UNIVERSIDADE LUSÓFONA

LUIS MIGUEL PIRES ALMEIDA

LITERACIA EM AVALIAÇÃO DE PROFESSORES

Desenvolvimento e Aplicação do Questionário de Aferição da Literacia em Avaliação (QALA)

Orientadora: Professora Doutora Glória de Magalhães Ramalho

Universidade Lusófona de Humanidades e Tecnologias

Faculdade de Ciências Sociais, Educação e Administração

Instituto de Educação

Lisboa

2021

LUIS MIGUEL PIRES ALMEIDA

LITERACIA EM AVALIAÇÃO DE PROFESSORES

Desenvolvimento e Aplicação do Questionário de Aferição da Literacia em Avaliação (QALA)

Tese defendida em Provas Públicas para a obtenção do Grau de Doutor no Curso de Doutoramento em Educação, conferido pela Universidade Lusófona de Humanidades e Tecnologias, no dia 17/12/2021, perante júri nomeado pelo Despacho de Nomeação n.º 294/2021 de 04 de novembro de 2021, com a seguinte composição:

Prof.ª Doutora Rosa Serradas Duarte (Presidente)
Universidade Lusófona de Humanidades e Tecnologias

Prof.ª Doutora Marília Cid (Arguente)
Universidade de Évora

Prof.ª Doutora Vera Monteiro (Arguente)
ISPA - Instituto Universitário

Prof.ª Doutora Elsa Estrela (Vogal)
Universidade Lusófona de Humanidades e Tecnologias

Prof.ª Doutora Constança Vasconcelos (Vogal)
Universidade Lusófona de Humanidades e Tecnologias

Prof. Doutor Sérgio Claudino (Vogal)
Universidade de Lisboa

Prof.ª Doutora Glória Ramalho (Orientadora)
Universidade Lusófona de Humanidades e Tecnologias

Universidade Lusófona de Humanidades e Tecnologias

Faculdade de Ciências Sociais, Educação e Administração

Instituto de Educação

Lisboa

2021

Dedicatória

Às mulheres da minha vida
Minha mãe,
Minhas irmãs,
Minha mulher,
Minha filha.

Agradecimentos

Muito embora a realização de uma tese de Doutoramento seja um caminho em grande parte solitário, há sempre pessoas que nos ajudam a seguir em frente.

Um especial agradecimento, em primeiro lugar, à Professora Doutora Glória Magalhães Ramalho, minha orientadora, pelo apoio e confiança demonstrada ao longo deste trajeto. O seu papel foi fundamental no domínio científico, mas também para me escutar e me guiar nas várias inquietações, dúvidas e inseguranças que marcaram este percurso.

A todos os professores do Instituto de Educação da Universidade Lusófona que, direta ou indiretamente, contribuiram para o sucesso do meu caminho, em especial a Professora Doutora Rosa Serradas que sempre me incentivou a concluir este doutoramento.

Aos meus colegas de Doutoramento, pelo apoio e incentivo, em especial o Professor Doutor João Robert Nogueira e a Dr.ª Liliane Barros.

A todos os diretores e professores dos estabelecimentos escolares do subsistema público e particular e cooperativo, em especial os meus colegas do Colégio Académico e do Colégio Atlântico, que permitiram a recolha de dados, fundamentais para a realização da presente investigação.

À minha mãe, Bia, mulher inspiradora que sempre acreditou e apoiou os meus sonhos. Ao meu padrasto, Beto, que sempre me tratou como um filho e sempre me apoiou para ir mais além.

Às minhas irmãs, Irina e Énia, pelo apoio e incentivo.

À minha mulher, Catarina, pelo apoio e carinho, fundamentais para o sucesso deste projeto.

À minha filha, Eva, pelas brincadeiras que não pudemos realizar. Mas, com ela e por ela, foi possível a conclusão deste Doutoramento.

Por fim, à Universidade Lusófona de Humanidades e Tecnologias que acreditou e financiou este projeto, por intermédio de uma bolsa de estudo.

Resumo

A avaliação das aprendizagens é uma das mais importantes responsabilidades dos professores, assim como uma das tarefas nas quais os professores despendem mais tempo (Mertler, 2003; Ramesal, 2011). Desta forma, as capacidades em avaliação são uma ferramenta fundamental que todo o professor deve dominar. Ao conjunto de capacidades em avaliação dá-se o nome de literacia em avaliação. O conceito de literacia em avaliação foi primeiramente apresentado por Stiggins (1991) como o conhecimento profundo das questões de avaliação. Do mesmo modo, em 1995, o mesmo autor refere que um educador/professor com literacia em avaliação sabe o que avaliar, a razão de avaliar, como avaliar, quais os possíveis problemas relacionados com a avaliação e como prevenir que esses problemas surjam no processo de ensino e aprendizagem. Para além disso, tem um conhecimento profundo dos efeitos negativos de uma má avaliação. Desta forma, a ausência de capacidades em avaliação, por parte do professor, é um fator que pode pôr em causa tanto a avaliação dos alunos como todo o processo de ensino e aprendizagem.

A presente investigação assentou em dois objetivos principais. Em primeiro lugar, pretendeu-se analisar as perceções que os professores, em exercício, tinham dos seus conhecimentos e capacidades em avaliação. Em segundo lugar, procurou-se aferir a literacia em avaliação desses mesmos professores. Destes dois objetivos principais derivaram uma série de objetivos específicos que possibilitaram uma

análise mais aprofundada da problemática em investigação. Para tal, foi criado um questionário desenvolvido para o efeito e ao qual designámos de QALA – Questionário de Aferição da Literacia em Avaliação. O QALA foi aplicado a um conjunto de 253 professores do ensino básico e secundário a lecionar na Zona Pedagógica de Lisboa e Península de Setúbal. A análise dos dados seguiu uma abordagem quantitativa do tipo *survey*. Procedeu-se, em primeiro lugar, a uma avaliação das propriedades psicométricas do QALA com recurso ao modelo Rasch. Posteriormente, foi realizada uma análise descritiva e inferencial, de forma a responder aos objetivos gerais e específicos propostos para a presente investigação.

Os resultados alcançados pela aplicação do modelo Rasch, parecem evidenciar boas qualidades psicométricas das 3 partes do QALA. Ficou também evidente que, embora os professores tenham valores positivos em relação às perceções sobre os seus conhecimentos e capacidades em avaliação (Parte 2), nos quatro domínios considerados, os resultados obtidos nas Partes 3 e 4 parecem demonstrar níveis inadequados de literacia em avaliação, o que pode ter implicações negativas em todo o processo de ensino e aprendizagem.

Palavras-chave: Literacia em Avaliação, Avaliação das Aprendizagens, Questionário de Aferição da Literacia em Avaliação, Professores em serviço, Modelo Rasch, Abordagem Quantitativa.

Abstract

Teachers' Assessment Literacy

Development and application of the Assessment Literacy Admeasurement Questionnaire (QALA)

Assessing students' learning is one of the most important responsibilities a teacher can have, as well as being one of the tasks teachers spend more time doing (Mertler, 2003; Ramesal, 2011). Thus, assessment skills are a primary tool that every teacher should master. The set of assessment skills is known as assessment literacy. The concept of assessment literacy was coined for the first time by Stiggins (1991) as the deep understanding of assessment issues. Similarly, in 1995, the same author refers that an educator/teacher with assessment literacy knows what to assess, the reason to do it, how to do it and the possible problems related with it and how to prevent that those problems occur in the teaching-learning process. Furthermore, the teacher has a deep knowledge about the negative effects of a poor assessment. This way, the lack of assessment skills is a factor that could lead to a poor judgment of the student's assessment, as well of all the teaching-learning process.

The present investigation was based in two main objectives. The first was to analyse the teachers' perceptions about their knowledge and assessment skills. The second was to measure the assessment literacy from those teachers. These two main

objectives originated a set of specific objectives, which enabled a deeper analysis in the investigation. For that purpose, a questionnaire was developed, called QALA - Assessment Literacy Admeasurement Questionnaire. It was implemented to a set of 253 teachers from the basic and secondary education levels teaching in the pedagogical zone of Lisbon and Setubal Peninsula. The data analysis followed a quantitative survey approach. Firstly, was initiated a psychometric assessment of the QALA properties, using the Rasch model. Subsequently, an explanatory inferential analysis was developed, as a way to respond to the specific objectives of the investigation.

The achieved results showed by the Rasch Model seem to demonstrate good psychometric skills in the 3 parts of QALA. It was also clear that, although teachers show positive results concerning the perceptions about their knowledge and assessment skills (part 2), in the four considered fields, the results concerning parts 3 and 4 seem to show inadequate levels of literacy skills, which can influence the teaching-learning process negatively.

Keywords: Assessment Literacy, Learning Assessment, Assessment Literacy Admeasurement Questionnaire, Inservice Teachers, Rasch Model, Quantitative Analysis.

Lista de Abreviaturas

ACPr	Análise de Componentes Principais dos Resíduos
AFE	Análise Fatorial Exploratória
AFT	American Federation of Teachers
ALI	Assessment Literacy Inventory
CALI	Classroom Assessment Literacy Inventory
CCI	Curva Característica dos Itens
DGS	Direção-Geral de Saúde
KMO	Teste de Keiser-Meyer-Olkin
MLQ	Measurement Literacy Questionnaire
NCME	National Council on Measurement in Education
NEA	National Education Association
OECD	Organização para a Cooperação e Desenvolvimento Económico

PCK	Pedagogical Content Knowledge (Conhecimento Pedagógico do Conteúdo)
QALA	Questionário de Aferição da Literacia em Avaliação
QZP	Quadro de Zona Pedagógica
SWOT	Strengths, Weaknesses, Opportunities and Threats
TALQ	Teacher Assessment Literacy Questionnaire
TCT	Teoria Clássica de Testes
TIC	Tecnologias de Informação e Comunicação
TRI	Teoria de Resposta ao Item
UMS	Unweighted Mean Squares
WMS	Weighted Mean Squares

Lista de Figuras

Lista de Tabelas

Conteúdo

Introdução

Contextualização

O conceito de literacia em avaliação foi apresentado pela primeira vez por Richard Stiggins, em 1991, como sendo o conhecimento que o professor tem sobre os vários aspetos relacionados com a avaliação das aprendizagens. A literacia em avaliação assume-se como uma característica fundamental de toda a prática pedagógica, uma vez que a avaliação dos alunos está numa estreita relação com todo o processo de ensino e aprendizagem, pelo que deverá ser considerado um elemento-chave na melhoria do ensino. Assim, a ausência de conhecimentos e capacidades em avaliação é um fator que pode colocar em causa tanto a avaliação dos alunos, como o processo de ensino e aprendizagem.

Vários estudos relacionados com a literacia em avaliação, mostram que uma grande parte dos professores apresenta uma fraca capacidade para desenvolver e aplicar métodos, técnicas e instrumentos de avaliação variados, bem como uma incapacidade para interpretar os resultados oriundos da aplicação desses mesmos instrumentos de avaliação (Koh, 2011). Estes aspetos podem originar falta de confiança na capacidade de avaliar dos professores, levando-os, em muitos casos, a avaliar os alunos da mesma forma em que foram avaliados quando eles mesmos

foram alunos (McGee & Colby, 2014).

A literatura especializada nesta temática sugere que, em geral, os professores possuem uma baixa literacia em avaliação pelo que estes continuam a utilizar formas de avaliação desajustadas e longe daquilo que são consideradas as boas práticas de avaliação das aprendizagens (Volante & Fazio, 2007; Xu & Brown, 2016). Tal facto pode ser explicado, em parte, devido a uma certa negligenciação, por parte das Universidades, uma vez que a formação inicial de professores não tem dado a devida atenção às questões da avaliação (McGee & Colby, 2014; Popham, 2011).

O estudo da literacia em avaliação dos professores assume uma especial relevância dado o espaço que a avaliação das aprendizagens assume no contexto de ensino e aprendizagem. Com efeito, professores com melhores níveis de literacia em avaliação têm mais competências em verificar e melhorar as aprendizagens dos alunos. Este aspeto reflete-se, por exemplo, na utilização de instrumentos de avaliação diversificados e adequados para cada contexto de avaliação e numa comunicação eficaz dos resultados de avaliação (*feedback*).

Objetivos do estudo

A problemática que nos propusemos analisar está assente em dois grandes objetivos. Em primeiro lugar, pretende-se analisar as percepções que os professores em exercício têm sobre os seus conhecimentos e capacidades em avaliação. Em segundo lugar, pretende-se aferir a literacia em avaliação dos professores do ensino básico e ensino secundário, nos quatro domínios em análise.

A partir destes dois objetivos principais derivam um conjunto de objetivos

específicos que permitem-nos compreender de uma forma mais aprofundada o tema em análise, nomeadamente:

1. Analisar as perceções sobre os conhecimentos e capacidades em avaliação dos professores em quatro domínios considerados, nomeadamente:

   (a) Conhecimentos sobre os objetivos e funções da avaliação;

   (b) Conhecimentos sobre currículo e sobre aquilo que é importante aprender e avaliar;

   (c) Conhecimentos sobre a utilização de instrumentos de avaliação diversificados;

   (d) Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação.

2. Aferir a literacia em avaliação dos professores do ensino básico e ensino secundário nos quatro domínios identificados no ponto anterior;

3. Analisar a relação entre as perceções sobre os conhecimentos e capacidades em avalição com algumas variáveis de contexto, nomeadamente sexo, idade, subsistema de ensino, tipo de habilitação, vínculo, experiência letiva, nível de ensino, área disciplinar e formação contínua em avaliação;

4. Analisar a relação entre a literacia em avaliação e algumas variáveis de contexto, nomeadamente sexo, idade, subsistema de ensino, tipo de habilitação, vínculo, experiência letiva, nível de ensino, área disciplinar e formação contínua em avaliação;

5. Analisar a eventual a relação entre a literacia em avaliação e as perceções sobre os conhecimentos e capacidades em avaliação;

Motivações

O interesse pela temática da avaliação das aprendizagens, em especial da literacia em avaliação, surgiu ainda no decorrer do meu mestrado em Ensino da História e da Geografia no 3°Ciclo do Ensino Básico e Secundário. No decorrer da minha formação inicial, senti que o programa curricular não deu uma especial relevância aos temas da avaliação, pelo que foi necessário um investimento pessoal extra para aprofundar os meus conhecimentos nestas matérias. De facto, uma sólida formação inicial é crucial para que o futuro professor possa adquirir e desenvolver competências essenciais para a sua prática pedagógica, pelo que a formação em avaliação deveria estar bem presente nos planos curriculares da formação inicial de professores.

De salientar, contudo, que tais lacunas, ao nível da formação em avaliação, não são uma realidade exclusivamente portuguesa. Estudos realizados em países tão diversos como Austrália, Canadá, Filipinas, Estados Unidos da América e Omã, evidenciam que a literacia em avaliação, tanto de professores em formação como professores em serviço, está muito abaixo daquilo que seria expectável e desejável. Segundo Popham (2018), baixos níveis de literacia em avaliação dos professores podem levá-los a cometer diversos erros dos quais o autor destaca a utilização de instrumentos de avaliação inadequados, a utilização incorreta de instrumentos de avaliação adequados e a não utilização de instrumentos de avaliação formativa. Estes aspetos, vêm reforçar a importância que a literacia em avaliação tem na prática pedagógica e no sucesso do processo de ensino e aprendizagem.

Estrutura da tese

A presente tese encontra-se organizada em quatro capítulos, subdivididos em vários subcapítulos e secções. O primeiro capítulo constitui-se como uma revisão da literatura das questões relacionadas com a avaliação das aprendizagens. Aqui são abordados assuntos como a distinção entre avaliação e classificação, a evolução das conceções teóricas em torno da avaliação, funções, modalidades, qualidade, métodos e instrumentos de avaliação e a avaliação no quadro legal português.

O segundo capítulo é uma revisão da literatura em torno da Literacia em Avaliação, em especial no conceito de literacia em avaliação, da sua importância para a melhoria da aprendizagem, dos domínios da literacia em avaliação e sobre alguns instrumentos que foram desenvolvidos com o objetivo de medir a literacia em avaliação.

O terceiro capítulo é dedicado às questões metodológicas. A presente investigação seguiu uma abordagem quantitativa do tipo *survey* que, segundo Gorard (2001), é especialmente indicada quando os dados necessários à investigação não existem e as questões de investigação em causa não são suscetíveis de uma verificação experimental. A recolha dos dados foi realizada por intermédio de um questionário desenvolvido para o efeito e o qual designámos de QALA – Questionário de Aferição da Literacia em Avaliação. O QALA foi desenvolvido em torno de quatro domínios da Literacia em Avaliação inspirados em Abell e Siegel (2011), nomeadamente:

- Domínio 1: Conhecimentos sobre os objetivos e funções da avaliação;
- Domínio 2: Conhecimentos sobre o Currículo e sobre aquilo que é importante aprender e avaliar;

- Domínio 3: Conhecimentos sobre a utilização de instrumentos de avaliação diversificados;

- Dimensão 4: Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação.

Embora existam alguns questionários desenvolvidos para aferir a Literacia em Avaliação, optámos por desenvolver um que pudesse ser aplicado ao contexto português e que tivesse como base as dimensões que se aproximassem ao nosso conceito de Literacia em Avaliação. O QALA encontra-se organizado em 4 partes. A primeira parte (Parte 1 - Informações Gerais) é constituída por itens que permitem caracterizar a amostra. A segunda parte (Parte 2 – Perceções sobre conhecimentos e capacidades em avaliação) é constituída por 20 itens do tipo *Likert* e procura recolher informações sobre a autoperceção dos professores face aos seus conhecimentos e capacidades em avaliação. A terceira parte (Parte 3 – Conhecimentos em Avaliação) é composta por 40 questões de verdadeiro e falso e procura recolher informações sobre os conhecimentos gerais relacionados com a avaliação das aprendizagens. Na quarta, e última, parte (Parte 4 – Cenários em contextos de avaliação) são colocadas 20 questões de escolha múltipla, organizadas em torno de 5 cenários hipotéticos em contexto escolar, relacionados com a avaliação. O QALA foi aplicado a um conjunto de 253 professores do ensino básico e secundário, a lecionar em estabelecimentos da rede pública e particular e cooperativa da Zona Pedagógica 7 – Lisboa e Península de Setúbal (QZP7). Embora estivesse prevista a recolha dos dados de forma presencial, dada a situação pandémica e as recomendações da Direção-Geral de Saúde (DGS), optou-se pela recolha à distância, através de uma versão online do QALA.

No quarto, e último, capítulo são apresentados e analisados os dados resultantes

da aplicação do QALA. Este capítulo subdivide-se em 3 partes. Na primeira parte procede-se à apresentação das Qualidades Psicométricas do QALA, determinadas a partir do modelo Rasch. Na segunda parte é realizada a análise descritiva dos dados, apresentando os resultados alcançados em cada uma das partes do QALA. A terceira parte é dedicada à análise inferencial dos dados, através de estatísticas não-paramétricas como o Teste *U* de Mann-Whitney, o Teste *H* de Kruskal-Wallis e a Correlação de Spearman (r_s), o que possibilita a análise das relações entre os resultados obtidos no QALA e algumas das variáveis de contexto recolhidas na Parte 1 e a existência de correlações entre as diferentes partes do QALA e os respetivos domínios que os constituem.

Capítulo 1

Avaliação das Aprendizagens

1.1 Clarificando os conceitos de avaliação e classificação

Em primeiro lugar, e recorrendo a um dicionário online da Porto Editora, o conceito de avaliação vem descrito como sendo *o estabelecimento do valor de algo* ou ainda como a *apreciação de competência ou progresso de um aluno*. Nestas duas definições há uma clara mistura entre os conceitos de avaliação e de classificação.

Proença (1989) faz uma distinção destes dois conceitos ao referir que a classificação *visa colocar um indivíduo numa escala adoptada, de acordo com os resultados que obteve nas provas a que foi submetido* (p.144). Já sobre a avaliação, a autora considera ser *um processo contínuo e sistemático que permite detectar em que medida os objectivos educacionais foram atingidos* (p.144). Desta forma, a classificação tem o objetivo de valorizar e seriar – ou seja, tem um caráter seletivo – ao invés da avaliação que tem um papel de descrever e informar – ou seja, tem um

caráter predominantemente formativo (p.145).

A definição de avaliação de Proença realça um aspeto importante, a definição de objetivos. Ralph Tyler considerado, por muitos, como o pai da avaliação educacional (Finder, 2004; Mathison, 2005; Stufflebeam, Madaus & Kallaghan, 2000), definiu avaliação como um processo que tem como propósito determinar se os objetivos educacionais estão a ser alcançados. Assim, a avaliação assumiu um caráter funcional, na medida em que ela se desenrola em função de um conjunto de objetivos previamente estabelecidos. Se até então a avaliação consistia na atribuição de notas consoante o grau dos alunos se aproximasse, ou não, do discurso do professor, com Tyler a avaliação passou a traduzir o grau de proximidade ou afastamento dos conhecimentos dos alunos face aos objetivos definidos, sendo este tipo de informação reinvestido no processo pedagógico (Valadares & Graça, 1998). Desta forma, o processo de avaliação correspondia à identificação de pontos fortes e fracos e à verificação da eficiência dos currículos escolares, procedendo-se a melhorias em caso de necessidade.

Benjamin Bloom, um dos pupilos de Tyler, define avaliação como um método de aquisição e processamento de evidências que permitam a melhoria das condições de ensino e aprendizagem (Bloom, Hastings & Madaus, 1971). Contudo, a averiguação dessas mesmas evidências não se deverá cingir aos testes sumativos, ou como designam os autores, pelo usual exame de papel e lápis. Desta forma, a avaliação assume-se como um sistema de controlo da qualidade, no qual pode ser determinada, a cada etapa do processo, o sucessso, ou não, das estratégias adotadas. Em caso de insucesso, poderão ser tomadas medidas que permitam remediar a situação, de forma a que tanto professores como alunos tenham consciência daquilo que poderão melhorar com vista ao sucesso escolar. Para tal o *feedback* é um elemento essencial. Ou seja, ao longo do processo de

ensino-aprendizagem, o professor deve realizar um balanço das aprendizagens adquiridas pelos alunos de forma a identificar possíveis fragilidades que existam. Esse balanço deverá permitir ao professor, por um lado, a adoção de medidas de remediação e, por outro, informar os alunos sobre o estado das suas aprendizagens, de forma a que possam corrigir eventuais dificuldades.

