eles são justos*" (id10) e também que "*no Moodle ou no papel são sempre justos*" (id57). Concordaram os alunos que "*se o aluno souber responder, tanto responde na escolha múltipla como em papel*" (id215) e que "*são momentos de avaliação iguais aos outros*" (id359).

- Teste aleatório

19 alunos referiram que os testes são justos porque são gerados de forma aleatória, isto é, as "*perguntas são atribuídas aleatoriamente*" (id9) e portanto "*a probabilidade de sair qualquer que seja a questão a uma pessoa é a mesma*" (id112).

- Avaliam os conteúdos lecionados

14 alunos referiram que os testes são justos porque avaliam os conteúdos lecionados. Disseram os alunos que os testes são justos porque "*os conteúdos questionados estão de acordo com o*

que foi lecionado e trabalhado na aula" (id35), mais ainda, "*porque são de acordo com o que damos nas aulas e estas preparam-nos bem*" (id410)

- Apresentam várias opções de resposta

12 alunos consideraram que os testes são justos porque apresentam várias opções de resposta. Disseram os alunos que "*o teste sendo de escolha múltipla ajuda a eliminarmos hipóteses quando sabemos que estas estão erradas*" (id244) e "*tendo as respostas é sempre uma ajuda*" (id301). Em relação a este indicador, saliente-se que apesar de haver mais alunos do sexo masculino a considerarem os testes justos (108 alunos do sexo masculino e 95 alunos do sexo feminino), houve mais alunos do sexo feminino a referirem este indicador (5 alunos do sexo masculino e 7 alunos do sexo feminino).

- Não há erros na correção

6 alunos indicaram que os testes são justos porque não há erros na correção, isto porque os testes "*são feitos e corrigidos por um programa, não podendo haver enganos*" (id50). Salienta-se neste indicador o facto de apenas ser referido por alunos do sexo feminino. Também foi referido apenas por alunos diurnos. Além disso, foi referido apenas por alunos que frequentam a UC pela primeira vez. Ou seja, todos os alunos que referiram este indicador são alunos diurnos do sexo feminino que frequentam a UC pela primeira vez.

- Melhor classificação

Houve ainda 3 alunos que consideram que os testes são justos porque obtêm melhor classificação. Salienta-se neste indicador o facto de apenas ter sido referido por alunos do sexo feminino. Também foi referido apenas por alunos diurnos. Ou seja, todos os alunos que referiram este indicador foram alunos diurnos do sexo feminino.

- Não apresenta motivo

Por fim, 24 alunos não apresentaram qualquer motivo pelo qual consideraram que os testes são justos.

Quanto aos indicadores dos motivos pelos quais os alunos consideraram que os testes não são justos, foram identificados os seguintes:

- Não se avalia o raciocínio

70 alunos identificaram o facto de não se contar o raciocínio necessário para chegar à resposta certa mas de se considerar apenas o resultado final, como uma das razões que os fizeram considerar os testes de escolha múltipla injustos. Disse um aluno que "*não acho justos porque*

se errarmos perdemos logo a pontuação toda, se fosse de desenvolvimento poderíamos ter sempre alguns pontos" (id28). Outro aluno afirmou que os testes não são justos "*porque se conseguirmos entender o raciocínio necessário e se o desenvolvermos mas a solução final estiver errada para além de não contabilizar o raciocínio ainda desconta, o que, a meu ver, não é muito justo*" (id36) e outro afirmou que "*uma vez que os testes são realizados no Moodle apenas são cotados os resultados finais, enquanto se fosse tudo feito manualmente os passos intermédios também teriam cotação. A meu ver seria mais justo*" (id42). É importante referir ainda que a "*Matemática tem uma componente de raciocínio ao qual o professor não tem acesso. Muitas vezes, a solução pode estar errada, no entanto, o seu raciocínio estava correto*" (id115) e ainda que "*não é um método de avaliação justo para uma disciplina como Matemática pois esta disciplina deve ser avaliada pelo desenvolvimento das questões e não só pelo seu resultado. Um resultado final errado não significa que o aluno não seja conhecedor da matéria em si, pode haver um simples engano nos cálculos que põe em causa todo o seu conhecimento acerca da matéria.*" (id375).

- Testes com níveis de dificuldade diferentes

57 alunos identificaram o facto os testes poderem ter níveis de dificuldade diferentes para os vários alunos, pois não são iguais para todos pois porque são gerados de forma aleatória, como o motivo pelo qual os testes não são justos. Disseram os alunos que os testes não são justos "*porque as perguntas não são iguais para todos, o que pode ser mais fácil para mim pode ser mais difícil para algum colega*" (id11). Além disso "*porque alguns alunos recebem questões mais fáceis do que outros*" (id15) e as questões "*são escolhidos aleatoriamente logo uns podem calhar mais fáceis do que outros*" (id350). Por fim referir que "*os testes diferentes causam desigualdades na avaliação uma vez que os alunos podem estar mais a vontade em determinado exercício que saiu no teste de um colega*" (id386).

- Penalizações são prejudiciais

28 alunos referiram que os testes não são justos porque as penalizações, que foram introduzidas para reduzir a possibilidade de os alunos tentarem acertar na resposta correta de forma aleatória, são prejudicais. Disse um aluno, "*não concordo que uma resposta errada anule parte de uma resposta certa*" (id25), acrescentando outro que "*desconta demasiado e por vezes torna-se muito prejudicial*" (id14) e outro que "*prejudicam um pouco os alunos pelo facto dos elevados descontos por resposta errada*" (id411). Por fim, refira-se que os testes foram considerados injustos "*Porque alunos que saibam o procedimento e não sejam precisos na resolução, serão penalizados injustamente*" (id100).

- Dificuldades com o *Moodle*

10 alunos afirmaram que os testes não são justos, devido a problemas relacionados com o *Moodle*. As dificuldades encontradas são de tipos distintos. Dizem os alunos que "*para além do*

stress de haver problemas com o computador" (id182), "*a contagem do tempo atrapalha um pouco" (id40), que "já me aconteceu o exame fechar no momento em que ia gravar*" (id157) e que "*enganei me no teste e pus a opção errada mas depois fui corrigir e não gravou*" (id161). Saliente-se que este indicador apenas é referido por alunos que estão a frequentar a UC pela 1ª vez. Além disso, apesar de haver mais alunos diurnos que consideraram que os testes não são justos (120 alunos diurnos e 47 alunos noturnos), houve mais alunos noturnos a referirem este indicador (4 alunos diurnos e 6 alunos noturnos).

- Possibilidade de adivinhar a resposta

7 alunos referiram que o facto de se poder acertar na resposta correta sem ter conhecimentos para tal é uma das razões para os testes não serem considerados justos. Disseram os alunos que "*até quem não sabe pode conseguir tirar boa nota*" (id268) e que tanto "*podemos ter sorte e acertar como nos enganar a escolher a resposta*" (id282). Assim sendo, os testes não são justos "*porque há alunos que não sabem a verdadeira resposta e podem ter a sorte de acertar*" (id400).

- Não avalia verdadeiramente os conhecimentos

7 alunos indicaram que os testes não são justos porque não avaliam verdadeiramente os seus conhecimentos.

- São difíceis

4 alunos referiram que os testes não são justos porque são difíceis. Saliente-se que este indicador apenas foi referido por alunos do sexo feminino e por alunos da UC Matemática e portanto da Licenciatura em Contabilidade a Administração. Ou seja, este indicador apenas foi identificado por alunos do sexo feminino da Licenciatura em Contabilidade e Administração.

É melhor o formato em papel ou o uso de novas tecnologias?

Verificamos que 226 (61%) dos alunos consideraram que é igual ter os testes em formato papel ou em formato eletrónico no *Moodle*, 84 (23%) dos alunos consideraram que é melhor em formato papel e 61 (16%) dos alunos consideraram que é pior em formato papel. Considerando as variáveis utilizadas para a caracterização dos alunos que responderam ao questionário, estas percentagens mantêm-se idênticas para todas as variáveis.

Quanto aos indicadores dos motivos pelos quais os alunos consideraram que os testes em papel são melhores do que os testes implementados no *Moodle*, foram identificados os seguintes:

- Não há problemas informáticos

29 alunos consideraram que os testes com QEM são melhores em papel porque assim não há problemas informáticos. Disseram os alunos que nesse caso "*não existiriam falhas de rede*"

(id23). "*Além disso, não haveria atrasos devido ao "loading" da página do Moodle*" (id66) e "*não teríamos de estar preocupados com o computador bloquear*" (id182), logo "*diminuiria o stress*" (id183). Para terminar, refira-se a opinião de um aluno sobre estas questões tecnológicas: "*As tecnologias têm muita tendência a falhas o que põe em causa a avaliação dos alunos o que é completamente absurdo pois trata-se da nossa avaliação, das nossas notas que podem influenciar todo o nosso percurso no ensino*" (id375).

- Tornava-se mais simples

5 alunos referiram que os testes em formato papel se tornam mais simples. Disseram os alunos que assim é "*mais fácil de ler, sublinhar aspetos importantes das perguntas*" (id65) e outro aluno referiu o seguinte: "*pessoalmente prefiro ver o enunciado no papel*" (214).

- O tempo não é cronometrado

5 alunos disseram que preferem os testes em formato papel porque o tempo não é cronometrado. Afirmou um aluno que "*no papel não teríamos o tempo cronometrado, o que nos deixaria menos nervosos*" (id26).

- Não seria necessário transportar computador

4 alunos referiram que com os testes em papel seria melhor porque não haveria necessidade de transportar os seus próprios computadores. Afirma um aluno que "*não teríamos de trazer os computadores para o ISCAP e corrermos o risco de os perdermos ou sermos assaltados*" (id387) e também "*não havia o peso de trazer o computador*" (id281). Quanto a este indicador, salienta-se o facto de apenas referirem este aspeto alunos que frequentaram a UC pela 1ª vez.

- Haveria melhores classificações

3 alunos referiram que é melhor o formato papel porque haveria melhores classificações. Disse um aluno que "*conseguiríamos obter melhores classificações*" (id319). Quanto a este indicador, salienta-se que apenas alunos do sexo feminino referiram este aspeto. Também, apenas alunos da UC Matemática I, isto é da Licenciatura em Comércio Internacional identificaram este indicador. Acresce que apenas alunos que frequentaram a UC pela 1ª vez referiram este indicador. Assim sendo, este indicador foi identificado pelos alunos do sexo feminino, da Licenciatura de Comércio Internacional, que frequentam esta UC pela 1ª vez.

- É mais justo

3 alunos referiram que os testes em papel são melhores porque são mais justos: "*Os testes seriam todos iguais, logo seriam mais justos*" (id387)).

- Rapidez na apresentação das classificações aos alunos

2 alunos referiram que é melhor o formato papel porque há maior rapidez na apresentação das classificações: "*é bom pois assim as notas saem rapidamente*" (id24).

- Não responde à questão/Resposta ambígua

Por fim, 36 alunos não responderam ou apresentaram uma resposta ambígua.

Quanto aos indicadores dos aspetos identificados pelos alunos como justificativos para que os testes em papel sejam piores do que os testes implementados no *Moodle*, foram identificados os seguintes:

- Piores classificações em papel

9 alunos referiram que os testes em formato papel são piores, pois obtêm piores classificações com os testes realizados nesse formato. Disseram os alunos que com os testes em formato papel "*as notas iriam ser mais baixas"* (id257) e "*o número de reprovações seria maior*" (id373).

- Demora na apresentação das classificações aos alunos, em papel

8 alunos referiram que os testes em formato papel são piores porque há mais demora na apresentação das classificações aos alunos: "*para se saber a nota o tempo demorado seria maior*" (id149).

- Mais confuso em papel

6 alunos referiram que os testes em formato papel são mais confusos. Disse um aluno que em papel "*ia gerar mais confusão*" (id10).

- Mais prático em computador

5 alunos afirmaram que os testes realizados em computador são mais práticos do que os realizados em papel. Disseram os alunos que "*a nível de organização ao elaborar o teste é melhor por este meio*" (id265) e "*acho que era mais difícil arranjar tanta variedade de perguntas*" (id419).

- Teste mais difícil em papel

5 alunos afirmaram que os testes em formato papel são mais difíceis. Disseram os alunos que em formato papel "*tornar-se-ia mais difícil para alguns aluno*" (id228) e que "*no computador é mais fácil*" (id145). Salienta-se que este indicador apenas foi indicado por alunos da UC Matemática, isto é, da Licenciatura em Contabilidade e Administração.

- Mais demorado em papel

5 alunos afirmaram que os testes são mais demorados quando realizados em formato papel: "*uma das desvantagens és o tempo*" (id407)

- Gasto desnecessário em papel

3 alunos evidenciaram que com os testes em formato papel há um gasto desnecessário em papel: "*é melhor fazer no computador do que no papel pois poupa-se folhas. Estamos em altura de crise*" (id267). Salienta-se que este indicador apenas foi indicado por alunos da UC Matemática, isto é, da Licenciatura em Contabilidade e Administração.

- Maior probabilidade de fraude

2 alunos afirmaram que é pior em papel devido à existência de uma maior probabilidade de cometer fraude: "*haveria mais hipóteses de copiar*" (id402). Salienta-se que este indicador apenas foi identificado por alunos do sexo masculino, da UC Matemática, isto é, da Licenciatura em Contabilidade e Administração e que frequentam a UC pela 1ª vez.

- Gosto pelas novas tecnologias

1 aluno apontou que considera pior o formato dos testes em papel devido ao seu gosto pessoal pelas novas tecnologias: "*gosto dos testes realizados no computador*" (id255).

- Maior probabilidade de o professor cometer erros

1 aluno apontou o facto de que com os testes em papel o professor poder cometer erros quer na correção que na elaboração dos testes: "*podem ocorrer enganos na correção ou na elaboração dos testes*" (id50).

- Não responde à questão/Resposta ambígua

Por fim, 21 alunos não responderam ou apresentaram uma resposta ambígua.

Os testes QEM influenciam as práticas educativas

Verifica-se que 328 (88,41%) dos alunos afirmaram que a existência dos testes com QEM não influenciou as suas práticas educativas e que 43 (11,59%) dos alunos afirmaram que houve influência dos testes com QEM nas suas práticas educativas. Considerando as variáveis utilizadas para a caracterização dos alunos que responderam ao questionário, estas percentagens mantêm-se idênticas para todas as variáveis.

Quanto aos indicadores dos aspetos em relação as quais os alunos consideram terem mudado nas suas práticas educativas, foram identificados os seguintes:

- Aborda de modo diferente a resolução de exercícios

15 alunos afirmaram que passaram a abordar de uma forma diferente a resolução dos exercícios. Disse um aluno que "*quando não sei a matéria baseio-me nas opções de resposta para responder*" (id15) e outro que "*não valorizo tanto o procedimento como dantes*" (id100) pois "*ao ver as 4 hipóteses a primeira coisa que faço logo é a exclusão de duas. Até podia não saber como chegar à resposta certa, mas ao ter lá as opções torna-se mais fácil*" (id224). Assim sendo, comecei a "*focar mais o resultado e menos o processo*" (id360). Saliente-se que apesar de haver mais alunos que frequentam a UC pela 1ª vez a afirmarem que alteraram as suas práticas educativas (27 alunos frequentam a UC pela 1ª vez e 16 alunos não frequentam esta UC pela 1ª vez), este indicador foi identificado mais vezes pelos alunos que não frequentam a UC pela 1ª vez (8 alunos não frequentam a UC pela 1ª vez e 7 alunos frequentam a UC pela 1ª vez).

- Estudar menos

10 alunos afirmaram que alteraram as suas práticas na medida em que estudam menos. Afirmaram eles que "*o estudo não foi tão intensivo*" (id55), que "*não me empenho tanto*" (id225) e ainda que "*não dediquei tantas horas de estudo à unidade Curricular, passando a canalizar a minha atenção para outras Unidades*" (id408).

- Estudar mais

Por outro lado houve 9 alunos que afirmam que estudam mais. Disse um aluno, "*resolvi mais exercícios de escolha múltipla*" (id227) e outro disse, "*tive que praticar mais*" (id261). Em relação ao facto de ter testes QEM outro aluno disse que "*fez com que eu estudasse mais para conseguir melhores notas*" (id407). Saliente-se que apesar de haver mais alunos que frequentam a UC pela 1ª vez a afirmarem que alteraram as suas práticas educativas (27 alunos frequentam a UC pela 1ª vez e 16 alunos não frequentam esta UC pela 1ª vez), este indicador foi identificado mais vezes pelos alunos que não frequentam a UC pela 1ª vez (5 alunos não frequentam a UC pela 1ª vez e 4 alunos frequentam a UC pela 1ª vez).

- Maior atenção aos detalhes

4 alunos realçaram que passaram a dar mais atenção aos detalhes. Disseram os alunos que "*tive mais cuidado com pequenos pormenores*" (id270) pois "*é necessária mais atenção, pois um pequeno deslize pode levar à resposta errada, enquanto que em resposta aberta o processo de resolução poderia ser cotado*" (id16). Salienta-se que este indicador apenas foi indicado por

alunos do sexo feminino e da UC Matemática, isto é, da Licenciatura em Contabilidade e Administração.

- Deixar de usar calculadora

Esta prática foi apontada por um aluno (id253).

- Diminuição da importância da linguagem Matemática escrita

Há ainda 1 aluno que sugeriu que deixou de dar importância à linguagem Matemática escrita: "*apenas preocupo-me com conceitos não com a escrita*" (id178).

- Não responde à questão/Resposta ambígua

Por fim, 4 alunos não responderam ou apresentaram uma resposta ambígua.

O regime de avaliação influencia a presença nas aulas

Verificamos que 244 (66%) dos alunos consideraram que o regime de avaliação teve influência no que diz respeito à sua presença regular nas aulas e que os restantes referiram que não teve qualquer influência. Considerando as variáveis utilizadas para a caracterização dos alunos que responderam ao questionário, estas percentagens mantêm-se idênticas para todas as variáveis.

Quanto aos indicadores da forma pela qual os alunos consideram que a sua presença regular nas aulas depende do regime de avaliação (contínua ou final) escolhido, foram identificados os seguintes:

- Avaliação contínua obriga a estar presente

70 alunos referiram que a sua presença regular nas aulas depende do regime de avaliação escolhido dado que a avaliação contínua obriga os alunos a estarem presentes nas aulas. Os alunos afirmaram que a influência se justifica "*pela "obrigação" da presença nas aulas, se estivesse em regime de avaliação final já não sentia essa obrigação*" (id50). Disseram ainda que "*se estivesse por final, muito facilmente faltaria a uma ou outra aula, pelo que estando em avaliação contínua sinto a obrigação de estar presente*" (id149) e que "o regime de avaliação contínua facilita a conclusão da cadeira, para beneficiar de avaliação continua é necessário estar presente a 75% das aulas, logo temos de ter presença regular".

- Estudo parcelar e organização pessoal de estudo mais fácil

36 alunos referiram que ao optarem pelo regime de avaliação participam mais regularmente nas aulas porque isso facilita o estudo parcelar e a organização do estudo pessoal. Disseram os

alunos "*Assim vou estando a par da matéria dada e assim fica mais fácil para estudar*" (id161), "*permite-me dividir as matérias por mini-testes (o método contínuo) e dedicar a minha atenção a cada uma em especial*" (id49) e "*uma pessoa tenta participar mais e não deixar tudo para a ultima*" (id159).

- Benefício da assiduidade e participação na classificação

16 alunos referiram que a sua presença regular nas aulas depende do regime de avaliação escolhido visto que há um benefício para a classificação final pela assiduidade e participação nas aulas. Disseram os alunos que "*uma vez que em avaliação contínua é-nos dado mais um ponto no fim do semestre pela assiduidade, esse ponto pode ser crucial para fazer a disciplina*" (id14). Disseram eles que optando pelo regime de avaliação contínua vão às aulas "*de forma a ter o valor adicional*" (id32), "*devido a bonificação pela presença*" (id56), "*por causa do valor extra*" (id56) e "*devido ao valor de presenças que nos é dado, motiva mais os alunos a escolher avaliação continua e ir às aulas*" (id266).

- Prefere Exame Final porque estuda sozinho

Há 1 aluno que referiu que opta pelo regime de avaliação final para não ir às aulas, porque prefere estudar sozinho.

- Não responde à questão/Resposta ambígua

Por fim, 8 alunos não responderam ou apresentaram uma resposta ambígua.

Quanto aos indicadores da forma pela qual os alunos consideram que a sua presença regular nas aulas não depende do regime de avaliação (contínua ou final) escolhido, foram identificados os seguintes:

- Para aprender/É importante

158 alunos referiram que vão às aulas para aprender, porque é importante frequentar as aulas e que isso não depende do regime de avaliação escolhido. referiram os alunos, "*frequento as aulas porque as considero essenciais para a minha formação, independentemente do regime que frequento*" (id16), "*porque tento estar em todas as aulas para exercitar, mesmo que alguém escolha final é necessário aulas para aprender a matéria*" (id25), "*porque tanto em avaliação contínua como final é necessário um acompanhamento sistemático da matéria para que exista uma melhor compreensão da mesma*" (id29), "*porque considero as aulas importantes tanto a esta como a outras disciplinas que se um aluno não frequentar não sabe o que fazer nos testes*" (id78), "*um aluno consciente e se tiver responsabilidade sabe que tem que ir às aulas para ter melhores resultados*" (id57), "*tiro rendimento das aulas, facilitando o estudo em casa e conseguindo conciliar com outras unidades curriculares*" (id143) e "*para quem*

já não tem Matemática há muito tempo as aulas são fundamentais para acompanhar a matéria" (id182).

- Vai sempre às aulas

70 alunos referiram que vão sempre às aulas independentemente do regime de avaliação escolhido. referiram os alunos que "*irei de qualquer das formas estar presente em todas as aulas*" (id9), "*penso que a presença regular não tem a ver com o regime de avaliação*" (id54), "*não importa qual o regime de avaliação em que estou, tenho de vir as aulas na mesma*" (id62) "*porque venho à faculdade para aprender, não para andar a faltar*" (id303).

- É importante a ajuda dos professores

22 alunos referiram que vão às aulas independentemente do regime de avaliação pelo qual optaram porque a ajuda dos professores é importante. Disseram os alunos que frequentam as aulas "*porque matemática é uma disciplina que não se consegue aprender sozinho*" (id77). Um outro aluno disse mesmo, "*não sou autodidata*" (id62). Referiram ainda que "*a Matemática é uma disciplina que deve ser estudada com acompanhamento do professor*" (id177), "*os professores ajudam-nos a preparar bem para os testes/exames*" (id135) e "*acho importante ir sempre às aulas porque a professora explica as coisas melhor do que se formos nós a querer perceber sozinhos*" (id265).

- Não responde à questão/Resposta ambígua

Por fim, 4 alunos não responderam ou apresentaram uma resposta ambígua.