Embora deva haver uma prevalência da avaliação em relação à classificação, já que esta dá melhores informações acerca da progressão dos alunos ao longo do processo de ensino-aprendizagem, o que acontece em muitos casos é exactamente o oposto. Segundo Crahay (1999, citado por Ferreira, 2007, p.12) isto acontece pela frequência dos testes (que se traduzem numa classificação), pelo seu caráter normativo e pela grande importância que lhe é atribuída pelos pais, professores e sociedade em geral. Não quer isto dizer que a classificação não tenha as suas vantagens. Harlen (2007), por exemplo, refere que a classificação é importante na medida em que:

> is required for keeping records of the progress of individual students, reporting to parents and students at regular intervals, passing information to other teachers on transfer from class to class, or guiding decisions about subjects for further study (p.53).

Há a salientar, contudo, que um sistema de ensino baseado na classificação apresenta diversas limitações e desvantagens. Destaque-se o facto de não informar acerca da aprendizagem dos alunos e dos aspetos em que têm maior ou menor dificuldade, o que não contribui para o sucesso escolar. Assim, um sistema de classificação não se constitui como uma medida clara de aproveitamento, visto reduzir a um símbolo toda uma gama de informação variada (Ribeiro, 1990, p.78).

De forma a complementar o sistema de classificação, o professor deve munir-se de várias formas e instrumentos de avaliação para que, como afirma Fernandes (2004), a avaliação não se reduza a *pouco mais do que a administração de um ou mais testes*

e à atribuição de uma classificação em períodos determinados (p.11).

A avaliação deve ser então encarada como um sistema que integra duas componentes basilares, uma componente qualitativa e uma componente quantitativa (classificação). Estas duas componentes não deverão ser consideradas opostas mas sim complementares, ou seja, para que a avaliação seja efetiva, ambas terão de estar presentes. Como afirma Demo (2008), não se trata de estabelecer uma polarização radical e estanque entre a componente qualitativa e quantitativa da avaliação, como se uma fosse a perversão da outra. Cada componente tem a sua própria razão de ser e de agir na realidade. Assim, são ambas necessárias para que haja um verdadeiro sistema de avaliação. Por um lado, temos a componente qualitativa que tem como principal finalidade o acompanhamento e potencialização do ensino e aprendizagem (Santos, 2016) e, por outro, a componente quantitativa que é a tradução para uma escala pré-estabelecida da componente qualitativa.

1.2 Evolução das conceções teóricas em torno da avaliação

A história da avaliação é quase tão antiga como a própria história da humanidade. Exemplo disso mesmo é a existência de registos de exames escritos realizados pelos chineses datados de 2000 a.C. Estes exames, eram utilizados no exército chinês para a seleção dos seus oficiais (Pinto & Santos, 2006). Também na antiga civilização grega, os mestres utilizavam o questionamento como metodologia primordial de ensino (Neves & Ferreira, 2015).

Durante a idade média, os aprendizes partilhavam o seu quotidiano com os seus

respetivos mestres o que permitia o desenvolvimento das suas aprendizagens. Estas partilhas eram analisadas com o objetivo de evitar fracassos futuros. Já na Universidade, segundo Neves e Ferreira (2015), privilegiavam-se as disputas intelectuais, o que possibilitava o desenvolvimento da argumentação e a avaliação das capacidades e conhecimentos a ela associados. A partir do século XVI, deu-se uma generalização da utilização dos exames por parte dos Jesuítas, uma vez que estes:

> preconizavam o ensino de muitos como se fossem um só, atingindo o seu apogeu no período de ascensão plena da burguesia ao controlo do poder em termos sociais, isto é, com a Revolução Francesa. Nos ideais de liberdade, fraternidade e igualdade, onde a Escola Pública se ancorou, os exames inscreviam-se num conjunto de práticas que procuravam combater os privilégios da aristocracia obtidos por nascimento e fortuna (Pinto & Santos, 2006, p.11-12).

No século XIX, surgiram diversos exames e diplomas, bem como um conjunto de reformas dos sistemas educativos o que levou ao surgimento das turmas. Também neste século, deu-se o desenvolvimento da docimologia[1] *como resposta a preocupações com a qualidade da informação recolhida – validade e fiabilidade das provas de exame, escritas e orais* (Neves & Ferreira, 2015, p.28).

Embora tenhamos verificado a presença da avaliação ao longo dos tempos, as conceções de avaliação, tal como as conhecemos hoje, têm uma história bastante recente (pouco mais de um século). Guba e Lincoln (1989) identificaram quatro gerações, ou conceções, de avaliação ao longo do século XX, a saber:

1. avaliação como medida;

2. avaliação como descrição;

[1] Ciência que tem por objeto o estudo sistemático dos exames, em particular do sistema de atribuição de notas e do comportamento dos examinadores e dos examinandos (Landsheere, 1976).

3. avaliação como juízo;

4. avaliação como negociação e construção.

1.2.1 Avaliação como Medida

Esta conceção de avaliação prevaleceu nos primeiros anos do século XX (Lucea, 2005). A sua principal finalidade era *medir los aprendizajes que los alumnos han hecho y que éstos puden manifestar através de la conducta o de outros procedimientos* (Lucea, 2005, p.22). Já Fernandes (2004) refere que a avaliação era uma questão essencialmente técnica que, através de testes bem construídos, permitia medir com rigor e isenção as aprendizagens escolares dos alunos.

Nesta perspetiva, a avaliação era considerada como um instrumento que media os conteúdos assimilados pelos alunos. Furlan (2007) salienta que, nesta abordagem, as questões passaram a valer pontos que somados e divididos davam a média de quanto o aluno apreendeu (e não aprendeu). A autora refere que as notas obtidas pelos alunos não refletem a aprendizagem, na medida em que o saber não é mensurável, não é algo que tenha tamanho, peso, volume ou quantidade. Depreende-se, das palavras da autora, que a utilização dos testes não é a melhor opção para avaliar as aprendizagens dos alunos já que estes refletem uma pequena parte das aprendizagens alcançadas sendo necessário, por isso, o recurso a outros tipos de instrumentos de avaliação.

Para Henderson (1978), esta abordagem promoveu uma forte dependência com a classificação e com outros índices susceptíveis de serem manipulados matematicamente ou estatisticamente. Assim, as variáveis que não podiam ser medidas tendiam a ser ignoradas, o que se traduz numa grave limitação à utilidade de tal conceção, dado que não avalia o desenvolvimento do aluno relativamente à

sua autonomia moral e cognitiva, nem a sua capacidade de convivência e interação (Furlan, 2007).

Fernandes (2004) reconhece que muitas das características desta conceção de avaliação se mantiveram e têm influência nos sistemas educativos atuais. Contudo, apresenta um conjunto de argumentos que comprovam as limitações desta abordagem, nomeadamente:

1. Prevalecem as funções sumativa, classificativa e selectiva da avaliação;
2. O único objeto da avaliação são os conhecimentos;
3. Há pouca, ou nenhuma, participação dos alunos no processo;
4. A avaliação é, em geral, descontextualizada;
5. Privilegia a quantificação das aprendizagens em busca da objetividade e da neutralidade do professor (avaliador);
6. A avaliação é referida a uma norma ou padrão [...] e, por isso, os resultados de cada aluno são comparados com os de outros grupos de alunos (Fernandes, 2004).

1.2.2 Avaliação como Descrição

A conceção de avaliação como descrição é considerada por Guba e Lincoln (1989) como *an approach characterized by description of patterns of strenghts and weaknesses with respect to certain stated objective* (p.28). Esta abordagem surgiu nos Estados Unidos da América nos anos 1950 e *procurou superar algumas das limitações detectadas nas avaliações da primeira geração* (Fernandes, 2004, p.11). Esta conceção de avaliação foi fortemente influenciada pelos trabalhos de Ralph Tyler que, através da definição de objetivos educacionais, permitiu aos professores/avaliadores descreverem as diferenças e semelhanças entre os resultados alcançados pelos alunos e os referidos objetivos educacionais definidos.

Desta forma, a avaliação deixou de se centrar unicamente nos resultados alcançados, passando a descrever em que medida os alunos realizavam as aprendizagens face a um conjunto de objetivos de aprendizagens pré-definido.

1.2.3 Avaliação como Juízo

Esta nova abordagem à avaliação surgiu no início dos anos 1960 com o intuito de, tal como no caso anterior, superar algumas falhas e pontos fracos existentes. Assim, *sentiu-se que se deveriam fazer esforços para que as avaliações permitissem formular juízos de valor acerca do objecto de avaliação* (Fernandes, 2004, p.11). Para Guba e Lincoln (1989), nesta abordagem

> [the] evaluation was characterized by efforts to reach judgements, and in which the evaluator assumed the role of judge, while retaining the earlier technical and descriptive functions as well (p. 30).

Para serem emitidos os juízos, a avaliação teve de assumir um caráter sistemático (Lucea, 2005). Este caráter sistemático foi necessário para que fosse possível comparar os objetivos definidos e os resultados alcançados. Para que pudessem ser observáveis e mensuráveis, os objetivos tinham de ser formulados sob a forma de comportamentos. Para a mensuração dos objetivos alcançados utilizavam-se os mesmos tipos de instrumentos que nas conceções anteriormente descritas, isto é, exames e testes, por exemplo.

É ao longo do período em que esta conceção predomina que se deram importantes avanços no domínio da avaliação das aprendizagens. Destaque-se, por exemplo, o desenvolvimento da taxonomia de Bloom que, através de objetivos bem delineados e devidamente hierarquizados, consoante o seu grau de complexidade, permitiu observar o comportamento dos alunos segundo três domínios: cognitivo,

afetivo e psicomotor. Foi desenvolvida assim uma *Pedagogia da Mestria* (Neves & Ferreira, 2015) que tinha como pressuposto que um aluno não saltasse etapas na sua aprendizagem sem que as anteriores estivessem plenamente dominadas.

Foi também neste período que, graças aos trabalhos de Scriven (1967), se dá a distinção entre avaliação sumativa e formativa, sendo que a primeira tinha como finalidade estudar os resultados alcançados e a segunda em recolher informações contínuas para se proceder às reformulações essenciais no sentido de regular as aprendizagens (Afonso, 2011).

1.2.4 Avaliação como Negociação e como Construção

A avaliação como Negociação e como Construção, definida por Guba e Lincoln (1989), procura romper com as conceções atrás descritas e estabelecer-se como uma verdadeira alternativa. Nesta abordagem, *começa-se a sobrevalorizar a avaliação formativa, com tendência para uma avaliação formativa alternativa em que se coloca grande realce em quem aprende* (Afonso, 2011, p.10). A ideia central desta conceção é o não estabelecimento, à priori, de parâmetros sendo que estes vão sendo definidos *através de um processo negociado e interactivo com aqueles que, de algum modo, estão envolvidos na avaliação* (Fernandes, 2004, p.13).

Sendo esta uma abordagem construtivista, a avaliação deve assentar num conjunto de princípios, entre os quais se destacam:

1. a partilha do poder de avaliar entre professores, alunos e outros intervenientes (*p.e.* Encarregados de Educação);

2. o predomínio da função formativa da avaliação, ao invés de um sistema que julga ou classifica os alunos numa escala;

3. a utilização do *feedback* nas suas mais variadas formas;

4. a utilização de métodos predominantemente qualitativos, embora não se coloquem de parte os métodos quantitativos.

Assim, nesta quarta geração, definida por Guba e Lincoln (1989), a avaliação assume uma função pedagógica uma vez que, segundo Marinho, Fernandes e Leite (2014), incide diretamente no processo de ensino e aprendizagem, tendo como função melhorar as aprendizagens dos alunos mais do que os classificar. Ainda segundo os mesmos autores, nesta conceção de avaliação, o poder de avaliar é partilhado entre professores e alunos que, numa base de constante *feedback*, fortalece as aprendizagens.

1.3 Funções da Avaliação

Em todo o processo de avaliação, há que haver uma consciência clara sobre quais as funções da avaliação. Tal como as conceções teóricas da avaliação, também as funções da avaliação se foram atualizando ao longo do tempo. São reconhecidas duas funções principais da avaliação: informar sobre o desempenho dos alunos no processo de ensino e aprendizagem e informar sobre a eficiência das práticas de ensino (Berry, 2008; Harlen, 2007).

Jean Cardinet (1983) propôs 3 funções distintas da avaliação, nomeadamente as funções de regulação, certificação e orientação/seleção, sendo que, para cada uma delas, deveriam existir ferramentas especifícas. A função de regulação deverá ser entendida como *um processo deliberado e intencional que visa controlar os processos de aprendizagem, para que se possa consolidar, desenvolver ou*

redirecionar essa mesma aprendizagem (Fernandes, 2005, p.67). Deste modo, Perrenoud (2001) sugere uma recolha constante de informações que permitam otimizar as estratégias de ensino. A ausência dessa recolha e reflexão, sobre a mesma, poderá levar à criação de clivagens, de díficil remediação, entre os alunos.

A função de certificação corresponde a um processo que visa condicionar os fluxos de entrada e de saída do sistema de escolar, bem como as passagens entre os diferentes subsistemas, classes e cursos (Afonso, 1998). Esta função parte da necessidade de *dar garantias válidas sobre o domínio das aprendizagens, através da hierarquização e da seleção dos alunos, e ainda, da respectiva certificação daqueles que cumprem os requisitos de excelência pretendidos* (Ferreira, 2007, p.19).

A avaliação como orientação/seleção visa compreender as aptidões de um aluno face a novas situações de aprendizagem. É a partir da análise da posse, ou não, de tais aptidões que os alunos poderão ser, por um lado, selecionados para determinadas instituições/cursos ou, por outro, serem orientados para outros percursos escolares. Esta característica da avaliação é explicada de uma forma muito interessante por Ted Wragg ao referir que a seleção:

> is a word with immense political significance. Yet assessment is often linked with selection, whether anyone likes it or not. Some pupils apply for entry to schools that have entrance tests; many schools have higher and lower sets based on ability in particular subject fields; children are picked for school teams, performances in concerts, parts in plays or other dramatic events; and teachers write references for former pupils who are applying for jobs. It may be done informally or semi-formally, but it exists (Wragg, 2001, p.29).

Contando com a função de orientação/seleção, Wragg (2001) considera que as funções da avaliação podem ser organizadas em seis categorias distintas conforme a Figura 1 (embora admita que essa lista não seja exaustiva).

A função de informar, ou de *feedback*, surge da necessidade e curiosidade que os

Figura 1: Categorias de Funções da Avaliação

(Baseado em: Wragg, 2001)

alunos têm em saber, por exemplo, o grau de compreensão de um conceito, princípio ou habilidade. Assim, o *feedback* pode ser entendido como sendo a informação com a qual um aluno pode confirmar, adicionar, reescrever, afinar ou reestruturar informações existentes na memória, mesmo que a informação seja do domínio do conhecimento, do conhecimento metacognitivo, de crenças sobre si mesmo e sobre tarefas e estratégias cognitivas (Lopes & Silva, 2010). Esta informação é, geralmente, fornecida pelos pares, professores, tutores e amigos, pode ser realizada de um modo formal ou informal (Irons, 2008) e deve permitir ao aluno se deslocar de onde se encontra para onde tem como objetivo ir. Assim, as informações disponibilizadas devem ter um duplo enfoque formativo, ou seja, deve envolver o factor cognitivo, de forma a que o aluno compreenda o ponto em que se encontra na sua aprendizagem e o que pode fazer para a melhorar, e o fator motivacional, de forma a que o aluno sinta que tem o controlo sobre a sua aprendizagem (Irons, 2008).

A função de apoio e encorajamento é utilizada em várias situações através, por exemplo, da estimulação dos alunos para um estudo mais aprofundado de um determinado conteúdo programático. Estamos também perante esta função quando um professor, de forma deliberada, inflaciona as notas de determinados alunos de forma a encorajá-los para as novas situações de aprendizagem que se avizinham. A avaliação pode ser ainda utilizada para um resultado oposto:

> Assessment may even be used cynically to achieve the opposite, that is, to undermine and demoralise. This may be a comparative rarity in education, but it is not unknown in other fields, like initial training in the armed services and in some professional sports, where ”taking somebody down a peg”is a part of the culture (Wragg, 2001, p.27).

Uma das principais funções da avaliação é o de motivar os alunos. Ou seja, os alunos deverão querer fazer um esforço e estarem dispostos a continuar, mesmo quando acham a aprendizagem difícil (Nezvalová, 2010). Por outras palavras, a avaliação aqui deverá ser encarada como incentivadora da aprendizagem, dando um maior ênfase aos progressos realizados ao invés dos fracassos. Stefanou e Parker (2003) referem ainda que, quando a avaliação é usada como uma forma de motivação, os alunos tedem a preocupar-se mais com as suas aprendizagens e acreditam que o esforço é um mecanismo fundamental para alcançar tais aprendizagens.

Há ainda a considerar uma função corretiva na medida em que avaliação formativa deve permitir corrigir os erros cometidos ao longo do processo de ensino e aprendizagem (Hadji, 2001). Trata-se, assim, de uma função pedagógica da avaliação que não visa a sanção e a punição do aluno, porque os seus erros são considerados normais no percurso de aprendizagem (Ferreira, 2007), bem como uma fonte de informação importante, quer para a própria aprendizagem dos alunos, quer para o diagnóstico de dificuldades (Amigues & Zerbato-Poudou, 1996).

Quanto às funções de diagnóstico e de medição/comparação da avaliação, e dada a sua relevância, serão analisadas no subcapítulo seguinte.

1.4 Modalidades da Avaliação

Em relação às modalidades de avaliação, são consideradas tradicionalmente a avaliação de diagnóstico, a avaliação formativa e a avaliação sumativa. Cada uma destas modalidades apresenta funções e finalidades distintas, bem como diferentes *momentos de avaliação que se podem distinguir entre antes, durante e depois do processo de aprendizagem* (Ferreira, 2007, p.23).

1.4.1 Diagnóstica

A avaliação diagnóstica, também designada de avaliação inicial ou de pré-requisitos (Nérici, 1983), visa a recolha de informações acerca da posição do aluno face a novas aprendizagens que lhe irão ser propostas (Ribeiro, 1993). Esta dimensão da avaliação, como o próprio nome indica, permite ao professor realizar um diagnóstico da situação e "prescrever" as medidas que se afigurem adequadas face aos objetivos que se pretendem atingir (Ribeiro & Ribeiro, 1990). No entanto, os resultados obtidos neste tipo de avaliação não podem servir para rotular os alunos (Luckesi, 2013), mas sim para se estabelecer como um ponto de partida a partir do qual os alunos e o professor, em conjunto, procurarão um progresso na aprendizagem (Ferreira, 2007). A ideia de que a avaliação diagnóstica deverá ser realizada no início do ano letivo é errada. Esta deverá ocorrer no início de novas aprendizagens, não estando ligada a qualquer período de tempo (Ribeiro, 1993).

1.4.2 Formativa

A avaliação formativa é, talvez, uma das temáticas mais abordadas no contexto da avaliação educacional (Bennet, 2011) e considerada por muitos como um elemento fundamental no processo de ensino e aprendizagem.

A noção de avaliação formativa foi proposta por Scriven, em 1967, no âmbito da avaliação de programas sociais (Alves, 2004). Mais tarde, o conceito de avaliação formativa foi aplicada ao contexto educacional, sendo então considerada como *[a] systematic evaluation in the process of curriculum construction, teaching, and learning for the purposes of improving any of these three processes* (Bloom, Hastings & Madaus, 1971, p.155). Desta forma, a avaliação formativa assume um papel preponderante no processo de ensino-aprendizagem, uma vez que permite um acompanhamento permanente da natureza e qualidade da aprendizagem de cada aluno, orientando a intervenção do professor de modo a dar-lhe a possibilidade de tomar as decisões adequadas às capacidades e necessidades dos alunos (OCDE, 2005).

Sadler (1989) afirma que a função formativa da avaliação incide sobre *how judgements about quality of students' responses can be used to shape and improve their competences by short-circuiting the randomness and inefficency of trial-and-error learning* (p.120). Na mesma linha de raciocínio Gipps (1994) refere que a avaliação formativa utiliza a informação recolhida, no processo de avaliação, para fornecer *feedback* (tanto aos professores como aos alunos) de forma a que essa informação possa ser reintroduzida no processo de ensino e aprendizagem.

Num artigo de referência, no contexto da avaliação formativa, Black e William (1998a) referem-se à avaliação formativa como sendo *all those activities undertaken*

by teachers (and by their students in assessing themselves), which provide feedback to shape and develop the teaching and learning activities in which both teachers and students are engaged (p.7). Black e William (1998a, 1998b), após uma análise a várias obras e trabalhos de investigação (onde se inclui uma em Portugal), concluiram que a avaliação formativa apresenta várias evidências firmes de que melhora significativamente a aprendizagem. Partindo deste pressuposto, Irons (2008) elencou, a partir da análise de vários autores, um conjunto de razões que validam a afirmação anterior. Assim, segundo Irons (2008), a avaliação formativa é um elemento fundamental na melhoria das aprendizagens dos alunos na medida em que:

1. fomenta um ambiente de diálogo entre professores e alunos, o que facilita a partilha de preocupações e dificuldades por parte dos alunos;
2. os alunos tendem a arriscar mais num contexto de avaliação formativa, uma vez que correm menos 'riscos' do que aquilo que acontece normalmente na avaliação sumativa;
3. cria a possibilidade de diálogo e negociação sobre as atividades formativas a realizar e sobre a forma de as adequar às necessidades dos alunos;
4. cria um ambiente propício para que os alunos melhorem os seus conhecimentos e compreensão, ao invés de se concentrarem apenas em passar na avaliação sumativa;
5. favorece a autonomia no processo de aprendizagem;
6. favorece a capacidade dos alunos em se auto-avaliarem;
7. contribui para a capacidade do aluno em desenvolver uma aprendizagem reflexiva[2];
8. cria um ambiente favorável à avaliação por pares e auto-avaliação.

Para desenvolver um ambiente formativo, considerando os aspetos elencados anteriormente, é necessário ter em conta um conjunto de fatores que Black e William (1998a) consideram fundamentais, como sejam:

1. o reconhecimento da importância do *feedback*;

[2]Sobre a aprendizagem reflexiva recomenda-se a leitura da obra *A Handbook of Reflective and Experiental Learning - Theory and Practice*, de Jennifer Moon.

2. o envolvimento dos alunos no processo de ensino e aprendizagem;
3. a adaptação do processo de ensino às necessidades e características dos alunos;
4. a compreensão do impacto que a avaliação tem na motivação dos alunos;
5. o reconhecimento da importância da auto-avaliação de forma a possibilitar aos alunos a identificação dos seus pontos fracos e necessidades.