O número de testes é importante na escolha do regime de avaliação

Verificamos que 324 (87%) dos alunos consideraram que o facto de existirem 3 testes não foi importante na opção pelo regime de avaliação, enquanto os restantes consideraram que sim. Considerando as variáveis utilizadas para a caracterização dos alunos que responderam ao questionário, estas percentagens mantêm-se idênticas para todas elas.

Quanto aos indicadores dos motivos pelos quais a existência de 3 testes não foi importante na opção pelo regime de avaliação, foram identificados os seguintes:

- Prefere sempre contínua

33 alunos indicaram que preferem sempre avaliação contínua. Disseram os alunos que "*opto sempre por avaliação contínua*" (id390), "*tencionava fazer em contínua de qualquer forma*" (id363), "*porque avaliação contínua é sempre melhor sendo dois, quatro ou dez testes*" (id306).

- Prefere contínua por ser mais acessível

7 alunos preferiram avaliação contínua porque consideram mais acessível, independentemente do número de testes. Dizem os alunos que preferem avaliação contínua de qualquer das formas, "*pois a matéria é dividida e assim mais fácil para se estudar*" (id179).

- Prefere contínua devido às dificuldades

4 alunos preferiram avaliação contínua devido às dificuldades da UC: "*tenho muitas dificuldades*" (id26)

- Prefere Contínua para não ir a Exame Final

2 alunos referiram que preferem avaliação contínua para não terem de ir a Exame Final: "*inscrevi me em avaliação contínua, para caso conseguisse passar não necessitaria de ir a Exame Final*" (id150).

- Aconselhamento do docente

1 aluno referiu que optou por avaliação contínua por aconselhamento do docente: "*Escolhi o regime de avaliação contínua por aconselhamento do docente*" (id408).

- Sem opinião

3 alunos não apresentaram a sua opinião neste domínio.

Opinião sobre o Teste de "Repescagem"

Todos os alunos foram inquiridos sobre a sua opinião quanto à existência do Teste de "Repescagem", já referido anteriormente. Foram identificados os seguintes os indicadores pelos alunos:

- Motivação, outra oportunidade para não desistir da avaliação contínua

179 alunos concordaram com a existência do Teste de "Repescagem" pois para eles é mais uma motivação, mais uma oportunidade para continuarem no regime de avaliação contínua. Disseram os alunos que "*assim não nos deixa logo de lado caso a nota seja menos boa, pois vamo-nos esforçar nos próximos para no Teste de "Repescagem" ir repetir aquele que correu menos bem*" (id14), "*boa oportunidade de terminar a avaliação contínua no caso de um dos testes ter corrido mal, visto toda a gente ter uns dias melhores e outros piores*" (id29), "*é ótimo pois isso deixa-nos um pouco mais tranquilos em relação ao tempo que temos para estudar, e dá-nos como uma segunda oportunidade*" (id48), "*é motivador para os alunos*

continuarem a assistir às aulas, e não desistirem tão facilmente da disciplina" (id193), "*permite que caso se esteja desconcentrado ou doente num dos testes, existe sempre a possibilidade de repescagem*" (id204) e "*faz com que os alunos se esforcem mais porque existe mais oportunidades e consequentemente uma avaliação mais justa face aos conhecimentos*" (id388).

- Ajuda alunos a ter positiva

78 alunos concordaram com o Teste de "Repescagem" porque ajuda os alunos a conseguirem alcançar uma classificação positiva no final do semestre. Disseram os alunos que "*é uma forma de ajudar os alunos a ter nota positiva a Matemática*" (id11), "*para aqueles alunos que não tenham nota para passar à disciplina é um grande alívio*" (id63) e "*é uma grande ajuda para conseguir uma nota para passar*" (id200).

- Apoio numa UC difícil

18 alunos referiram que concordam com o Teste de "Repescagem" porque é um apoio numa UC que é difícil. Disseram os alunos que "*parece-me um conceito interessante numa unidade curricular na qual os alunos têm muitas dificuldades*" (id9), "*pois dado não ter bases de secundário existem sempre algumas matérias mais complicadas, o que leva a possível substituição de um teste ser fundamental para a aprovação da disciplina*" (id134) e que é "*bastante bom porque ajuda imenso aqueles que não têm tantas facilidades na Matemática*" (id387).

- Devia ser também para melhorar notas

10 alunos não concordam totalmente com o Teste de "Repescagem", porque consideram que também deveria ser utilizado para os alunos que pretendem obter melhores classificações. Disseram os alunos que "*na minha opinião o teste de repescagem não deveria ser só para alunos com nota negativa, também poderia ser para melhoria de notas*" (id34), "*acho bem mas acho que a repescagem deveria ser também para aqueles que embora já tenham positiva fizessem uma melhoria" (id181) e "não deveria ser de repescagem mas sim de melhoramento*" (id370).

- Só concorda

86 alunos apenas afirmaram concordarem com o Teste de "Repescagem", sem apresentarem qualquer motivo adicional.

- Não concorda

5 alunos não concordaram com o Teste de "Repescagem". Disseram eles que "*não penso que seja necessário*" (id262), "*o teste de repescagem não é justo*" (id370) e "*não deveria existir. Nas outras disciplinas também não existe*" (id400).

- Não sabe da existência deste teste

1 aluno não sabe da existência do Teste de "Repescagem": "*não tenho opinião pois não sabia*" (id368).

- Não responde à questão/Resposta ambígua

Por fim, 1 aluno não respondeu ou apresenta uma resposta ambígua.

Quais as vantagens das QEM para os alunos

Todos os alunos foram inquiridos sobre a sua opinião quanto às vantagens dos testes com QEM implementados no *Moodle*, do ponto de vista dos alunos. Com exceção de 70 alunos que referiram a ausência de vantagens, os restantes apontaram respostas que se englobam nos seguintes indicadores:

- Opções de resposta ajudam a encontrar solução

153 alunos referiram que uma das vantagens prende-se com a existência das opções de resposta que ajudam a encontrar a resposta correta. Afirmaram os alunos que "*é vantajoso, pois sabemos que terá que ser uma das 4 soluções*" (id26), "*se a resposta que obtiver através dos seus cálculos não for nenhuma das opções é que alguma coisa está errada*" (id25), "*sabemos que uma das respostas é a correta o que nos permite, em certos casos, ir por tentativas*" (id55), "*é uma pequena ajuda, pois algumas vezes não temos certeza de que a nossa resposta está certa e quando olhamos para as respostas podemos logo excluir algumas e confirmar se a nossa é a correta ou não*" (id60) e "*penso que a vantagem é em algumas alíneas não todas claro, por vezes existem questões em que temos a hipótese de substituir as respostas nas alíneas que nos são dadas no próprio enunciado. O que muitas vezes facilita o trabalho e a gestão do tempo dos alunos*" (id106).

- Menor preocupação com a resolução e menor tempo de resposta

46 alunos referiram que há menor preocupação com a resolução e consequentemente é necessário menor tempo para a resposta. Disseram os alunos, "*responder a escolha múltipla é mais rápido pois podemos apenas fazer um esboço no papel a explicar como chegamos à resposta*" (id17), "*não há necessidade de escrever tanto pelo que se poupa tempo*" (id76) e "*não é preciso fazer todos os cálculos, e explicar o raciocínio, poupa-se tempo no teste*" (id13).

- Tentar acertar à sorte

42 alunos referiram que a possibilidade de tentar acertar à sorte na resposta correta é uma vantagem dos testes. Disseram os alunos que "*podem sempre responder à sorte quando não sabem*" (id15), "*podem ter a sorte de mesmo não sabendo a matéria acertarem na resposta*" (id72), "*há sempre a possibilidade de, em caso de dúvida, poder arriscar numa oportunidade*" (id398) e "*a vantagem é que os alunos com sorte, ao selecionarem as respostas aleatoriamente têm aprovação por vezes sem sequer saberem bem o que estão a fazer*" (id139).

- Mais fáceis

37 alunos referiram que os factos de os testes serem mais fáceis é uma vantagem. Disseram os alunos que os testes são "*mais intuitivos e de certa forma mais fáceis" (id 304) e que "as perguntas têm tendência a ser mais fáceis*" (id349).

- Correção e classificações mais rápidas

10 alunos apontaram como vantagem o facto de a correção e consequente obtenção das classificações ser mais rápida. Disseram os alunos que os testes "*são corrigidos rapidamente e por consequência as notas também saem rapidamente*" (id24), que há "*rapidez na obtenção das notas*" (id201) e permitem "*saber as notas mais cedo*" (id148).

- Justiça na correção

4 alunos referiram que uma das vantagens é haver mais justiça na correção. Referiram os alunos que "*diminuindo as diferenças entre correções de professores*" (id203), "*não haverá injustiça na correção, porque ou está certa a opção ou não está*" (id63).

- Ter uma ideia da classificação final

2 alunos apresentaram como vantagem o poderem ter uma ideia da classificação final: "*a vantagem é de o aluno ficar com uma ideia mais clara da classificação que poderá obter*" (id151).

- Não há repetição de perguntas de aluno para aluno

2 alunos disseram que não haver repetição de perguntas de aluno para aluno é uma vantagem. "*A vantagem que posso enumerar é precisamente a não repetição de perguntas de aluno para aluno*" (id61) e por isso "*não se copia*" (id206).

- Poder copiar

2 alunos referiram que poder copiar é uma vantagem.

- Poupança de papel

Um aluno referiu que uma vantagem é a poupança de papel: "*É engraçado e gastamos menos folhas de papel, logo, protegemos o ambiente!*" (id62).

- Não responde à questão/Resposta ambígua

Por fim, 23 alunos não responderam ou apresentaram uma resposta ambígua.

Quais as desvantagens dos testes com QEM para os alunos

Todos os alunos foram inquiridos sobre a sua opinião quanto às desvantagens dos testes com QEM implementados no *Moodle*, do ponto de vista dos alunos. Com exceção de 69 alunos, que referiram a ausência de desvantagens, os restantes apresentaram respostas que foram classificadas nos seguintes indicadores:

- Raciocínio não ser considerado

145 alunos apresentaram que uma das desvantagens é o raciocínio não ser considerado. Disseram os alunos que "*se fossem respostas de desenvolvimento podíamos ter sempre uns pontinhos por algumas resoluções*" (id17), "*para um aluno que tenha dificuldades mas que saiba fazer os exercícios, é uma desvantagem pois por vezes sabemos fazer os exercícios mas não conseguimos obter o resultado*" (id45), "*o aluno não consegue justificar o seu raciocínio, logo se errar a resposta, os seus cálculos não serão avaliados, mesmo que a sua linha de raciocínio estivesse correta*" (id66) e "*não existe cotações intermédias, basicamente os nosso cálculos não contam, só conta o resultado final, e caso nos enganemos numa coisa mínima e tenha lá essa resposta temos mal e se fosse de escrita tínhamos alguma cotação pelos passos que fizemos*" (id112).

- Penalização por escolha errada

77 alunos referiram que as penalizações existentes são uma desvantagem. Afirmaram os alunos que "*quando se erra uma pergunta não só se fica sem a cotação mas como nos é retirada uma percentagem da mesma*" (id43) e existem "*Descontos elevados por resposta errada*" (id411).

- Opções de respostas semelhantes - causa confusão

45 alunos referiram que há opções de respostas muito semelhantes o que causa confusão e isso é uma desvantagem. Disseram os alunos que "*as opções por vezes são muito parecidas o que pode levar a enganos*" (id40).

- Possibilidade de acertar à sorte

10 alunos entenderam a possibilidade de acertar numa resposta certa "à sorte" como uma desvantagem: "*os alunos que não estão bem preparados ao ficar com dúvidas nos resultados, arriscam na resposta, podendo esta estar errada*" (id156), ou seja, "*este tipo de teste pode ser desvantajoso para os alunos que sabem a matéria e tiram a mesma nota que os que não sabem a matéria e têm a mesma nota dos que são bons a esta disciplina*" (id158).

- O grau de dificuldade das questões não é igual para todos

7 alunos apontaram como desvantagem o facto de o grau de dificuldade dos testes poder não ser igual para todos: "*o grau de dificuldade por vezes pode não ser exatamente o mesmo*" (id301).

- Pouco tempo de resolução

4 alunos referiram que têm pouco tempo para a resolução do teste: "*única desvantagem dos testes de escolha múltipla é o tempo*" (id355).

- Problemas relacionados com sistema informático

4 alunos referiram a desvantagem de poderem existir problemas com o sistema informático: "*É correr o risco de a Internet do computador ir a baixo ou de a página atualizar, pois caso isso aconteça já não há mais hipóteses para voltar atrás*" (id24).

- Testes mais difíceis

4 alunos apresentaram como desvantagem o facto de os testes serem mais difíceis: "*as questões acabam por ser mais exigentes do que num teste de desenvolvimento escrito*" (id19).

- Carregar o computador

3 alunos apontaram como desvantagem o facto de terem de transportar o computador com eles para a realização do teste. Disseram eles que é uma desvantagem "*ter que andar com o portátil às costas*" (id212).

- Cria ideia de facilidade

2 alunos referiram que os testes criam uma ideia de facilidade e que isso é uma desvantagem: "*a meu ver, cria uma ideia de facilidade no aluno que não é benéfica ao seu desenvolvimento intelectual*" (id408).

- Testes fáceis

Numa linha de pensamento idêntico, 2 alunos referiram que os testes serem mais fáceis é uma desvantagem: "*a desvantagem principal é o facto de não ser necessário desenvolver problemas mais complexos e que exigem mais tempo*" (id349).

- Expressões matemáticas confusas

2 alunos indicaram que as expressões matemáticas apresentadas pelo *Moodle* são um pouco confusas. Relativamente às expressões apresentadas no ecrã disseram os alunos que é desvantajoso pois verifica-se "*a confusão entre alíneas por vezes, só no sinal de - ou de + , ou mesmo no denominador, depende*" (id11).

- Apontamentos sem QEM

1 aluno referiu que uma desvantagem é os apontamentos da UC não terem exercícios com QEM: "*o estudo torna-se complicado uma vez que os exercícios propostos não são em escolha múltipla*" (id65).

- É difícil cometer fraude

1 aluno referiu que o facto de não poder cometer fraude é uma desvantagem: "*os testes são muito diferentes, não dá pra copiar*" (id62).

- Opções a mais

1 aluno referiu que existirem opções a mais pode ser uma desvantagem: "*excesso de respostas pode dificultar*" (id23).

- Não responde à questão/Resposta ambígua

Por fim, 12 alunos não responderam ou apresentaram uma resposta ambígua

6.6.3. Síntese da opinião dos estudantes sobre o processo de *e-assessment* implementado

Em termos de síntese, podemos comprovar que os alunos referiram que as QEM são justas, e em particular uniformes, simples e acessíveis, sendo mais difícil cometer "fraude". Notaram ainda que as opções de respostas os ajudam a encontrar a solução. Os alunos que não consideram as QEM justas apresentam como principais razões a não avaliação do raciocínio e os testes aleatórios terem níveis de dificuldade diferentes, bem como a existência de penalização por uma resposta errada.

Quando questionados sobre se preferiam o teste em papel ou no *Moodle*, a maioria dos alunos considera não ter preferência por uma das formas. Daqueles que preferem os testes em papel, sugerem que a única vantagem seria não ocorrerem problemas informáticos. Referem, contudo, que em papel poderiam ter piores classificações e que os testes poderiam ser mais confusos e mais difíceis.

Quanto às mudanças que este processo provocou nas suas práticas, os alunos realçaram o facto de terem passado a abordar de modo diferente a forma como resolviam os exercícios e estarem mais atentos aos pormenores. Mais importante ainda, foi o facto de considerarem que a sua presença nas aulas se tornou mais assídua e que tiveram que estudar de forma mais parcelar de modo a não deixar o trabalho somente para o final. O número de testes ao longo do semestre não era importante desde que houvesse avaliação contínua. Finalmente, o teste de "Repescagem" surge como um fator importante para não desistirem da avaliação continua.

CAPÍTULO 7. DISCUSSÃO

O problema principal que levou à realização deste trabalho prende-se com a implementação do Processo de Bolonha, o qual apontava para novas metodologias no ensino e em especial para a mudança do antigo regime de avaliação. Um dos aspetos considerados era que, para a generalidade das instituições de Ensino Superior, o sistema de avaliação assentava somente num único Exame Final (Pereira & Flores, 2012, p. 535). O Processo de Bolonha apontava não só para a necessidade de realizar avaliação contínua ao longo do(s) semestre(s), mas também para a necessidade de englobar metodologias diversificadas (Boticki & Milasinovic, 2008; Llamas-Nistal et al., 2013; Mora et al., 2012; Rod et al., 2010). A implementação do Processo de Bolonha na instituição onde o presente estudo foi desenvolvido levou a uma redução da carga horária destinada à lecionação, devido à reestruturação dos cursos e, além disso, houve também um aumento no número de alunos por turma. Esta situação resultou no problema de investigação desta tese: como implementar uma estratégia de avaliação contínua, com turmas numerosas, numa instituição de Ensino Superior, em tempo útil e utilizando os recursos disponibilizados pela instituição?

Na literatura refere-se que o *e-assessment* permite aliviar o trabalho que representa para o professor avaliar um elevado número de alunos (Blanco & Ginovart, 2012; Boticki & Milasinovic, 2008; Bull & Danson, 2001; Jordan, 2013; Mora et al., 2012; Moscinska & Rutkowski, 2012; Rust, 2001; Yorke, 2001). Além disso, refere-se também que a utilização de questões de escolha múltipla (QEM), nomeadamente em formato de *e-assessment*, apresenta ainda uma maior facilidade na avaliação de um elevado número de alunos, em testes de grande escala, permitindo poupança de tempo e de recursos e sendo mais fácil de gerir (Bible et al., 2008; Brown, 2001; Burton et al., 1991; Camilo & Silva, 2008; Clegg & Cashin, 1986; Douglas et al., 2012; Ferrão, 2010; Green & Mitchell, 2009; Haladyna et al., 2002; Heron & Lerpiniere, 2013; Jordan, 2013; Liu et al., 2011; Nicol, 2007; Wild et al., 1997). A estratégia de avaliação contínua abordada nesta tese veio de encontro a estas indicações recolhidas na literatura e, além disso, foram-se introduzindo novas metodologias na avaliação contínua que ao longo dos anos se mostraram positivas.

Cumpre-nos, assim, avaliar o primeiro objetivo específico desta investigação: perceber como o *e-assessment* pode influenciar o processo de ensino-aprendizagem por parte dos alunos.

Verificámos que houve uma melhoria nos resultados das classificações dos estudantes entre o momento prévio à introdução da avaliação com *e-assessment* e o final da implementação deste trabalho. Em termos globais, poder-se-á afirmar que, quer por anos letivos, quer por ciclos de IA, houve uma evolução positiva das classificações, sendo esta mais acentuada nas UC do 1.º semestre. Esta diferença poderá dever-se ao facto de os conteúdos lecionados nas UC do 1.º semestre serem mais básicos do que aqueles que são lecionados nas UC do 2.º semestre, para além de outros aspetos que referiremos adiante.

A análise mais detalhada da evolução das classificações relativamente às UC do 1.º semestre, entre 2008 e 2010, anos que corresponderam ao 1.º ciclo de IA, permite constatar um decréscimo das classificações, as quais foram, além disso, muito baixas. Este decréscimo poderá ter origem em diversos fatores. Um desses fatores poderá estar relacionado com o facto de os testes terem sido realizados como trabalhos de casa. Apesar de haver limitações em relação ao tempo disponível para os alunos responderem às questões em casa, descobrimos no 2.º ano da experiência (2009), mas mais no 3.º ano (2010, final do 1.º ciclo de IA) que os alunos estariam a ser ajudados por colegas ou até explicadores, o que levaria a que não estudassem por eles próprios os conteúdos da UC, refletindo-se depois essa falta de estudo nas classificações obtidas nos testes realizados na escola. Apesar de tudo, estes trabalhos de casa foram importantes para a familiarização com a plataforma *Moodle*, e também para aprendizagem quer por parte dos docentes, quer por parte dos alunos. Outro fator que poderá ter influenciado o decréscimo nas classificações tem a ver com o número de testes que foram realizados em avaliação contínua nestes primeiros anos, relativos ao 1.º Ciclo de IA, que foram somente dois. Assim sendo, caso os alunos tivessem uma classificação baixa no primeiro teste, teriam tendência a desistir do regime de avaliação contínua e quase de imediato da frequência às sessões de contacto, o que teria quase sempre como consequência a obtenção de uma classificação negativa. Outro fator, que poderá ter influenciado o nível tão baixo destas classificações, estará relacionado com a forma de realização dos testes neste período temporal. Estes testes, com QEM em formato papel, foram realizados fora do período letivo habitual, geralmente às quartas-feiras de tarde, em três ou quatro turnos. Apesar de os alunos serem previamente avisados dos turnos em que deveriam realizar os seus testes e dos respetivos horários, havia sempre quem não estivesse devidamente informado, provocando alguma perturbação no início e no final de cada turno, devido ao elevado número de alunos. Para alguns estudantes este era um fator de ansiedade adicional e também, para muitos deles, de quebra de concentração.

No final do 1.º ciclo de IA, constatou-se que os resultados não estavam a melhorar e daí a decisão de introduzir algumas mudanças, as quais vieram a ter reflexos no 2.º ciclo de IA. A principal mudança neste 2.º ciclo de IA, que corresponde aos anos 2011 e 2012, foi a mudança dos testes em formato papel para formato digital, implementados no *Moodle*. Além disso, os testes passaram a ser realizados durante o tempo letivo de cada turma, embora em ambiente fora da sala de aula. Esta alteração possibilitou a criação de um ambiente mais tranquilo durante a realização dos testes, visto que foi possível proceder a uma gestão mais efetiva dos espaços de realização dos testes e do número de alunos a avaliar, o que, na nossa opinião, influenciou de forma positiva todo o processo. Outra mudança consistiu no número de testes a realizar em avaliação contínua. Atendendo ao número de aulas por semestre e à distribuição dos conteúdos em cada UC, considerou-se que o número de testes mais adequado seria de três por semestre. O que se foi verificando na prática, com a realização de três testes, e sendo o peso de cada um deles na avaliação mais ou menos equilibrado, foi uma tendência para os alunos não abandonarem as aulas nem a avaliação contínua, até realizarem o último teste (e

não logo no final do primeiro teste, caso este tivesse corrido mal, como anteriormente acontecia, pois deixavam de ter a possibilidade de recuperar a sua classificação quando tinham apenas dois testes). Este aspeto está patente nos questionários realizados aos alunos, visto que a grande maioria reconheceu que o número de testes influenciou a escolha do regime de avaliação contínua. Uma outra alteração que consideramos importante consistiu na introdução de um parâmetro adicional na avaliação contínua, o qual envolvia a assiduidade dos alunos e a sua participação nas aulas. Assim sendo, para além dos testes realizados no *Moodle*, os alunos assíduos e participativos tinham uma pequena bonificação na classificação final. A grande falta de assiduidade dos alunos era, sem dúvida, um dos principais problemas que se verificavam nas UC de Matemática. Esta alteração ajudou a provocar mudanças nos comportamentos dos alunos, os quais passaram a frequentar as aulas com maior regularidade e a participarem mais nas mesmas. Este facto é evidente tanto no discurso dos estudantes como no dos docentes.