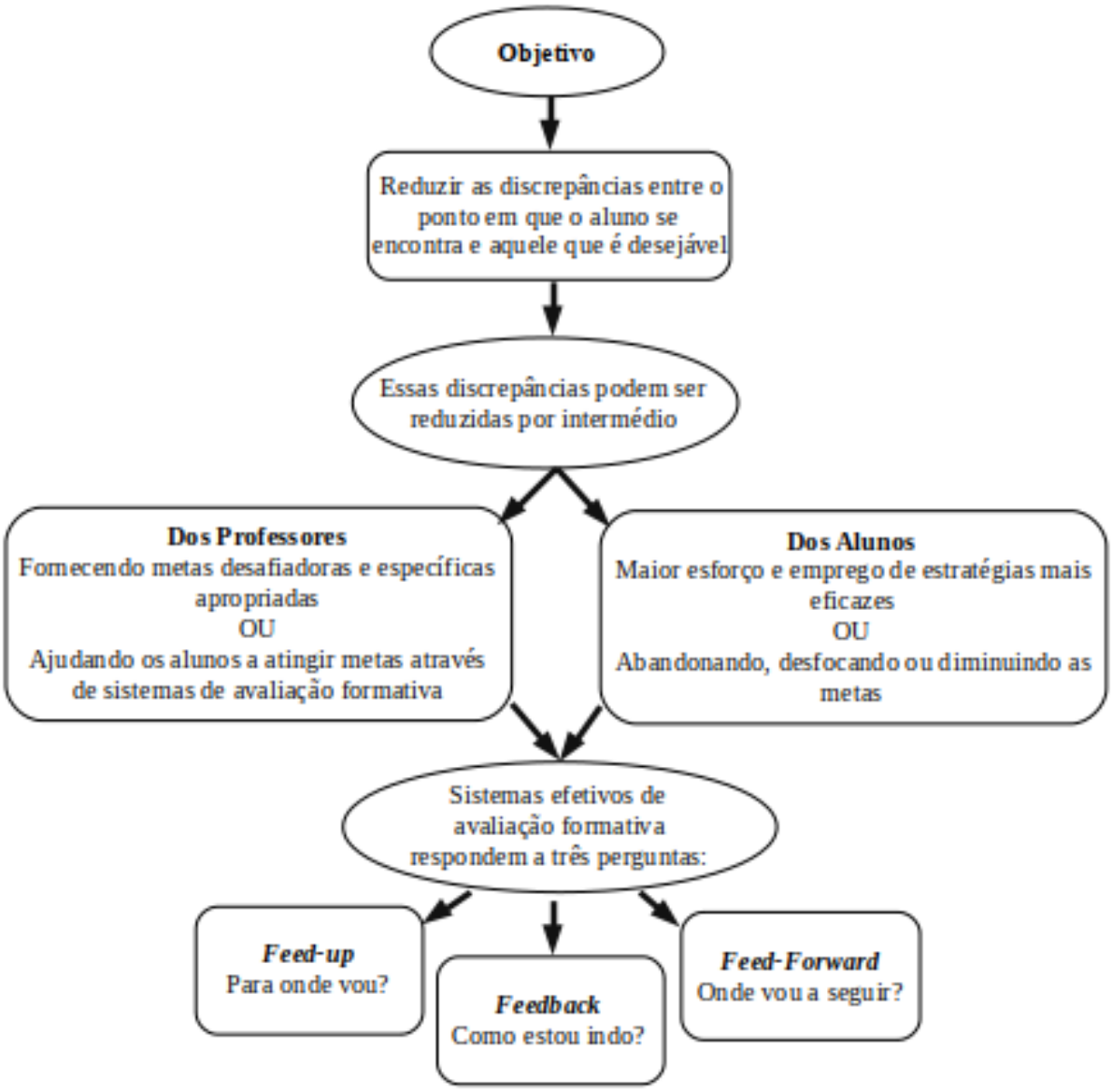
Ao contrário do que acontece na avaliação sumativa, a avaliação formativa incide sobre o processo de aprendizagem, sendo uma forma de recolha de informação que permite ao professor percecionar se os alunos atingiram o objetivos educacionais propostos. Assim, o professor possui uma ferramenta importante para que possa adequar os seus métodos de ensino, de forma a ir ao encontro das necessidades dos alunos. Para além disso, Black e William (1998b) verificaram que o reforço da avaliação formativa possibilita a melhoria das aprendizagens dos alunos. Este importante contributo foi verificado através da análise dos resultados dos alunos nos testes de avaliação. Assim, os alunos que beneficiaram de um ambiente de avaliação formativa, em comparação com o grupo que não beneficiou de tal ambiente, alcançou médias mais elevadas, mesmo no caso dos alunos que vinham apresentando algum insucesso escolar ou com dificuldades de aprendizagem.

Seguindo esta perspetiva, Luckesi considera a avaliação formativa como um verdadeiro ato de amor (Luckesi, 2005; Neto & Aquino, 2009) na medida em que é um ato acolhedor, integrativo e inclusivo que acompanha permanentemente o aluno na sua trajetória de construção do conhecimento. Assim, este ato amoroso da avaliação formativa é visível no facto de o professor se disponibilizar para observar constantemente os alunos, não para os julgar, mas sim para *criar estratégias de superação dos limites e ampliação das possibilidades* (Neto & Aquino, 2009, p.224) de forma a garantir as suas aprendizagens.

Para um verdadeiro sistema de avaliação formativa existir em sala de aula, Frey e Fisher (2011), a partir do trabalho de Hattie e Timperley (2007), afirmam que é necessária a presença de três componentes fundamentais: o *feed-up*, o *feedback* e o *feed-forward* (ver Figura 2). O *feed-up* deverá procurar garantir que os alunos compreendem os objetivos das tarefas propostas ou das aulas propriamente ditas, incluindo a forma como eles serão avaliados. O *feedback* deverá fornecer aos alunos (e aos professores) informações sobre os seus sucessos e necessidades. Por último, o *feed-forward* deverá orientar a aprendizagem dos alunos, tendo em consideração os dados recolhidos acerca do seu desempenho. Segundo Frey e Fisher (2011), a presença destas três componentes contribui para níveis elevados de aprendizagem dos alunos. Já a ausência de uma ou mais componentes pode colocar a aprendizagem em risco. Quando, por exemplo, um aluno não entende o objetivo de uma aula ou de uma tarefa proposta (*feed-up*), o empenho tende a ser significativamente inferior uma vez que, sem um objetivo claro, os alunos não são motivados e não vêem a relevância do conteúdo que deve ser dominado (Frey & Fisher, 2011). A ausência de *feedback* leva a que os alunos não tenham a certeza sobre o seu desempenho. Desta forma, tendem a assumir que o seu percurso está correto, não procedendo às devidas correções no seu processo de aprendizagem. Por último, quando os professores deixam de planear as aulas e as tarefas tendo em consideração o desempenho dos alunos (*feed-forward*), os conceitos erróneos são reforçados, os erros não são abordados e as lacunas no conhecimento persistem.

Figura 2: Um Sistema de Avaliação Formativa

(Adaptado de Frey e Fisher, 2011)



Os vários fatores referenciados anteriormente demonstram que o *feedback* se assume como um elemento fundamental dentro da função formativa da avaliação. Hattie (1999) afirma mesmo que uma 'grande dose' de *feedback*, entre professores e alunos, é a melhor prescrição para melhorar tanto o ensino como a aprendizagem. Para que isso aconteça, Hattie (1999) revela que é necessária uma combinação eficaz entre metas e *feedback*. Quando tal acontece, as metas podem, aos poucos, tornar-se mais exigentes. Assim, e existindo a tal combinação eficaz entre os dois

fatores referidos, há uma maior probabilidade de os alunos tomarem a iniciativa de procurar, receber e assimilar as informações de *feedback*.

Em *The Power of Feedback*, Hattie e Timperley (2007) identificam 4 níveis de *feedback*: Quanto à tarefa, quanto ao processo, como auto-regulação e como consciência pessoal. O primeiro nível de *feedback* (quanto à tarefa), procura dar informações sobre o quão bem uma determinada tarefa está a ser realizada. Neste nível, deverão ser fornecidas informações que permitam ao aluno compreender se a forma como desempenhou as tarefas sugeridas foi ou não bem sucedida. Desta forma, é dada a possibilidade ao aluno de melhorar ou corrigir a forma como desempenhou as tarefas propostas. O *feedback* quanto ao processo incide no procedimento utilizado para criar um produto ou completar uma tarefa (Hattie & Timperley, 2007). Segundo Lopes e Silva (2010), o *feedback* quanto à tarefa:

> É um tipo mais directamente orientado para o processamento da informação ou processo de aprendizagem necessário para compreender ou completar a tarefa. Deve fornecer pistas ou utilizar perguntas para ajudar os alunos a desenvolver o seu processo de pensamento ou uma estratégia, por exemplo, para a prossecução de uma determinada hipótese. (p.52)

Relativamente ao *feedback* como auto-regulação, este deve centrar-se, conforme o próprio nome indica, no apoio ao aluno na sua auto-regulação. Isto significa que as informações que são transmitidas aos alunos devem possibilitar um maior desenvolvimento das suas capacidades em se auto-avaliarem, bem como dar-lhes uma maior confiança para que se envolvam na tarefa ou tarefas propostas. Por último, o *feedback* como consciência pessoal apela aos atributos pessoais e, normalmente, é realizado sob a forma de elogio sendo, por isso, de menor eficácia na melhoria das aprendizagens quando comparado com os outros três níveis (Hattie & Timperley, 2007; Lopes & Silva,2010). Mais do que fornecer informações sobre a

tarefa a desenvolver, este nível de *feedback* dirige-se mais ao *Self* dos alunos, servindo muitas vezes como uma forma de motivação para o seu envolvimento nas tarefas propostas. Podemos concluir então que é um lado mais afetivo do *feedback*. No entanto, quando esta componente afetiva do *feedback* é negligenciada pode levar os alunos, a resultados indesejáveis e fazer aumentar o medo do fracasso (Fonseca *et al.*, 2015).

Já o *feedback* assente apenas na comunicação dos resultados, sejam eles de caráter quantitativo ou qualitativo, são muito pouco eficazes, na medida em que, como refere Black *et al.* (2004), os alunos utilizam-no apenas para estabelecerem comparações com os colegas e não para melhorarem as suas aprendizagens.

Por último, há que frisar que os professores não são, nem devem ser, a única fonte de *feedback* (Ozan & Kincal, 2018), também os próprios alunos deverão ser integrados nesse processo, seja através de processos de auto-avaliação ou de avaliação entre pares. A auto-avaliação pode ser aqui entendida como um processo onde os alunos refletem e criticam o próprio trabalho, tendo em consideração as suas expectativas e os objetivos ou critérios definidos, possibilitando ao aluno rever o seu trabalho e promovendo, dessa forma, a sua aprendizagem (Andrade, 2019; Ozan & Kincal, 2018). Já a avaliação entre pares possibilita aos alunos comentarem os trabalhos e realizações um dos outros, contribuindo, dessa forma, para a criação de uma cultura mais participativa dentro do ambiente de aprendizagem (Ozan & Kincal, 2018). Estudos levados a cabo por Black e William (2001 *apud* Monteiro & Fragoso, 2005) mostraram que práticas assentes em processos de auto-avaliação e de avaliação entre pares podem trazer benefícios para o processo de ensino e aprendizagem já que possibilitam aos alunos: 1) aprenderem a avaliar o seu trabalho e o trabalho dos outros, desenvolvendo critérios para julgarem a sua qualidade; 2) desenvolverem hábitos e capacidades de colaboração na aprendizagem; 3)

tornarem-se participantes, e não vítimas, no processo de avaliação.

1.4.3 Sumativa

A avaliação sumativa foi, e em muitos casos continua a ser, a principal modalidade de avaliação que os professores utilizam em sala de aula (Fautley & Savage, 2008; Harlen, 2007). Este distingue-se dos demais tipos de avaliação atrás descritos quer pela intenção que lhe preside, quer pela estrutura que apresentam os instrumentos que se enquadram neste tipo de avaliação (Ribeiro & Ribeiro, 1990). Tal como na avaliação formativa, o conceito de avaliação sumativa foi proposto por Scriven e, mais tarde, aplicada por Bloom ao contexto da avaliação das aprendizagens, sendo considerada uma forma de avaliação muito geral e de servir como suporte à atribuição de classificações, ou seja, tem um caráter predominantemente quantitativo (Bloom, Hastings & Madaus, 1971).

No seguimento da definição anterior, Irons (2008) entende a avaliação sumativa como sendo:

> Any assessment activity which results in a mark or grade which is subsequently used as a judgement on student performance. Ultimately judgements using summative assessment marks will be used to determine the classification of award at the end of a course or programme (Irons, 2008, p.7).

A definição apresentada tem como elemento central uma das principais finalidades atribuídas à avaliação sumativa, o de ajuizar o progresso realizado pelos alunos no final de uma unidade de aprendizagem. Para além desta finalidade, Irons (2008), a partir de uma revisão de literatura, aponta para um conjunto de funções que podem ser atribuídas à avaliação sumativa, tais como:

- Medir a capacidade de compreensão do aluno;

- Observar o comportamento dos alunos e produzir dados para serem utilizados no julgamento dos alunos face ao que eles sabem ou não sabem;
- Fornecer informações à comunidade escolar (professores, pais, alunos, administradores, entre outros) que lhes permitam tomar medidas e decisões face à realidade em causa;
- Ser utilizado como uma medida de sucesso da aprendizagem e do ensino, servindo também, em muitos casos, para julgar a competência dos professores e, em último caso, das escolas como um todo;
- Motivar a aprendizagem dos alunos, uma vez que ao terem consciência da existência de uma avaliação sumativa, os alunos dedicam-se mais nas tarefas escolares para que, no momento de avaliação, possam alcançar melhores resultados;
- Preparar para a vida adulta, na medida em que para alcançarem determinados empregos são sujeitos as provas de estrutura mais ou menos semelhante às realizadas em contexto escolar.

Embora sejam reconhecidas estas funções à avaliação sumativa, na verdade são apontadas diversas fragilidades a esta modalidade de avaliação. Biggs (1996) sugere mesmo que a avaliação sumativa nem sempre promove uma boa aprendizagem e, por vezes, pode mesmo ter efeitos prejudiciais. Também Black e Wiliam (1998a) afirmam que a avaliação sumativa não é a melhor forma para aferir as aprendizagens dos alunos. Os autores referem que práticas assentes na avaliação sumativa dão uma importância excessiva à função de classificação, deixando para segundo plano a função da aprendizagem. Black e Wiliam (1998a) defendem igualmente que práticas assentes na função sumativa promovem uma maior competição dos alunos, ao invés de criar um ambiente de entreajuda e aprimoramento pessoal. Com tais práticas, alunos que alcançam classificações mais baixas, tendem a assumir que têm menos habilidade, de modo que ficam desmotivados e perdem a confiança na sua própria capacidade de aprender.

Falchikov (2005) aponta vários problemas relacionados com a avaliação sumativa. O primeiro é que uma avaliação que assenta na avaliação sumativa tende a dar um

ênfase demasiado grande aos testes. O segundo aspeto apontado por Falchikov está relacionados com os problemas ao nível da fiabilidade dos instrumentos de avaliação sumativa utilizados, bem como problemas ao nível da subjetividade nos momentos de classificação. A autora sugere ainda que esta modalidade de avaliação não contribui para uma motivação positiva do aluno, levando muitas vezes a situações de *stress*. Por último, é referido que a avaliação sumativa promove uma aprendizagem superficial dos conteúdos, ao invés de uma aprendizagem aprofundada dos mesmos. Quanto a este último aspeto, Irons (2008) acrescenta que a avaliação sumativa é incapaz de se focar num conjunto de aspetos mais complexos e considerados importantes relacionados com as aprendizagens e competências adquiridas.

Para além de algumas das críticias apontadas anteriormente, Harlen e Crick (2002) elencam também um conjunto de fragilidades relacionadas com a avaliação sumativa. Uma das principais evidências encontradas pelos autores foi o impacto que a avaliação sumativa tinha na auto-estima dos estudantes. Verificou-se que alunos que normalmente atingem resultados escolares mais baixos tendem a apresentar níveis muito inferiores aos alunos que alcançam melhores resultados. Outra das conclusões do estudo foi de que, quando há uma forte aposta na avaliação sumativa (seja por parte dos professores ou das próprias direções escolares, por exemplo), os docentes tendem a adotar estilos de ensino focados na transmissão de conhecimentos, favorecendo alunos mais passivos na sua aprendizagem ao invés de alunos que apresentam características mais ativas e/ou criativas. Este aspeto pode levar a uma baixa auto-estima destes últimos face aos primeiros. Harlen e Crick (2002) salientam ainda o facto de que quando há um ambiente em sala de aula assente numa avaliação sumativa, há uma tendência para os alunos se concentrarem fundamentalmente nos momentos formais de avaliação e menos noutro tipo de tarefas que possam ser promovidas em sala de aula. Por último, os

autores referem que a avaliação sumativa promove um clima de competição que favorece, à partida, os estudantes que apresentam uma certa predisposição para a aprendizagem, aumentando ainda mais o fosso entre aqueles que apresentam essa mesma predisposição e os alunos com mais dificuldades na sua aprendizagem.

De salientar ainda que, segundo Pinto e Santos (2006), a avaliação sumativa causa sentimentos de insatisfação ou inquietação nos professores face ao rigor e à justiça das classificações atribuídas e às suas consequências nos alunos. Estes sentimentos, devem-se ao facto de a avaliação sumativa se centrar num juízo avaliativo traduzido numa 'nota', sendo que essa mesma nota servirá de base a decisões de retenção e transição de ano. Referem os autores que isto deve-se ao peso que é dado à avaliação sumativa, uma vez que *[...] a avaliação sumativa tende a impor-se em toda a ação de avaliação, confundindo-se mesmo com a própria avaliação* (Pinto & Santos, 2006, p.98).

Neste sentido, a avaliação sumativa deverá ser utilizada de forma complementar às outras modalidades de avaliação. Ribeiro (1993) refere que a avaliação sumativa deverá ser utilizada no sentido de aferir resultados já recolhidos em avaliações de caráter formativo. Assim, este tipo de avaliação corresponde a um balanço final e permite uma visão de conjunto relativamente a um todo ao qual, até então, apenas se fizera juízos parcelares (Ribeiro, 1993). A avaliação sumativa assume-se assim como um elemento complementar à avaliação de diagnóstico e à avaliação formativa, na medida em que contribui para uma apreciação mais equilibrada do trabalho realizado pelos alunos. Uma das articulações possíveis é a utilização formativa da avaliação sumativa (Black *et al.*, 2004) que, quando corretamente utilizada, facilita o entendimento do aluno face aos resultados alcançados, levando-o a orientar e reorganizar a sua participação, bem como as produções subsequentes.

Apesar de se reconhecer a necessidade de articulação entre a avaliação sumativa e formativa, Santiago *et al.* (2012) referem que no caso português, muito embora seja dado especial relevo e importância à avaliação formativa nas políticas educacionais[3], as práticas em sala de aula dão maior ênfase à componente sumativa do que à componente formativa. Este aspeto, segundo os autores, tem um efeito negativo no papel formativo dos professores, em particular, e da avaliação, em geral. Tal facto reflete-se numa atenção obsessiva pelos resultados, patente em muitas situações como, por exemplo, na propaganda dos *mídia* em torno dos resultados dos exames nacionais, nas práticas pedagógicas assentes na preparação dos exames e na utilização exagerada de testes.

Num artigo publicado em 2016, Leonor Santos aponta algumas razões que explicam o maior recurso a práticas de avaliação sumativa ao invés de formativa. Mesmo reconhecendo a importância que a avaliação formativa tem no processo de ensino e aprendizagem, um grande leque de professores olha para ela numa perspetiva de mais trabalho a adicionar àquele já existente o que, segundo a autora leva-os a confrontarem-se com restrições de tempo para, por exemplo, cumprir o programa curricular. (Santos, 2016). Muito embora este seja um fator importante a considerar quando se fala nas dificuldades em articular a avaliação formativa e sumativa, há outros aspetos importantes a ter em consideração e que são igualmente identificados neste artigo, como sejam a extensão das turmas, dos curriculos e a dificuldade em encontrar desafios adequados para as necessidades dos alunos. No entanto, há um aspeto, a nosso ver, muito importante e que está diretamente relacionado com os objetivos desta tese e que se prende com falta de conhecimentos que os professores revelam sobre as questões relacionadas com a avaliação, isto é, com baixos níveis de literacia em avaliação (que daremos especial atenção no

[3]Conforme se poderá constatar no subcapítulo 1.7.

segundo capítulo).

1.5 Qualidade em Avaliação

Quando se elabora e analisa um determinado instrumento de avaliação, há que ter em consideração um conjunto de fatores que têm uma influência direta ou indireta no referido instrumento a utilizar e, por consequência, na qualidade da informação que é recolhida por intermédio desse mesmo instrumento. Dois desses fatores são a validade e a fiabilidade, ou seja, é necessário, por um lado, que o instrumento avalie aquilo que tem a avaliar e, por outro, que os dados recolhidos sejam fiáveis. Ambos os conceitos estão intrinsecamente relacionados entre si e são fatores interdependentes, na medida em que:

> não são analisáveis isoladamente, podendo mesmo afirmar-se que a fiabilidade da informação recolhida é condição necessária mas não suficiente da validade dessa mesma informação, isto é, mesmo que a informação recolhida apresente fiabilidade tal não garante que seja válida (Neves & Ferreira, 2015, pp.55-56).

Muito embora a validade e a fiabilidade sejam tradicionalmente os fatores mais abordados, quando se fala em qualidade em avaliação, há um conjunto de outros fatores determinantes, como veremos mais adiante.

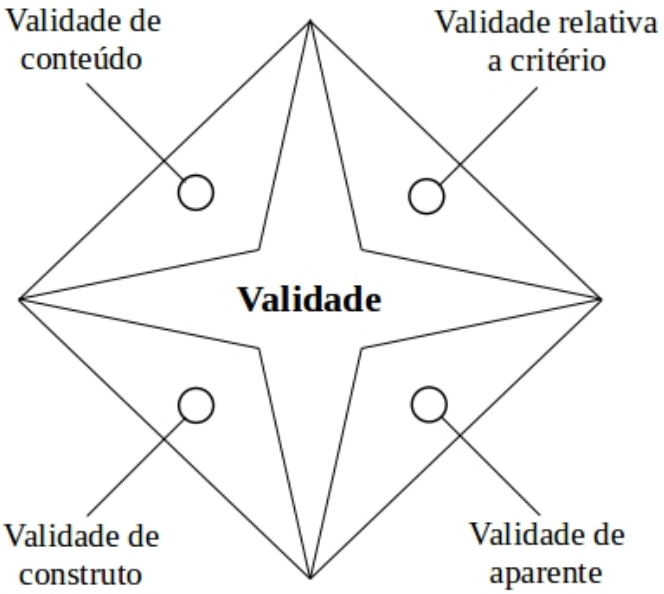
1.5.1 Validade

A validade, de um determinado instrumento de avaliação, está diretamente relacionada com o quão bem o que está sendo avaliado corresponde ao comportamento ou aos resultados da aprendizagem que se pretende que sejam

avaliados (Harlen, 2007). Nas palavras de Valadares e Graça (1998), trata-se de saber em que medida um instrumento de avaliação corresponde à função para que foi concebido. São normalmente considerados quatro tipos de validade, a validade de conteúdo, a validade relativa a um critério, a validade de construto e a validade aparente.

A propósito dos vários tipos de validade, Gareis e Grant (2015) fizeram uma analogia com as faces de um diamante (Figura 3). Consideram os autores que o valor do diamante varia consoante o valor de cada uma das suas faces, ou seja, que a validade da avaliação é função das múltiplas faces dos quatro tipos de validação e que o grau de validade global é tanto maior quanto maior for o grau de validade de cada uma das faces que a constituem.

Figura 3: As quatro faces da validade
(Adaptado de Gareis e Grant, 2015)



A validade aparente *corresponde ao que a prova, pelo tipo de questões ou de situações apresentadas, aparenta avaliar* (Bessa, 2007, p.120). Embora este tipo de

validade já tenha sido posta em causa e até descredibilizada, ela poderá ser muito importante na medida em que se o instrumento parecer, aos olhos dos respondentes irrelevante ou inadequado toda a recolha de informação poderá estar comprometida (Martins, 2006).

A validade de conteúdo consiste em compreender em que medida os itens selecionados, para aferir uma construção teórica, representam bem todos os aspetos importantes do conteúdo a ser medido (Alexandre & Coluci, 2011). De outro modo, a validade de conteúdo:

> concerns the coverage of appropriate and necessary content i.e. does the test cover the skills necessary for good performance, or all the aspects of the subject taught? Content validity tends to be based on professional judgments about the relevance of the test content to the content of a particular domain (Gipps, 1994, p.59)

A principal preocupação a ter, para garantir este tipo de validade, é assegurar que as questões e/ou tarefas selecionadas para a avaliação respeita o currículo ensinado[4] e que as mesmas se apresentam nas proporções adequadas. Desta forma, poderá ser útil elencar os objetivos a serem alvo de avaliação e, a partir deles, construir uma tabela de especificações para cada um dos objetivos (Gareis & Grant, 2015). Assim, poderá ser mais fácil garantir que outro tipo de objetivos não sejam avaliados de forma não intencional.

Quando se fala em validade relativa a um critério, de uma forma genérica, procura-se saber em que medida o(s) instrumento(s) de avaliação permite(m) *prever o desempenho relativo a um dado critério* (Fernandes, 2005, p.113). Em alternativa, a

[4]Neves e Ferreira (2015) distinguem entre currículo oficial, currículo ensinado e currículo apreendido. O currículo oficial corresponde aos planos de estudo e programas oficiais aprovados pelas autoridades nacionais e que reflete aquilo que, em cada momento, se considera dever ser aprendido pelos alunos. O currículo ensinado resulta da interpretação, por parte dos professores, do currículo oficial e que é transmitido em contexto de sala de aula. Por último, o currículo apreendido corresponde ao conjunto das aprendizagens efetivamente realizadas pelos alunos.

validade relativa a um critério pode ser definida como o:

> processo de determinar em que medida o grau de consecução num teste [ou outro tipo de instrumento de avaliação] está relacionado com a medida de uma outra performance diferente. A esta outra medida, que tanto pode ser efetuada no futuro como no presente, chama-se critério. (Valadares & Graça, 1998, p.140).

Na definição apresentada anteriormente estão subententidas duas subcategorias deste tipo de validade: a validade preditiva e a validade concorrente. A validade preditiva diz-nos em que medida a informação recolhida permite prevêr o desempenho futuro dos alunos, ao passo que a validade concorrente permite estimar o desempenho presente de forma correspondente ao estimado por um instrumento aplicado anteriormente (Neves & Ferreira, 2015).

Assim, a validade relativa a um critério é uma combinação da validade concorrente e preditiva já que, segundo Gipps (1994), ambas se relacionam com a previsão de desempenho em algum critério, quer no presente quer no futuro.

Por último, a validade de construto verifica em que medida os instrumentos de avaliação são uma medida adequada do construto. Segundo autores como Angoff (1988) e Gareis e Grant (2015), a validade de construto é a mais importante entre as quatro faces da validade. Segundo Lucie Ribeiro (1993) a validade de construto procura verificar:

> até que ponto um teste [ou outro instrumento de avaliação] tem capacidade para revelar uma característica dos respondentes não diretamente analisada (representa apenas um conceito teórico formulado) mas que acaba por ser confirmada (ou não) pela consistência de resultados obtidos noutros testes que se relacionam com tal característica ou factor (Ribeiro, 1993, p.119).

Assim, a validade de construto diz respeito à precisão com que o instrumento de avaliação se alinha com o conceito teórico ou com os objetivos de aprendizagem que se

pretendem avaliar (Gareis & Grant, 2015). Dito de outra forma, a validade de construto verifica em que medida o instrumento mede aquilo que pretende medir.

Considerando os vários tipos de validade abordados, e de forma a garantir que a validade dos instrumentos desenvolvidos seja tanto maior quanto possível, tem de haver a preocupação em evitar um conjunto de erros que ponham em causa essa mesma validade. A tipologia de fatores que podem pôr em causa a validade dos instrumentos de avaliação podem, como indica Valadares e Graça (1998), estar relacionados com o próprio instrumento e o modo como ele está construído, com o modo como o processo de ensino decorreu, da forma como o instrumento de avaliação foi administrado, corrigido e aplicado. Há também fatores ligados ao modo como os estudantes respondem às questões colocadas e à natureza do grupo ao qual o instrumento é aplicado. Por consequência, Valadares e Graça (1998) destacam um conjunto de erros comuns que todo o professor deve evitar de forma a não colocar em causa a validade dos seus instrumentos de avaliação, nomeadamente:

1. Instruções pouco claras;
2. Vocabulário e nível de construção literária inadequados;
3. Questões com nível de dificuldade inapropriado;
4. Questões mal construídas;
5. Questões ambíguas;
6. Inadequação das questões aos resultados de aprendizagem a avaliar;
7. Limites de tempo inadequados;
8. Instrumentos de avaliação com um número de questões insuficiente;
9. Distribuição inadequada de questões e um padrão de respostas adivinhável;
10. Instrumentos com ênfase em conteúdos pouco explorados em contexto de sala de aula;

11. Diferentes critérios de correção e classificação.

1.5.2 Fiabilidade

A fiabilidade de um instrumento de avaliação está diretamente relacionada com a consistência dos resultados obtidos na sua aplicação. Gipps (1994) afirma que as questões que estão subjacentes à fiabilidade são o de saber se uma avaliação produzirá resultados semelhantes em ocasiões distintas, ou caso seja aplicada por avaliadores distintos.

Para aferir a fiabilidade de um instrumento de avaliação, Gipps (1994) destaca alguns procedimentos que podem ser adotados, nomeadamente:

1. Aplicar o mesmo instrumento com alguns dias de diferença, é o designado de procedimento de teste-reteste;
2. Utilizar formas alternativas do mesmo teste para comparar o desempenho de populações semelhantes (formas paralelas);
3. Caso seja possível, subdividir o instrumento de avaliação em duas partes equivalentes e analisar a consistência dos resultados em ambas as partes, é o designado de *split-half reliability*;
4. Determinar, a partir de uma análise estatística todas as possíveis relações resultando num coeficiente de consistência interna.

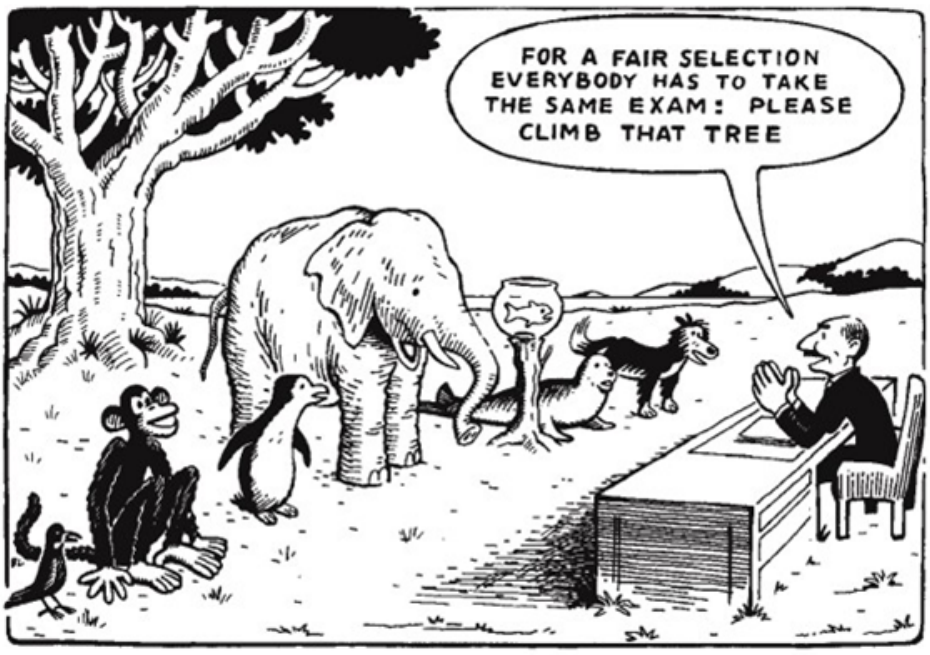
São múltiplos os fatores que concorrem para a fiabilidade de um instrumento de avaliação. Valadares e Graça (1998) sugerem um conjunto de aspetos que relevam para o aumento da fiabilidade, nomeadamente a extensão do instrumento de avaliação, a homogeneidade dos itens que o constituem, o poder discriminante dos itens que o compõem e, por fim, a amplitude de variação de conhecimentos e capacidades dos alunos aos quais é aplicado o instrumento de avaliação.

1.5.3 Outros fatores

Embora, como já referido, os fatores mais abordados sejam a validade e a fiabilidade, são vários os fatores que concorrem para a qualidade em avaliação. Um desses fatores é o da imparcialidade. A imparcialidade de uma ferramenta de avaliação refere-se à ausência de elementos tendenciosos. A avaliação deve ser adequada a todos os elementos do grupo avaliado, independentemente da raça, religião, sexo ou idade. Assim, a avaliação não deve prejudicar qualquer elemento, ou grupo, em qualquer outra base além da falta de conhecimentos e habilidades que a ferramenta de avaliação é destinada a medir.

Figura 4: A (im)parcialidade na avaliação

(Autor desconhecido)



A figura 4 é ilustrativa da introdução de elementos tendenciosos na avaliação, colocando em causa a sua imparcialidade. Como se pode constatar, apenas alguns elementos do grupo são fisicamente aptos a realizar a tarefa proposta, pelo que não se teve qualquer consideração pelas características dos restantes elementos do grupo, prejudicando-os claramente no processo de avaliação.

Subjacente às questões da imparcialidade na avaliação estão fatores como a justiça e a equidade. Gipps e Stobart (2009) argumentam que a avaliação no século XXI precisa cada vez mais de considerar os contextos sociais da avaliação e continuar o movimento longe de ver a justiça como uma mera preocupação técnica na construção de elementos de avaliação. Para tal, é necessário que os professores deixem de manifestar *preocupação relativamente à definição e aplicação de critérios iguais aos mesmos alunos, nas mesmas circunstâncias* (Neves & Ferreira, 2015, p.37) de forma a garantir que seja tratados com justiça. Gipps e Stobart referem mesmo que:

> The dilemma is that different groups will have different qualities and experiences, so fairness in assessment cannot be judged in terms of equal scores or outcomes. Differences in performance on a test may be due to differing access to learning, or because the test is biased in favour of one group (Gipps & Stobart, 2009, p.105).

Há então que respeitar as diferenças existentes entre os alunos, os seus ritmos, as suas necessidades, os seus contextos sociais. Assim, e no sentido de combater a injustiça que a igualdade pode criar e aumentar a qualidade do processo de avaliação, esta deve-se pautar pelo príncipio da equidade (Gipps, 1994; Gipps & Stobart, 2009; Neves & Ferreira, 2015), adaptando os instrumentos de avaliação e respeitando as diferenças e necessidades de cada um e de cada grupo.

Outro aspeto importante que se deve considerar para garantir a qualidade em avaliação é a transparência. Uma forma de entender a transparência neste contexto, é que esta deverá corresponder ao conhecimento que os alunos têm daquilo que será avaliado, ou seja, não deverão ocorrer, como indica Race (2009), surpresas desagradáveis. Por outras palavras, a transparência:

> involves ensuring that information is available about what learning is expected, what criteria will be used to judge student learning, and what

> rules are being applied when decisions are made about learning (OECD, 2011, p.53).

Assim, para que a transparência da avaliação esteja garantida, pelo menos duas condições deverão ser verificadas. Por um lado, os alunos deverão ter conhecimento dos conteúdos sobre os quais versará a avaliação e, por outro, os critérios de avaliação deverão ser claros.

Destacar, por último, outro fator bastante referido na literatura e que é fundamental nas melhores práticas em avaliação, a autenticidade. Uma das dimensões da autenticidade remete-nos, segundo Fletcher (2000), para duas questões importantes. A primeira, procura resposta à preocupação de saber se um determinado resultado foi atingido por um determinado aluno. Já a segunda, visa saber se o aluno atingiu um determinado resultado fruto de um trabalho individual ou como parte de um grupo. Estas questões prendem-se com a necessidade que os professores têm em saber se os resultados alcançados pertencem efetivamente a um determinado aluno, e não a outro, e se foram atingidos de forma legítima (por exemplo, sem que tenha havido plágio). A segunda dimensão da autenticidade visa assegurar que a avaliação é realizada nos contextos adequados (Race *et al*., 2005). O exemplo apresentado pelas autoras é o de que para avaliar de forma autêntica a performance num determinado domínio artístico (dança, por exemplo), esta deverá ocorrer num contexto de performance e não a partir, por exemplo, de um exame escrito, onde as competências no domínio em causa dificilmente serão postas em evidência.

1.6 Métodos e Instrumentos ao serviço da avaliação

Nas suas atividades letivas, muitos professores utilizam tanto métodos formais, como métodos informais de avaliação dos seus alunos. No entanto, e embora seja um paradigma que se tem alterado nos últimos anos, a literatura em avaliação tem-se centrado, sobretudo, nos métodos formais como sejam, por exemplo, os testes *standardizados*.

No dia a dia, como refere Wragg (2001), a avaliação centra-se, sobretudo, em situações informais de avaliação, sendo os momentos formais muito limitados no tempo (geralmente no fim de uma unidade temática). O mesmo autor nota que, tal como pode acontecer com alguns métodos formais de avaliação, os métodos informais são uma forma eficaz de avaliar todos os aspetos relacionados com o conhecimento, compreensão, habilidades, atitudes e comportamentos, tendo a vantagem, em muitos casos, de serem de mais fácil aplicação e interpretação.

Seguidamente, serão analisados alguns métodos e instrumentos formais e informais de avaliação em sala de aula que nos ajudarão a compreender a diversidade de formas de avaliação, não estando, por isso, os professores e os alunos sujeitos apenas aos vulgos testes de papel e caneta.

1.6.1 Observação Direta

A observação em sala de aula é uma das mais utilizadas formas de avaliação, embora, em muitos casos, utilizada de forma pouco sistemática. Esta forma de avaliação *permite a recolha de informações sobre o modo como os alunos vão desempenhando as suas tarefas e quais as competências e as atitudes*

desenvolvidas enquanto decorre o processo de ensino-aprendizagem (Valadares & Graça, 1998, p.106). No mesmo sentido, Harlen (2007) afirma que os métodos de avaliação assentes na observação direta, permitem a recolha de informações sobre os processos de aprendizagem e não apenas no produto da aprendizagem. Neves e Ferreira (2018) afirmam mesmo que métodos assentes na observação são, entre todas as técnicas, aquelas em que a avaliação se encontra mais integrada no processo de ensino e aprendizagem e que permite a *identificação imediata das dificuldades e a resposta às necessidades de cada aprendente* (p.73).

De forma a sistematizar o recurso à observação direta em sala de aula é necessário, por um lado, a sua devida planificação e, por outro, instrumentos desenvolvidos especialmente para esse efeito (Ferreira, 2015). Segue-se um conjunto de instrumentos que se podem desenvolver e utilizar em articulação com a observação:

- **Anedotários**: Os anedotários, também designados de registos de incidentes críticos, podem ser entendidos como *um conjunto de notas nos quais os professores podem descrever, de modo breve, factos importantes, significativos e relevantes ao processo diário de sala de aula* (Monteiro & Pissaia, 2018, p. 109). Os anedotários assumem-se assim como uma ferramenta que possibilita a descrição de comportamentos pouco habituais, tenham eles um caráter positivo ou negativo, e devem contribuir para aumentar o conhecimento dos sujeitos avaliados (Neves & Ferreira, 2015), auxiliando a reflexão e a tomada de decisão mais consciente por parte do professor (Monteiro & Pissaia, 2018).

- **Listas de verificação**: são compostos por ações, comportamentos ou procedimentos que o professor deseja observar nos alunos na realização das tarefas para os quais foram concebidos (Neves & Ferreira, 2015; Ferreira,

2018). São especialmente úteis para registar a presença ou ausência de um comportamento ou de um resultado de aprendizagem (Valadares & Graça, 1998). As listas de verificação podem incluir tanto comportamentos desejados, como erros tipificados permitindo *verificar a presença do que é esperado como detectar aquilo que ainda necessita de ser corrigido* (Neves & Ferreira, 2015, p.77).

- **Grelhas de observação**: à semelhança das listas de verificação, as grelhas de observação são listas compostas por ações, comportamentos ou procedimentos que se desejam observar no aluno ao longo da execução de uma tarefa. No entanto, e ao contrário das listas de verificação, as grelhas de observação possibilitam um registo descritivo sobre a progressão e as características dos comportamentos observados.

- **Escalas de classificação**: são constituídas por um conjunto de características ou qualidades, distribuídas por níveis, os quais indicam o grau de cada atributo (Neves & Ferreira, 2015). Por outras palavras, as escalas de classificação são usadas para atribuír um nível de desempenho ao comportamento ou ação esperada de acordo com a qualidade desse mesmo comportamento ou ação, numa escala pré-estabelecida.

1.6.2 Relatórios

Os relatórios constituem-se como produções escritas dos alunos que, segundo Sant'Anna (1995), têm como finalidade informar, relatar, fornecer resultados, dados e experiências, por parte dos alunos, ao professor e a todos os envolvidos no processo de ensino e aprendizagem. Já Valadares e Graça (1998) definem os relatórios como

produções escritas mais ou menos extensas, sobre problemas, atividades de investigação ou projetos em que os alunos trabalharam. Desta forma, e segundo o mesmo autor, este tipo de produções constitui-se tanto como um instrumento de aprendizagem como um instrumento de avaliação.

Este tipo de instrumentos, procura avaliar níveis de complexidade de ordem superior já que, segundo Valadares e Graça (1998), são tarefas fortemente associadas à aplicação do conhecimento a situações novas. Para além disso, e segundo os mesmos autores, os relatórios têm um grande potencial formativo, visto possibilitarem o desenvolvimento de competências e atitudes, como a resolução de problemas, raciocínio, interpretação e comunicação, mas também o gosto pela pesquisa e a responsabilidade.

Para a realização dos relatórios, é fundamental que os alunos tenham conhecimento dos critérios de avaliação (Ferreira, 2018; Rampazzo, 2011), de forma a poderem refletir e indicar as aprendizagems realizadas, assim como as dificuldades sentidas e aspetos a melhorar.

1.6.3 Questionamento em sala de aula

O questionamento em sala de aula é uma das múltiplas estratégias de avaliação informal e uma das formas mais eficientes para recolha de informações de forma contínua no processo de ensino e aprendizagem (Ruiz-Primo, Solano-Flores & Li, 2014). Segundo Wragg (2001), o questionamento em sala de aula é muitas vezes utilizado pelos professores como uma forma de verificar os conhecimentos adquiridos ou a compreensão de determinadas situações, sendo por isso uma ferramenta fundamental no diagnóstico das dificuldades sentidas por parte dos alunos. Este

aspeto é fundamental para que o professor possa atuar de imediato, no sentido de reencaminhar o processo de ensino e aprendizagem na direção que pretende (Ferreira, 2007).

No dia a dia da sala de aula, os professores realizam inúmeras questões dirigidas aos seus alunos. No entanto, conforme Silva e Lopes (2015), a maioria são de nível cognitivo baixo, uma vez que muitas delas exigem apenas recordar factos. A utilização deste tipo de questões tem como finalidade corrigir ou consolidar a aprendizagem. De forma a estimular o pensamento do aluno e promover um ambiente de interação entre professor e colegas, Silva e Lopes (2015) referem que as questões formuladas deverão ser abertas e de nível cognitivo mais elevado, ou seja, que exijam análise, síntese e/ou avaliação.

Os professores utilizam o questionamento em sala de aula por várias razões. Brualdi (1998) elenca alguns exemplos, tais como:

- o questionamento ajuda o professor a manter os alunos envolvidos na aula;
- através do questionamento, os alunos têm oportunidade de expressarem as suas opiniões;
- através do questionamento, os alunos podem ouvir diferentes explicações e ideias produzidas pelos seus pares;
- o questionamento em sala de aula permite um melhor acompanhamento das aulas para além de ajudar o professor a moderar o comportamento dos alunos;
- o questionamento oral ajuda o professor a avaliar as aprendizagens dos alunos e a rever as suas aulas conforme a necessidade.

Já Santos e Pinto (2018) assumem que o questionamento é potenciador de uma avaliação formativa, por diversas razões. A primeira razão é o facto do questionamento oral ocorrer a par com as experiências de aprendizagem, permitindo, dessa forma, uma

regulação no momento. A segunda prende-se com o facto de recorrer a uma forma mais habitual de comunicação entre professor e alunos, ou seja, a forma oral. Em terceiro, e último lugar, por permitir deslocar a responsabilidade do professor para o aluno.