Apesar de a existência de três testes ter melhorado quer a assiduidade às aulas quer à avaliação contínua no último ciclo de IA, correspondente ao ano a 2014, foi introduzida a possibilidade de os alunos que tivessem uma classificação final negativa após a realização do último teste, realizarem ainda um teste suplementar, o Teste de "Repescagem". Este teste possibilitava que alunos com classificação final negativa pudessem selecionar um dos testes, de entre aqueles que já tinham realizado, tendo assim oportunidade de melhorar a sua classificação de modo a poderem concluir a UC com aproveitamento. Este teste veio motivar mais os alunos a não abandonarem a avaliação contínua, pois sabiam que no final ainda lhes restava uma oportunidade para poderem realizar com sucesso a avaliação à UC. Assim, além dos 3 testes que já ajudavam a que os alunos não abandonassem, nem as aulas nem a avaliação contínua, o Teste de "Repescagem" veio fortalecer mais este aspeto. Compreende-se, assim, que os estudantes tenham mostrado opiniões bastante positivas sobre o Teste de "Repescagem".

Apesar das classificações terem melhorado desde 2010, o grande salto nessa melhoria ocorreu a partir de 2013. A melhoria da assiduidade e do número de alunos a concluir as UC por avaliação contínua é, na nossa opinião, uma das grandes mais-valias desta nova estratégia de avaliação contínua e é, no nosso entender, a principal responsável pela melhoria das classificações finais dos alunos que se verificou ao longo de todo o processo de implementação desta estratégia de *e-assessment*. Este aspeto está também patente nas entrevistas realizadas aos docentes.

Relativamente à evolução das classificações nas UC do 2.º semestre, podem ser salientados, em linhas gerais, aspetos semelhantes aos já referidos para o 1.º semestre. No entanto, as melhorias nas classificações que se verificaram no 1.º semestre não se observaram de forma tão evidente no 2.º semestre. Como já referimos acima, um dos motivos poderá prender-se com o facto de os conteúdos programáticos do 1.º semestre serem mais básicos. No entanto, pensamos que este não será o único fator responsável pelo facto de as classificações no 2.º semestre não serem tão positivas como as do 1.º semestre. Na verdade, o 2.º semestre foi sempre mais sujeito a turbulências do que o 1.º semestre. Em 2010, por exemplo, não se realizou avaliação contínua, dado não existirem condições para tal na instituição. Devido à

implementação do Processo de Bolonha, a presidência, também considerando a contestação existente por parte da Associação de Estudantes, exercia algum tipo de pressão para que as UC tivessem o regime de avaliação contínua mas, por outro lado, não eram criadas as condições necessárias. Em 2012, excecionalmente, apenas se realizaram dois testes no 2º semestre. Este facto ocorreu devido a problemas técnicos relacionados com o servidor onde o *Moodle* está alojado, o qual não tinha capacidade para que um tão grande número de alunos pudesse aceder à plataforma em simultâneo (após o sucedido, foi necessário alojar o *Moodle* noutro servidor). Em 2013 fez-se a tentativa de alterar o número de testes para quatro, mas a gestão em termos da calendarização das aulas foi muito complicada e os benefícios obtidos não foram relevantes.

Apesar de tudo, os bons resultados obtidos nas classificações, quer no 1.º quer no 2.º semestre, confirmam que as medidas introduzidas em cada um dos ciclos de IA produziram bons frutos desde o começo até ao final deste trabalho.

A maioria dos estudantes que participaram no estudo afirmaram que não houve mudanças nas suas práticas com a implementação do *e-assessment*. No entanto, numa análise mais aprofundada aos depoimentos fornecidos, acaba por ser evidente que houve algumas mudanças, entre as quais se destaca o aumento da assiduidade às sessões de contacto e à avaliação contínua. Esta mudança foi fundamental para que as UC de Matemática passassem a apresentar melhores classificações e deixassem de ser vistas como UC demasiado difíceis por parte dos alunos, o que limitava, logo à partida, todas as suas expectativas.

Outra mudança passou pela incorporação do uso das TIC, constatando-se que a grande maioria dos estudantes já olha para a sua utilização com muita naturalidade. Uma das principais dificuldades referidas pelos estudantes tem a ver com a utilização do *Moodle* propriamente dita. Estas dificuldades apenas foram referidas por estudantes que frequentavam a UC pela primeira vez ou por alunos noturnos. Estes últimos eram, no geral, alunos mais velhos (todos os alunos com mais de 36 anos de idade eram alunos noturnos), os quais poderão não estar tão familiarizados com as TIC como os mais jovens. Quanto aos primeiros, os que estavam a frequentar a UC pela primeira vez, eram estudantes recém-ingressantes no Ensino Superior, vindos do Ensino Secundário, e que estavam no 1.º semestre do 1.º ano do curso. O processo de transição para o Ensino Superior implica, na maioria das situações, mudanças significativas na vida pessoal, social e académica dos jovens, com novas exigências no seu percurso académico e desenvolvimental, pelo que tem sido entendido como um momento importante, onde níveis superiores de maturidade, autonomia e autoeficácia parecem ser desafios na qualidade da adaptação ao novo contexto académico (Soares, Guisande, & Almeida, 2007). Assim, as dificuldades relatadas por este grupo particular de estudantes podem estar relacionadas com esta etapa particular da sua adaptação ao Ensino Superior, não estando ainda devidamente familiarizados com toda a sua envolvente. Neste seguimento, saliente-se que fez parte da estratégia de *e-assessment* implementada, a realização de uma sessão, antes do primeiro teste, na qual os alunos realizavam um teste modelo. Esta sessão destinava-se precisamente a preparar os estudantes para a utilização do *Moodle*, tendo sido devidamente divulgada,

preparada e acompanhada por docentes e técnicos de informática do ISCAP, em todos os anos e semestres letivos em que decorreu este processo, pelo que estas dificuldades não deveriam acontecer. Contudo, importa salientar que no início do 2.º ciclo de IA observaram-se alguns constrangimentos, pois o servidor onde estava alojado o *Moodle* tinha pouca capacidade, o que provocou várias dificuldades aos alunos no acesso aos testes e na sua finalização. No entanto, esta situação ficou resolvida na etapa final do estudo e os problemas durante a realização dos testes já eram praticamente inexistentes nessa altura. As dificuldades ainda detetadas estavam normalmente relacionadas com a má utilização dos computadores por parte dos alunos, o que também se foi conseguindo resolver com sucesso. Uma outra situação que, de alguma forma, suscitou alguma surpresa, tem a ver com o facto de alguns (quatro) alunos apontarem como dificuldade a necessidade de transportarem os seus computadores para o teste. No entanto, a partir da sua experiência como docente, o autor da tese pode afirmar, sem sombra de dúvida, que o número de alunos com problemas quanto ao uso das TIC foi diminuindo bastante, sendo neste momento praticamente nulo.

O segundo objetivo da investigação consistia em perceber como o *e-assessment* pode influenciar o processo de ensino-aprendizagem por parte dos docentes.

Tal como os alunos, quase todos os docentes afirmaram, nas entrevistas que foram realizadas, não ter havido mudanças nas suas práticas educativas. Apesar disso, os docentes acabaram por reconhecer que melhoraram no que diz respeito à elaboração das QEM e que passaram a ter uma atenção redobrada sobre a elaboração das questões e sobre a forma de lecionar as aulas. Outro aspeto de mudança tem a ver com a introdução das TIC no processo de ensino-aprendizagem, as quais foram incorporadas de forma natural depois de todo o processo de aprendizagem que houve ao longo de toda a implementação. Uma dificuldade sentida pelos docentes durante a criação do banco de questões, e que foi referida nas entrevistas, prendeu-se com a utilização do *TeX* para a escrita das expressões matemáticas a colocar no *Moodle*. Aliás, para professores que iniciam a escrita de caracteres Matemáticos em *e-assement* é uma dificuldade acrescida, pois nem sempre o *software* utilizado para a construção das QEM permite uma escrita rápida (Brito et al., 2009, p. 167). Como já foi referido, a maioria dos docentes não dominava o *TeX* e portanto foi utilizado o *software TeXaide* para ajudar os docentes com mais dificuldades. Ainda assim, os docentes tiveram dificuldades na escrita das expressões matemáticas. Nos últimos anos a experiência dos docentes no uso do TeX já permite ultrapassar este problema. Além disso, as últimas versões do *Moodle* também têm um bom editor de *TeX* incorporado, o qual permite a inserção das fórmulas matemáticas diretamente no *Moodle*, para quem não domina o *TeX*. Mesmo com a evolução e melhoria na introdução e interpretação do TeX por parte do *Moodle*, alguns alunos ainda referiram uma certa dificuldade no que diz respeito à leitura de algumas das fórmulas, principalmente quando elas eram muito semelhantes. Contudo, nos últimos anos, estes problemas surgiram mais a quem usava *Tablets* com tamanhos de ecrã pequeno. Pensamos que com a implementação da última versão do *Moodle* este problema deverá desaparecer. Neste momento, todos os docentes estão bastante

familiarizados com a sua utilização, sendo capazes de resolver os poucos problemas que vão surgindo, principalmente durante a realização dos testes. Além disso, neste momento já todos os docentes conseguem, sem qualquer tipo de dificuldade, introduzir as QEM no *Moodle*.

Houve, no entanto, três docentes que reconheceram mudanças, mas referiram que as mesmas não foram provocadas pela implementação do *e-assessment*. Para estes docentes, a alteração na avaliação foi, ela sim, resultado de um processo de mudança global. Estas mudanças globais mencionadas pelos docentes referem-se a todas as mudanças institucionais necessárias à implementação do Processo de Bolonha, nomeadamente os equipamentos instalados e todo o investimento feito no processo de avaliação. O autor da tese é, no entanto, de opinião que o processo de mudança global ao qual se referem os colegas apenas foi possível devido à utilização das QEM na avaliação, as quais serviram como catalisador dessa mudança.

Podemos assim concluir que a alteração na forma de avaliação provocou mudanças nas práticas quer dos alunos, quer dos docentes, como documentado na literatura, na qual se refere que a introdução de sistemas de avaliação diferentes poderá provocar impacto importante em todo o processo educativo (Botticki & Milasinovic, 2008; Brown, 2001; Bull & Danson, 2001; Frankland, 2007a; Garfield & Ben-Zvi, 2008; Holmes, 2015; Jacob et al., 2006; Jarvis et al., 2003; JISC, 2007; Redecker & Johannessen, 2013; Scouller, 1998; Smith et al., 1996; Stödberg, 2012; Wild et al., 1997).

Definir boas práticas para o desenvolvimento de QEM na área da Matemática foi o terceiro objetivo desta investigação.

Este objetivo está diretamente ligado ao desenvolvimento do banco de questões que foi elaborado ao longo do período de tempo da investigação. Tal como referido na literatura, os bancos de questões podem contribuir para assegurar a validade e a fiabilidade do processo de avaliação (Bull & Danson, 2001; McAlpine, 2002b), consistindo já em si numa boa prática para o desenvolvimento de QEM. Este foi o trabalho mais demorado, quer em termos de elaboração das questões, quer em termos da sua colocação no *Moodle*. Verificou-se que o esforço necessário para o desenvolvimento do banco de questões, principalmente no início, foi bastante elevado, tal como é referido na literatura (Burton et al., 1991; Clegg & Cashin, 1986; Ferrão, 2010; Guo et al., 2014; Jordan, 2013; Liu et al., 2011).

Criar um banco de questões de modo a permitir que praticamente todos os alunos de uma mesma sala tivessem testes diferentes, levava a que o número de questões a implementar fosse grande. De facto, no final do 1.º Ciclo de IA havia 742 questões para as UC do 1.º semestre e 756 para as UC do 2.º semestre. Com este número de questões, à partida, estaria assegurado que os testes gerados de forma aleatória fossem diferentes de aluno para aluno. No entanto, poderia não estar garantido que o grau de dificuldade dos testes fosse idêntico para todos os alunos, apesar dos esforços para que tal fosse possível. Este foi, aliás, um dos aspetos mais apontados pelos alunos como motivo para não considerarem os testes justos. Salientamos que,

3 alunos referiram que os testes realizados em papel seriam mais justos porque seriam todos iguais. Parece-nos que terá havido alguma confusão por parte dos alunos, dado que os testes com QEM em formato papel já tinham várias versões. Assim sendo, parece-nos que os alunos terão dado esta resposta por estarem a pensar em testes em papel com questões de resposta aberta. Durante as aulas no 1º. Ciclo de IA, alguns alunos confirmaram aos docentes que um ou outro teste não teria o mesmo grau de dificuldade. De qualquer das formas, os docentes tiveram, desde o princípio desta investigação, o cuidado de assegurar que os testes fossem o mais justos possível. Conscientes destes problemas, tiveram sempre em mente esta preocupação. Com o avançar da investigação foi-se aos poucos conseguindo assegurar que as diferentes versões dos testes gerados pelo *Moodle* tivessem graus de dificuldade idênticos para todos os alunos. Houve desde o princípio uma preocupação com o problema da igualdade da dificuldade entre questões incluídas na mesma categoria. Assim, um grupo de professores elaborava as questões para uma mesma categoria, um outro grupo elaborava para outra categoria, etc., de modo a que a dificuldade das questões incluídas em cada uma das categorias fosse idêntica. As questões foram posteriormente analisadas pelo coordenador das UC e no seguimento desta análise foram criadas diferentes categorias ou subcategorias considerando graus de dificuldade diferentes, conforme os conteúdos programáticos. Este processo foi sempre executado em todos os ciclos de IA. Foi implementado um processo de revisão rigoroso, o que se revelou como muito importante, tal como é apontado na literatura (Haladyna, 2004), e que podemos afirmar ser uma boa prática para o desenvolvimento de QEM.

Na literatura encontramos uma outra boa prática para o desenvolvimento de QEM que consiste em seguir um conjunto de linhas de orientação (Burton et al., 1991; Camilo & Silva, 2008; Clegg & Cashin, 1986; Haladyna, 2004; Haladyna et al., 2002). De entre estas, considerámos as apresentadas por Haladyna e colaboradores (2002) como sendo as mais relevantes, tendo sido elaborado um questionário aos docentes relativo a estas linhas de orientação. Em primeiro lugar reconhecemos que os resultados obtidos não poderão ser objeto de qualquer generalização. No entanto, apresentamos algumas reflexões que nos parecem importantes.

Em primeiro lugar, a resposta a este questionário foi importante para os docentes, dado que os levou a refletir sobre as linhas de orientação apresentadas, confrontando as suas práticas com as que são propostas pelos especialistas, o que permitiu uma reflexão sobre os aspetos aos quais era dado maior relevância na elaboração das QEM. Apesar de, na generalidade, os docentes concordarem com todas as linhas de orientação, o grau de importância de cada uma das regras não é coincidente com o grau de importância verificado no estudo realizado por Haladyna e colaboradores (2002): apenas a linha de orientação "Utilizar humor, se ele é compatível com o professor e com o ambiente de aprendizagem" coincide como fazendo parte da lista das linhas de orientação menos relevantes em ambos os casos. Acresce ainda que existem 4 linhas de orientação que coincidem como sendo as mais importantes, a saber. "Incluir a ideia central no enunciado", "Garantir que todos os distratores são plausíveis", "Utilizar materiais inovadores para testar aprendizagens de nível elevado …" e "Certificar-se que as instruções no enunciado

são claras". No que diz respeito às linhas de orientação que geram mais controvérsia, apenas duas destas linhas de orientação são coincidentes com as do estudo de Haladyna e colaboradores (2002): "Utilizar cuidadosamente nenhum dos anteriores" e "Escrever o enunciado na forma afirmativa".

Saliente-se, dos resultados deste questionário, que os docentes mostraram uma grande preocupação com o facto de ser necessário que as questões elaboradas sejam claras e facilmente compreendidas pelos alunos, tal como é evidente quando consideramos as linhas de orientação que estes docentes apontaram como as mais relevantes. Surge como natural o facto de as linhas de orientação "Colocar as opções por ordem, lógica ou numérica" e "Desenvolver tantas opções eficazes quantas seja possível, mas a investigação sugere que 3 é adequado" terem sido consideradas as menos importantes por parte dos docentes. No que diz respeito à primeira, os docentes escolheram como alternativa "Misturar as opções aleatoriamente" nas definições do teste no *Moodle*, por forma a dificultar a fraude por parte dos alunos, logo as QEM foram elaboradas de modo a não haver qualquer tipo de ordem nas opções. Quanto à segunda, dado que foi acordado pelos docentes a existência de 4 opções, é natural que não tivesse existido concordância com esta linha de orientação. Poderemos considerar que algumas das linhas de orientação que terão muita importância noutras áreas do saber, não a terão na área da Matemática e vice-versa. De qualquer das formas, não era esse o objetivo do questionário. Há algumas linhas de orientação que levantaram alguma controvérsia entre os docentes, pelo que nos parece que, apesar de já ter passado algum tempo sobre a implementação deste questionário, poderá ser importante voltar a levantar estas questões e propor aos docentes uma nova reflexão, mais centrada nestas linhas de orientação que levantaram mais controvérsia. Será de considerar a versão atualizada destas linhas de orientação (Haladyna, 2004).

Um outro aspeto que se pode considerar como uma boa prática, e que é muito apontado pelos alunos nas entrevistas por questionário, tem a ver com as penalizações que são atribuídas pela indicação de uma resposta errada. Muitos deles não concordaram com essas penalizações, por descontarem na cotação das respostas que estavam corretas. Parece-nos que, neste caso, alguns dos alunos poderão ainda não ter percebido que esta penalização tem como objetivo desencorajar as tentativas de os alunos acertarem na resposta correta de forma aleatória, sem terem efetivo domínio das competências necessárias para tal, como é sugerido na literatura (Bush, 2015; Haladyna, 2004; Triantis & Ventouras, 2012). Este aspeto ficou mais claro a partir de algumas respostas que os alunos deram nos questionários, afirmando que uma das vantagens dos testes com QEM é que podem tentar acertar na resposta quando não sabem qual é a correta. No entanto, alguns alunos referiram que as penalizações não lhes permitem tentar acertar na resposta de forma aleatória. A este respeito, consideramos interessante que que alguns alunos consideraram-no como uma desvantagem e outros como vantagem. Os primeiros, porque assim não teriam a possibilidade de obter uma resposta correta mesmo sem terem competências para tal; os segundos consideraram que assim os testes seriam mais justos porque apenas os alunos

com competências obteriam a resposta correta. De qualquer forma, talvez seja conveniente, no futuro, esclarecer melhor os alunos em relação a este aspeto das penalizações.

Por fim, o quarto objetivo de investigação consistia em descobrir formas adequadas de análise das QEM de modo a fomentar uma avaliação tão justa quanto possível para os alunos.

Com o intuito de tornar os testes gerados a partir das QEM contidas no banco de questões o mais possível justos, realizou-se a análise das questões utilizando as teorias TCT e TRI. Assim, verificamos que mesmo com toda a preocupação que existiu na forma como as questões foram elaboradas e revistas, e apesar dos resultados serem animadores, havia ainda muitas questões que apresentavam problemas graves. Um dos principais problemas, no nosso caso, foi não ter sido possível aplicar a TRI e obter resultados que permitissem tirar conclusões no que diz respeito à dificuldade das questões que se encontravam no banco de questões, devido ao número elevado de questões por categoria. Este número elevado de questões, que por um lado é benéfico pois permite um elevado número de versões diferentes dos testes, provoca que o número de respostas por questão seja baixo tendo sido essa a causa principal para não ter sido possível aplicar a TRI. Este problema é apontado na literatura (Haladyna, 2004; Hambleton & Jones, 1993; Zickar & Broadfoot, 2009). Contudo, a TCT permitiu retirar algumas conclusões interessantes e, apesar de não ter sido possível aplicar a TRI, quando se analisaram todas as questões em simultâneo, tanto para o 1.º como para o 2.º semestre, obtiveram-se valores para o alfa de Cronbach de 0.953 e 0.943 respetivamente. Estes valores, sendo superiores a 0.8, indicam que, de certa forma, pelo menos no seu conjunto o banco de questões apresenta alguma consistência interna e, consequentemente, alguma fiabilidade. Isto é, podemos afirmar que as QEM medem o que pretendem medir. Pode-se dizer que apesar de algumas QEM poderem conter alguns problemas, no cômputo geral, os testes não serão assim tão injustos como alguns, poucos, alunos referiram. No entanto, salientamos que houve mais alunos a afirmarem que os testes são justos do que os que afirmaram o contrário, apesar de a diferença não ter sido muito grande.

É interessante observar que os alunos que não consideraram os testes justos apresentaram como razões principais o "facto" de não se avaliar o raciocínio e o facto de existirem penalizações, sendo que somente uma pequena parte, 15% dos alunos, referiu que os testes poderão apresentar níveis de dificuldade diferentes. A questão da existência de penalizações já foi discutida anteriormente, por isso, iremos discutir agora a questão relacionada com a capacidade (ou não) de os testes com QEM avaliarem efetivamente o raciocínio dos alunos e terem como limitação a impossibilidade de explicação das respostas dadas. Este aspeto foi particularmente salientado pelos alunos nas entrevistas por questionário e houve também alguns docentes que fizeram referência a este aspeto. Também na revisão de literatura, esta é uma das desvantagens assinaladas por alguns autores (Bible et al., 2008; Brown, 2001; Burton et al., 1991; Douglas et al., 2012; Ferrão, 2010; Green & Mitchell, 2009; Guo et al., 2014; Heron & Lerpiniere, 2013; Jordan, 2013; Lee et al., 2011; Liu et al., 2011; Nicol, 2007; Rod et al., 2010; Wild et al., 1997). Apesar de se tratar de um aspeto bastante controverso, o autor da

tese não concorda totalmente com a opinião destes alunos, indo mais de encontro ao que é expresso na literatura, na qual se aponta que as QEM "têm potencial para medir a compreensão, a análise, a capacidade de resolução de problemas e a capacidade de cálculo", o que chega mesmo a ser apontado como uma das vantagens desta tipologia de testes (Brown, 2001; Burton et al., 1991; Clegg & Cashin, 1986; Kim et al., 2012; Nicol, 2007). Na verdade, os alunos para responderem a uma questão terão de situá-la no conteúdo ou conteúdos respetivos, terão de a analisar, terão de aplicar o raciocínio ou raciocínios mais adequados e efetuar cálculos caso seja necessário. Para fomentar nos alunos a necessidade de realizar estes procedimentos, foi-lhes sempre entregue uma folha de rascunho, salientando-se que ela se destinava a esse efeito. Logo, quando os alunos respondem corretamente à questão, está-se a avaliar o raciocínio empregue. O único aspeto em que poderemos concordar com os alunos tem a ver com o facto de que, numa questão de resposta aberta pode-se contabilizar uma parte da resposta, o que no caso das QEM não acontece. No entanto, sabemos que alguns alunos, mas poucos, usam uma estratégia de exclusão por partes das opções de resposta, chegando mais rapidamente à solução, aspeto este que também é referido na literatura (Bible et al., 2008). No entanto, os alunos ao fazerem esta abordagem, não deixam de desenvolver um tipo de raciocínio, completamente válido ainda que diferente, para chegar à solução, o que muitas vezes possibilita a obtenção mais rápida da resposta correta. É evidente que este assunto é demasiado importante, merecendo sem dúvida uma análise mais aprofundada no futuro. Neste sentido, algumas ideias para que seja possível obter as condições para que se possa aplicar a TRI serão deixadas na conclusão desta tese.