1.6.4 Portfólios

Nas últimas décadas os portfólios têm-se assumido como uma forma de avaliação alternativa aos métodos tradicionais. Entre as várias definições possíveis, podemos entender os portfólios como uma coleção sistemática e organizada de evidências, usadas tanto por professores como por alunos, de forma a monitorizar a evolução dos conhecimentos, capacidades e atitudes dos alunos (Kronowitz, 2004).

O principal propósito da avaliação por portfólio é, segundo Lam (2018):

> [...] enhancing teaching and learning in specific subject domains, since it can flexibly serve as an innovative pedagogy, a catalyst to promote quality learning or a downright assessment instrument which generates quantitative and qualitative learning evidence. (p.2)

Assim, com a utilização dos portfólios, como instrumento de avaliação, é possivel ter um registo das aprendizagens que não se limita a um momento em específico, como acontece, por exemplo, nas fichas de avaliação. Esse registo deve-se ao facto de que nos portfólios constam, entre outros possíveis elementos, uma seleção de trabalhos realizados pelos alunos ao longo do processo de ensino e aprendizagem.

Silva e Souza (2007) elencam um conjunto de vantagens associadas à utilização dos portfólios em sala de aulas dos quais se destacam:

- Permite o *feedback* imediato, na medida em que faculta, tanto a professores como

a alunos, informações sobre os sucessos e dificuldades dos alunos, permitindo delinear formas eficazes para a superação das dificuldades identificadas;

- É elaborado pelo próprio aluno o que lhe permite organizar as suas componentes de forma a refletir sobre as suas estratégias pessoais para a resolução de problemas. Dito de outra forma, a aprendizagem é centrada no aluno e não no professor;

- Relacionado com o item anterior, os portfólios respeitam a individualidade dos alunos, pois neles constam não somente as tarefas consideradas mais interessantes, como também críticas e reflexões pessoais;

- Favorece a autoavaliação pois, ao longo da sua elaboração, o aluno tem a possibilidade de analisar e refletir sobre as próprias aprendizagens;

- Promove a utilização de múltiplas linguagens, uma vez que os alunos podem expressar os seus conhecimentos de várias formas (*p.e.* texto, poemas, desenhos, audiovisuais, entre outros).

Já a principal desvantagem dos portfólios, segundo Silva e Souza (2007) prende-se, sobretudo, com a exigência pelo envolvimento e compromisso dos alunos no desenvolvimento das tarefas relacionadas com o portfólio. Os portfólios são instrumentos que exigem tempo e dedicação, pelo que há o risco de os alunos acumularem tarefas e começarem a abster-se de algumas reflexões. É necessário, portanto, que os alunos se sintam motivados e predispostos para desenvolverem as suas aprendizagens desta forma, pois os portfólios exigem tempo, preserverança e constância.

1.6.5 Testes

Os testes são, ainda hoje, uma das principais ferramentas de avaliação utilizadas pelos professores para recolher informações sobre o aproveitamento dos alunos, quanto às capacidades desenvolvidas e aos conhecimentos adquiridos numa perspetiva do seu desempenho máximo (Neves & Ferreira, 2015). Este lugar de destaque deve-se não só ao peso significativo que as aprendizagens do domínio cognitivo têm tradicionalmente apresentado em quase todas as disciplinas, como também ao facto do ensino ser, ainda em muitos casos, centrado no professor (Lemos, Neves, Campos, Conceição & Alaiz, 1993). Fernandes (2005) refere mesmo que os professores revelam uma preocupação primordial com a atribuição de classificações, facto a que não será alheia a utilização privilegiada, ou mesmo exclusiva, de testes para avaliar as aprendizagens.

Os testes são, normalmente, utilizados como base para a atribuição de classificações, pois correspondem à avaliação sumativa e devem realizar-se, regra geral, no fim de cada unidade programática (Cabral, 2001). Segundo Ribeiro (1990) este tipo de testes incide numa maior gama de objetivos, pelo que o grau de profundidade da avaliação, relativamente a cada objetivo, tem de ser menor no que acontece na avaliação formativa. Assim, e segundo o autor, estamos perante um instrumento de malha larga, que incide sobre uma vasta extensão de conteúdos. Desta forma, e uma vez que não se podem testar todos os objetivos, sob pena do teste não ser exequível, deverá ser selecionada uma amostra relevante que incida sobre conhecimentos fundamentais do universo testado, sendo razoável que o aluno, se adquiriu esses, adquiriu outros com ele relacionados.

Importa aqui clarificar dois tipos de testes: os testes referentes a normas e os testes referentes a critério. Segundo Arends (1995):

> os testes referidos a uma norma medem o desempenho de um aluno em relação a outros alunos, os testes referidos a um critério medem esse desempenho em relação a um nível ou um critério de desempenho pré-estabelecido (p.235).

O mesmo autor apresenta um exemplo prático de como se podem distinguir estes dois tipos de testes. Imagine-se uma prova de velocidade de 100 metros em que o aluno percorre a distância em 13 segundos. No caso de o avaliador utilizar o teste referido à norma, pode afirmar que o aluno foi, por exemplo, o 3º mais rápido num universo de 50 alunos, o que seria um resultado bastante positivo. Caso o avaliador utilizasse um teste referido a um critério, em que o critério definido para esta corrida fosse 12 segundos, concluir-se-ia que o aluno não tinha atingido o objetivo definido. Desta forma, segundo Ribeiro (1990), os testes normativos assumem um caráter seletivo e muito ligado à classificação, permitindo interpretações como "quais são os melhores?" e "quem é o pior?". Ao passo que os testes referentes a critérios não têm a finalidade de comparar resultados entre os alunos mas sim avaliar o desempenho dos alunos face a um conjunto de conteúdos e objetivos propostos.

A construção de testes deve ser uma tarefa bem ponderada e deverá respeitar alguns aspetos importantes. DeBlassie (1974), defende que para a realização de um teste deverão ser respeitadas as seguintes condições:

- O teste deverá ser escrito e duplicado tantas vezes quantas as necessárias para que cada aluno tenha uma cópia;
- O aluno deverá ter todas as indicações necessárias à realização do teste;
- O teste deverá ser programado de forma atender ao tempo disponível para a sua realização;
- O teste deverá ser facilmente corrigido e classificado;
- Deverá haver um cuidado no planeamento do teste, de forma a minimizar o tempo necessário para a sua construção, duplicação e correcção (pp.75-76).

Já Cosme *et al.* (2020) apresentam algumas sugestões importantes que deverão

nortear a construção de um teste, nomeadamente:

- Apresentação legível;
- Linguagem percetível, clara, objetiva e utilizada durante o processo de ensino-aprendizagem;
- Possibilidade de realização por todos os alunos;
- Conhecimento prévio dos critérios de avaliação por parte dos alunos;
- Presença de cotação de cada um dos itens no documento;
- A tipologia de resposta estar definida e separada da pergunta (p.150).

Outros dois aspetos, já abordados, que os testes deverão respeitar são a validade e a fiabilidade. A validade do teste refere-se ao facto de ele avaliar aquilo que é suposto avaliar. Assim, o teste tem de ser representativo dos conteúdos e objetivos que foram sendo transmitidos ao longo das aulas. Para facilitar a averiguação da validade de um teste, deverá ser construída uma tabela de especificações[5] onde constem os vários itens e objetivos a que dizem respeito, de forma a garantir que o teste *mede aquilo que se propõe a medir* (Arends, 1995, p.240). Já um teste diz-se fidedigno quando *produz resultados consistentes para as pessoas que o realizarem mais do que uma vez num determinado período de tempo* (Arends, 1995, p.240).

1.7 A Avaliação no quadro legal português

Atualmente, a avaliação das aprendizagens encontra-se regulamentada, em Portugal, pelo Decreto-Lei nº 55/2018, de 6 de julho, e pelas Portarias nº 223-A/2018, de 3 de agosto, e nº 226-A/2018, de 7 de agosto. São definidos dois tipos de avaliação, o de natureza interna e o de natureza externa. Relativamente à avaliação

[5]Sobre a importância das tabelas de especificações ver, por exemplo, Fives e DiDonato-Barnes (2013).

interna, são contempladas as modalidades de avaliação formativa e avaliação sumativa.

A avaliação formativa, assumida como sendo a principal modalidade de avaliação, integra o processo de ensino e de aprendizagem, fundamentando o seu desenvolvimento. Recorre a uma grande diversidade de formas de recolha de informação, ou seja, deverão ser utilizadas várias técnicas e instrumentos de avaliação conforme as respetivas finalidades. Esta utilização de métodos diversificados permite obter informações de maior qualidade, o que torna a avaliação formativa a modalidade, por excelência, para conhecer a forma como se ensina e como se aprende, fundamentando a adoção e o ajustamento de medidas e estratégias pedagógicas. A avaliação formativa ocorre ao longo de todo o ano letivo assumindo, por isso, um caráter contínuo e sistemático.

Já a avaliação sumativa apresenta-se como uma modalidade de avaliação fundamental para a formulação de um juízo globalizante sobre a aprendizagem dos alunos, o que permite, de igual modo, tomar decisões sobre o seu percurso escolar (*p.e.* decisões sobre a progressão, a retenção ou ainda a reorientação do percurso educativo). Outro aspeto relevante, é que a avaliação sumativa constitui-se como uma forma de informar, tanto os alunos como os encarregados de educação, acerca do desenvolvimento das aprendizagens definidas para cada uma das disciplinas. Referir ainda que é a partir da avaliação sumativa que se procede à verificação das condições de admissão aos exames nacionais ensino secundário. Para os alunos que reunam um conjunto de requisitos, a avaliação sumativa pode processar-se através da realização de provas de equivalência à frequência.

A avaliação externa gera informações de cariz formativo (no caso das provas de aferição) e de cariz sumativo (no caso das provas finais do ensino básico e dos

exames nacionais). A avaliação externa incide nas Aprendizagens Essenciais, com especial enfoque nas áreas de competências inscritas no Perfil dos Alunos à Saída da Escolaridade Obrigatória.

As provas de aferição ocorrem no final do ano letivo, nos 2°, 5°e 8°anos de escolaridade, sendo que no 2°ano são avaliados os conteúdos relativos às disciplinas de Português, Matemática, Estudo do Meio, Expressões Artísticas e Físico-Motoras. No 5°e 8°anos de escolaridade as provas de aferição incidem sobre as disciplinas de Português ou Matemática e, de forma rotativa, uma das outras disciplinas que fazem parte da estrutura curricular. As provas finais de ciclo realizam-se no 9°ano de escolaridade e destinam-se a todos os alunos do ensino básico[6]. Estas provas incidem sobre as disciplinas de Português[7] e Matemática e realizam-se em duas fases com uma chamada cada. Já no ensino secundário, mais concretamente nos cursos científico-humanísticos, os exames nacionais realizam-se no ano terminal das respetivas disciplinas.

[6]Excluem-se aqui os alunos que frequentem os Percursos curriculares alternativos, os Cursos de ensino vocacional, os Cursos de educação e formação, os Programas integrados de educação e formação, os Cursos de educação e formação de adultos e outras ofertas específicas.

[7]Português Língua Não Materna para alunos com nível de proficiência linguística de iniciação ou intermédia ou Português Língua Segunda para alunos com surdez severa a profunda

Capítulo 2

Literacia em Avaliação

2.1 Em torno do conceito de Literacia em Avaliação

A literacia em avaliação dos professores é um termo recorrentemente utilizado mas poucas vezes definido (Willis, Adie & Klenowski, 2013). Assim, neste ponto serão abordados alguns contributos teóricos que relevaram para a definição de literacia em avaliação.

A avaliação das aprendizagens é uma das mais importantes responsabilidades dos professores, assim como uma das tarefas nas quais os professores despendem mais tempo (Ramesal, 2011). Esta ideia é igualmente defendida por Craig Mertler ao referir que *assessing student performance is one of the most critical aspects of the job of a classroom teacher. It impacts nearly everything that teachers do* (Mertler, 2003, p. 3).

Embora a formação de professores possa dar competências aos professores para ensinarem os respetivos currículos, Perrenoud (1996), usando como exemplo os professores do ensino primário, levanta um conjunto de questões que procuram saber

se esses mesmos professores têm igualmente competências ao nível da avaliação:

> Supongamos que el curriculum formal ya la formación de un maestro de enseñanza primaria que inicia su prática docente fueran suficientes para indicarle lo que debe enseñar. ¿Sabrá con eso le ayudará a manejar la casi ilimitada diversidad de normas de excelencia que figuran, de modo implícito, en el curriculum formal? ¿És preciso corresponder una evaluación distinta para cada objetivo, para cada aprendizaje de valor? ¿Cómo saber lo que se puede y se debe exigir en un curso determinado? ¿Cuál es el grado mínimo de excelencia? ¿En qué términos? ¿De acuerdo con que código? ¿A quién hay que comunicarlos? ¿De qué modo? ¿Como efectuar las síntesis de las que dependerá el juicio global de éxito o fracaso? (Perrenoud, 1996, p.120)

Desta forma, compreende-se a necessidade dos professores dominarem as várias competências em avaliação. Ao domínio dessas competência dá-se o nome de literacia em avaliação.

O conceito de literacia em avaliação foi primeiramente apresentado por Richard Stiggins (1991) como o conhecimento profundo das questões de avaliação. Mais tarde, Stiggins (1995) sugere que um educador/professor com literacia em avaliação sabe o que avaliar, a razão de avaliar, como avaliar, quais os possíveis problemas relacionados com a avaliação e como prevenir que esses problemas surjam no processo de ensino e aprendizagem. Para além disso, têm um conhecimento profundo dos efeitos negativos de uma má avaliação.

Na mesma linha de raciocínio, Paterno (2001) refere que a literacia em avaliação diz respeito à posse de conhecimentos sobre os princípios básicos da avaliação, nomeadamente a terminologia, o desenvolvimento e aplicação de metodologias e técnicas de avaliação, a familiaridade com os *standards* de qualidade e com alternativas aos instrumentos tradicionais de medida das aprendizagens, vulgarmente conhecidos como testes de papel e caneta.

Mertler (2003), baseando-se nos trabalhos do *Center for School Improvement and*

Policy Studies da Boise State University elenca um conjunto de características que um professor com literacia em avaliação deve ter, nomeadamente:

> • Assessment literate educators recognize sound assessment, evaluation, communication practices;
>
> • They understand which assessment methods to use to gather dependable information and student achievement;
>
> • Communicate assessment results effectively, whether using report card grades, tests scores, portfolios or conferences;
>
> • Can use assessment to maximize student motivation and learning by involving students as full partners in assessment, record keeping, and communication. (Mertler, 2013, p.10)

Popham (2011) refere que a literacia em avaliação trata do entendimento dos conceitos e procedimentos de avaliação fundamentais susceptíveis de influenciar decisões educacionais. Outra proposta de definição considera a literacia em avaliação como a habilidade de desenhar, selecionar, interpretar e utilizar os dados resultantes da avaliação de modo apropriado e que permita a tomada de decisões educacionais adequadas (Brown, 2008; Quilter & Gallini, 2000).

As várias definições de literacia em avaliação apresentadas até agora centram-se num conjunto de conhecimentos e competências que o professor deve dominar para colocar em ação as melhores práticas em avaliação. Stiggins (1991, 1995) deu um maior destaque ao conhecimento sobre os objetivos e funções da avaliação, bem como à capacidade que os professores devem ter na definição daquilo que deve ser alvo de avaliação e na forma como deve ser avaliado. Para além disso, para Richard Stiggins é fundamental que um professsor saiba quais os efeitos da avaliação e de que forma pode minimizar os efeitos negativos. Já a definição de Paterno (2001) centra a sua definição de literacia em avaliação no conhecimento dos conceitos relacionados com a avaliação, mas também na competência em desenvolver

métodos e técnicas de avaliação diversificados. Embora as definições de Stiggins e Paterno sejam bastante semelhantes, elas distinguem-se no facto de Stiggins se centrar mais nas competências, como se verifica na utilização de expressões como 'sabe o que avaliar' e 'sabe como avalia' e Paterno se centrar mais nos conhecimento teóricos sobre os princípios e a terminologia em avaliação.

A perspetiva de Mertler (2003), para além de considerar os aspetos abordados pelos autores anteriores, acrescenta a necessidade dos professores em terem competências ao nível da comunicação com os alunos em contexto de avaliação. De facto, e como vimos no subcapítulo 1.4.2, a comunicação, em especial o *feedback*, é um elemento fundamental em avaliação, em especial na avaliação formativa.

Já nas abordagens de Quilter e Gallini (2000), Brown (2008) e Popham (2011) o conceito de literacia em avaliação está diretamente relacionado com o domínio dos conhecimentos e dos procedimentos em avaliação que tenham uma influência direta na tomada de decisões educacionais.

Na presente investigação, e ponderando os contributos teóricos dos autores analisados, consideraremos a literacia em avaliação como o conjunto de conhecimentos e capacidades que um professor tem dos vários aspetos relacionados com a avaliação, nomeadamente dos princípios e funções da avaliação, da construção e utilização de instrumentos de avaliação diversificados e adequados, da interpretação e utilização da informação recolhida no processo de avaliação.

2.2 A importância da literacia em avaliação e a sua relação com a aprendizagem

A investigação que se tem debruçado sobre as questões relacionadas com a Literacia em Avaliação, tem revelado dois aspetos cruciais. Por um lado, verifica-se a existência de uma preparação inadequada dos professores face à tarefa de avaliar eficazmente a aprendizagem dos alunos (DeLuca & Klinger, 2010; Koh, 2011; Xu & Brown, 2016) e, por outro, que os professores, seja no ínicio da carreira ou com vários anos de serviço, não se sentem confiantes na sua capacidade de avaliar os alunos com precisão e de forma adequada (Koh, 2011; Volante & Fazio, 2007; Yamtim & Wongwanich, 2014). Isto acontece devido a uma *limited preservice assessment education and a lack of research on the pedagogies that support teacher candidate learning in this area* (DeLuca *et al.*, 2013, p.128). Devido a esta falta de preparação, uma franja alargada de professores tem revelado, segundo Koh (2011), uma fraca capacidade para desenvolver e aplicar as mais variadas formas de avaliação, bem como uma incapacidade para interpretar os resultados oriundos da aplicação dos instrumentos de avaliação.

Stiggins (2002) sugere mesmo que os professores em formação raramente frequentam programas que os ensinem, por exemplo, sobre qual o papel da avaliação no processo de ensino e aprendizagem ou abordagens que tenham impactos significativos nas aprendizagens. Na mesma linha, Xu e Brown (2016), reforçam que *sadly, many pre-service teacher programs [...] only offer a one-semester assessment course that provides a general introduction to assessment [...] or else do not have such a course at all* (p.153). Tais lacunas na formação inicial implicam que, para além dos problemas identificados anteriormente, os professores não tenham a confiança na sua capacidade de avaliar, levando-os a *assess their students in a similar manner*

to the way they were assessed in schools (McGee e Colby, 2014, p.523).

No entanto, a tarefa de avaliar é, como já foi referido, uma das principais responsabilidades dos professores já que é fundamental para verificar e melhorar as aprendizagens dos alunos (Hailaya *et al.*, 2014, McGee & Colby, 2014). A Literacia em Avaliação assume-se assim como uma das principais características que todos os professores deverão desenvolver, mesmo antes do início das sua carreira docente, ou seja, a partir da sua formação inicial.

Newfields (2006) destaca três razões pelas quais a literacia em avaliação é tão importante. A primeira respeita-se à universalização da avaliação em contexto escolar, ou seja, a avaliação está presente na grande maioria dos sistemas escolares mundiais. Este fator leva a que os professores, em todo o mundo, consumam grande parte do seu tempo em atividades ligadas, direta ou indiretamente, à avaliação. Para além disso, o autor refere que [*in*] *many schools, a good portion of the budget also goes into formal testing* (Newfields, 2006, p.46). Em segundo lugar, o autor destaca a necessidade de compreender a literatura educacional vocacionada para as questões relacionadas com a avaliação. Uma maior familiaridade com os conceitos e com os processos estatísticos inerente às avaliação[8], permite aos professores uma maior facilidade em se manterem atualizados nesses domínios, tendo uma maior capacidade em introduzir novos métodos que melhorem a aprendizagem dos alunos e, por consequência, a sua avaliação. Por último, o autor destaca que um professor com literacia em avaliação consegue comunicar, de forma mais eficaz, os resultados escolares aos alunos. Recorde-se que uma comunicação eficaz, seja de cariz formativo ou sumativo, desempenha um importante papel no processo de ensino e aprendizagem.

[8] *p.e.* índices de dificuldade, índices de descriminação, aferição da validade e fiabilidade de instrumentos de avaliação

Gottheiner e Siegel (2012) destacam outro aspeto que permite uma melhor perceção da importância da literacia em avaliação. Os autores afirmam que a utilização de ferramentas diversificadas de avaliação deverá assumir-se como umas das principais características de um professor com literacia em avaliação e que esse facto *helps teachers to select the most relevant and powerful instruments for particular learning goals* (Gottheiner & Siegel, 2012, p.534).

Malone (2013) refere ainda que uma avaliação forte e devidamente implementada fornece a professores, a alunos, e a todas as partes interessadas, informações importantes sobre o desempenho dos alunos e sobre em que medida os objetivos educacionais estão, ou não, a ser cumpridos. Assim, a avaliação pode e deve integrar-se com o ensino, formando uma relação na qual informa e melhora o ensino e vice-versa. Contudo, esta relação de reciprocidade não pode florescer quando os professores não dispõem de formação suficiente para realizar todas as ações implicadas numa boa avaliação. Consequentemente, um baixo nível de literacia em avaliação põe em causa tanto a avaliação dos alunos, como todo o processo de ensino e aprendizagem.

Quanto melhor os professores dominarem as noções e os processos que conduzem à tomada de decisões, no respeitante à avaliação dos alunos, melhor serão as escolhas que o professor fará em benefício deles. Popham (2018) refere mesmo que, à partida, o sucesso de um professor aumenta quanto maior for a sua literacia em avaliação, pois evita erros típicos que normalmente são cometidos pelos professores que possuem baixos níveis de literacia em avaliação. Os erros típicos, a que se refere Popham (2018), encaixam-se normalmente nas seguintes categorias:

1. Utilização de instrumentos de avaliação inadequados;
2. Utilização incorreta de instrumentos de avaliação adequados;
3. Não utilização de instrumentos de avaliação formativa.

A utilização de instrumentos de avaliação inadequados assume-se como um dos erros mais graves cometidos pelos professores com baixos níveis de literacia em avaliação. Popham (2018) refere que um erro comum é a utilização de testes *standardizados* para avaliar as aprendizagens dos alunos já que, segundo o autor *most of today's standardized tests are accompanied by no evidence that those tests are suitable for such an important evaluative mission* (p.8).