Não estando diretamente relacionado com os objetivos da tese, há ainda um aspeto que não queremos deixar de referir nesta discussão. Este aspeto tem a ver com o facto de dois alunos terem referido que as notas em formato papel seriam conhecidas de forma mais rápida do que com o *Moodle*, o que aparentemente é uma contradição. O que se passou foi que os alunos não tiveram acesso imediato às suas classificações após terem realizado o teste no *Moodle*, por decisão dos docentes. Esta decisão prendeu-se com a necessidade de controlar problemas que pudessem vir a existir, tais como, erros em alguma das questões ou a possibilidade de não haver condições para a realização do teste em alguma das turmas. Assim sendo, foi considerado conveniente revelar as classificações aos alunos apenas após a realização dos testes em todas as turmas. Esta divulgação ocorreu, tipicamente, 2 a 4 dias após a realização do teste pela primeira turma. No entanto, houve um ano no qual se verificaram alguns problemas e essa divulgação ocorreu com algum atraso.

Ainda neste seguimento, alguns alunos, embora poucos, referiram que nos testes com QEM há uma diminuição da importância atribuída à linguagem Matemática escrita. Foi algo surpreendente para o autor da tese que os alunos tivessem essa preocupação, dado que seria mais expectável à partida que fossem os docentes a evidenciá-la. O autor da tese considera que esta é uma preocupação a valorizar e que deve ser objeto de análise no futuro.

Como conclusão desta discussão, é de referir que a estratégia de *e-assessment* implementada pode ser considerada um grande sucesso, apesar das limitações identificadas. Um dos fatores que reforça esta opinião é que alguns docentes de outras UC da instituição, vendo a forma como o trabalho foi feito e reconhecendo o seu sucesso, começaram também a utilizar o *e-assessment* para avaliação contínua sumativa.

CONCLUSÃO

Esta tese apresenta o processo de implementação de uma estratégia de avaliação contínua sumativa, utilizando *e-assessment*, com testes baseados em questões de escolha múltipla (QEM). Esta estratégia foi concebida com a finalidade de corresponder à necessidade de aplicar avaliação contínua sumativa, fortemente recomendada pelo Processo de Bolonha na instituição de Ensino Superior na qual a investigação foi desenvolvida. Foi implementado um banco de QEM no *Moodle* contendo um número considerável de questões. Para a implementação deste banco de questões foi seguido um rigoroso processo de revisão e de organização do trabalho. Os docentes refletiram sobre um conjunto de linhas de orientação, de modo a obter QEM de qualidade. Utilizando questões incluídas no banco de questões, foram implementados testes gerados aleatoriamente pelo *Moodle* a partir da seleção aleatória das questões.

Numa primeira fase, estes testes foram utilizados como trabalhos de casa opcionais, a saber, testar a utilização das QEM desenvolvidas na avaliação e ajudar os alunos a prepararem os testes de avaliação sumativa, tomando assim uma vertente formativa. Apesar de nesta primeira fase os testes serem opcionais, a adesão dos alunos foi bastante elevada, verificando-se que o número de alunos a realizarem os testes foi significativo.

Numa segunda fase, os testes passaram a ser utilizados para avaliação contínua sumativa, mas fora do ambiente de sala de aula. Foram criadas as condições tecnológicas necessárias para a implementação dos testes com o apoio dos serviços técnicos da instituição de Ensino Superior. Houve especial cuidado com as questões relacionadas com a segurança e com a fraude. As QEM do banco de questões foram analisadas para aferir a sua qualidade, utilizando a Teoria Clássica dos Testes (TCT) e a Teoria da Resposta ao Item (TRI), tendo sido obtidos resultados animadores quanto à qualidade das questões existentes no banco de questões.

Na terceira fase, foi possível implementar os testes para avaliação contínua sumativa em ambiente de sala de aula. Foram efetuadas entrevistas aos docentes e um questionário aos alunos com o objetivo principal de aferir mudanças no processo de ensino-aprendizagem. Foi assim possível implementar uma estratégia de avaliação contínua sumativa em Matemática no Ensino Superior utilizando *e-assessment* com testes contendo QEM. Verificou-se ainda que as classificações dos alunos tiveram uma evolução positiva ao longo de todo o processo. Estando conscientes que a estratégia implementada não foi o único fator responsável por esta melhoria, consideramos que ela teve um papel relevante dado que possibilitou a melhoria significativa da assiduidade dos alunos nas aulas. Assim sendo, este estudo pode ser visto como uma contribuição para a melhoria da "prestação das Instituições de Ensino Superior ao nível das suas taxas de insucesso e absentismo" (Flores, 2006, p. 10).

Consideramos que a principal contribuição desta tese é mostrar que é possível implementar avaliação contínua sumativa em Matemática no Ensino Superior recorrendo a *e-assessment* com testes contendo QEM. Acresce que esta implementação foi conseguida a custos reduzidos. Tanto quanto é do nosso conhecimento, a partir da pesquisa bibliográfica nas principais fontes de

referências científicas, não existem trabalhos de investigação que se debrucem sobre esta problemática da utilização de QEM na avaliação contínua sumativa em Matemática no Ensino Superior.

Outra contribuição importante é o conjunto de procedimentos que foram postos em prática durante todo o processo de implementação da estratégia de avaliação a que podemos chamar conjunto de boas práticas. Podemos afirmar que este conjunto de procedimentos pode ser seguido noutros contextos, desde que devidamente adaptados, dado que os resultados obtidos são bastante animadores, quer em termos da qualidade das questões do banco de questões, quer em termos do grau de satisfação dos docentes e dos alunos. Assim sendo, pode-se dizer que este processo poderá ser replicado e disseminado a outros contextos.

Esta tese contribui também como ajuda na compreensão de alguns aspetos importantes relacionados com a utilização de QEM na avaliação de alunos no Ensino Superior, devido ao facto de abordar aspetos diversificados no âmbito dessa temática.

No seguimento destas contribuições e no que diz respeito aos objetivos definidos no início deste trabalho, podemos afirmar que foram atingidos. Voltaremos a esta questão adiante, nomeadamente no que diz respeito às mudanças introduzidas.

Apesar do sucesso alcançado, foram identificadas algumas limitações. Uma das limitações tem a ver com o facto de não ter sido possível efetuar uma análise efetiva das QEM do banco de questões utilizando a TRI. Esta impossibilidade deve-se ao facto de o volume de dados não ser suficiente para o modelo convergir de forma conveniente, isto é, o número de respostas dadas pelos alunos a cada uma das questões é ainda muito reduzido. Assim sendo, pretende-se no futuro mudar o paradigma, que até ao momento era o de ter o maior número possível de questões no banco de questões, para um paradigma com o qual se pretende obter um maior número de respostas para as diversas questões, de modo a poderem ser analisadas com a TRI. De qualquer das formas, a análise efetuada com TCT já permitiu realizar uma boa análise das questões, a qual possibilitou tirar algumas conclusões quanto à qualidade das questões.

Outra limitação tem a ver com o facto de a análise não poder ser realizada diretamente no *Moodle*. Dado que os testes são gerados aleatoriamente pelo *Moodle*, a 1.ª questão no teste não é a mesma para todos os alunos, o mesmo se passando para todas as restantes questões, o que impossibilita a análise das questões diretamente no *Moodle*, dado que este efetua esta análise por teste. Neste trabalho a análise foi realizada no banco de questões, tendo havido a necessidade de extrair os dados necessários para aplicações externas. No futuro, pretende-se desenvolver trabalho no sentido de possibilitar a implementação da análise efetuada nesta tese diretamente no *Moodle*.

Uma limitação importante, não só a nível deste trabalho, mas globalmente na utilização de QEM na avaliação sumativa, tem a ver com a possibilidade de os alunos poderem acertar na resposta de forma aleatória, sem terem realmente os conhecimentos e as competências necessárias. As penalizações atribuídas pela seleção de uma resposta errada limitam esta

situação, mas na verdade não a erradicam, tal como é claro no questionário feito aos alunos. Assim sendo, mais trabalho tem de ser feito a este respeito, o qual poderá passar por definir estratégias adicionais para a avaliação, nomeadamente adaptar as questões existentes a novos tipos de QEM, tal como foi apresentado no capítulo 3, ou formas de penalização diferentes, ou outras estratégias consideradas pertinentes.

Outra limitação tem a ver com o facto de, apesar de as questões estarem alinhadas com os resultados de aprendizagem, estes não terem sido ainda classificados através de uma taxonomia adequada, como por exemplo a Taxonomia de Bloom. É importante que este trabalho seja realizado no futuro.

A metodologia Investigação–Ação foi, sem dúvida, adequada para a resolução do problema inicialmente identificado, que era a necessidade de implementar avaliação contínua sumativa numa UC de Matemática, numa instituição de Ensino Superior, com elevado número de alunos. Além de ter sido possível proceder à implementação de uma avaliação contínua sumativa, provocaram-se mudanças importantes ao nível da organização, tendo sido possível obter o apoio dos órgãos de gestão para a criação das condições necessárias à implementação de toda esta estratégia de avaliação, apesar de este apoio não ter sido imediato. Ao nível dos docentes, verificou-se que também houve mudanças, nomeadamente ao nível da capacidade de trabalho em equipa, da autoaprendizagem e da co–aprendizagem verificada ao longo dos anos de implementação do projeto. Estes aspetos foram reconhecidos pelos docentes nas entrevistas realizadas. Quanto aos alunos, verificou-se, através da análise realizada ao questionário, que se verificaram mudanças nomeadamente ao nível da assiduidade às aulas, da necessidade de estudo regular e de estarem mais atentos aos pormenores e da importância do acompanhamento regular pelos docentes.

Consideramos que o trabalho desenvolvido nesta tese, o qual resultou numa análise cuidada e criteriosa da utilização das QEM para avaliação contínua sumativa, é uma contribuição importante para uma melhor compreensão dos aspetos envolvidos. Consequentemente, representa uma contribuição importante para a melhoria de todo o processo e pode contribuir de forma eficaz para a credibilização desta estratégia de avaliação.

REFERÊNCIAS

Acosta-Gonzaga, E., & Walet, N. R. (2013). An investigation of the attitudes of instructors and students to on-line assessment in mathematical subjects. In *Proceedings of the 19th International Conference on Distributed Multimedia Systems* (pp. 112--117). Brighton - Seafront.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., … Wittrock, M. C. (2000). *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York Longman* (Vol. Complete e). New York: Pearson, Allyn & Bacon. http://doi.org/10.1207/s15430421tip4104_2

Ávila, C., & Torrubia, R. (2004). Personality, expectations, and response strategies in multiple-choice question examinations in university students: a test of Gray's hypotheses. *European Journal of Personality*, *18*(1), 45–59. http://doi.org/10.1002/per.506

Azevedo, J. (2015). E-assessment in mathematics courses with multiple-choice questions tests. In *Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015)* (pp. 260–266). Lisboa. http://doi.org/10.5220/0005452702600266

Azevedo, J., Torres, C., Lopes, A. P., & Babo, L. (2009). Enhancing math skills with moodle. In *Proceedings of ICERI 2009 - International Conference of Education, Research and Innovation* (pp. 2367–2377). Madrid.

Aziz, A., Salleh, T. S. A., Khatimin, N., & Zaharim, A. (2013). Evaluating multiple choice items in determining quality of test. In *TALE2013 - IEEE International COnference on Teaching, Assessment and Learning for Engineering* (pp. 565–569). http://doi.org/10.1109/TALE.2013.6654501

Babo, L., Azevedo, J., & Lopes, A. P. (2008). The active mathematics project at ISCAP. In *Proceedings of ENMA 2008 - International Conference on Engineering and Mathematics* (pp. 27–34). Bilbao.

Babo, L., Azevedo, J., Torres, C., & Lopes, A. P. (2010a). Moodle and multiple-choice tests. In *Proceedings of INTED 2010 - 4th International Technology, Education and Development Conference* (pp. 296–303). Valencia.

Babo, L., Azevedo, J., Torres, C., & Lopes, A. P. (2010b). New challenges in mathematics for the european higher education. In *Proceedings of ICERI 2010 - International Conference of Education, Research and Innovation* (pp. 4971–4980). Madrid.

Baker, F. (2001). *The basics of item response theory*. University of Maryland, College Park, MD: ERIC: Clearinghouse on Assessment and Evaluation.

Ball, G., Stephenson, B., Smith, G., Wood, L., Coupland, M., & Crawford, K. (1998). Creating a diversity of mathematical experiences for tertiary students. *International Journal of Mathematical Education in Science and Technology*, *29*(6), 827–841.

http://doi.org/10.1080/0020739980290605

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, *57*(1), 289–300. http://doi.org/10.2307/2346101

Bennie, K. (2013). The MATH taxonomy as a tool for analysing course material in mathematics: a study of its usefulness and its potential as a tool for curriculum development. *African Journal of Research in Mathematics, Science and Technology Education*, *9*(2), 81–95. http://doi.org/10.1080/10288457.2005.10740580

Bible, L., Simkin, M. G., & Kuechler, W. L. (2008). Using multiple-choice tests to evaluate students' understanding of accounting. *Accounting Education*, *17*(sup1), S55–S68. http://doi.org/10.1080/09639280802009249

Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: the SOLO taxonomy (structure of the observed learning outcome)*. New York: Academic Press.

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university: what the student does* (4th ed.). New York: McGraw Hill.

Blanco, M., & Ginovart, M. (2012). On how moodle quizzes can contribute to the formative e-assessment of first-year engineering students in mathematics courses. *RUSC Universities and Knowledge Society Journal*, *9*(1), 354–370. http://doi.org/10.7238/rusc.v9i1.1277

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, handbook I: the cognitive domain*. New York: David McKay Company, Inc.

Borba, M. de C., & Penteado, M. G. (2001). *Informática e educação matemática*. Belo Horizonte: Autêntica.

Boticki, I., & Milasinovic, B. (2008). Knowledge assessment at the faculty of electrical engineering and computing. In *Proceedings of the ITI 2008 - 30th Int. Conf. on Information Techonology Interfaces* (pp. 111–116). Cavtat. http://doi.org/10.1109/ITI.2008.4588392

Brito, E. B. C. (2012). *As implicações do processo de bolonha na formação de professores (Tese de Doutoramento)*. Universidade da Beira Interior. Retrieved from Ubi Thesis - Conhecimento Online (http://hdl.handle.net/10400.6/2593)

Brito, I., Figueiredo, J., Flores, M., Jesus, A., Machado, G., Malheiro, T., … Vaz, E. (2009). Using e-learning to self regulate the learning process of mathematics for engineering students. In N. Bulucea, CA and Mladenov, V and Pop, E and Leba, M and Mastorakis (Ed.), *Recent Advances in Applied Mathematics* (pp. 165–169). ATHENS: WORLD SCIENTIFIC AND ENGINEERING ACAD AND SOC.

Brown, G. (2001). *Assessment series n.º 3 - assessment: a guide for lectures*. York: Learning and Teaching Support Network (LTNS).

Brydon-Miller, M., Greenwood, D., & Maguire, P. (2003). Why action research?. *Action Research*, *1*(1), 9–28. http://doi.org/10.1177/14767503030011002

Bull, J., & Danson, M. (2001). *Assessment series N.º 14 - computer-assisted assessment (CAA)*. York: Learning and Teaching Support Network (LTNS).

Burns, A. (2007). Action research: contributions and future directions in elt. In J. Cummins & C. Davison (Eds.), *International Handbook of English Language Teaching*. Berlin, Heidelberg: Springer-Verlag.

Burrow, M., Evdorides, H., Hallam, B., & Freer-hewish, R. (2005). Developing formative assessment for postgraduate students in engineering. *European Journal of Engineering Education*, *30*(2), 255–263. http://doi.org/10.1080/03043790500087563

Burton, S., Sudweeks, R., Merrill, P., & Wood, B. (1991). *How to prepare better multiple-choice test items: guidelines for university faculty. Brigham Young University Testing Services and The Department of Instructional Science*. Retrieved from http://testing.byu.edu/info/handbooks/betteritems.pdf

Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education*, *40*(2), 218–231. http://doi.org/10.1080/02602938.2014.902192

Camilo, H., & Silva, J. A. P. da. (2008). *Os testes de escolha múltipla (TEM). Essências EDUcare*. Departamento de Educação Médica da Faculdade de Medicina - Universidade de Coimbra.

Capobianco, B. M., & Ní Ríordáin, M. (2015). Navigating layers of teacher uncertainty among preservice science and mathematics teachers engaged in action research. *Educational Action Research*, *23*(4), 581–598. http://doi.org/10.1080/09650792.2015.1045537

Clarke, P. A. J., & Fournillier, J. B. (2012). Action research, pedagogy, and activity theory: Tools facilitating two instructors' interpretations of the professional development of four preservice teachers. *Teaching and Teacher Education*, *28*(5), 649–660. http://doi.org/10.1016/j.tate.2012.01.013

Clegg, V. L., & Cashin, W. E. (1986). *Improving multiple-choice tests*. Kansas State University: Center for Faculty Evaluation & Development.

Coghlan, D., & Brydon-Miller, M. (2014). *The SAGE encyclopedia of action research* (Vol. 1–2). London: SAGE Publications. http://doi.org/10.4135/9781446294406

Cook, J., & Jenkins, V. (2010). Getting started with e-assessment. Retrieved from http://opus.bath.ac.uk/17712/1/Getting_started_with_e-assessment_14Jan2010.pdf

Cramer, D., & Howitt, D. (2004). *The SAGE dictionary of statistics. Statistics*. London: SAGE Publications. http://doi.org/10.4135/9780857020123

Curtis, D. A., Lind, S. L., Boscardin, C. K., & Dellinges, M. (2013). Does student confidence on multiple-choice question assessments provide useful information?. *Medical Education*,

47(6), 578–584. http://doi.org/10.1111/medu.12147

Darlington, E. (2014). Contrasts in mathematical challenges in A-level mathematics and further mathematics, and undergraduate mathematics examinations. *Teaching Mathematics and Its Applications*, *33*(4), 213–229. http://doi.org/10.1093/teamat/hru021

Dascalu, M., & Bodea, C. (2010). Challenges in building e-assessment services from project management knowledge perspective. *International Journal of Global Management Studies Professional*, *2*(1), 35–50.

Dick, B., Stringer, E., & Huxham, C. (2009). Theory in action research. *Action Research*, *7*(1), 5–12. http://doi.org/10.1177/1476750308099594

Douglas, M., Wilson, J., & Ennis, S. (2012). Multiple-choice question tests: a convenient, flexible and effective learning tool? A case study. *Innovations in Education and Teaching International*, *49*(2), 111–121. http://doi.org/10.1080/14703297.2012.677596

Elliott, J. (2007). Assessing the quality of action research. *Research Papers in Education*, *22*(2), 229–246. http://doi.org/10.1080/02671520701296205

Ferrão, M. (2010). E-assessment within the bologna paradigm: evidence from Portugal. *Assessment & Evaluation in Higher Education*, *35*(7), 819–830. http://doi.org/10.1080/02602930903060990

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London: SAGE Publications.

Flores, M. A. (Coord). (2006). *Perspectivas e estratégias de formação de docentes do ensino superior - um estudo na universidade do minho (Relatório de investigação)*. Braga. Retrieved from http://www.gaqe.uminho.pt/uploads/relatório Final Dez 2006.pdf

Flores, M. A., Simão, A. M. V., Barros, A., & Pereira, D. (2015). Perceptions of effectiveness, fairness and feedback of assessment methods: a study in higher education. *Studies in Higher Education*, *40*(9), 1523–1534. article. http://doi.org/10.1080/03075079.2014.881348

Frankland, S. (2007a). *Enhancing teaching and learning through assessment*. Dordrecth: Springer.

Frankland, S. (2007b). Peer assessment among students in a problem-based learning format. In S. Frankland (Ed.), *Enhancing Teaching and Learning Through Assessment* (pp. 144–155). Dordrecth: Springer.

Frankland, S. (2007c). Perspectives of teachers and students towards assessment. In S. Frankland (Ed.), *Enhancing Teaching and Learning Through Assessment* (pp. 64–76). Dordrecth: Springer.

Garfield, J. B., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: connecting research and teaching practice*. Dordrecht: Springer. http://doi.org/10.1007/978-1-4020-

8383-9

Gelade, S., & Fursenko, F. (2007). Can intrinsic graduate qualities be developed through assessment? Mapping assessment practices in it degree programs. In S. Frankland (Ed.), *Assessment series: Enhancing Teaching and Learning Trough Assessment* (pp. 476–487). Dordrecth: Springer.

Given, L. M. (2008). *The SAGE encyclopedia of qualitative research methods*. Thousand Oaks, California: SAGE Publications. http://doi.org/10.4135/9781412963909

Green, A., & Mitchell, C. (2009). E-assessment: opportunities and challenges for the sports marketing and educator. In *Proceedings of the 2nd International Conference of Teaching and Learning (ICTL 2009)* (pp. 1–9). Kuching.

Gruttmann, S., Böhm, D., & Kuchen, H. (2008). E-assessment of mathematical proofs: chances and challenges for students and tutors. In *2008 International Conference on Computer Science and Software Engineering (CSSE 2008)* (pp. 612–615). http://doi.org/10.1109/CSSE.2008.95

Guimarães, R. C., & Cabral, J. S. (2007). *Estatística* (2.ª). Lisboa, Portugal: McGraw Hill.

Guo, R., Palmer-Brown, D., Lee, S. W., & Cai, F. F. (2014). Intelligent diagnostic feedback for online multiple-choice questions. *Artificial Intelligence Review*, *42*(3), 369–383. http://doi.org/10.1007/s10462-013-9419-6

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items - third edition* (3rd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates. http://doi.org/10.1177/0146621605280143

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333. http://doi.org/10.1207/S15324818AME1503_5

Hall, R. J., Jung, E., & Pilant, M. S. (2012). Comprehensive statistical analysis of a mathematics placment test. In *Proceeding of SITE 2012 - International Conference of the Society for Information Technology & Teacher Education* (pp. 4432–4439). Austin, Texas.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement Issues and Practice*, *12*(3), 39–47. http://doi.org/10.1097/01.mlr.0000245426.10853.30

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item resposnse theory*. Newburry Park, California: Sage Publications.