O segundo erro identificado por James Popham ocorre quando instrumentos de avaliação desenvolvidos, para um determinado propósito, são utilizados para outros fins. Embora nada impeça um professor de encontrar novos usos para um instrumento de avaliação, é necessário garantir que esse instrumento é adequado ao fim a que se destina, caso contrário a informação recolhida poderá ser enviesada. Um exemplo ilustrativo deste tipo de erro poderia ser a aplicação de um teste a um aluno com necessidades educativas especiais que não tenha em conta as suas características e dificuldades. Embora o teste possa estar correto e adequado à generalidade dos alunos, pode não o ser para o aluno em questão.

A terceira categoria de erro está intimamente relacionada com a avaliação formativa. Sendo reconhecido que a avaliação formativa é a que mais contribui para o desenvolvimento das aprendizagens dos alunos, quando não é aplicada, ou é incorretamente utilizada, não produz os efeitos que deveria produzir. Professores com altos níveis de literacia em avaliação conhecem o valor e a utilidade da avaliação formativa, pelo que tomam melhores decisões sobre que instrumentos deverão utilizar com vista ao desenvolvimento das aprendizagens dos alunos. Já os professores com baixos níveis de literacia em avaliação tendem a não utilizar, ou a usar de forma incorreta, este tipo de avaliação.

2.3 Dimensões da Literacia em Avaliação

Sendo a avaliação um conceito multidimensional, a própria Literacia em Avaliação também o é. Tal facto é facilmente constatado quando se analisam as várias definições propostas para o conceito de Literacia em Avaliação. Todas as definições analisadas remetem para um conjunto diversificado de características que os professores devem possuir para que se possam ser considerados letrados no domínio avaliação.

Já em 1990, uma equipa estadunidense constituída por elementos da *American Federation of Teachers (AFT), do National Council on Measurement in Education (NCME) e do National Education Association (NEA)* apresentou os *Standards for Teacher Competence in Educational Assessment of Students*. A elaboração destes *Standards* surgiu da visão dos vários organismos envolvidos de que a avaliação dos alunos é uma parte essencial do processo de ensino e que este é indissociável de uma boa avaliação (AFT, NCME & NEA, 1990). Para que a avaliação tenha a eficácia desejada, os autores reconhecem que a formação inicial é uma etapa fundamental para o desenvolvimento de competências em avaliação. Adicionalmente, deverá existir uma oferta abrangente de formação (ao nível da formação contínua) que permita o aprofundar de conhecimentos nesta área tão sensível.

Os *Standards* definidos correspondem a uma expectativa de conhecimentos ou habilidades que um professor deve possuir para que possa desenvolver uma boa prática em avaliação. Por outras palavras, *all seven standards apply to teachers' development and use of classroom assessment of instructional goals and objectives that form basis for classroom instruction* (Mertler, 2003, p.9).

São sete os *Standards* propostos pela AFT, NCME e NEA (1990), nomeadamente:

1. Teachers should be skilled in choosing assessment methods appropriate

for instructional decisions.

2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
3. Teachers should be skilled in administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods.
4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
5. Teachers should be skilled in developing valid pupil grading procedures that use pupil assessments.
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

Estes *Standards* foram uma importante base para a investigação que se desenvolveu a partir de então na área da Literacia em Avaliação. Para além disso, serviram para uma mudança ao nível dos currículos da Formação de Professores estadunidense (Brookhart, 2011). Muito embora estes *Standards* tenham tido um impacto bastante positivo, quer ao nível da investigação, quer numa mudança de paradigma na formação de professores, autores como Brookhart (2011) e Alkharusi *et al.* (2012) afirmam que elas não consideram as conceções atuais de conhecimento e habilidades de avaliação formativa e que este é um aspeto fundamental na Literacia em Avaliação dos professores.

Stiggins (1991), na sua definição de Literacia em avaliação, remete-nos para 5 domínios principais da literacia em avaliação, nomeadamente:

(i) Capacidade de apresentar propostas claras de avaliação;

(ii) Conhecimento de métodos de avaliação apropriados para diferentes metas de

realização;

(iii) Compreensão sobre a importância de avaliar diferentes metas interrelacionadas;

(iv) Recolha de informações de desempenho dos alunos com base em tarefas propostas;

(v) Capacidade para evitar formas de avaliação distorcidas e erros técnicos.

Segundo Koh (2011) os cinco domínios defendidos por Richard Stiggins correspondem às ideias e princípios da avaliação autêntica. Por conseguinte, e segundo o mesmo autor, *assessment literacy involves being prepared to define, teach and assess the different kinds of competencies that match the higher order instructional goals for the twenty-first century* (Koh, 2011, p.257).

Inspirados pelos seus estudos empíricos sobre a literacia em avaliação de professores de Ciências, Abell e Siegel (2011) desenvolveram um modelo que propõe 4 categorias de conhecimento em avaliação, nomeadamente:

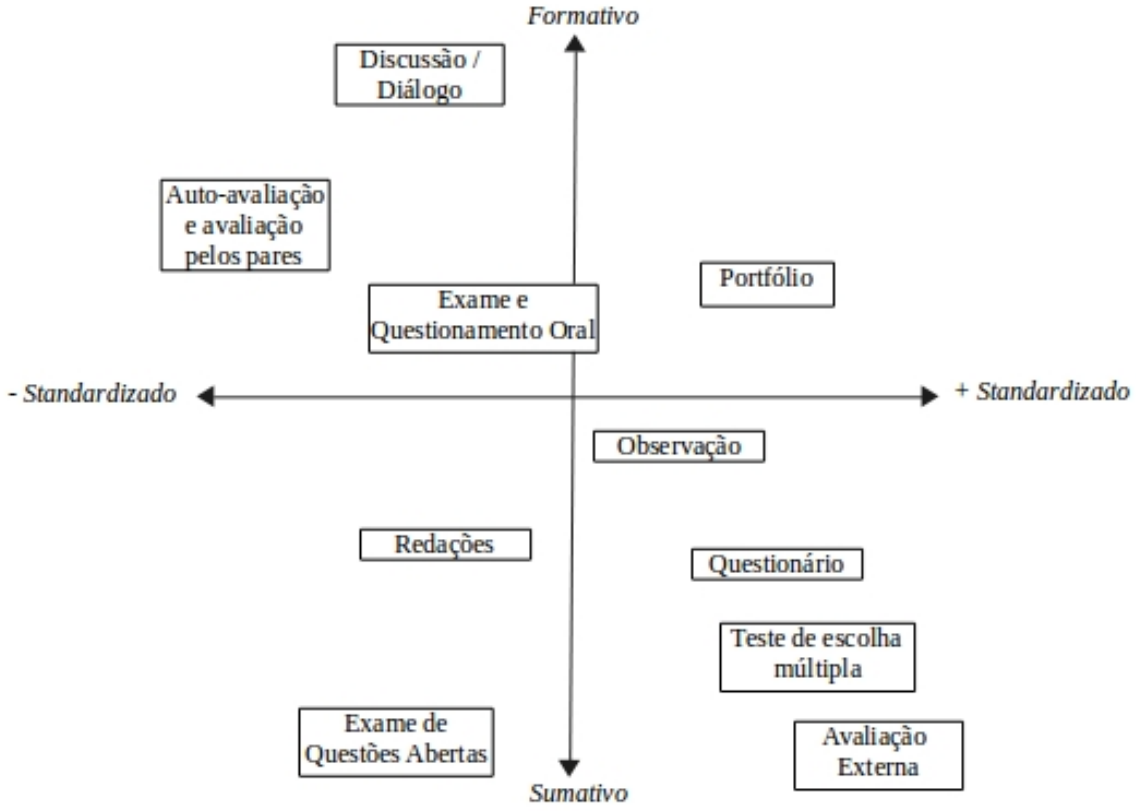
1. **Conhecimento sobre os objetivos da avaliação**: Nesta dimensão é importante verificar até que ponto os professores estão familiarizados com os objetivos da avaliação. Ou seja, para uma boa avaliação é fundamental que os professores saibam porque avaliam, como devem avaliar e quando devem avaliar. Este aspeto está intimanente relacionado com as funções da avaliação, em especial com a avaliação de diagnóstico, com a avaliação formativa e com a avaliação sumativa, pelo que a familiariedade com as características inerentes a cada uma das funções é preponderante. Abell e Siegel (2011) acrescentam ainda a função metacognitiva da avaliação, pois consideram que ajuda os alunos a monitorizarem a sua própria aprendizagem.

2. **Conhecimento sobre o que avaliar**: Esta categoria está intimamente relacionada com o conhecimento sobre o curriculo e as metas de aprendizagem das respetivas disciplinas (no caso português seriam as Aprendizagens Essenciais e o Perfil do Aluno à saída do Ensino Obrigatório). Abell e Siegel (2011) reforçam que *what to assess is related to curricular goals and to values of what is important to learn and how learning occurs* (p.214). Assim, esta dimensão é fundamental, dado que se um professor dominar o currículo e os objetivos da sua área curricular, também a avaliação se focará nos aspetos fundamentais reforçando a validade, fiabilidade e transparência da mesma.

3. **Conhecimento sobre estratégias em avaliação**: o conhecimento das estratégias de avaliação refere-se à forma como um professor avalia a aprendizagem dos alunos numa determinada unidade de curricular. Dessa forma, é necessário que os professores conheçam e saibam aplicar estratégias formais de avaliação, designadamente na avaliação sumativa, bem como estratégias informais, utilizadas especialmente na avaliação formativa. A figura 5 fornece um esboço das formas mais comuns de avaliação considerando, por um lado, as funções da avaliação e, por outro, o seu nível de padronização ou *standardização* (Bayer, Klieme & Jude, 2016). Abell e Siegel (2011) incluem ainda nesta dimensão o conhecimento de formas eficazes de inclusão dos estudantes no processo de avaliação (por exemplo de autoavaliação e de avaliação pelos pares) e o uso eficaz do *feedback* nas suas múltiplas formas.

Figura 5: Algumas formas de avaliação

(Adaptado de Bayer, Klieme e Jude, 2016)



4. **Conhecimento sobre interpretação e utilização da informação recolhida no processo de avaliação**: Autores como Südkamp, Kaiser e Moller (2014) argumentam que os julgamentos dos professores sobre o desempenho dos alunos tem um impacto considerável nas suas experiências de aprendizagem e trajetórias educacionais. Além disso, muitas decisões de instrução são determinadas pelos julgamentos subjetivos dos professores sobre a realização dos seus alunos. A capacidade de avaliar com precisão os resultados dos alunos é, portanto, uma das características-chave de um bom professor. Este tipo de argumentos levaram a que Abell e Siegel (2011) defendessem que, para além do quê, do quando, do como e do porquê de avaliar, os professores tenham competências ao nível do que fazer com a informação recolhida no

processo de avaliação, nomeadamente ao nível da interpretação dos resultados e a sua utilização, à semelhança das propostas de Xu e Brown (2016) e dos *Standards for Teacher Competence in Educational Assessment of Students* (AFT, NCME & NEA, 1990), mas também para uma mudança de práticas que acompanhem e se adaptem às necessidades dos alunos.

Inspirado pelos *Standards for Teacher Competence in Educational Assessment of Students* e pelos mais recentes estudos na área avaliação, Xu e Brown (2016) elencaram um conjunto de áreas de os professores devem ter conhecimento de forma a poderem desenvolver melhores práticas ao nível da avaliação, nomeadamente:

1. **Conhecimento Pedagógico do Conteúdo (PCK)**: O Conhecimento Pedagógico do Conteúdo, ou PCK (*Pedagogical Content Knowledge*), constitui um tipo de conhecimento exclusivo dos professores e corresponde ao cruzamento entre o seu conhecimento pedagógico (o que sabem sobre o ensino) com o que eles sabem sobre o que ensinam (Cochram, King & DeRuiter, 1991, p.5). Lee Shulman (1986), que introduziu pela primeira vez o conceito, refere que o PCK:

> *embodies the aspects of content most germane to its teachability. Within the category of pedagogical content knowledge I include, for the most regularly taught topics in one's subject area, the most useful forms of representation of those ideas, the most powerful analogies, illustrations, examples, explanations, and demonstrations - in a word, the ways of representing and formulating the subject that make it comprehensible to others . . . [It] also includes an understanding of what makes the learning of specific concepts easy or difficult: the conceptions and preconceptions that students of different ages and backgrounds bring with them to the learning* (Shulman, 1986, p. 9).

Este fator assume assim uma especial relevância dado que a avaliação das aprendizagens se centra no conteúdo do currículo ensinado. Assim, como afirma Xu e Brown (2016), o conhecimento das disciplinas e como ensinar esse

conteúdo não pode ser excluído da base dos conhecimentos em avaliação;

2. **Conhecimento das finalidades da avaliação, do conteúdo e dos métodos**: Para uma boa avaliação é fundamental que os professores compreendam a razão de avaliarem. Para além disso, é necessário que compreendam de que forma os diferentes métodos de avaliação se podem relacionar com os objetivos de aprendizagem das respetivas disciplinas;

3. **Conhecimento sobre escalas de classificação**: Um aspeto determinante para a avaliação é o conhecimento que os professores necessitam ter sobre os fundamentos das escalas de classificação, bem como os princípios e métodos para a criação das mesmas;

4. **Conhecimentos sobre o *feedback***: O *feedback* é um elemento crucial nas boas práticas em avaliação. Assim, é muito importante os professores conhecerem os fundamentos, princípios, formas, funções e características do *feedback* para que, dessa forma, possam contribuir para uma facilitação no processo de ensino e aprendizagem;

5. **Conhecimento sobre a interpretação e comunicação de resultados da avaliação**: Xu e Brown (2016), tal como outros autores já referidos, reforçam a importância dos professores conhecerem os melhores métodos para extraírem todas as informações relevantes do processo de avaliação. Para além disso, os professores deverão conhecer métodos para comunicarem esses mesmos resultados às várias partes interessadas (alunos, pais, encarregados de educação, entre outros).

6. **Conhecimentos sobre o envolvimento dos estudantes no processo de avaliação**: Os autores destacam também a importância dos professores

entenderem as vantagens de envolver os alunos no processo de avaliação. É igualmente necessário que promovam mais e melhores estratégias de autoavaliação e de avaliação por pares, de forma a incentivar uma participação mais ativa dos alunos no processo de avaliação.

7. **Conhecimentos sobre a ética em avaliação**: Por último, Xu e Brown (2016), referem que os professores precisam compreender as responsabilidades legais e éticas relativas ao uso, armazenamento e disseminação dos resultados da avaliação. Além disso, é necessário que os professores saibam trabalhar em prol da equidade, da não discriminação, da inclusão e da justiça social.

Dos autores analisados, verifica-se que, embora se distingam pelo número de categorias que consideram na definição da literacia em avaliação, existem vários pontos em comum. Destes destacam-se a necessidade de conhecer as funções e objetivos em avaliação, a importância do conhecimento do currículo (de forma a reconhecer aquilo que é realmente importante aprender e avaliar), a necessidade do conhecimento de diferentes estratégias de avaliação, consoante o contexto de ensino e aprendizagem, e o conhecimento sobre como interpretar e comunicar a informação recolhida no processo de avaliação e como essa informação deve ser reinvestida de forma a melhorar as aprendizagens dos alunos.

2.4 Medir a Literacia em Avaliação: Alguns Instrumentos

Vários estudos de natureza quantitativa foram conduzidos com o objetivo de medir a literacia em avaliação dos professores, tanto em contexto de formação inicial,

como em exercício de funções. Desta forma, procuraremos analisar, neste subcapítulo, alguns desses instrumentos tendo em consideração as suas características, estudos em que foram utilizados, resultados alcançados e potencialidades e limitações da sua administração.

2.4.1 TALQ - *Teacher Assessment Literacy Questionnaire*

A partir dos *Standards for Teacher Competence in Educational Assessment of Students* e reconhecendo a necessidade dos professores em dominarem uma série de questões relacionadas com a avaliação das aprendizagens, Plake e Impara (1992) desenvolveram um intrumento de forma a avaliar os conhecimentos dos professores neste domínio. O instrumento, designado de *Teacher Assessment Literacy Questionnaire* (TALQ) foi aplicado em 1991 a um conjunto de 555 professores em serviço de 48 Estados norteamericanos (Plake, Impara & Fager, 1993).

O TALQ é constituído por duas partes distintas. A primeira apresenta um conjunto de 35 questões de escolha múltipla com o objetivo de *assess teachers' knowledge in the competency areas identified in the Standards* (Plake, Impara & Fager, 1993, p.10). Assim, para cada um dos sete *standards* foram desenvolvidas cinco questões distintas, pelo que a cotação a cada um dos *standards* poderia variar entre os 0 e os 5 pontos. Já a segunda parte é constituída por itens relacionados com os dados gerais dos respondentes. Os resultados da aplicação do TALQ encontram-se resumidos na Tabela 1.

Tabela 1: Resumo dos resultados do TALQ

(Plake, Impara e Fager, 1993)

Standards	**Amostra**	**Média**	**Desvio-Padrão**
1. Escolher métodos apropriados	555	3,46	0,93
2. Desenvolver métodos apropriados	555	3,22	0,80
3. Aplicar, classificar e interpretar os resultados	555	3,96	0,90
4. Utilizar e tomar decisões com os resultados	555	3,40	1,11
5. Desenvolver processos de classificação	555	3,19	0,78
6. Comunicar os resultados da avaliação	555	2,70	1,21
7. Ética na avaliação	555	3,26	0,78
Total	**555**	**23,20**	**3,33**

Conforme se poderá verificar pela Tabela 1, o valor total médio de respostas corretas foi de apenas 23,20 em 35, o que representa pouco mais de 65% de acertos. Verifica-se também que a área em que a amostra obteve piores resultados foi na comunicação dos resultados da avaliação (média de 2,70 em 5) e que, por outro lado, os melhores resultados foram conseguidos no domínio da aplicação, classificação e interpretação dos resultados (média de 3,96 em 5). Embora a consistência interna tenha sido relativamente baixa (KR-20=0,54), os resultados apresentados sugerem as fragilidades existentes ao nível das competências em avaliação. Os autores verificaram também que os professores que haviam tido formação em avaliação obtiveram resultados substancialmente mais elevados que os professores que não haviam tido esse tipo de formação. Tal facto, parece ser revelador da necessidade em apostar na formação em avaliação logo a partir da formação inicial, de forma a minimizar estas lacunas, mas também em contexto de formação contínua.

Vários estudos foram realizados com recurso ao TALQ. Breziat e Coleman (2015) aplicaram o TALQ a um pequeno grupo de professores em formação de uma Universidade americana em dois momentos distintos, no início e no fim de uma disciplina de psicologia educacional, oferecida pela referida universidade, que abordava algumas questões relacionadas com a avaliação. Os resultados foram de

17,92 no primeiro momento e de 18,15 no segundo momento, o que corresponde a uma percentagem de acertos de 51,2% e 51,9%, respetivamente. Em ambos os momentos, o domínio que obteve melhores resultados foi o da escolha apropriada dos métodos de avaliação (com 3,08 e 3,35 respetivamente) e o pior foi, à semelhança do estudo desenvolvido por Plake, Impara e Fager (1993), o da comunicação dos resultados da avaliação (com 1,5 e 1,46 respetivamente). Verifica-se, através destes resultados, que a evolução foi positiva, mas pouco significativa (variação de apenas 0,7 pontos percentuais). Outra conclusão interessante que o estudo revelou foi de que os professores (em formação) do ensino secundário (*secondary*) apresentaram melhores resultados que os do ensino básico (*elementary*) e estes melhores resultados que os de educação pré-escolar (*early childhood*). De referir ainda que neste estudo o TALQ apresentou uma consistência interna maior (KR-20=0,77) que no estudo anterior, embora a amostra não permita realizar grandes generalizações. Contudo, parece-nos evidente que a literacia em avaliação é relativamente superior nos professores em exercício do que nos professores em formação, o que sugere, tal como defende Sohlberg *et al.* (2007), que a aprendizagem da profissão de docente se dá, sobretudo, em contexto de trabalho e não na formação inicial.

Alguns estudos foram também realizados por Alkharusi com recurso ao TALQ (*e.g.* Alkharusi, 2011a; Alkharusi *et al.*, 2012), especialmente em Omã. Num artigo publicado em 2012, Alkharusi e a sua equipa tinham, como um dos principais objetivos, descrever o conhecimento dos professores relativamente à avaliação educacional (Alkharusi *et al.*, 2012) utilizando, entre outros instrumentos, o TALQ[9]. Esta versão do TALQ foi aplicada a uma amostra aleatória de 165 professores em serviço na Governação Educativa de Muscat (capital e maior área metropolitana de

[9]A versão utilizada do TALQ foi ligeiramente alterada, ao invés das tradicionais 35 questões foram utilizadas 32 de forma a ir ao encontro da realidade omanense.

Omã). Os resultados obtidos, à semelhança dos estudos descritos anteriormente, foram baixos, demonstrando, mais uma vez, as fragilidades que os professores apresentam ao nível da literacia em avaliação já que o valor médio foi de apenas 12,42 (num máximo de 32), o que representa uma taxa de acertos na ordem dos 39%. De referir ainda que a consistência interna foi mais uma vez baixa (KR-20=0,62).

2.4.2 *Assessment Literacy Inventory* (ALI) e *Classroom Assessment Literacy Inventory* (CALI)

Campbell *et al.* (2002) procederam a pequenas alterações ao TALQ e renomearan-no como *Assessment Literacy Inventory* (ALI). Este instrumento foi aplicado a um conjunto de 220 professores, na sua formação inicial, que se encontrava a frequentar uma disciplina relacionada com a avaliação (Mertler, 2003). A média alcançada foi de cerca de 21 (num máximo de 35), tendo ficado ligeiramente abaixo dos resultados obtidos por Plake, Impara e Fager (1993) que, recorde-se, foi de 23,2 num máximo de 35. No entanto, há a salientar que a consistência interna foi significativamente superior (α =0,74).