Harris, H. L., Walsh, L., Tayyaba, S., Harris, A., Wilson, J., & Smith, E. (2015). A novel student-led approach to multiple-choice question generation and online database creation, with targeted clinician input. *Teaching & Learning in Medicine*, *27*(2), 182–189. http://doi.org/10.1080/10401334.2015.1011651

Hauk, S., Powers, R. A., & Segalla, A. (2015). A comparison of web-based and paper-and-pencil homework on student performance in college algebra. *PRIMUS: Problems, Resouces, and Issues in Mathematics Undergraduate Studies*, *25*(1), 61–79. http://doi.org/10.1080/10511970.2014.906006

Heller, F. (2004). Action research and research action: a family of methods. In C. Cassell & G. Symon (Eds.), *Essential Guide to Qualitative Methods in Organizational Research* (pp. 349–360). Thousand Oaks, California: SAGE Publications. http://doi.org/10.4135/9781446280119.n28

Helskog, G. H. (2014). Justifying action research. *Educational Action Research*, *22*(1), 4–20. http://doi.org/10.1080/09650792.2013.856769

Hernández, R. (2007). The impact of innovative assessment practices on students' learning. In S. Frankland (Ed.), *Assessment series: Enhancing Teaching and Learning Trough Assessment* (pp. 266–278). Dordrecth: Springer.

Heron, G., & Lerpiniere, J. (2013). Re-engineering the multiple choice question exam for social work. *European Journal of Social Work*, *16*(4), 521–535. http://doi.org/10.1080/13691457.2012.691873

Herr, K., & Anderson, G. L. (2005). *The action research dissertation: A guide for students and faculty*. Thousand Oaks, California: SAGE Publications. http://doi.org/10.4135/9781452226644

Holmes, N. (2015). Student perceptions of their learning engagement in response to the use of a continuous e-assement in a undergratuate module. *Assessment & Evaluation in Higher Education*, *40*(1), 1–14. http://doi.org/10.1080/02602938.2014.881978

Hughes, I. (2008). Action research in healthcare. In P. Reason & H. Bradbury (Eds.), *The Sage Handbook of Action Research Participative Inquiry and Practice* (pp. 381–393). Thousand Oaks, California: Sage Publications. http://doi.org/10.4135/9781446288696

Huntley, B., Engelbrecht, J., & Harding, A. (2009). Can multiple choice questions be successfully used as an assessment format in undergraduate mathematics?. *Pythagoras*, *0*(69), 3–16. http://doi.org/10.4102/pythagoras.v0i69.41

Ialongo, C. (2016). Lessons in biostatistics understanding the effect size and its measures. *Biochemia Medica*, *26*(2), 150–163. http://doi.org/10.11613/BM.2016.015

Imrie, B. W. (1995). Assessment for learning: quality and taxonomies. *Assessment & Evaluation in Higher Education*, *20*(2), 175–189. http://doi.org/10.1080/02602939508565719

Ivankova, N. V. (2015). *Mixed methods applications in action research*. Thousand Oaks, California: SAGE Publications.

Jacob, S. M., Issac, B., & Sebastian, Y. (2006). Impact on student learning from traditional continuous assessment and an e-assessment proposal. In *Proceedings of the PACIS 2006 - The 10th Pacific Asia Conference on Information Systems* (pp. 1482–1496). Kuala Lumpur.

Jarvis, P., Holford, J., & Griffin, C. (2003). *Theory & practice of learning - 2nd edition* (2nd ed.). New York: Routledge Falmer.

JISC. (2006). E-assessment glossary (extended). Retrieved September 15, 2014, from http://www.jisc.ac.uk/media/documents/themes/elearning/eassess_glossary_extended v101.pdf

JISC. (2007). Effective practice with e-assessment: an overview of technologies, policies and practice in further and higher education. Retrieved September 15, 2014, from http://www.jisc.ac.uk/media/documents/themes/elearning/effpraceassess.pdf

Jordan, S. (2013). E-assessment: past, present and future. *New Directions*, *9*(1), 87–106.

Khiat, H., Chia, H. T., Tan-Yeoh, A. C., & Kok-Mak, C. P. (2011). The perspectives of lecturers on the action research journey in the mathematics and science department of singapore polytechnic. *Educational Research for Policy and Practice*, *10*(1), 29–52. http://doi.org/10.1007/s10671-010-9092-3

Kim, M. K., Patel, R. A., Uchizono, J. A., & Beck, L. (2012). Incorporation of Bloom's taxonomy into multiple-choice examination questions for a pharmacotherapeutics course. *American Journal of Pharmaceutical Education*, *76*(6), 114. http://doi.org/10.5688/ajpe766114

Kitchen, J., & Stevens, D. (2008). Action research in teacher education two teacher-educators practice action research as they introduce action research to preservice teachers. *Action Research*, *6*(1), 7–28. http://doi.org/10.1177/1476750307083716

Knight, P. (2001). *Assessment series n.º 7 - a briefing on key concepts: formative and summative, criterion and norm-referenced assessment*. York: Learning and Teaching Support Network (LTNS).

Kotrlik, J. W., & Williams, H. A. (2003). The incorporation of effect size in information technology, learning, and performance research. *Information Technology, Learning, and Performance Journal*, *21*(1), 1–7. http://doi.org/10.1.1.331.4489

Langlois, L., Lapointe, C., Valois, P., & de Leeuw, A. (2014). Development and validity of the ethical leadership questionnaire. *Journal of Educational Administration*, *52*(3), 310–331. http://doi.org/10.1108/JEA-10-2012-0110

Larkin, K., Jamieson-Proctor, R., & Finger, G. (2012). TPACK and pre-service teacher mathematics education: defining a signature pedagogy for mathematics education using ICT and based on the metaphor "mathematics is a language." *Computers in the Schools*, *29*(1–2), 207–226. http://doi.org/10.1080/07380569.2012.651424

Lee, H.-S., Liu, L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, *24*(2), 115–136. http://doi.org/10.1080/08957347.2011.554604

Lei n.º 49/2005, de 30 de agosto. Segunda alteração à lei de bases do sistema educativo e primeira alteração à lei de bases do financiamento do snsino superior, Diário da República:

I Série - A 5122 (2005). Retrieved from https://dre.pt/application/file/245260

Leung, C. F. (2000). Assessment for learning: using the SOLO taxonomy to measure design performance of design & technology students. *International Journal of Technology and Design Education*, *10*(2), 149–161. http://doi.org/10.1023/A:1008937007674

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*(4), 612–625. http://doi.org/10.1093/hcr/28.4.612

Lewin, K. (1946). Action research and minority problems. *Journal of Social Issues*, *2*(4), 34–46. http://doi.org/10.1111/j.1540-4560.1946.tb02295.x

Liu, O. L., Lee, H.-S., & Linn, M. C. (2011). An investigation of explanation multiple-choice items in science assessment. *Educational Assessment*, *16*(3), 164–184. http://doi.org/10.1080/10627197.2011.611702

Llamas-Nistal, M., Fernández-Iglesias, M. J., González-Tato, J., & Mikic-Fonte, F. A. (2013). Blended e-assessment: migrating classical exams to the digital world. *Computers & Education*, *62*(1), 72–87. http://doi.org/10.1016/j.compedu.2012.10.021

Lopes, A. P., Babo, L., & Azevedo, J. (2008). Teaching and learning mathematics using moodle. In *Proceedings of INTED 2008 - 2nd International Technology, Education and Development Conference* (p. NA). Valencia.

Lopes, A. P., Babo, L., Azevedo, J., & Torres, C. (2010). Multiple-choice tests - a tool in assessing knowledge. In *Proceedings of INTED 2010 - 4th International Technology, Education and Development Conference* (pp. 256–265). Valencia.

Lopes, A. P., Babo, L., Azevedo, J., & Torres, C. (2011). Innovating mathematics in the european higher education. In *Proceedings of INTED 2011 - 5th International Technology, Education and Development Conference* (pp. 1215–1222). Valencia.

Maroco, J., & Garcia-Marques, T. (2006). Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas?. *Laboratório de Psicologia*, *4*(1), 65–90.

Mathai, E., & Olsen, D. (2013). Studying the effectiveness of online homework for different skill levels in a college algebra course. *PRIMUS: Problems, Resouces, and Issues in Mathematics Undergraduate Studies*, *23*(8), 671–682. http://doi.org/10.1080/10511970.2013.782479

Matos, R., Torrão, S., & Vieira, T. (2012). Moodlewatcher: detection and prevention of fraud when using moodle quizzes. In *INTED 2012* (pp. 4997–5001). Valencia.

McAlpine, M. (2002a). *A summary of methods of item analysis*. Leicestershire: The CAA Centre TLTP Project.

McAlpine, M. (2002b). *Design requirements of a databank*. Leicestershire: The CAA Centre TLTP Project.

McAlpine, M. (2002c). *Principles of assessment*. Bedfordshire: The CAA Centre TLTP Project.

McGuire, G. R., Youngson, M. A., Korabinski, A. A., & McMillan, D. (2002). Partial credit in mathematics exams - a comparison of traditional and CAA exams. In *Proceedings of the 6th CAA Conference* (pp. 223–230). Loughborough: Loughborough University.

Melo, A. L. P. do S. (2012). *O impacto do processo de bolonha na formação de professores de educação visual e tecnológica (Tese de Doutoramento)*. Universidade da Beira Interior. Retrieved from Ubi Thesis - Conhecimento Online (http://hdl.handle.net/10400.6/2592

Mora, M. C., Sancho-Bru, J. L., Iserte, J. L., & Sánchez, F. T. (2012). An e-assessment approach for evaluation in engineering overcrowded groups. *Computers & Education*, *59*(2), 732–740.

Moreno, G. A. (2015). Making meaning about educational experiences through participatory action research: a project conducted with adults enrolled in a community college mathematics course. *Educational Action Research*, *23*(2), 178–193. http://doi.org/10.1080/09650792.2014.980285

Moscinska, K., & Rutkowski, J. (2012). Rethinking e-assessment in a core engineering course. In *Global Engineering Education Conference (EDUCON)* (pp. 1–4). 2012 IEEE. http://doi.org/10.1109/EDUCON.2012.6201136

Mostofo, J., & Zambo, R. (2015). Improving instruction in the mathematics methods classroom through action research. *Educational Action Research*, *23*(4), 497–513. http://doi.org/10.1080/09650792.2015.1019903

Munzenmaier, C., & Rubin, N. (2013). Bloom's taxonomy: what's old is new again. *Perspectives*, 1–47. Retrieved from http://www.elearningguild.com/insights/index.cfm?id=164&action=viewonly&utm_campaign=research-blm13&utm_medium=email&utm_source=elg-insider

Neilsen, E. H. (2006). But let us not forget John Collier commentary on David Bargal's "Personal and intellectual influences leading to Lewin"s paradigm on action research'. *Action Research*, *4*(4), 389–399. http://doi.org/10.1177/1476750306070102

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education*, *31*(1), 53–64. http://doi.org/10.1080/03098770601167922

O'Toole, G. (2007). Can assessment of student attitudes assist both the teaching and learning process as well as ultimate performance in professional practice. In S. Frankland (Ed.), *Enhancing Teaching and Learning through Assessment: Deriving an Appropriate Model* (pp. 468–474). Dordrecth: Springer.

Oldham, J., Freeman, A., Chamberlain, S., & Ricketts, C. (2007). Formative assessment for progress tests of applied medical knowledge. In S. Frankland (Ed.), *Enhancing Teaching and Learning through Assessment: Deriving an Appropriate Model* (pp. 32–39). Dordrecht: Springer.

Pereira, D. R., & Flores, M. A. (2012). Percepções dos estudantes universitários sobre a

avaliação das aprendizagens: um estudo exploratório. *Avaliação (Campinas)*, *17*(2), 529–556. http://doi.org/10.1590/S1414-40772012000200012

Poitras, S.-C., Guay, F., & Ratelle, C. F. (2012). Using the self-directed search in research: selecting a representative pool of items to measure vocational interests. *Journal of Career Development*, *39*(2), 186–207. http://doi.org/10.1177/0894845310384593

Ponte, J. P. (2002). Investigar a nossa própria prática. In G.-G. de T. de Matemática (Ed.), *Reflectir e investigar sobre a prática profissional* (pp. 5–28). Lisboa: Associação de Professores de Matemática.

Ponte, J. P. (2008). Investigar a nossa própria prática: uma estratégia de formação e de construção do conhecimento profissional. *PNA*, *2*(4), 153–180.

Race, P. (2001). *Assessment series no. 9 a briefing on self, peer and group assessment*. York: Learning and Teaching Support Network (LTNS). Retrieved from http://phil-race.co.uk/wp-content/uploads/Self,_peer_and_group_assessment.pdf

Reason, P., & Bradbury, H. (2008). *The SAGE handbook of action research: participative inquiry and practice - 2nd edition*. Thousand Oaks, California: SAGE Publications.

Redecker, C. (2013). *The use of ICT for the assessment of key competences*. Luxembourg: European Union.

Redecker, C., & Johannessen, Ø. (2013). Changing assessment - towards a new assessment paradigm using ICT. *European Journal of Education*, *48*(1), 79–96. http://doi.org/10.1111/ejed.12018

Reese, S. (2015). "Knowing is not enough; we must apply": reflections on a failed action learning application. *Action Learning: Research and Practice*, *12*(1), 78–84. http://doi.org/10.1080/14767333.2015.1006912

Resolução do conselho de ministros n.º 137/2007. (2007). *Diário Da República, I série*(N.º 180 de 18 de Setembro de 2007), 6563 a 6577. Retrieved from https://dre.pt/application/file/642102

Resolução do conselho de ministros n.º 51/2008. (2008). *Diário Da República, I série*(N.º 56 de 19 de Março de 2008), 1619 e 1620. Retrieved from https://dre.pt/application/file/246469

Rice, M., & Campbell, C. (2007). Using online environments to promote assessment as a learning enhancement process. In S. Frankland (Ed.), *Enhancing Teaching and Learning through Assessment: Deriving an Appropriate Model* (pp. 418–430). Dordrecht: Springer. http://doi.org/10.1007/978-1-4020-6226-1_9

Rod, J. K., Eiksund, S., & Fjaer, O. (2010). Assessment based on exercise work and multiple-choice tests. *Journal of Geography in Higher Education*, *34*(1), 141–153. http://doi.org/10.1080/03098260903062039

Rust, C. (2001). *Assessment Series n.º 12 - a briefing on assessment of large groups*. York: Learning and Teaching Support Network (LTNS).

Salas-Morera, L., Cubero-Atienza, A. J., Redel-Macías, M. D., Arauzo-Azofra, A., & García-Hernández, L. (2012). Effective use of e-learning for improving students' skills. In R. Babo & A. Azevedo (Eds.), *Higher Education Institutions and Learning Management Systems* (pp. 292–314). Hershey, PA: IGI Global. http://doi.org/10.4018/978-1-60960-884-2.ch014

Salleh, H. (2006). Action research in Singapore education: constraints and sustainability. *Educational Action Research*, *14*(4), 513–523. http://doi.org/10.1080/09650790600975684

Scouller, K. (1998). The influence of assessment method on students' learning approaches: multiple choice question examinations versus assignment essay. *Higher Education*, *35*(4), 453–472. http://doi.org/10.1023/A:1003196224280

Serrazina, L., & Oliveira, I. (2002). O professor como investigador: Leitura crítica de investigações em educação matemática. In Grupo de Trabalho sobre Investigação (Ed.), *Reflectir e Investigar sobre a Prática Profissional* (pp. 283–308). Lisboa: Associação de Professores de Matemática.

Smith, G., & Wood, L. (2000). Assessment of learning in university mathematics. *International Journal of Mathematical Education in Science and Technology*, *31*(1), 125–132. http://doi.org/10.1080/002073900287444

Smith, G., Wood, L., Coupland, M., Stephenson, B., Crawford, K., & Ball, G. (1996). Constructing mathematical examinations to access a range of knowledge and skills. *International Journal of Mathematical Education in Science and Technology*, *27*(1), 65–77. http://doi.org/10.1080/0020739960270109

Soares, A. P., Guisande, M. A., & Almeida, L. S. (2007). Autonomia y ajustamiento académico: un estudio con estudiantes portugueses de primer año. *International Journal of Clinical and Health Psychology*, *7*(3), 753–765.

Sommer, R. (2009). Dissemination in action research. *Action Research*, *7*(2), 227–236. http://doi.org/10.1177/1476750308097028

Sorensen, E. (2013). Implementation and student perceptions of e-assessment in a chemical engineering module. *European Journal of Engineering Education*, *38*(2), 172–185. http://doi.org/10.1080/03043797.2012.760533

Sousa, I. (2011). *Processo de bolonha e mudanças na educação superior: um estudo no ensino superior politécnico português (Tese de Doutoramento)*. Universidade Lusófona de Humanidades e Tecnologias. Retrieved from RECIPP - Repositório Científico do Instituto Politécnico do Porto (http://hdl.handle.net/10400.22/4608)

Sousa, M. J., & Baptista, C. S. (2011). *Como fazer investigação, dissertações, teses e relatórios segundo bolonha*. Lisboa: Pactor.

Stödberg, U. (2012). A research review of e-assessment. *Assessment & Evaluation in Higher Education*, *37*(5), 591–604. http://doi.org/10.1080/02602938.2011.557496

Stoline, M. R. (1981). The status of multiple comparisons: simultaneous estimation of all pairwise comparisons in one-way ANOVA designs. *The American Statistician*, *35*(3), 134–141. http://doi.org/10.2307/2683979

Torres, C., Lopes, A. P., Babo, L., & Azevedo, J. (2009). Developing multiple-choice questions in mathematics. In *Proceedings of ICERI 2009 - International Conference of Education, Research and Innovation* (pp. 6218–6229). Madrid.

Torres, C., Lopes, A. P., Babo, L., & Azevedo, J. (2011). Improving multiple-choice questions. *US-China Education Review*, *B*(1), 1–11. http://doi.org/10.1212/01.CON.0000394686.28362.cc

Triantis, D., & Ventouras, E. (2012). Enhancing electronic examinations through advanced multiple-choice questionnaires. In *Higher Education Institutions and Learning Management Systems: Adoption and Standardization* (pp. 178–198). http://doi.org/10.4018/978-1-60960-884-2.ch009

Valois, P., Houssemand, C., Germain, S., & Abdous, B. (2011). An open source tool to verify the psychometric properties of an evaluation instrument. *Procedia - Social and Behavioral Sciences*, *15*, 552–556. http://doi.org/10.1016/j.sbspro.2011.03.140

Vora, S. S., & Shinde, S. A. (2014). A service oriented approach for an e-assessment system. *International Journal of Engineering Research & Technology*, *3*(5), 1468–1474.

Watters, A. (2015). Multiple choice and testing machines: a history. Retrieved September 30, 2015, from http://hackeducation.com/2015/01/27/multiple-choice-testing-machines

Wild, C., Triggs, C., & Pfannkuch, M. (1997). Assessment on a budget: using traditional methods imaginatively. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 205–220). Amsterdam: IOS Press.

Wilson, K., Boyd, C., Chen, L., & Jamal, S. (2011). Improving student performance in a first-year geography course: examining the importance of computer-assisted formative assessment. *Computers and Education*, *57*(2), 1493–1500. http://doi.org/10.1016/j.compedu.2011.02.011

Wong, C. (2007). Views on the adoption and implementation of the SOLO taxonomy. In S. Frankland (Ed.), *Enhancing Teaching and Learning through Assessment: Deriving an Appropriate Model* (pp. 4–15). Dordrecht: Springer.

Yin, R. K. (2011). *Qualitative research from start to finish*. *Qualitative research from start to finish*. New York: The Guilford Press. http://doi.org/10.2307/41305080

Yonker, J. E. (2011). The relationship of deep and surface study approaches on factual and applied test-bank multiple-choice question performance. *Assessment & Evaluation in Higher Education*, *36*(6), 673–686. http://doi.org/10.1080/02602938.2010.481041

Yorke, M. (2001). *Assessment series n.º 1 - assessment: a guide for senior managers*. York: Learning and Teaching Support Network (LTNS).

Zaiontz, C. (2015). Real statistics using MS Excel™. Retrieved March 1, 2015, from www.real-statistics.com

Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. L. R. J. Vandenberg (Ed.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 37–59). CHAP, New York, NY: Routledge/Taylor & Francis Group.

ANEXO A – QUESTIONÁRIO AOS DOCENTES NO 1.º CICLO DE IA

Linhas de Orientação para a escrita de Questões de Escolha Múltipla

Na bibliografia podem ser encontradas várias **linhas de orientação** para a **escrita de questões de escolha múltipla**.

Este questionário pretende saber a sua **opinião sobre a importância de cada uma destas** linhas de orientação no **contexto específico da Matemática**. Assim sendo, **tendo em conta a sua experiência** na escrita de questões de escolha múltipla na área da Matemática, responda, por favor, às questões seguintes.

Obrigado pela sua colaboração!

Existem 7 perguntas neste inquérito

Opinião sobre as linhas de orientação para a escrita de perguntas de E.M.

1 [1.]

CUIDADOS COM O CONTEÚDO

Indique o seu grau de concordância com cada uma das seguintes linhas de orientação, tendo em conta o contexto específico da Matemática.

*

Por favor, seleccione uma resposta apropriada para cada item:

	1 - Discordo Total-mente	2 - Discordo	3 - Nem concordo nem discordo	4 - Concordo	5 - Concordo Total-mente
Cada questão deve reflectir conteúdo específico e um único comportamento mental concreto, tal como preconizado nas especificações dos testes.	○	○	○	○	○
Fundamentar cada questão em termos de conteúdos de aprendizagem importantes; evitar conteúdo trivial.	○	○	○	○	○
Utilizar materiais inovadores para testar aprendizagens de nível mais elevado. Reescrever a linguagem utilizada no livro de apoio ou a linguagem utilizada durante as aulas, quando incluídas nas questões de um teste, de modo a evitar testes apenas de memorização.	○	○	○	○	○
Manter o conteúdo de cada questão independente do conteúdo de outras questões do teste.	○	○	○	○	○
Evitar conteúdos demasiado específicos ou demasiado genéricos ao escrever as questões.	○	○	○	○	○
Evitar questões baseadas em opiniões.	○	○	○	○	○
Evitar questões com artimanhas.	○	○	○	○	○
Manter o vocabulário simples, tendo em conta o grupo de alunos que está a ser testado.	○	○	○	○	○

2 [2.]

CUIDADOS COM A FORMATAÇÃO

Indique o seu grau de concordância com cada uma das seguintes linhas de orientação, tendo em conta o contexto específico da Matemática.

*

Por favor, seleccione uma resposta apropriada para cada item:

	1 - Discordo Total-mente	2 - Discordo	3 - Nem concordo nem discordo	4 - Concordo	5 - Concordo Total-mente
Formatar a questão verticalmente e não horizontalmente.	○	○	○	○	○

3 [3.]

CUIDADOS COM O ESTILO

Indique o seu grau de concordância com cada uma das seguintes linhas de orientação, tendo em conta o contexto específico da Matemática.

*

Por favor, seleccione uma resposta apropriada para cada item:

	1 - Discordo Total-mente	2 - Discordo	3 - Nem concordo nem discordo	4 - Concordo	5 - Concordo Total-mente
Editar e rever as questões.	○	○	○	○	○
Usar correctamente a gramática, a pontuação, as letras maiúsculas e a ortografia.	○	○	○	○	○
Minimizar a quantidade de leitura necessária em cada questão.	○	○	○	○	○

4 [4.]

ESCREVENDO O ENUNCIADO DA QUESTÃO

Indique o seu grau de concordância com cada uma das seguintes linhas de orientação, tendo em conta o contexto específico da Matemática.

*

Por favor, seleccione uma resposta apropriada para cada item:

	1 - Discordo Total-mente	2 - Discordo	3 - Nem concordo nem discordo	4 - Concordo	5 - Concordo Total-mente
Certificar-se que as instruções no enunciado são muito claras.	○	○	○	○	○
Incluir a ideia central no enunciado ao invés de nas opções.	○	○	○	○	○
Evitar palavreado excessivo.	○	○	○	○	○
Escrever o enunciado na forma afirmativa, evitando negações tais como NÃO ou EXCEPTO. Se forem utilizadas negações, usar as palavras com cautela e garantir sempre que a palavra aparece em maiúsculas e em negrito.	○	○	○	○	○

5 [5.]

ESCREVENDO AS OPÇÕES DA QUESTÃO

Indique o seu grau de concordância com cada uma das seguintes linhas de orientação, tendo em conta o contexto específico da Matemática.

*

Por favor, seleccione uma resposta apropriada para cada item:

	1 - Discordo Total-mente	2 - Discordo	3 - Nem concordo nem discordo	4 - Concordo	5 - Concordo Total-mente
Desenvolver tantas opções eficazes quantas seja possível, mas a investigação sugere que três é adequado.	○	○	○	○	○
Certificar-se que apenas uma dessas opções é a resposta correcta.	○	○	○	○	○
Variar a localização da resposta correta de acordo com o número de opções.	○	○	○	○	○
Colocar as opções por ordem, lógica ou numérica.	○	○	○	○	○
Garantir opções independentes; as opções não devem ter elementos comuns.	○	○	○	○	○
Garantir opções homogéneas, quer em termos de conteúdo quer em termos de estrutura gramatical.	○	○	○	○	○
Manter o tamanho das opções aproximadamente igual.	○	○	○	○	○
Utilizar cuidadosamente "Nenhum dos anteriores".	○	○	○	○	○
Evitar utilizar "Todos os anteriores".	○	○	○	○	○
Escrever as opções na forma afirmativa; evitar negações tais como NÃO.	○	○	○	○	○

	1 - Discordo Total-mente	2 - Discordo	3 - Nem concordo nem discordo	4 - Concordo	5 - Concordo Total-mente
Evitar dar dicas para a resposta correcta, tais como: **a) Determinantes específicos incluindo sempre, nunca, completamente e absolutamente;** **b) Associações de palavras com sons idênticos, escolhas idênticas ou parecidas com termos utilizados no enunciado;** **c) Incoerências gramaticais que dêem pistas ao aluno sobre a resposta correta.** **d) Resposta correta evidente;** **e) Pares ou tripletos de opções que irão indicar ao aluno a resposta correta;** **f) Opções ostensivamente absurdas ou ridículas.**	○	○	○	○	○
Garantir que todos os distratores são plausíveis.	○	○	○	○	○
Usar erros típicos dos alunos para escrever os distratores.	○	○	○	○	○
Utilizar humor, se ele é compatível com o professor e com o ambiente de aprendizagem.	○	○	○	○	○

6 [6.]Caso considere que alguma(s) linha(s) de orientação não está(ão) incluída(s), por favor, indique-a(s):

Por favor, escreva aqui a sua resposta:

7 [7.]

Comentários adicionais:

Por favor, escreva aqui a sua resposta:

Submeter o seu inquérito
Obrigado por ter concluído este inquérito.

ANEXO B – QUESTIONÁRIO AOS ALUNOS NO 3.º CICLO DE IA

Matemática - Testes de Escolha Múltipla - Questionário

Este questionário destina-se a avaliar a opinião dos alunos sobre os testes de escolha múltipla realizados no Moodle nas Unidades Curriculares de Matemática e Matemática I.

A sua opinião é muito importante, pelo que pedimos que responda com toda a sinceridade.

O questionário é anónimo.

Existem 24 perguntas neste inquérito

Grupo I - Caraterização

1 [1.1]Género: *

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ Feminino
- ○ Masculino

2 [1.2]Idade: *

Por favor, escreva aqui a sua resposta:

3 [1.3]Aluno: *

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ Diurno
- ○ Noturno

4 [1.4]Trabalhador-Estudante? *

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ Sim
- ○ Não

5 [1.5]Qual a Unidade Curricular? *

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ Matemática
- ○ Matemática I

6 [1.6]Está a frequentar esta Unidade Curricular pela primeira vez? *

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ Sim
- ○ Não

7 [1.6.1]Qual o 1º ano letivo em que frequentou a U.C.? *

Responda a esta pergunta apenas se as seguintes condições são verdadeiras:
° Resposta era N'Não' na pergunta '6 [1.6]' (Está a frequentar esta Unidade Curricular pela primeira vez?)

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ 2013/14
- ○ 2012/13
- ○ 2011/12
- ○ Anterior

Grupo II - Opinião sobre os testes

8 [2.1]Considera que os testes de escolha múltipla realizados no Moodle, na Unidade Curricular são justos? *

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ Sim
- ○ Não

9 [2.1.1]Porque não os considera justos? *

Responda a esta pergunta apenas se as seguintes condições são verdadeiras:
° Resposta era N'Não' na pergunta '8 [2.1]' (Considera que os testes de escolha múltipla realizados no Moodle, na Unidade Curricular são justos?)

Por favor, escreva aqui a sua resposta:

10 [2.1.2]Porque os considera justos? *

Responda a esta pergunta apenas se as seguintes condições são verdadeiras:
° Resposta era Y'Sim' na pergunta '8 [2.1]' (Considera que os testes de escolha múltipla realizados no Moodle, na Unidade Curricular são justos?)

Por favor, escreva aqui a sua resposta:

11 [2.2]Considera que se estes testes (escolha múltipla) fossem realizados em papel em vez de serem realizados no Moodle, seriam: *

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ Melhor
- ○ Igual
- ○ Pior

12 [2.2.1]Melhor, em que aspectos? *

Responda a esta pergunta apenas se as seguintes condições são verdadeiras:
° Resposta era 1'Melhor' na pergunta '11 [2.2]' (Considera que se estes testes (escolha múltipla) fossem realizados em papel em vez de serem realizados no Moodle, seriam:)

Por favor, escreva aqui a sua resposta:

13 [2.2.2]Pior, em que aspectos? *

Responda a esta pergunta apenas se as seguintes condições são verdadeiras:
° Resposta era 3'Pior' na pergunta '11 [2.2]' (Considera que se estes testes (escolha múltipla) fossem realizados em papel em vez de serem realizados no Moodle, seriam:)

Por favor, escreva aqui a sua resposta:

14 [2.3]O facto de os testes serem de escolha múltipla alterou de alguma forma o modo como estudou? *

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ Sim
- ○ Não

15 [2.3.1]Alterou em que aspectos? *

Responda a esta pergunta apenas se as seguintes condições são verdadeiras:
° Resposta era Y'Sim' na pergunta '14 [2.3]' (O facto de os testes serem de escolha múltipla alterou de alguma forma o modo como estudou?)

Por favor, escreva aqui a sua resposta:

16 [2.4]A sua presença regular nas aulas depende do regime Avaliação (Contínua ou Final) escolhido? *

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ Sim
- ○ Não

17 [2.4.1]De que forma? *

Responda a esta pergunta apenas se as seguintes condições são verdadeiras:
° Resposta era Y'Sim' na pergunta '16 [2.4]' (A sua presença regular nas aulas depende do regime Avaliação (Contínua ou Final) escolhido?)

Por favor, escreva aqui a sua resposta:

18 [2.4.2]Porquê? *

Responda a esta pergunta apenas se as seguintes condições são verdadeiras:
° Resposta era N'Não' na pergunta '16 [2.4]' (A sua presença regular nas aulas depende do regime Avaliação (Contínua ou Final) escolhido?)

Por favor, escreva aqui a sua resposta:

19 [2.5]O facto de existirem 3 testes foi importante para que escolhesse o Regime de Avaliação Contínua? *

Por favor, seleccione **apenas uma** das seguintes opções:

- ○ Sim
- ○ Não

20 [2.5.1]Porquê? *

Responda a esta pergunta apenas se as seguintes condições são verdadeiras:
° Resposta era N'Não' na pergunta '19 [2.5]' (O facto de existirem 3 testes foi importante para que escolhesse o Regime de Avaliação Contínua?)

Por favor, escreva aqui a sua resposta:

21 [2.6]Qual a sua opinião sobre a existência de um teste de Repescagem? *

Por favor, escreva aqui a sua resposta:

22 [2.7]Na sua opinião quais são as vantagens, para os alunos, dos testes de escolha múltipla? *

Por favor, escreva aqui a sua resposta:

23 [2.8]Na sua opinião quais são as desvantagens, para o aluno, dos testes de escolha múltipla? *

Por favor, escreva aqui a sua resposta:

24 [2.9]Comentários adicionais:

Por favor, escreva aqui a sua resposta:

Submeter o seu inquérito
Obrigado por ter concluído este inquérito.

ANEXO C – GUIÃO DA ENTREVISTA AOS DOCENTES NO 3.º CICLO DE IA

PLANIFICAÇÃO DA ENTREVISTA

A entrevista é semiestruturada, de modo a permitir uma melhor organização dos tópicos a abordar e ainda assim dar liberdade aos entrevistados para expressar livremente as suas ideias.

Tema

Utilização de *e-assessment* com questões de escolha múltipla, para avaliação contínua em Matemática.

Objetivos

- Refletir sobre o *e-assessment* implementado.
- Verificar a existência de mudanças nas práticas pedagógicas, por parte dos docentes.
- Aferir quais as vantagens e desvantagens para o docente deste tipo de avaliação.
- Verificar a existência de mudanças nas práticas educativas, por parte dos alunos.

Dimensões

As dimensões coincidem com as questões formuladas.

1 Guião de Entrevista

Esta entrevista destina-se a uma reflexão sobre o *e-assessment* que tem vindo a ser implementada nas disciplinas da área científica de Matemática do ISCAP, nomeadamente Matemática, Matemática I, Matemática Aplicada e Matemática II. O *e-assessment* consistiu na utilização de testes com questões de escolha múltipla implementados na plataforma *Moodle*, para realização da avaliação contínua.
Para facilitar a transcrição da entrevista, peço autorização para a gravar.
Antes de começar agradeço a colaboração e o tempo que vais disponibilizar nesta entrevista.

Nome:
Idade:
Área disciplinar:
Tempo de serviço no ISCAP:

Tópicos a abordar:

1. **NO GERAL, QUAL A SUA OPINIÃO SOBRE ESTA FORMA DE AVALIAÇÃO *E-ASSESSMENT*?**
2. **QUAIS FORAM AS PRINCIPAIS DIFICULDADES ENCONTRADAS NA IMPLEMENTAÇÃO DESTA FORMA DE AVALIAÇÃO?**
3. **ESTA FORMA DE AVALIAÇÃO MUDOU DE ALGUMA FORMA AS SUAS PRÁTICAS PEDAGÓGICAS? EM QUE MEDIDA?**
4. **QUAIS AS VANTAGENS QUE ESTA FORMA DE AVALIAÇÃO TEM, DO PONTO DE VISTA DO DOCENTE?**
5. **QUAIS AS DESVANTAGENS QUE ESTA FORMA DE AVALIAÇÃO TEM, DO PONTO DE VISTA DO DOCENTE?**
6. **PARECE-LHE QUE OS ALUNOS MODIFICARAM DE ALGUMA FORMA AS SUAS PRÁTICAS EDUCATIVAS? EM QUE MEDIDA?**
7. **PRETENDE ACRESCENTAR MAIS ALGUMA INFORMAÇÃO E/OU COMENTÁRIO?**

ANEXO D – PROGRAMA DAS UC DE MATEMÁTICA E MATEMÁTICA I

1. **FUNÇÕES REAIS DE VARIÁVEL REAL.**

Definição e Notação.

Domínio.

Operações com Funções: Aritméticas e Composição.

Representação Geométrica e Ferramentas Gráficas.

Funções Especiais, seus Gráficos e Aplicações.

Funções Polinomiais.

Funções Racionais e outras Funções Algébricas envolvendo raízes.

Funções Definidas por Ramos (Função Módulo).

Funções Exponenciais e Logarítmicas.

Função Inversa.

Limites.

Introdução e Notação.

Definição.

Propriedades.

Cálculo.

Continuidade.

Definição.

Pontos de Descontinuidade.

Operações com Funções Contínuas.

Continuidade de algumas Funções Elementares (Polinomiais, Racionais).

Propriedades.

Cálculo Diferencial.

Introdução.

Taxa de Variação Média (num Intervalo) e Taxa de Variação Instantânea (num Ponto).

Definição de Derivada e sua Interpretação.

Notação.

Função Derivada.

Regras de Derivação.

Derivadas de uma Constante, uma Potência, Somas, Diferenças.

Regras do Produto e do Quociente.

Derivadas de Ordem Superior.

Regra da Cadeia.

Derivação da Função Implícita.

Diferenciais.

Incrementos e Diferenciais.

Aproximação Linear da Função na Vizinhança de um Ponto.

Aplicação do Cálculo Diferencial ao Estudo de Funções.

Monotonia e Extremos.

Concavidade e Pontos de Inflexão.

Assímptotas.

Esboço do Gráfico de Funções.

Teorema de Cauchy e expressões indeterminadas

Aplicação do Cálculo Diferencial em Ciências Empresariais.

Terminologia. Função de Custo e Lucro Marginais.

Exemplos de Optimização.

2. FUNÇÕES REAIS DE VÁRIAS VARIÁVEIS REAIS.

Definição e Notação.

Domínio. Representação Geométrica

Limites.

Definição de Limite de uma Função.

Propriedades.

Continuidade.

Definição de Função Contínua num Ponto e numa Região.

Derivadas Parciais.

Definição. Notação.

Interpretação Geométrica.

Derivadas Parciais de Ordem Superior.

Diferenciais e Diferencial Total.

Extremos.

Definição de Máximos e Mínimos Relativos e Absolutos.

Sua Determinação no Caso de Funções de Duas Variáveis.

Aplicações em Ciências Empresariais.

Custo Conjunto e Custo Marginal.

Funções de Produção.

Funções de Procura.

Lucro Máximo.

3. ÁLGEBRA LINEAR.

Matrizes.

Definição de matriz.

Aplicações.

Definições de Matriz Linha, Matriz Coluna, Ordem, Igualdade de Matrizes, Matriz Quadrada, Matriz Identidade.

Operações com Matrizes.

Transposição.

Adição.

Multiplicação por Escalar.

Multiplicação de Matrizes.

Resolução de Sistemas de Equações Lineares.

Matriz Ampliada e Matriz dos Coeficientes.

Operações Elementares sobre Linhas.

Método de Eliminação de Gauss-Jordan.

Sistemas com Soluções Únicas.

Sistemas com Soluções Não-Únicas.

Inversa de uma Matriz Quadrada.

Equações Matriciais.

Determinantes.

Definição de Determinante de uma Matriz Quadrada.

Menor Complementar, Complemento Algébrico e Expansão de Laplace.

Propriedades e Cálculo de Determinantes.

Resolução de Sistemas de Equações Lineares pela Regra de Cramer.

Determinantes e Matrizes Inversas.

ANEXO E – PROGRAMA DAS UC DE MATEMÁTICA II E MATEMÁTICA APLICADA

1. CÁLCULO INTEGRAL.

Integral Indefinido.

Primitiva.

Definição de Integral Indefinido e Interpretação Geométrica.

Propriedades e Regras Básicas de Integração.

Métodos de Integração.

Integração por Partes.

Integração por Mudança de Variável.

Integração de Funções Racionais.

Integral Definido.

Definição de Integral Definido e Interpretação Geométrica.

Propriedades.

Primeiro Teorema Fundamental do Cálculo Integral.

Teorema do Valor Médio. Média Aritmética de uma Função.

Segundo Teorema Fundamental do Cálculo Integral.

Cálculo de Áreas.

Integrais Impróprios.

Integrais Impróprios de 1.ª e 2.ª espécie.

Integral Múltiplo.

Integração Parcial.

Integral Duplo.

Problema do Cálculo de Volumes.

Cálculo de Integrais Duplos sobre Regiões Rectangulares e não Rectangulares.

Aplicação do Cálculo Integral na Resolução de Problemas de Economia.

2. ANÁLISE COMBINATÓRIA.

Introdução.

Terminologia e Conceitos Básicos.

Experiência Aleatória. Espaço Amostral. Evento.

Operações com eventos.

Intersecção. Reunião. Diferença. Propriedades.

Partição do espaço amostral.

Métodos de Contagem.

Princípio Fundamental de Contagem.

Factorial de um número natural.

Permutações.

Combinações.

Triângulo de Pascal. Binómio de Newton.

3. SÉRIES NUMÉRICAS REAIS.

Sucessões.

Definição, Representação Geométrica e Determinação do Termo Geral.

Limite de uma sucessão. Infinitésimos e Infinitamente Grandes.

Teoremas sobre Sucessões Convergentes.

Progressões Aritméticas e Geométricas.

Séries e Convergência.

Definição.

Séries Convergentes e Divergentes. Soma de uma Série.

Séries Geométricas, Telescópicas e de Riemann.

Propriedades.

Séries de Termos não Negativos.

Critérios de Convergência.

Condição Necessária.

Critérios de Comparação.

Critério D'Alembert.

Critério de Cauchy.

Aplicações.

ANEXO F – INSTRUÇÕES PARA TESTE DE SIMULAÇÃO

Testes de
MATEMÁTICA (CA), MATEMÁTICA I (CI), MATEMÁTICA APLICADA (CA) e MATEMÁTICA II (CI)

INSTRUÇÕES

1. REDE SEM FIOS

- Verifique se a rede sem fios (rede *wireless*) está ligada.
- Procure as redes **AVALIA1, AVALIA2**, …, **AVALIA15.**
- Ligue-se preferencialmente àquela que tem sinal mais forte.

2. *MOODLE*

- Abra um dos navegadores, ***Mozilla Firefox*** (recomendado) ou ***Internet Explorer***.
- Na barra de endereços, escreva online.iscap.ipp.pt/*Moodle*21 e pressione a tecla ***Enter***.
- Clique em ***Entrar***, coloque em ***Nome de utilizador*** e ***Senha*** as suas credenciais da Secretaria OnLine do ISCAP e volte a clicar em ***Entrar***.
- Em "Grupo de disciplinas" clique em "**Disciplinas da área da Matemática (4)**".
- Clique no nome da Unidade Curricular em que vai fazer o teste. Caso não se tenha inscrito previamente nessa UC clique em ***Enroll me***.

3. REALIZAÇÃO DO TESTE

3.1. CUIDADOS GERAIS

- Durante toda a prova a **única janela que pode estar aberta** é a do navegador.
- **Responda a cada questão, clicando na opção pretendida só quando tiver a certeza de que é essa a opção que pretende selecionar** (selecionada uma opção de resposta jamais poderá não responder a essa questão).
- **Utilize apenas o rato. Não utilize as setas de cursor,** dado que inadvertidamente poderá estar a alterar a opção de resposta que seleccionou no teste.
- **Vá controlando o tempo** no relógio do *Moodle* e **no seu relógio** (se sair do teste por instantes e tornar a entrar, o tempo continuará a ser contabilizado).

3.2. INICIAR O TESTE

- Depois de aceder à Unidade Curricular, escolha o turno a que pertence para iniciar o teste.
- Selecione ***Tentar resolver agora o teste*** e de seguida a opção ***Começar a tentativa***.
- Introduza a ***Senha*** que o Professor escrever no quadro e inicie o teste.

3.3. DURANTE O TESTE

- Vá gravando as suas respostas, clicando no botão ***Próximo*** no final da página. Sugere-se que este procedimento seja feito pelo menos uma vez, 6 minutos antes de terminar o teste.

- Surgir-lhe-á uma tabela indicando a situação de cada pergunta; **clique** agora **no número de qualquer pergunta** para VOLTAR AO TESTE.
- **Nunca** use a opção ***retroceder/recuar*** nem a opção ***avançar*** do Navegador/*Browser*.
- **Não use** a tecla ***Back Space***.

3.4. TERMINAR O TESTE

- Para terminar, clique em ***Próximo*** e de seguida em ***Enviar tudo e terminar***.
- Aparece uma janela de **confirmação** se quer mesmo ***Enviar tudo e terminar*** ou **cancelar** e rever alguma opção.
- Não se esqueça de **enviar o teste dentro do tempo limite**. **Caso não o faça poderá ter zero na classificação.**

ANEXO G – PRIMEIRO RELATÓRIO RELATIVO AO PRIMEIRO TESTE DE SIMULAÇÃO

RELATÓRIO
de
Simulação de provas de avaliação no *Moodle* em 26 de Outubro de 2011.

1. Introdução

Conforme previamente programado, no dia supracitado procedeu-se à simulação de um teste de avaliação contínua nas salas 221 (1 e 2) e 223 (1 e 2).

Pretendia saber-se se a infraestrutura informática e o *Moodle* estavam preparados para o nº de alunos que as salas suportavam.

Dividiu-se a simulação de provas em 4 turnos de alunos. Todos os turnos estavam planeados para durarem no máximo 20 minutos, mas na prática foram sempre ultrapassados devido a problemas que foram surgindo. Estava planeado que o primeiro turno começasse às 14h30m, o segundo às 15h, o terceiro às 15:30 e o último às 18 horas. Este último turno envolvia apenas alunos nocturnos.

Realizaram a simulação de teste, aproximadamente 327 alunos. O nº de alunos poderia ser maior, não tivessem alguns deles desistido devido ao atraso no início dos segundos e terceiros turnos. Dos 327 alunos, 236 eram diurnos e 91 noturnos. Esperávamos mais alunos noturnos do que aqueles que realmente apareceram.

Foi dada uma folha A4 aos alunos com as principais indicações para se orientarem. Continha, entre outras indicações: (i) os nomes das redes e qual a que deveriam escolher, (ii) indicação do endereço *web* para se ligarem ao *Moodle* 2.1.2, (iii) inscrição na disciplina e atualização de perfil, (iv) escolha do teste (turma) e respetiva senha e (v) sugestões e cuidados a ter na realização do teste.