Também Mertler (2003) utilizou uma versão adaptada por si do TALQ à qual designou de *Classroom Assessment Literacy Inventory* (CALI), com o objetivo de comparar a literacia em avaliação em dois grupos distintos, professores em formação e professores em serviço. O grupo de professores em formação, constituído por 67 elementos, obteve uma média ligeiramente inferior a 19 (em 35 possíveis) e um valor aceitável de consistência interna (α =0,74). O *standard* no qual este grupo de professores obteve melhores resultados foi o da escolha apropriada de métodos de avaliação, com uma média de 3,25 num máximo possível de 5, já os piores resultados foram obtidos no domínio do desenvolvimento de processos de

classificação, com uma média de 2,06 num máximo possível de 5. Já o grupo de professores em serviço, constituído por 197 elementos, obteve uma média ligeiramente superior ao grupo anterior, cerca de 22 em 35 possíveis. No entanto, o nível de consistência interna do instrumento aplicado a este grupo foi bastante baixo (α =0,57). Os melhores resultados deste grupo foram alcançados no domínio da aplicação, classificação e interpretação de resultados (média de 3,95 em 5 possíveis) e os piores no domínio do desenvolvimento de processos de classificação (com uma média de 2,06 em 5 possíveis).

Os estudos desenvolvidos tanto por Campbell *et al.* (2002) como por Mertler (2003) vieram, como refere Hailaya *et al.* (2014), pôr em evidência as fragilidades psicométricas do TALQ, em especial, devido à sua baixa fiabilidade, evidenciada nos baixos valores de consistência interna. Para além disso, afirmam os autores que *the original instrument was difficult to read, extremely lengthy, and contained items that were presented in a decontextualized way* (Mertler e Campbell, 2005, pp.8-9). Tais críticas levaram Mertler e Campbell a desenvolver um novo instrumento para determinar a literacia em avaliação dos professores, nascendo assim o novo *Assessment Literacy Inventory* (ALI) (Mertler & Campbell, 2005).

2.4.3 O novo *Assessment Literacy Inventory*

Tanto a estrutura como as questões colocadas no novo ALI diferem do TALQ. O número de questões é o mesmo (35 questões) mas estas encontram-se organizadas em torno de cinco cenários, com sete questões cada, sendo que os respondentes deverão escolher uma das quatro opções que reflete a sua decisão perante aquele cenário que é proposto (Mertler & Campbell, 2005). Cada cenário reflete uma situação em sala de aula, levando o professor a desenvolver determinadas ações ou tomar

decisões relacionadas com a avaliação. De referir ainda que, à semelhança do TALQ, o ALI está alinhado com os sete *Standards for Teacher Competence in Educational Assessment of Students* (daí a existência de sete questões para cada cenário).

A aplicação do ALI foi realizada em duas fases. Na primeira fase foi realizado um teste piloto a 152 professores em formação. A coerência interna (KR-20) foi de 0,75, o índice de dificuldade médio 0,64 e o índice de discriminação médio foi de 0,32. Estes valores *indicate that the ALI appeared to function reasonably well, from a psychometric perspective* (Mertler & Campbell, 2005, p.10). Posteriormente, e após algumas alterações, procedeu-se à segunda fase do teste piloto, com uma amostra de 250 professores em formação de duas instituições onde a frequência de uma disciplina diretamente relacionada com as questões da avaliação era obrigatória. Os resultados foram muito semelhantes aos da primeira fase. A coerência interna (KR-20) foi de 0,74, o índice de dificuldade médio de 0,682 e o índice de discriminação médio de 0,313. Estes resultados vieram uma vez mais reforçar as qualidades psicométricas do ALI. Tais qualidades, segundo Mertler e Campbell (2005), validam a utilização do ALI enquanto uma ferramenta aceitável de medida da literacia em avaliação dos professores.

Após a publicação do ALI por Mertler e Campbell, vários foram os estudos realizados com recurso a este instrumento. Hailaya *et al.* (2014) publicou um artigo onde o principal objetivo era analisar a utilidade e portabilidade do ALI nos sistemas educativos da região Ásia-Pacífico, em especial no contexto filipino, onde o estudo foi conduzido. Foram realizadas algumas adaptações à versão original do ALI, de forma a enquadrar-se no contexto em que a investigação ocorreu. Tais adaptações foram realizadas, sobretudo, ao nível dos nomes utilizados e alguns esclarecimentos adicionais foram dados aos vários cenários descritos. Apesar das alterações introduzidas, os autores procuraram preservar a integridade do instrumento através

de uma validação levada a cabo pelos próprios autores e por dois especialistas. A partir da aplicação do ALI a um conjunto de 582 professores de 128 escolas da província de Tawi-Tawi, nas Filipinas, os autores recorreram a um conjunto de métodos estatísticos, como o modelo Rasch e a Análise Fatorial Confirmatória (CFA) para aferirem as qualidades psicométricas do ALI no contexto em questão. Embora Hailaya *et al.* (2014) reconheçam algumas qualidades psicométricas do ALI, o que o torna um instrumento útil para aferir a literacia em avaliação dos professores, os resultados mostraram, por um lado, a necessidade de aplicar o ALI em outras contextos, que não aquele em que Mertler e Campbell (2005) aplicaram, bem como, por outro, o de utilizar outros métodos de validação. Para além disso, os autores concluiram que para que o ALI possa ser utilizado em diferentes contextos, deverá ser alvo de uma revisão profunda, nomeadamente ao nível da sua estrutura e da clarificação dos vários itens que a constituem.

2.4.4 MLQ - *Measurement Literacy Questionnaire*

Embora os instrumentos já mencionados sejam os mais amplamente utilizados, ao longo das últimas décadas foram desenvolvidos vários outros que merecem o nosso interesse e a nossa análise. Destacaremos aqui mais um exemplo de instrumento desenvolvido para a aferição da Literacia em avaliação. Esse destaque, deve-se a dois fatores principais: o primeiro prende-se com a sua estrutura e o segundo com o facto de não ter sido desenvolvido com base nos *Standards for Teacher Competence in Educational Assessment of Students*.

O *Measurement Literacy Questionnaire* (MLQ) foi desenvolvido por Larry Daniel e Debra King (1998) com o objetivo de (a) determinar a literacia dos professores do ensino básico e secundário no respeitante ao desenvolvimento de testes e outros

instrumentos de medida, (b) examinar de que forma são aplicados os conceitos de testes e medição em contexto de avaliação em sala de aula e (c) determinar se as estratégias de avaliação variam entre os professores de diferentes níveis de ensino.

Para além disso, os diversos itens que constituem o questionário foram desenvolvidos com base na literatura especializada na área da avaliação onde se destacam autores como Gullickson (1984), Kubiszyn e Borich (1996) e Popham (1995). Ao contrário dos vários instrumentos analisados até aqui e que se inspiraram nos *Standards for Teacher Competence in Educational Assessment of Students*, o MLQ debruçou-se sobre 3 domínios: os conhecimentos em avaliação, as competências na interpretação dos resultados da avaliação e nas competências de comunicação desses mesmos resultados.

O MLQ é constituído por 67 itens organizados em três partes distintas: 7 itens da informações gerais sobre os participantes, 30 itens de verdadeiro e falso para aferir conhecimentos sobre testes e outros instrumentos de medida e 30 itens do tipo *Likert* relacionados com a utilização de métodos e técnicas em avaliação. O instrumento foi aplicado a um conjunto de 95 professores em serviço do ensino básico (*elementary*) e do ensino secundário (*secondary*) de duas escolas do sul do Mississipi (E.U.A.). Os resultados alcançados neste estudo vieram confirmar estudos anteriores (Gullickson, 1985; Scales, 1993; Schafer e Lissitz, 1987; Stiggins e Bridgeford, 1985; Salmon-Cox, 1982) nomeadamente no que diz respeito à falta de conhecimentos adequados sobre os procedimentos em avaliação. Outro aspeto importante que os autores concluiram foi que a aplicação de testes é a forma de avaliação mais frequente utilizadas pelos professores.

2.5 Síntese

Neste capítulo procurou-se fazer uma fundamentação teórica em torno da problemática da literacia em avaliação. Esta fundamentação focou-se em 4 aspetos considerados essenciais para a compreensão da problemática que nos propomos analisar no desenvolvimento da Tese: A definição de Literacia em Avaliação e a sua importância, as dimensões consideradas quando se analisa a literacia em avaliação e os instrumentos já desenvolvidos por outros autores que possibilitaram a medição da literacia em avaliação.

Ao nível da conceptualização foram analisados autores como Stiggins (1991, 1995), Quilter e Gallini (2000), Paterno (2001), Mertler (2004) e Popham (2011). A partir destes autores, delineou-se uma definição de literacia em avaliação que norteou o desenvolvimento da presente tese. Na definição apresentada considera-se a literacia em avaliação como um conhecimento aprofundado dos aspetos relacionados com a avaliação. Destes, destacaram-se, numa primeira fase, 3 domínios que se consideraram fundamentais:

- O conhecimento dos príncipios e funções da avaliação;

- O conhecimento sobre a utilização de instrumentos de avaliação diversificados;

- O conhecimento sobre a utilização e interpretação dos resultados obtidos no processo de avaliação.

Numa análise posterior, e a partir da leitura de autores como Xu e Brown (2016), Abell e Siegel (2011), Cochram *et al*. (1991) e Shulman (1986), considerou-se fundamental a inclusão de uma quarta dimensão: o conhecimento sobre o currículo e sobre aquilo que é importante aprender e avaliar. Esta quarta dimensão está

diretamente relacionada com o conceito de conhecimento pedagógico do conteúdo, ou seja, o cruzamento entre o conhecimento científico e o conhecimento pedagógico.

Ao nível da importância da literacia em avaliação, e da sua relação com a melhoria da aprendizagem dos alunos, analisaram-se alguns dos principais autores que desenvolveram trabalho neste domínio dos quais se destacam Popham (2018), Xu e Brown (2016), McGee e Colby (2014), Malone (2013), Gottheiner e Siegel (2012), Newfields (2006) e Stiggins (2002). Da análise destes autores conclui-se que há uma preparação inadequada dos professores face à tarefa de avaliar. Esta preparação inadequada deve-se sobretudo a uma carência de disciplinas de avaliação nos programas de formação inicial de professores. Tais lacunas na formação inicial, têm um impacto direto na capacidade dos professores em desenvolver e aplicar formas de avaliar diversificadas e na capacidade para interpretar e utilizar os resultados oriundos da aplicação dos instrumentos de avaliação. Agrava-se ainda mais quando os professores são responsáveis por erros na utilização da avaliação, pondo em causa as próprias aprendizagens dos alunos.

Quanto à análise das dimensões da literacia em avaliação, consideraram-se os contributos teóricos de Xu e Brown (2016), Abel e Siegel (2011), Stiggins (1991) e AFT, NCME e NEA (1990). Tais contributos foram fundamentais para a definição das dimensões da Literacia em Avaliação a analisar na tese e que já foram referidos anteriormente.

Por último, analisaram-se os principais instrumentos de aferição da literacia em avaliação, em especial o Teacher Assessment Literacy Questionnaire (AFT, NCME & NEA, 1990), o Assessment Literacy Inventory (Campbel *et. al*, 2002), o Classroom Assessment Literacy Inventory (Mertler, 2003), o novo Assessment Literacy Inventory (Mertler & Campbell, 2005) e o Measurement Literacy Questionnaire (Daniel & King,

1998). A análise destes instrumentos foi fundamental para a construção do questionário que foi aplicado na nossa investigação, o Questionário de Aferição da Literacia em Avaliação (QALA).

Capítulo 3

Metodologia do Estudo Empírico

Neste capítulo serão desenvolvidas as opções metodológicas adotadas relativamente aos objetivos a que nos propusemos na presente investigação. Assim, e num primeiro momento, analisar-se-ão alguns aspetos teóricos relacionados com a metodologia quantitativa em educação, visto ser essa a abordagem em que a presente investigação assenta. Num segundo momento, abordar-se-ão os fundamentos teóricos relacionados com o design do estudo, neste caso, a pesquisa por *survey* interseccional. A opção por este tipo de design deveu-se ao facto de ser nosso objetivo a recolha de dados quantitativos, num momento apenas e a uma determinada amostra, pelo que o design adotado se assume como a melhor opção. Neste capítulo serão também abordados os aspetos relacionados com a delimitação da amostra e a caracterização do instrumento a ser aplicado, neste caso o QALA.

3.1 Breve enquadramento da abordagem quantitativa em educação

Os métodos de investigação em ciências sociais, no geral, e em educação, em particular, assentam sobretudo em dois tipos: Qualitativo e Quantitativo (Muijs, 2004). Estando a presente tese assente em métodos de investigação quantitativa, abordaremos aqui alguns aspetos relacionados com este método de investigação.

Aliaga e Gunderson (2002) definem investigação quantitativa como a explicação de fenómenos a partir da recolha de dados numéricos analisados com recurso a métodos matemáticos (em especial estatísticos). Segundo Guilford e Frutcher (1978), os dados númericos inserem-se, regra geral, em duas categorias: coisas que podem ser contadas, estabelecendo-se frequências, e coisas que podem ser medidas, constituindo as métricas ou as escalas. Em ambos os casos é possível a utilização de procedimentos estatísticos para a análise dos dados (Guilford & Frutcher, 1978).

Partindo da definição de investigação quantitativa de Aliaga e Gunderson (2002) apresentada no parágrafo anterior, Muijs (2004) procedeu a uma análise passo a passo. A primeira componente da definição é a de 'Explicar fenómenos' o que, segundo o autor é:

> a key element of all research, be it quantitative or qualitative, when we set out do some research, we are always looking to explain something. In education this could be questions like *why do teachers leave teaching?', 'what factors influence pupil achievement?* and so on. (Muijs, 2004, p.1).

Em segundo lugar, a investigação quantitativa assenta na recolha de 'dados numéricos' (Muijs, 2004; Vieira, 2009), sendo esta uma das principais especificidades que a distinguem da investigação qualitativa. Este segundo aspeto está intimamente

relacionado com a última parte da definição apresentada por Aliaga e Gunderson (2002) que destaca a utilização de métodos matemáticos (em especial estatísticos) para a análise dos dados recolhidos.

A recolha de dados numéricos, para a investigação quantitativa em educação, nem sempre é imediata, uma vez que são poucos os fenómenos que geram dados quantitativos de forma *natural* (Muijs, 2004). No entanto, e como refere o autor:

> Many data that do not naturally appear in quantitative form can be collected in a quantitative way. We do this by designing research instruments aimed specifically at converting phenomena that don't naturally exist in quantitative form into quantitative data, which we can analyse statistically. Examples of this are attitudes and beliefs. We might want to collect data on pupils' attitudes to their school and their teachers. These attitudes obviously do not naturally exist in quantitative form (we don't form our attitudes in the shape of numerical scales!). Yet we can develop a questionnaire that asks pupils to rate a number of statements (for example, 'I think school is boring') as either agree strongly, agree, disagree or disagree strongly, and give the answers a number (*e.g.* 1 for disagree strongly, 4 for agree strongly) (Muijs, 2004, p.2).

Esta é uma das várias técnicas que se pode utilizar para obter dados quantitativos que, à partida, não existem. Assim, a partir da utilização de questionários e testes, é possível recolher dados quantitativos de um conjunto mais alargado de fenómenos ligados à educação.

De um modo geral, o processo de investigação, segundo Creswell (2012), desenrola-se em seis etapas, nomeadamente a identificação da problemática, a revisão de literatura, a definição dos objetivos/questões de investigação, a recolha de dados, a análise e interpretação dos dados recolhidos e a escrita do relatório e avaliação da investigação. Analisemos agora as características da investigação quantitativa à luz de cada uma das etapas identificadas.

Na investigação quantitativa, o investigador procura identificar a sua problemática

a partir das tendências no seu campo de estudo ou a partir da necessidade de explicar alguns fenómenos. Para Creswell (2012), descrever uma tendência significa que o problema da pesquisa pode ser melhor respondido por um estudo no qual o investigador procura estabelecer uma tendência geral das respostas dos indivíduos e verificar de que forma essa tendência varia entre as pessoas. A investigação quantitativa pode também incidir sobre de que forma determinadas variáveis afetam outras variáveis. Através da análise da relação de variáveis, o investigador poderá determinar de que forma um determinado grupo de variáveis pode (ou não) estar correlacionado com outro grupo de variáveis.

A revisão de literatura nos estudos quantitativos, segundo Creswell (2012), tem uma grande importância sobretudo no início do estudo. Segundo o autor, esta importância revela-se de duas formas: por um lado, justifica a necessidade daquele estudo e daquela problemática e, por outro, sugere ao autor novas abordagens de investigação bem como pode sugerir novas questões de investigação. A revisão de literatura assume especial relevância na investigação quantitativa, uma vez que é um elemento crucial para a identificação de potenciais variáveis, relações entre variáveis e tendências (Creswell, 2012) pelo que se assume como um importante auxiliar na formulação tanto das questões de investigação como das hipóteses.

As questões de investigação, na abordagem quantitativa, deverão ser limitadas e possibilitar a identificação dos dados a recolher, sendo que estes deverão ser quantificáveis e observáveis. As questões deverão ainda, segundo Neuman (2007), contemplar a relação de um pequeno número de variáveis. Neuman (2007) sugere ainda outra forma de formular uma questão de investigação a partir da identificação do universo à qual a resposta poderá ser generalizada. Ainda segundo o autor:

> All research questions, hypotheses, and studies apply to some group or category of people, organizations, or other units. The universe is the set of

> all units that research covers, or to which it can be generalized (Neuman, 2007, p.87).

Para a recolha de dados quantitativos, são utilizados instrumentos que permitem mensurar as variáveis em estudo ou estabelecer frequências. Estes instrumentos, segundo Creswell (2012), contemplam um conjunto de questões e possibilidades de respostas preestabelecidas. Exemplos de instrumentos de recolha de dados quantitativos são os questionários, os testes padronizados e as listas de verificação que, quando aplicados, possibilitam a recolha de dados sob a forma de números.

A análise dos dados, tal como referido na definição de Aliaga e Gunderson (2002), é realizada a partir de processos matemáticos, em especial especial com recurso à estatística. Esta análise:

> consist of breaking down the data into parts to answer the research questions. Statistical procedures such as comparing groups or relating scores for individuals provide information to address the research questions or hypotheses (Creswell, 2012, p.15).

Por último, o relatório e avaliação da investigação quantitativa obedece, regra geral a uma estrutura tipo: introdução, revisão de literatura, metodologia, resultados e discussão (Creswell, 2012). Sendo uma estrutura padrão, facilita a análise e a avaliação da qualidade do estudo, uma vez que os elementos que os constituem são facilmente localizados. Para além disso, as características da investigação quantitativa levam a que os resultados tenham uma menor predisposição para serem alvo dos julgamentos pessoais do investigador, uma vez que são utilizados diversos procedimentos de validação que dificultam que tal aconteça.

Referir ainda que, dentro da abordagem quantitativa, são considerados normalmente dois tipos de *design* de métodos de investigação: Os métodos experimentais e quasi-experimentais e os métodos não-experimentais (Edmonds &

Kennedy, 2017; Mertens, 2010; Muijs, 2004). Relativamente a estes dois métodos, Muijs (2004) refere que:

> Experimental designs are sometimes known as 'the scientific method' due to their popularity in scientific research where they originated. Non-experimental research is sometimes [...] equated with survey research and is very common in the social sciences (Muijs, 2004, p.13).

A presente investigação enquadra-se nos métodos de investigação não-experimentais, nomeadamente nos *survey designs*. No ponto seguinte, abordaremos de forma mais detalhada o *design* do presente projeto de investigação que se insere na pesquisa interseccional, um dos vários exemplos de métodos não-experimentais da investigação quantitativa.

3.2 A pesquisa por *survey* como desenho de investigação

Entre as várias abordagens quantitativas de investigação, as do tipo survey são, talvez, as mais utilizadas. Muijs (2004), afirma mesmo que são a forma mais popular de fazer investigação em Ciências Sociais uma vez que são mais flexíveis e podem assumir diferentes formas.

Kerlinger (1979) definiu a pesquisa por *survey* como um estudo aplicado a grandes ou pequenas populações através da seleção de amostras escolhidas da população que se pretende estudar e, através delas, descubrir a incidência, a distribuição e a relação de fenómenos.

Já Creswell (2012) refere que a investigação por *survey* consiste num conjunto de procedimentos, enquadrados na investigação quantitativa, onde o investigador aplica

um *survey* a uma amostra, ou à totalidade da população, de forma a descrever as suas atitudes, opiniões, comportamentos ou características. Neuman (2006) elenca um conjunto de possibilidades de categorias (Tabela 2) em que um *survey* pode incidir, sendo que, dependendo da natureza da investigação, poderão ser analisadas mais do que uma categoria.

Tabela 2: Categorias de Questões de um *survey* (Adaptado de Neumam, 2006)

Categoria	**Exemplos de Questões**
Comportamentos	Com que frequência escova os dentes?Votou nas últimas eleições legislativas? Quando visitou um parente próximo pela última vez?
Atitudes, Crenças e Opiniões	Acredita que o Presidente da República está a fazer um bom trabalho?Pensa que as outras pessoas referem mais aspetos negativos a seu respeito quando não está presente?
Características	Qual a sua idade? Qual o seu estado civil? Quantas horas dorme, em média, por noite?
Expectativas	Planeia comprar automóvel novo nos próximos 12 meses? Como pensa que irá evoluir a população da sua cidade nos próximos 10 anos?
Auto-classificação	Considera-se religioso? Politicamente, enquadra-se à esquerda, centro ou direita?
Conhecimento	Que partido político obteve mais votos nas últimas eleições legislativas? Qual a percentagem de estrangeiros a residir na sua cidade?

A utilização dos *surveys* é especialmente indicada quando os dados necessários à investigação não existem e as questões de investigação não são suscetíveis de uma verificação experimental (Gorard, 2001). Para além disso, são uma das melhores formas de obtenção de dados sobre factos simples ou de comportamentos. Assim, com a aplicação deste tipo de instrumentos, o investigador obtém vários indicadores que lhe permitem medir variáveis e, dessa forma, testar múltiplas hipóteses. Como refere Creswell (2012):

> researchers collect quantitative, numbered data using questionnaires [...] or interviews [...] and statistically analyze the data to describe trends about responses to questions and to test research questions or hypotheses. They

> also interpret the meaning of the data by relating results of the statistical test back to past research studies (p.376)

Dependendo do tempo em que os dados são coletados, os *survey* podem ser classificados de interseccionais ou longitudinais. No caso de um *survey* interseccional, os dados são recolhidos num certo momento, de uma amostra selecionada, para descrever alguma população na mesma ocasião (Babbie, 2003), mas também para determinar relações entre variáveis. Já a pesquisa longitudinal tem como característica fundamental o facto de envolver a recolha de dados de uma amostra em, pelo menos, dois momentos distintos. São considerados três tipos de pesquisa longitudinal: Estudos de tendência, estudos de coorte e estudos de painel (Creswell, 2012; Edmonds & Kennedy, 2017):

- Estudos de Tendência: Neste tipo de estudo é identificada uma população de forma a examinar as mudanças ao longo do tempo. A amostra utilizada não é necessariamente a mesma ao longo do tempo:

- Estudo de coorte. Estas pesquisas incidem numa subpopulação baseada em características específicas. No entanto, tal como no caso dos estudos de tendência, as amostras utilizadas, ao longo do tempo, não são necessariamente as mesmas.