Tivemos o apoio Dra. Luciana do PAOL e dos senhores Joaquim Silva (desde o início) e Bruno Sousa (pontualmente) do HelpDesk.

Apresentam-se de seguida os resultados desta simulação.

2. Portáteis dos alunos

Quase todos os alunos se apresentaram com o respectivo portátil. Poucos alunos (menos de 10) não apareceram com o portátil, mas disseram que providenciariam um para o dia do teste de avaliação contínua.

Os primeiros problemas surgiram com a ligação à rede. Alguns portáteis não reconheciam as redes. Estes problemas foram todos resolvidos. Uns pelos professores de Matemática, outros pela Dra. Luciana e pela equipa do HelpDesk.

Os principais problemas detetados prenderam-se com as propriedades da placa de rede sem fios. Algumas não tinham o protocolo IP versão 4, de modo a obter o endereço IP automaticamente. Outros tinham nas definições da rede local, dos navegadores (IE ou Firefox, etc) servidores de "proxy" ativos, pelo que tivemos que os desativar. Alguns com o sistema operativo Windows Vista, foram os mais difíceis de resolver; tendo que se criar manualmente as redes sem fios. Para resolver este problema, a ajuda de Bruno Sousa foi decisiva.

3. Infraestrutura Informática

Os grandes problemas surgiram na infraestrutura informática.

Apesar de cada aluno ter a indicação da rede a que se devia ligar, a maioria não o conseguiu fazer.
Dos alunos que conseguiram ligar-se à rede sem fios, muitos não puderam entrar no *Moodle*. E outros conseguiram-no através de uma ligação muito lenta.
As imagens do teste não apareciam, aparecendo somente o texto em alguns casos e noutros partes das imagens.
Para resolver este problema pedimos aos alunos que tinham acedido ao teste sem problemas, para se desligarem. Desta forma verificou-se que outros que não tinham conseguido ligar-se puderam agora aceder ao teste.
Este problema repetiu-se nos vários turnos. Aliás foi necessário fazer (Joaquim) vários *Reset* ao *Routers*.
Parece-nos que o número máximo de alunos que conseguem estar ligados simultaneamente não ultrapassa os 20.

4. Moodle

Um dos problemas detetados prende-se com a limitação de acesso por IP ao teste por parte dos alunos.
Verificou-se que os alunos acederam por outras vias que não unicamente os *Routers* colocados nas salas.

Tínhamos pedido que o acesso ao *Moodle* fosse unicamente acedido pelos *Routers* ligados nas salas. Foi-nos garantido na reunião que tivemos na presidência que o acesso nunca poderia ser feito de outra forma.

Este problema foi detetado no *Moodle*, mas parece-nos que se poderá resolver através da estrutura informática.

5. Conclusões

Parece-nos que reside na infraestrutura informática os principais constrangimentos.

1 - Os dois *Routers* que estiveram a suportar a primeira ligação à rede, devem ter problemas ou então não estão ajustados ao número de utilizadores que se pretende que estejam simultaneamente ligados.

2 - A rede de ligação entre os *Routers* e o servidor do *Moodle* tem graves limitações. Não sabendo se trata de problemas de *hardware* ou de outro tipo.

3 - O servidor onde se encontra alojado o *Moodle* não tem capacidade para ter muitas sessões simultâneas.

4 - É possível aceder ao *Moodle* fora da rede interna criada unicamente para a avaliação de matemática.

Tendo em conta a calendarização para disciplina de Matemática, em que os primeiros testes se realizam entre 15 e 18 de Novembro, agradece-se brevidade na resolução destes problemas. Se tal não acontecer, a avaliação contínua em Matemática pode ser posta em causa.

S. Mamede de Infesta, 3 de Novembro de 2011

Sr. PA

José Manuel Azevedo

XXXXX

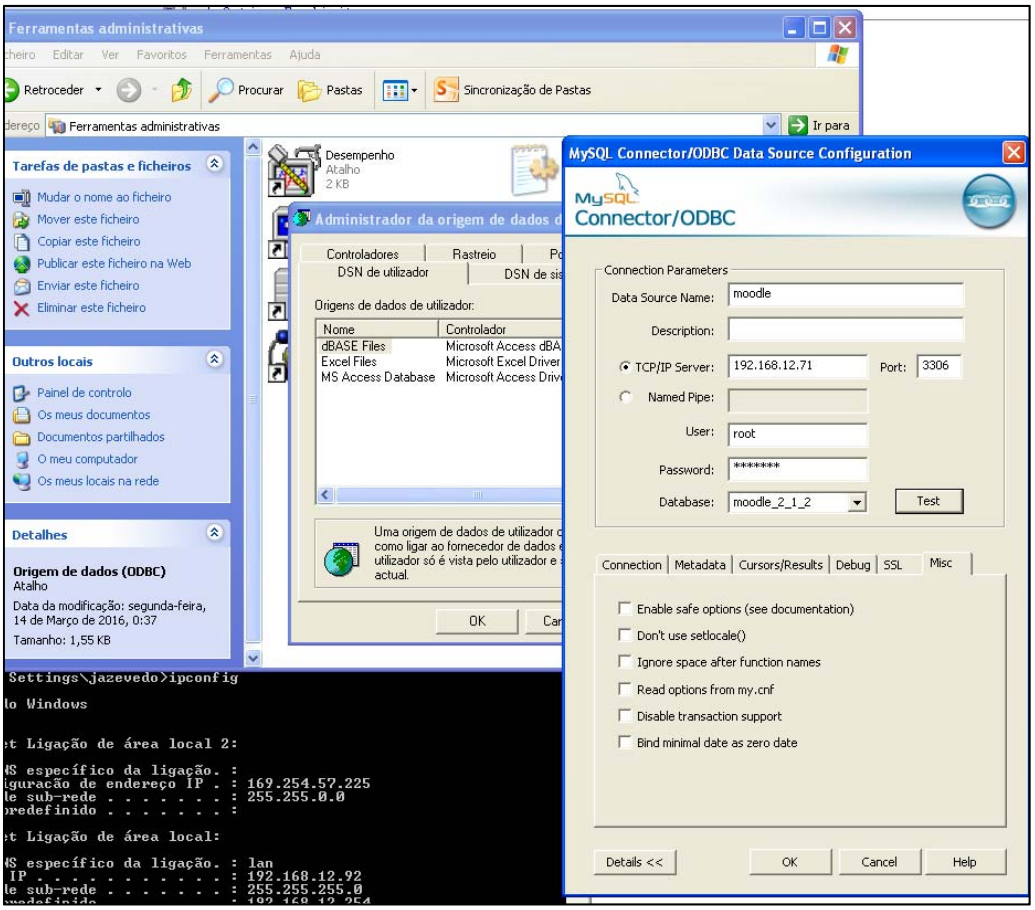
XXXXX

XXXXX

XXXXX

Sr FJ

ANEXO H – CONFIGURAÇÃO EM WINDOWS DA LIGAÇÃO ODBC



ANEXO I – TAMANHO DO EFEITO COMO COMPLEMENTO A ALGUNS TESTES ESTATÍSTICOS

No Anova a um fator as medidas mais utlizadas para verificar o tamanho do efeito dos resultados encontrados são a medida "Omega Sq" (ómega quadrado), a medida "*Cohen d*" e a medida "ψ de Steiger (RMSSE)". Existe alguma polémica acerca dos intervalos de valores e sua validade para determinar a real magnitude do efeito. Pode-se encontrar em (Kotrlik & Williams, 2003, p. 5) uma tabela com várias sugestões para as medidas de magnitude do efeito, bem como o teste estatístico associado que mais se adequa. Em (Ialongo, 2016) pode-se encontrar um conjunto mais alargado de todas as medidas do efeito conhecidas até ao momento e ainda encontrar a indicação de qual ou quais as medidas do efeito mais adequadas a cada teste estatístico e respetivas indicações de medida. Contudo, outros autores são mais precisos na indicação das medidas que se devem utilizar e sugerem um conjunto de procedimentos para o seu uso. (Levine & Hullett, 2002, p. 620) sugere em primeiro lugar que "Os investigadores devem mais frequentemente usar o *eta quadrado*, *ómega quadrado* ou *epsilon quadrado* em vez do *eta quadrado parcial*". Este último é usual no SPSS.

As medidas de efeitos que usamos neste trabalho, além de outros apresentados, vão ser *Cohen d*, *RMSSE* e a *Ómega Quadrado*. A *Cohen d*, é mais adequada quando se comparam duas amostras. Por isso, na nossa análise ANOVA usamos a *RMSSE* e a *Ómega quadrado* e a *Cohen d* nos Contrastes. As tabelas com valores indicativos para avaliação da magnitude do efeito dos resultados do teste ANOVA para estas três medidas, encontram-se na **Error! Reference source not found.**.

Na tabela seguinte encontram-se os valores indicativos para avaliar a magnitude do efeito dos resultados do teste ANOVA:

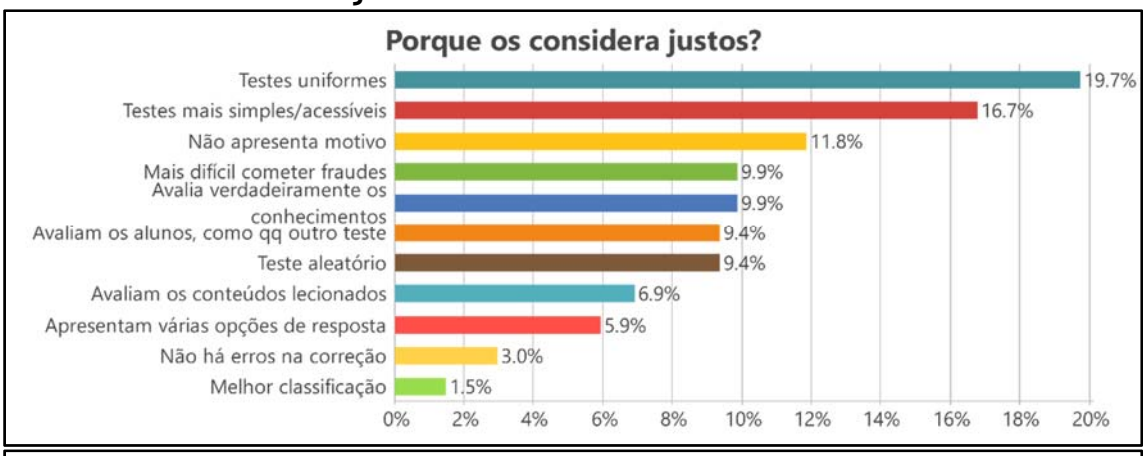
Teste de Cohen (d)	
Tamanho do Efeito	d
Pequeno	0.20
Moderado	0.50
Elevado	0.80

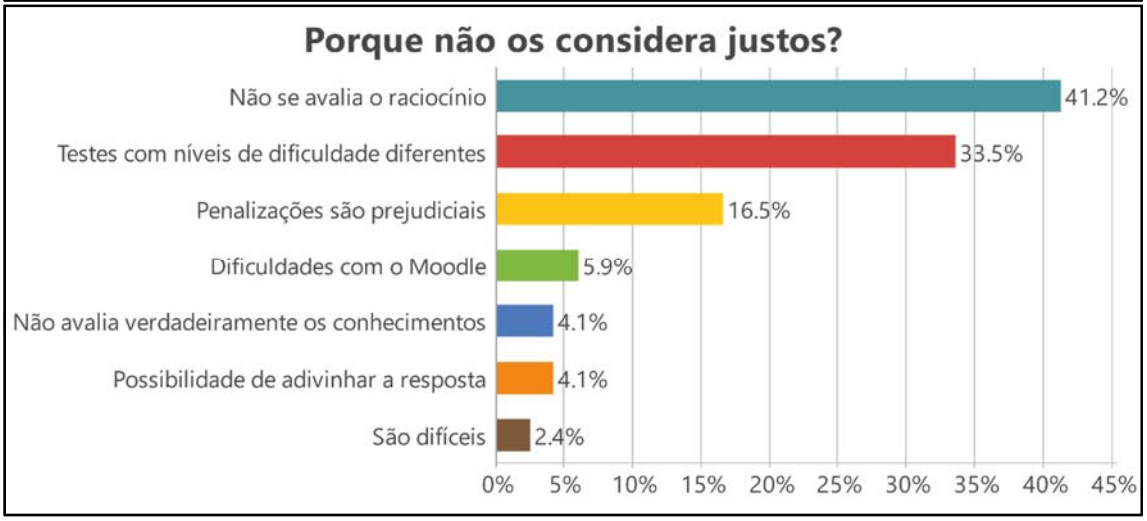
Teste Omega quadrado (Omega Sq)	
Tamanho do Efeito	ω^2
Pequeno	0.010
Moderado	0.059
Elevado	0.138

ψ de Steiger (RMSSE)	
Tamanho do Efeito	ψ
Pequeno	0.10
Moderado	0.25
Elevado	0.40

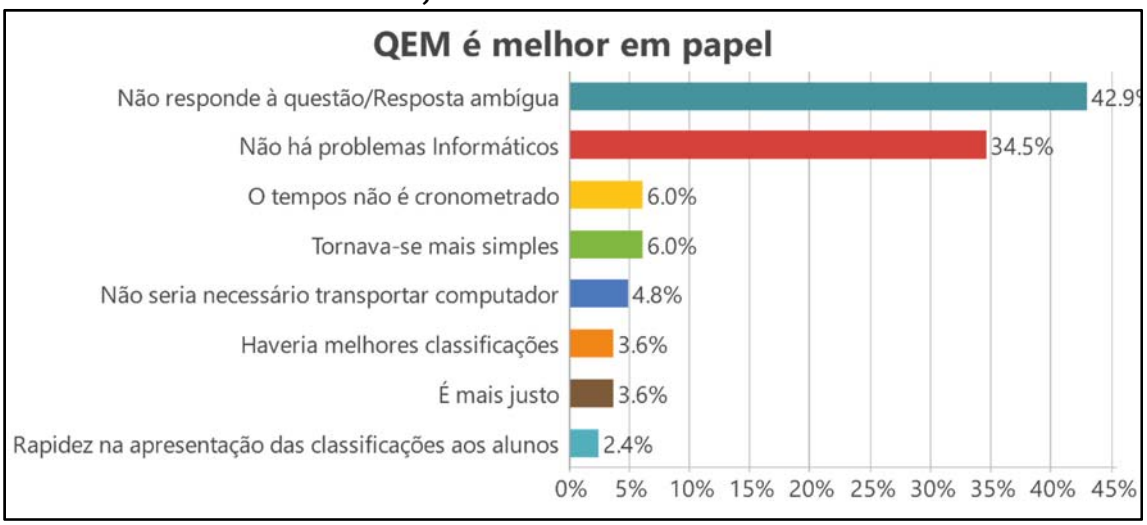
ANEXO J – ESTATÍSTICAS DOS INDICADORES

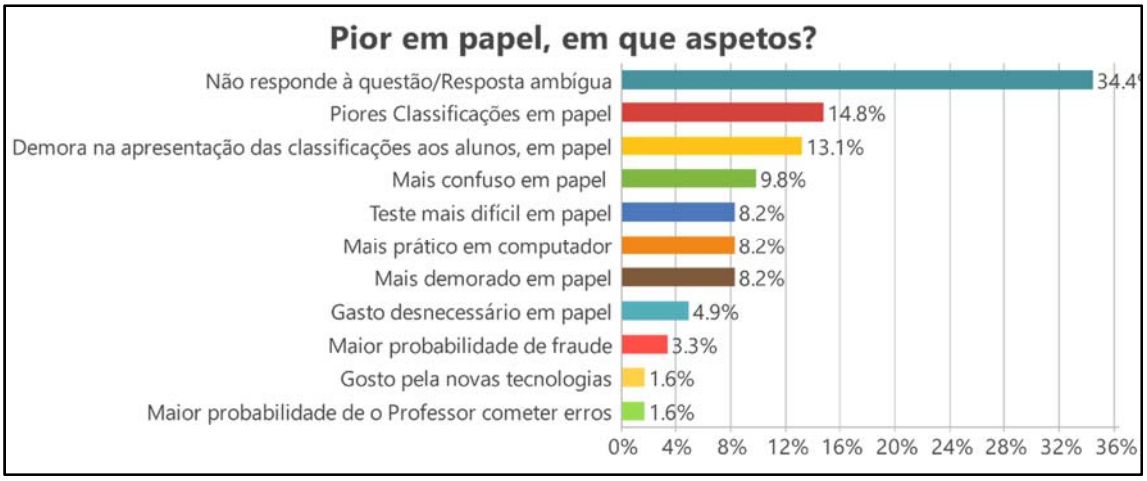
Considera que os testes de escolha múltipla realizados no *Moodle*, na Unidade Curricular são justos?



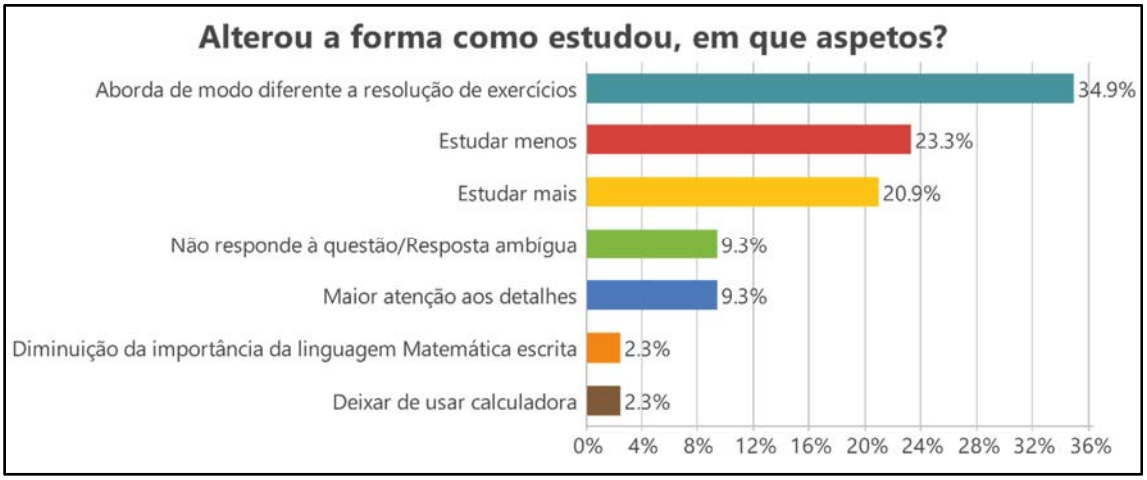


Considera que se estes testes (EM) fossem realizados em papel em vez de serem realizados no *Moodle*, seriam:

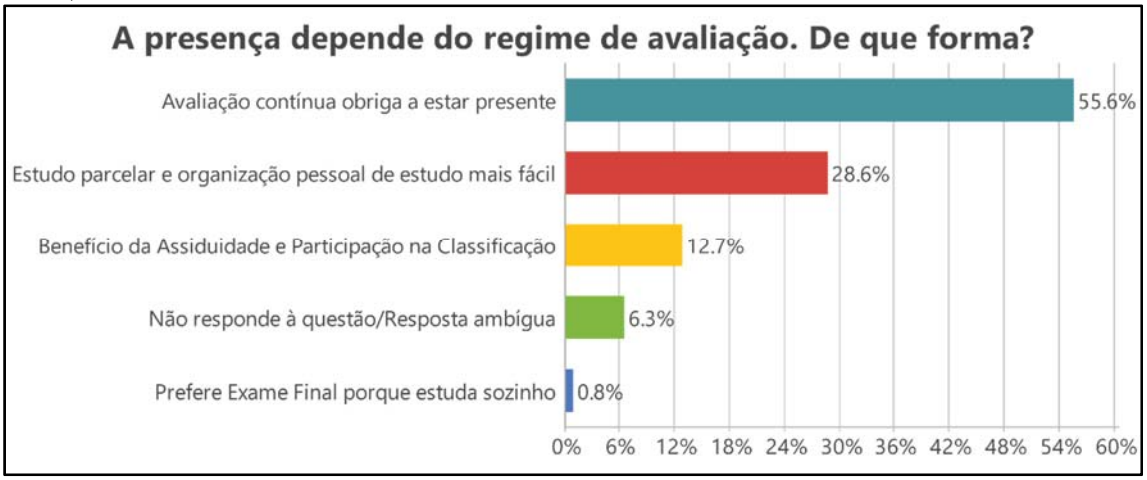


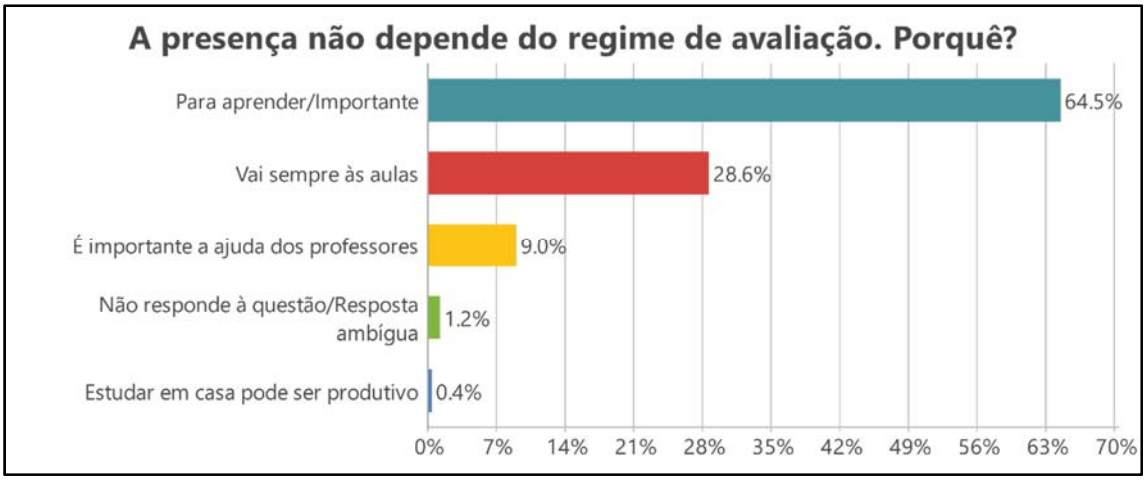


O facto de os testes serem de escolha múltipla alterou de alguma forma o modo como estudou?

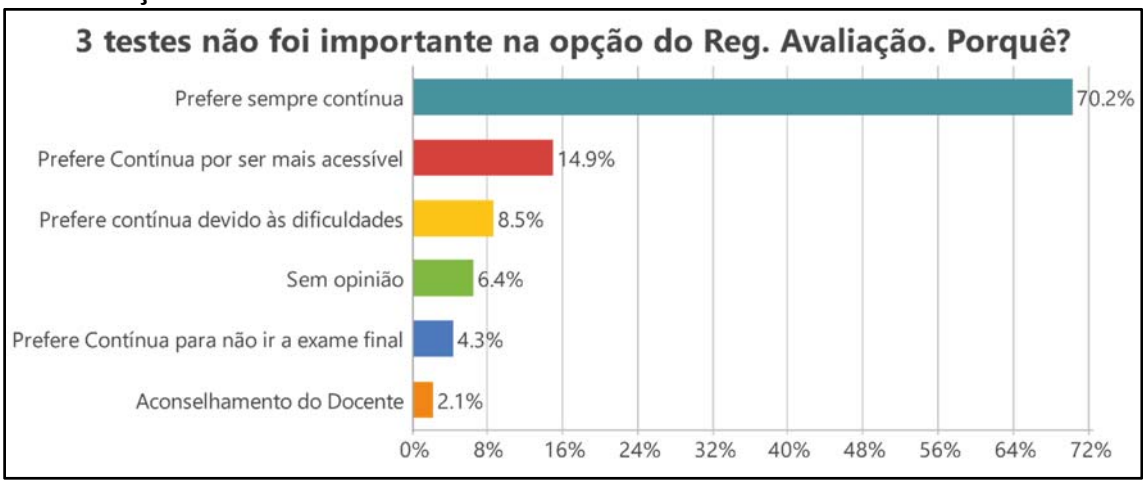


A sua presença regular nas aulas depende do regime avaliação (contínua ou final) escolhido?

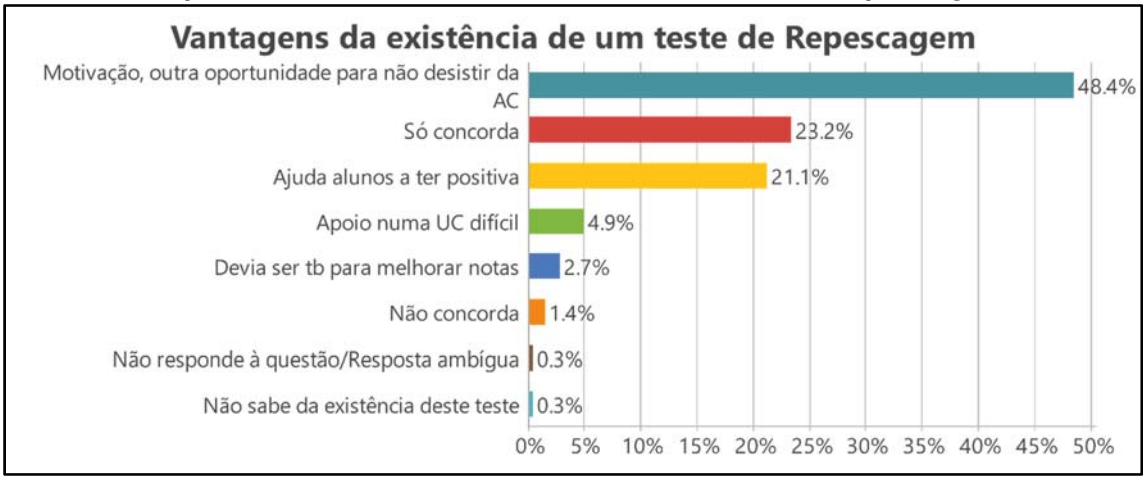




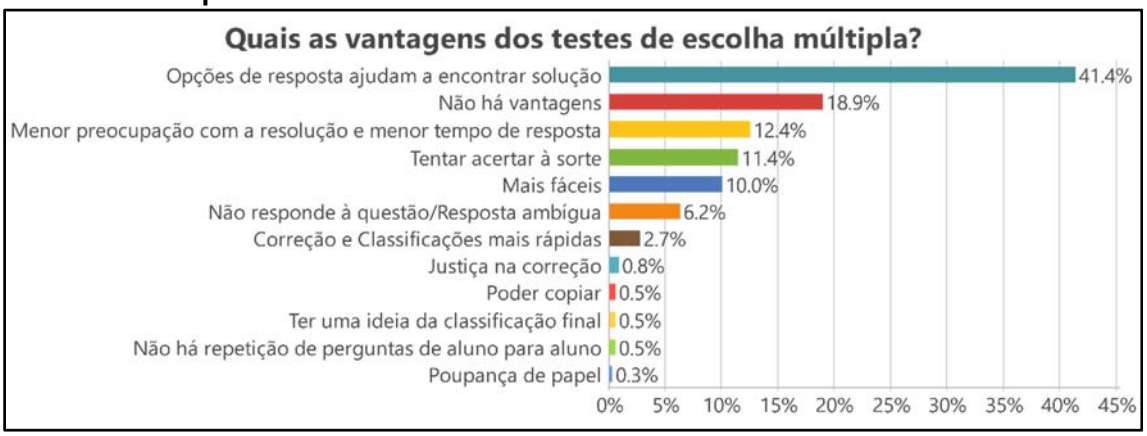
O facto de existirem 3 testes foi importante para que escolhesse o regime de avaliação contínua?



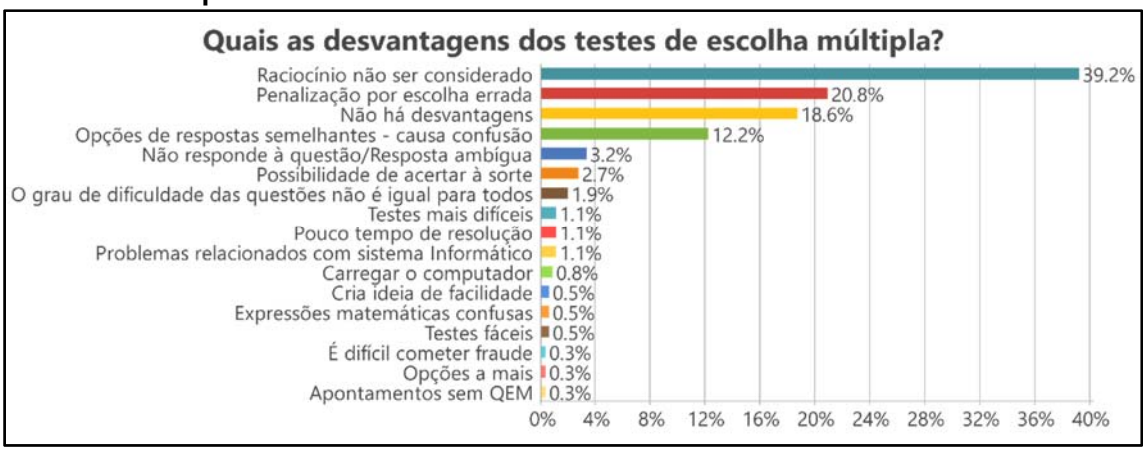
Qual a sua opinião sobre a existência de um Teste de "Repescagem"?



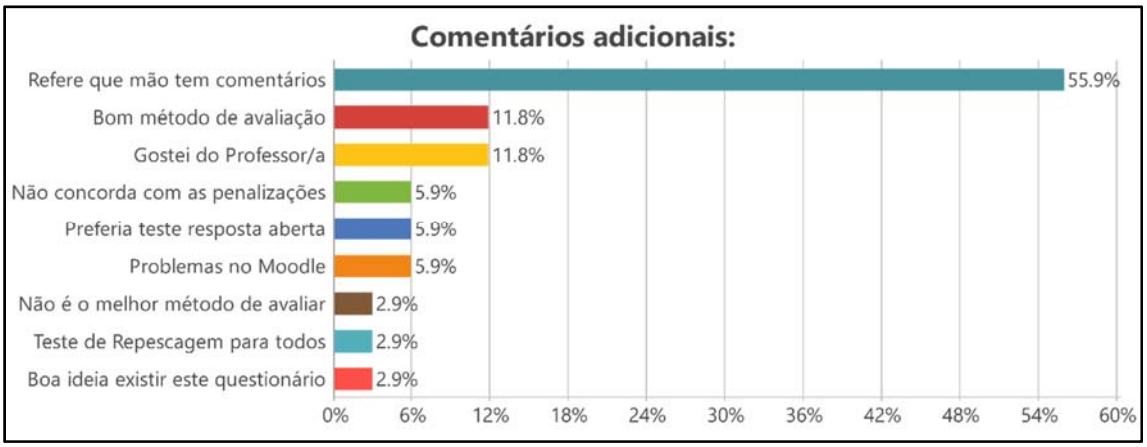
Na sua opinião quais são as vantagens, para os alunos, dos testes de escolha múltipla?



Na sua opinião quais são as desvantagens, para o aluno, dos testes de escolha múltipla?



Comentários adicionais:



ANEXO K - TABELAS DE CONTINGÊNCIA ENVOLVENDO AS DIMENSÕES E INDICADORES COM VARIÁVEIS DE

Considera que os testes de escolha múltipla realizados no Moodle, na Unidade Curricular são justos?										
	id2	**Masculino**	**Feminino**	**Diurno**	**Noturno**	**Matemática**	**Matemática I**	**UC 1.ª vez? Não**	**UC 1.ª vez? Sim**	**QEM são justos? Sim**
Porque os considera justos?	**203**	**108**	**95**	**140**	**63**	**160**	**43**	**63**	**140**	**203**
Avalia verdadeiramente os conhecimentos	20	11	9	12	8	14	6	8	12	20
Não há erros na correção	6	0	6	6	0	5	1	0	6	6
Avaliam os conteúdos lecionados	14	10	4	11	3	10	4	3	11	14
Não apresenta motivo	24	14	10	15	9	19	5	8	16	24
Testes mais simples/acessíveis	34	17	17	17	17	27	7	12	22	34
Mais difícil cometer fraudes	20	8	12	15	5	15	5	4	16	20
Melhor classificação	3	0	3	0	3	2	1	2	1	3
Apresentam várias opções de resposta	12	5	7	9	3	9	3	4	8	12
Testes uniformes	40	25	15	34	6	34	6	10	30	40
Avaliam os alunos como qq outro teste	19	12	7	13	6	16	3	5	14	19
Teste aleatório	19	10	9	14	5	14	5	9	10	19
SOMA	211	112	99	146	65	165	46	65	146	211
N (Documentos)	371	168	203	261	110	297	74	107	264	203
	id2	**Masculino**	**Feminino**	**Diurno**	**Noturno**	**Matemática**	**Matemática I**	**UC 1.ª vez? Não**	**UC 1.ª vez? Sim**	**QEM são justos? Não**
Porque não os considera justos?	167	60	107	120	47	136	31	44	123	168
São difíceis	4	0	4	2	2	4	0	3	1	4
Possibilidade de adivinhar a resposta	7	2	5	5	2	6	1	2	5	7
Dificuldades com o *Moodle*	10	3	7	4	6	8	2	0	10	10
Não avalia verdadeiramente os conhecimentos	7	3	4	4	3	5	2	3	4	7
Penalizações são prejudiciais	28	10	18	22	6	21	7	9	19	26
Testes com níveis de dificuldade diferentes	57	24	33	46	11	48	9	12	45	57

Não se avalia o raciocínio	70	23	47	49	21	54	16	19	51	69
SOMA	183	65	118	132	51	146	37	48	135	180
N (Documentos)	371	168	203	261	110	297	74	107	264	168

Considera que se estes testes (escolha múltipla) fossem realizados em papel em vez de serem realizados no Moodle, seriam:

	id2	Masculino	Feminino	Diurno	Noturno	Matemática	Matemática I	UC 1.ª vez? Não	UC 1.ª vez? Sim	Seriam: Melhor
QEM é melhor em papel	**84**	**29**	**55**	**58**	**26**	**66**	**18**	**22**	**62**	**84**
Não seria necessário transportar computador	4	3	1	3	1	3	1	0	4	4
Haveria melhores classificações	3	0	3	2	1	3	0	0	3	3
É mais justo	3	2	1	2	1	1	2	2	1	3
Tornava-se mais simples	5	1	4	4	1	4	1	2	3	5
Não responde à questão/Resposta ambígua	36	13	23	30	6	27	9	12	24	35
O tempo não é cronometrado	5	2	3	2	3	4	1	2	3	5
Rapidez na apresentação das classificações aos alunos	2	1	1	1	1	2	0	1	1	2
Não há problemas Informáticos	29	9	20	18	11	25	4	3	26	29
SOMA	87	31	56	62	25	69	18	22	65	86
N (Documentos)	371	168	203	261	110	297	74	107	264	84

	id2	Masculino	Feminino	Diurno	Noturno	Matemática	Matemática I	UC 1.ª vez? Não	UC 1.ª vez? Sim	Seriam: Pior
Pior em papel, em que aspetos?	**61**	**32**	**29**	**42**	**19**	**50**	**11**	**21**	**40**	**61**
Gosto pelas novas tecnologias	1	0	1	1	0	1	0	1	0	1
Mais prático em computador	5	3	2	4	1	4	1	1	4	5
Teste mais difícil em papel	5	2	3	4	1	5	0	2	3	5
Gasto desnecessário em papel	3	2	1	2	1	3	0	2	1	3
Maior probabilidade de o Professor cometer erros	1	0	1	1	0	1	0	0	1	1

Demora na apresentação das classificações aos alunos, em papel	8	4	4	6	2	5	3	3	5	8
Mais demorado em papel	5	1	4	3	2	4	1	1	4	5
Não responde à questão/Resposta ambígua	21	11	10	15	6	18	3	7	14	20
Maior probabilidade de fraude	2	2	0	1	1	2	0	0	2	2
Piores Classificações em papel	9	7	2	5	4	8	1	3	6	9
Mais confuso em papel	6	4	2	3	3	4	2	3	3	6
SOMA	66	36	30	45	21	55	11	23	43	65
N (Documentos)	371	168	203	261	110	297	74	107	264	61

O facto de os testes serem de escolha múltipla alterou de alguma forma o modo como estudou?

	id2	Masculino	Feminino	Diurno	Noturno	Matemática	Matemática I	UC 1.ª vez? Não	UC 1.ª vez? Sim	ALTEROU? SIM
Alterou a forma como estudou, em que aspetos?	**43**	**16**	**27**	**32**	**11**	**36**	**7**	**16**	**27**	**43**
Deixar de usar calculadora	1	1	0	1	0	1	0	0	1	1
Diminuição da importância da linguagem Matemática escrita	1	1	0	0	1	1	0	0	1	1
Não responde à questão/Resposta ambígua	4	2	2	3	1	4	0	0	4	4
Estudar menos	10	4	6	9	1	7	3	2	8	10
Maior atenção aos detalhes	4	0	4	3	1	4	0	2	2	4
Aborda de modo diferente a resolução de exercícios	15	5	10	10	5	12	3	8	7	15
Estudar mais	9	3	6	7	2	8	1	5	4	9
SOMA	44	16	28	33	11	37	7	17	27	44
N (Documentos)	371	168	203	261	110	297	74	107	264	43

A sua presença regular nas aulas depende do regime avaliação (contínua ou final) escolhido?

	id2	Masculino	Feminino	Diurno	Noturno	Matemática	Matemática I	UC 1.ª vez? Não	UC 1.ª vez? Sim	Depende do regime
A presença depende do regime de avaliação. De que forma?	**126**	**62**	**64**	**84**	**42**	**109**	**17**	**34**	**92**	**127**
Prefere Exame Final porque estuda sozinho	1	1	0	1	0	1	0	0	1	1
Estudo parcelar e organização pessoal de estudo mais fácil	36	21	15	18	18	29	7	7	29	36
Não responde à questão/Resposta ambígua	8	3	5	7	1	8	0	3	5	8
Avaliação contínua obriga a estar presente	70	33	37	51	19	59	11	21	49	70
Benefício da Assiduidade e Participação na Classificação	16	8	8	11	5	15	1	3	13	15
SOMA	131	66	65	88	43	112	19	34	97	130
N (Documentos)	371	168	203	261	110	297	74	107	264	127

	id2	Masculino	Feminino	Diurno	Noturno	Matemática	Matemática I	UC 1.ª vez? Não	UC 1.ª vez? Sim	Não depende do regime
A presença não depende do regime de avaliação. Porquê?	**257**	**113**	**144**	**183**	**74**	**198**	**59**	**76**	**181**	**244**
Não responde à questão/Resposta ambígua	4	3	0	2	1	3	0	2	1	3
É importante a ajuda dos professores	22	11	11	16	6	17	5	6	16	22
Para aprender/Importante	158	65	93	117	41	120	38	45	113	157
Vai sempre às aulas	70	32	38	51	19	57	13	22	48	69
SOMA	254	111	143	186	68	197	57	75	179	252
N (Documentos)	371	168	203	261	110	297	74	107	264	244

O facto de existirem 3 testes foi importante para que escolhesse o regime de avaliação contínua?

	id2	Masculino	Feminino	Diurno	Noturno	Matemática	Matemática I	UC 1.ª vez? Não	UC 1.ª vez? Sim	Não Foi Importante
3 testes não foi importante na opção do reg. avaliação. Porquê?	**47**	**25**	**24**	**29**	**16**	**40**	**7**	**10**	**33**	**45**
Aconselhamento do Docente	1	1	0	1	0	0	1	0	1	1

Prefere Contínua por ser mais acessível	7	4	3	3	4	6	1	4	3	7
Prefere Contínua para não ir a Exame Final	2	1	1	1	1	2	0	1	1	2
Sem opinião	3	0	3	2	1	3	0	1	2	3
Prefere contínua devido às dificuldades	4	2	2	3	1	3	1	0	4	4
Prefere sempre contínua	33	18	15	23	10	28	5	8	25	33
SOMA	50	26	24	33	17	42	8	14	36	50
N (Documentos)	371	168	203	261	110	297	74	107	264	47

Qual a sua opinião sobre a existência de um Teste de "Repescagem"?									
	id2	**Masculino**	**Feminino**	**Diurno**	**Noturno**	**Matemática**	**Matemática I**	**UC 1.ª vez? Não**	**UC 1.ª vez? Sim**
Vantagens da existência de um Teste de "Repescagem"	**371**	**168**	**203**	**261**	**110**	**297**	**74**	**107**	**264**
Não responde à questão/Resposta ambígua	1	1	0	1	0	0	1	0	1
Não sabe da existência deste teste	1	1	0	1	0	1	0	0	1
Não concorda	5	2	3	4	1	5	0	0	5
Devia ser tb para melhorar notas	10	2	8	8	2	7	3	0	10
Só concorda	86	41	45	61	25	70	16	34	52
Motivação, outra oportunidade para não desistir da AC	179	92	87	120	59	147	32	48	131
Ajuda alunos a ter positiva	78	26	52	58	20	60	18	22	56
Apoio numa UC difícil	18	5	13	14	4	12	6	4	14
SOMA	378	170	208	267	111	302	76	108	270
N (Documentos)	371	168	203	261	110	297	74	107	264

Na sua opinião quais são as vantagens, para os alunos, dos testes de escolha múltipla?									
	id2	**Masculino**	**Feminino**	**Diurno**	**Noturno**	**Matemática**	**Matemática I**	**UC 1.ª vez? Não**	**UC 1.ª vez? Sim**
Quais as vantagens dos testes de escolha múltipla?	**370**	**167**	**203**	**260**	**110**	**296**	**74**	**107**	**263**
Ter uma ideia da classificação final	2	1	1	1	1	2	0	1	1

Não responde à questão/Resposta ambígua	23	11	12	15	8	19	4	9	14
Mais fáceis	37	20	17	22	15	29	8	14	23
Justiça na correção	3	0	3	2	1	2	1	1	2
Poupança de papel	1	0	1	1	0	0	1	0	1
Não há repetição de perguntas de aluno para aluno	2	1	1	0	2	1	1	2	0
Poder copiar	2	1	1	2	0	2	0	2	0
Correção e Classificações mais rápidas	10	4	6	4	6	4	6	4	6
Não há vantagens	70	28	42	47	23	56	14	19	51
Tentar acertar à sorte	42	21	21	36	6	37	5	13	29
Menor preocupação com a resolução e menor tempo de resposta	46	21	25	35	11	41	5	15	31
Opções de resposta ajudam a encontrar solução	153	69	84	111	42	118	35	38	115
SOMA	391	177	214	276	115	311	80	118	273
N (Documentos)	371	168	203	261	110	297	74	107	264

Na sua opinião quais são as desvantagens, para o aluno, dos testes de escolha múltipla?									
	id2	**Masculino**	**Feminino**	**Diurno**	**Noturno**	**Matemática**	**Matemática I**	**UC 1.ª vez? Não**	**UC 1.ª vez? Sim**
Quais as desvantagens dos testes de escolha múltipla?	**371**	**168**	**203**	**261**	**110**	**297**	**74**	**107**	**264**
Cria ideia de facilidade	2	2	0	2	0	1	1	0	2
Testes fáceis	2	1	1	2	0	2	0	2	0
Pouco tempo de resolução	4	3	1	1	3	3	1	2	2
Carregar o computador	3	2	1	1	2	1	2	2	1
Possibilidade de acertar à sorte	10	3	7	6	4	9	1	6	4
Apontamentos sem QEM	1	0	1	1	0	1	0	1	0
É difícil cometer fraude	1	0	1	1	0	0	1	0	1
Não responde à questão/Resposta ambígua	12	7	5	9	3	10	2	5	7
Problemas relacionados com sistema Informático	4	3	1	4	0	4	0	1	3
Opções a mais	1	0	1	1	0	1	0	1	0
Testes mais difíceis	4	1	3	3	1	2	2	1	3

O grau de dificuldade das questões não é igual para todos	7	2	5	6	1	3	4	3	4
Expressões matemáticas confusas	2	0	2	0	2	2	0	2	0
Opções de respostas semelhantes - causa confusão	45	24	21	28	17	34	11	15	30
Não há desvantagens	69	39	30	51	18	56	13	11	58
Raciocínio não ser considerado	145	54	91	101	44	119	26	36	109
Penalização por escolha errada	77	34	43	62	15	62	15	25	52
SOMA	389	175	214	279	110	310	79	113	276
N (Documentos)	371	168	203	261	110	297	74	107	264

Comentários adicionais:									
	id2	**Masculino**	**Feminino**	**Diurno**	**Noturno**	**Matemática**	**Matemática I**	**UC 1.ª vez? Não**	**UC 1.ª vez? Sim**
Comentários adicionais:	**34**	**20**	**14**	**22**	**12**	**24**	**10**	**12**	**22**
Não é o melhor método de avaliar	1	0	1	0	1	1	0	0	1
Bom método de avaliação	4	2	2	2	2	3	1	3	1
Não concorda com as penalizações	2	0	2	1	1	1	1	2	0
Gostei do Professor/a	4	3	1	3	1	2	2	2	2
Teste de “Repescagem” para todos	1	0	1	1	0	1	0	0	1
Preferia teste resposta aberta	2	0	2	1	1	1	1	1	1
Problemas no Moodle	2	1	1	2	0	1	1	0	2
Boa ideia existir este questionário	1	1	0	1	0	1	0	0	1
Refere que mão tem comentários	19	14	5	12	7	14	5	5	14
SOMA	36	21	15	23	13	25	11	13	23
N (Documentos)	371	168	203	261	110	297	74	107	264