- Estudo de Painel: Neste caso, a amostra utilizada é sempre a mesma. A grande vantagem deste tipo de estudo, segundo Creswell (2012), é possibilitar ao investigador determinar as alterações que ocorrem dentro de um grupo específico de indivíduos, pelo que este tipo de estudo é considerado o mais rigoroso dos três tipos de pesquisa longitudinal (Creswell, 2012; Neuman, 2007).

A presente investigação enquadra-se, desta forma, na pesquisa por *survey* interseccional. Para o efeito, foi desenvolvido e aplicado um questionário a

professores do ensino básico e secundário, a lecionar em estabelecimentos do ensino público e privado na Zona Pedagógica de Lisboa e Península de Setúbal. O questionário desenvolvido tem por base dois grandes objetivos. Por um lado, visa recolher dados sobre as perceções que os professores têm sobre os conhecimentos e competências que possuem em avaliação (Parte 2 do Questionário de Aferição da Literacia em Avaliação - QALA). Por outro, visa avaliar os conhecimentos que os professores têm sobre vários aspetos relacionados com a avaliação, permitindo, dessa forma, aferir a literacia em avaliação dos referidos professores (Parte 3 e 4 do QALA).

3.3 Delimitação e caracterização dos Participantes

A investigação teve como população-alvo os professores do ensino básico e secundário que lecionavam, aquando da aplicação do QALA, na Zona Pedagógica 7 (figura 6) que engloba os seguintes concelhos:

- Concelho da Amadora
- Concelho do Barreiro
- Concelho da Moita
- Concelho de Alcochete
- Concelho de Almada
- Concelho de Cascais
- Concelho de Lisboa
- Concelho de Loures
- Concelho do Montijo
- Concelho de Odivelas
- Concelho de Oeiras

- Concelho de Palmela
- Concelho do Seixal
- Concelho de Sesimbra
- Concelho de Setúbal
- Concelho de Sintra
- Concelho de Vila Franca de Xira

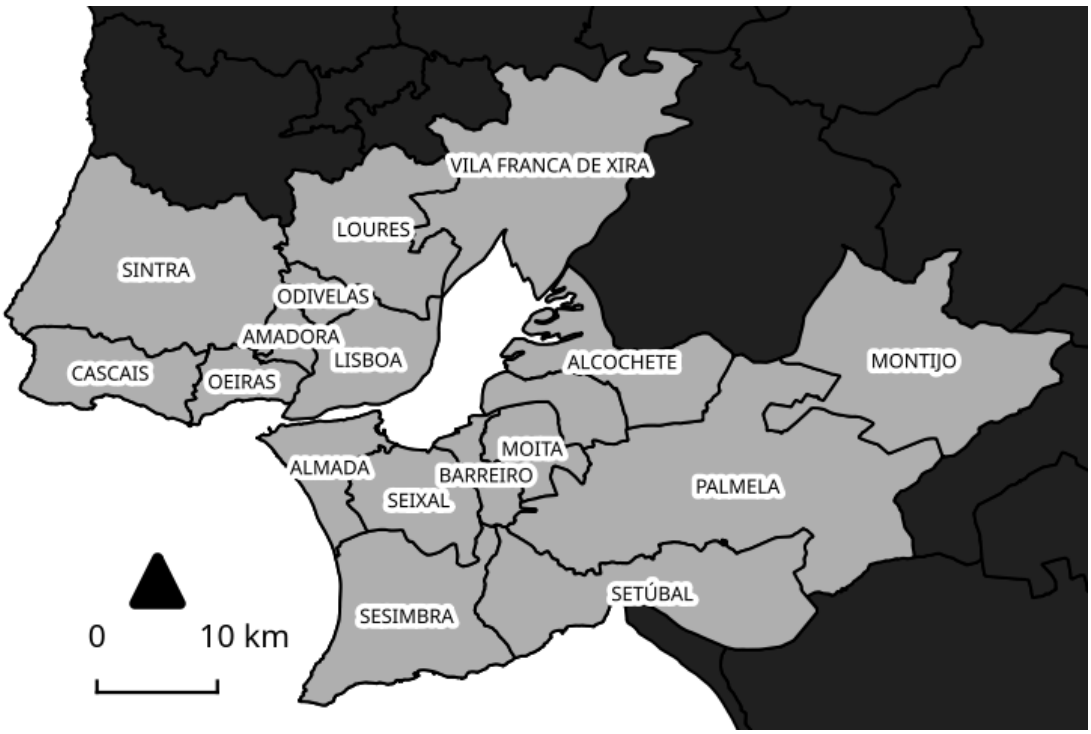


Figura 6: Zona Pedagógica 7 - Lisboa e Península de Setúbal

Embora se reconheça a importância de alargar uma investigação desta natureza a todos os professores do território nacional, optámos por nos cingir a uma unidade territorial de menor dimensão, a Zona Pedagógica 7 - Lisboa e Península de Setúbal. Vários autores, que analisaram a literacia em avaliação dos professores, também optaram por delimitar os seus estudos a territórios de menor dimensão. Alguns exemplos que podemos destacar são Hailaya, Alagumalai e Ben (2014), que aplicaram o ALI na divisão de Tawi-Tawi (Filipinas), Yamtim e Wongwanich (2014), que aplicaram o CALI no distrito de Suphan Buri (Tailândia), Alkharusi *et. al* (2012)

aplicaram o TALQ em Muscat (Omã) e Daniel e King (1998), que aplicaram o MLQ no sul do estado do Mississipi (Estados Unidos da América).

A escolha da Zona Pedagógica 7 como território para a recolha de dados deveu-se a vários fatores dos quais se destacam:

- familiaridade do investigador com a área de estudo;
- proximidade do investigador à área de estudo;
- conhecimento, por parte do investigador, de alguns agentes educativos, o que facilitou o acesso a mais professores que se dispuseram a participar na investigação;
- território inserido na Área Metropolitana de Lisboa, agregando um elevado número de escolas públicas e privadas/cooperativas nos vários níveis de ensino (1°/2°/3°ciclos e secundário) e, consequentemente, um elevado número de professores.

Foi utilizado um plano de amostra não-probabilístico, neste caso, inserido na amostragem por conveniência. Ou seja, a amostra é formada pelos participantes que demonstraram interesse em participar na recolha de dados (Creswell, 2011; Mertens, 2010). Embora se reconheçam as limitações inerentes a este tipo de amostragem, entendemos que foi a única forma viável de recolher os dados, uma vez que teremos sempre de estar sujeitos à disponibilidade e vontade em participar por parte dos professores. Desta forma, e conscientes da possibilidade de a amostragem por conveniência poder incluir vieses, em especial na análise inferêncial, recomendamos alguma prudência na interpretação e generalização dos resultados alcançados.

Dadas das recomendações da Direção-Geral de Saúde (DGS) de distanciamento

social, derivado da pandemia associada ao COVID-19, não foi possível a recolha de dados de forma presencial, conforme inicialmente previsto. Assim, optou-se por recolher os dados à distância através de uma versão online do QALA. Foi enviado um e-mail (conforme Anexo 1) às direções das várias instituições de ensino do setor público e privado da Zona Pedagógica 7 que tivessem pelo menos um dos ciclos de estudos considerados na presente investigação.

De salientar que, face aos objetivos da investigação, os docentes do grupos 100 (Pré-escolar), 910, 920 e 930 (Educação Especial) não foram contemplados, visto que as práticas e instrumentos de avaliação diferem daqueles que são utilizados nos demais Grupos de Recrutamento.

De forma a minimizar os efeitos da desejabilidade social[10], foi garantido o anonimato dos respondentes, não sendo solicitadas quaisquer informações que os pudessem identificar. Adicionalmente, os participantes foram informados que as suas respostas seriam agrupadas em áreas disciplinares (ver Anexo 2) e não em Grupos de Recrutamento, de forma a reforçar ainda mais o anonimato dos dados recolhidos.

A recolha de dados ocorreu nos meses de junho e julho de 2020. Responderam ao questionário um total de 253 professores do Ensino Básico e Ensino Secundário. A grande maioria dos respondentes é do sexo feminino (Tabela 3) e o escalão etário (Figura 7) mais representativo é o de entre os 41 e 50 anos (n=96; 37.945%).

[10]Entenda-se desejabilidade social como uma tendência presente nos sujeitos para atribuírem a si próprios atitudes ou comportamentos com valores socialmente desejáveis e para rejeitarem em si mesmos a presença de atitudes ou comportamentos com valores socialmente indesejáveis, quando respondem aos questionários de personalidade e às escalas de atitudes (Almiro, 2017).

Tabela 3: Total de professores participantes por sexo

Sexo	Frequência	Percentagem
Feminino	201	79.447
Masculino	52	20.553
Total	253	100.000

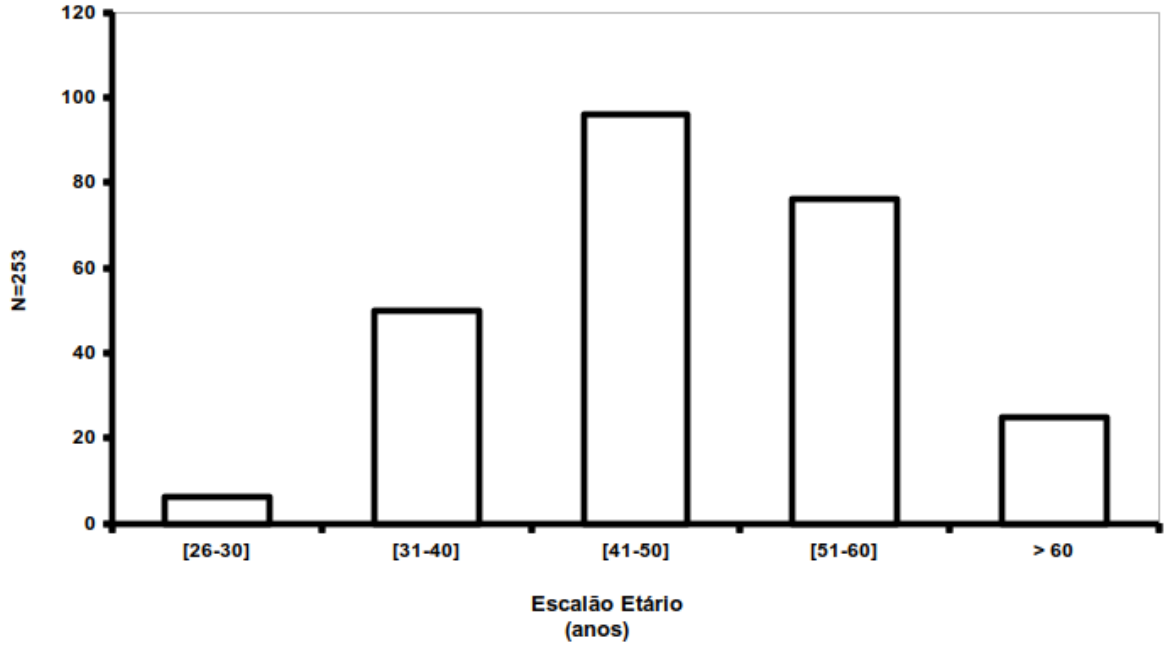


Figura 7: Total de professores participantes por escalão etário

Quanto à distribuição dos participantes por subsistema de ensino (Tabela 4), verifica-se uma predominância de professores do Ensino Público (n=191, 75.494%) comparativamente ao Ensino Particular e Cooperativo (n=55, 21.739%). Sete dos participantes lecionam em ambos os subsistemas de ensino. Mais de 75% dos participantes têm um vínculo laboral estável (Tabela 5), visto pertencerem aos quadros das escolas em que lecionam. Já os restantes (n=60) possuem um vínculo laboral precário já que são contratados (23.320%) ou prestadores de serviços (0.395%).

Tabela 4: Total de professores participantes por subsistema de ensino

Subsistema de ensino	Frequência	Percentagem
Público	191	75.494
Particular/Cooperativo	55	21.739
Ambos	7	2.767
Total	253	100.000

Tabela 5: Total de professores participantes por vínculo contratual

Vínculo contratual	Frequência	Percentagem
Contratado	59	23.320
Quadro	193	76.285
Prestação de serviços	1	0.395
Total	253	100.000

Relativamente à habilitação para a docência (Tabela 6), a esmagadora maioria dos participantes declarou possuir habilitação profissional (n=199; 78.656%) tendo os restantes (n=54; 21.344%) habilitação própria.

Tabela 6: Total de professores participantes por tipo de habilitação para a docência

Habilitação para a docência	Frequência	Percentagem
Própria	54	21.344
Profissional	199	78.656
Total	253	100.000

Conforme se poderá verificar na Tabela 7, a maioria dos professores (n=181, 71.542%) tem como habilitação literária a licenciatura (anterior ao Processo de Bolonha). Já os detentores de Mestrado Pós-Bolonha (que habilita atualmente para o desempenho de funções docentes), representam 9.091% do total de respondentes.

Tabela 7: Total de professores participantes por tipo de habilitações literárias

Habilitações literárias	Frequência	Percentagem
Bacharelato	3	1.186
Licenciatura (Pré-Bolonha)	181	71.542
Licenciatura (Pós-Bolonha)	7	2.767
Mestrado (Pré-Bolonha)	35	13.834
Mestrado (Pós-Bolonha)	23	9.091
Doutoramento	2	0.791
Outro	2	0.791
Total	253	100.000

Quanto à experiência letiva (Figura 8), a maioria possui entre 7 e 25 anos (n=134; 52.963%), seguindo-se os participantes com entre 26 e 35 anos de serviço (n=72; 28.458%) e com mais de 35 (n=27; 10.672%). Este aspeto, aliado à idade dos docentes, vem reforçar ainda mais que a classe docente se encontra envelhecida, sobretudo os professores do Ensino Público.

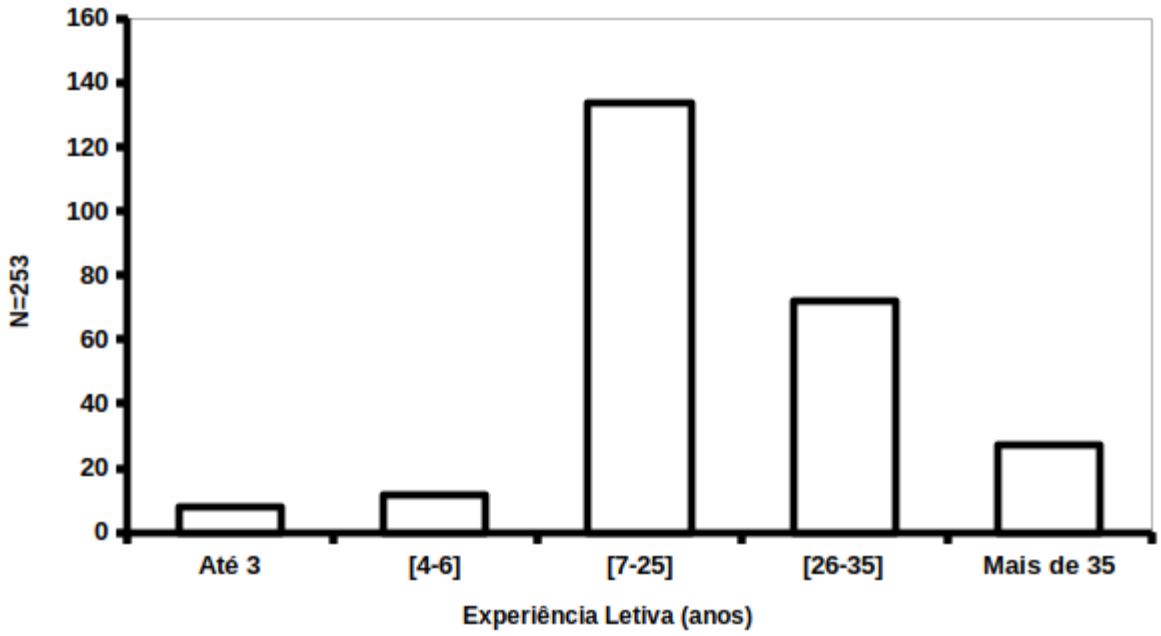


Figura 8: Experiência letiva dos participantes

Na distribuição dos participantes por nível de ensino (Tabela 8) verifica-se uma maior representatividade de professores dos 3°Ciclo e Secundário (n=150; 51.020%),

seguindo-se o 2°Ciclo (n=80; 27.211%) e, por fim, o 1°Ciclo (n=64; 21.769%). Note-se que o total, neste caso, é superior a 253 na medida em que há professores que lecionam em mais que um nível de ensino.

Tabela 8: Total de professores participantes por nível de ensino

	1°Ciclo	2°Ciclo	3°Ciclo e Secundário
Frequência	64	80	150
Percentagem	21.769	27.211	51.020

Já na distribuição dos participantes por áreas disciplinares (Tabela 9) verifica-se que as Línguas têm uma maior representatividade (n=79, 27.241%), seguindo-se o 1°Ciclo (n=64; 22.069%), a Matemática e as Ciências Experimentais (n=61; 21,035%), as Ciências Sociais e Humanas (n=44; 15.172%) e, por fim, Expressões (n=36; 14.5%). Tal como no caso anterior, também aqui o total é superior a 253 na medida em que há professores que lecionam disciplinas em mais que uma área disciplinar.

Tabela 9: Total de professores participantes por área disciplinar

	1°Ciclo	MCE(1)	CSH(2)	Línguas	Expressões
Frequência	64	61	44	79	42
Percentagem	22.069	21.035	15.172	27.241	14.483

(1) Matemática e Ciências Experimentais (2) Ciênciais Sociais e Humanas

De salientar, por último, que a maioria dos participantes (n=185; 73.123%) afirmou que frequentou algum tipo de formação em avaliação após a sua formação inicial (Tabela 10).

Tabela 10: Formação contínua em avaliação

	Frequência	Percentagem
Não	68	26.877
Sim	185	73.123
Total	216	100.000

3.4 O Questionário de Aferição da Literacia em Avaliação como instrumento de pesquisa

No capítulo 2.4. analisaram-se alguns dos instrumentos utilizados para a aferição da literacia em avaliação, em especial o *Teacher Asessment Literacy Questionnaire* (TALQ), o *Classroom Assessment Literacy Inventory* (CALI), o *Assessment Literacy Inventory* (ALI) e o *Measurement Literacy Questionnaire* (MLQ). Dessa análise, verificou-se que tanto o TALQ como o CALI, embora tenham sido amplamente utilizados, possuiam qualidades psicométricas muito baixas, pelo que a sua utilização neste estudo poderia pôr em causa algumas das conclusões da presente investigação. Para além disso, vimos que de acordo com as críticas, no caso do TALQ, era de difícil leitura, extenso e era constituído por vários itens que eram apresentados de forma descontextualizada

Quanto ao ALI, o instrumento foi traduzido para português (por uma professora de Português e Inglês do 3°Ciclo e Secundário) e foram realizadas pequenas alterações de forma a poder ser aplicável ao contexto português (alteraram-se disciplinas, nomes, escalas de classificação e uma questão que, pela sua natureza, estava claramente direcionada para os professores norteamericanos). Posteriormente, o instrumento foi aplicado a um conjuntos de 20 pessoas, com algum tipo de ligação ao investigador, dos quais 10 eram professores em exercício de diversas áreas

disciplinares e os restantes tinham pelo menos uma licenciatura mas sem qualquer formação de professores ou experiência letiva. Os resultados obtidos neste pequeno *ensaio* foram muito baixos e, inclusivé, as médias alcançadas pelos 10 participantes que não tinham habilitação para a docência foram substancialmente superiores aos participantes que tinham habilitação. Embora se reconheça que a amostra utilizada não permita fazer grandes inferências, receámos que a utilização do ALI pusessse em causa os resultados. Para além deste aspeto, os participantes neste *ensaio* foram unânimes em classificar o instrumento de confuso e demasiado extenso. Outro aspeto que nos fez não optar por nenhum dos instrumentos enunciados, foi o facto de estes estarem alinhados com os *Standards for Competence in Educational Assessment of Students*, ou seja, são instrumentos que foram construídos considerando o contexto norte-americano, pelo que algumas das questões não se adequavam ao contexto português. Para além destes aspetos Brookhart (2011), refere que os *Standards for Competence in Educational Assessment of Students* estão desatualizados e refletem as práticas avaliativas do início da década de 90 do século passado. Xu e Brown (2017) referem ainda que tanto o TALQ como os instrumentos que dele derivaram (ALI, CALI e novo ALI) não dão a necessária importância à avaliação formativa, o que por si só se assume como uma importante limitação.

Já o MLQ é um instrumento que se centra sobretudo na verificação de conhecimentos e competências ao nível dos testes e outros instrumentos de medida das aprendizagens, ou seja, em métodos e técnicas de avaliação sumativa, o que não é adequado ao paradigma atual que atribui especial importância à avaliação formativa.

Outro aspeto importante, que nos fez não optar por nenhum dos instrumentos referidos, foi que era objetivo da presente tese analisar as perceções dos professores

face à tarefa de avaliar, por um lado, e recolher informações sobre os níveis de literacia em avaliação, por outro. Com esta informação, pretendíamos verificar de que forma se relacionavam estas duas variáveis nos respetivos domínios. Ora, os instrumentos referidos não permitiam realizar este tipo de análise pelo que a construção de um novo questionário foi fundamental para a prossecução deste objetivo.

Assim, optámos pelo desenvolvimento de um questionário que pudesse ser aplicado ao contexto português e que tivesse como base os domínios que se aproximasssem da nossa definição de literacia em avaliação. Foi construído assim o *Questionário de Aferição da Literacia em Avaliação* (QALA), dirigido a professores do ensino básico e secundário e organizado em torno de 4 domínios inspirados na proposta de Abell e Siegel (2011). Assim, os domínios considerados são:

1. **Conhecimentos sobre objetivos e funções da avaliação:** Procura-se verificar os conhecimentos sobre os objetivos e funções da avaliação, em geral, e da avaliação de diagnóstico, formativa e sumativa em particular. Inclui-se ainda nesta dimensão o conhecimento sobre as diferenças entre avaliação criterial e normativa;

2. **Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar:** Nesta dimensão importa verificar o conhecimento dos professores sobre os diferentes tipos de currículo, as Aprendizagens Essenciais e o Perfil do Aluno à Saída do Ensino Obrigatório, a legislação em vigor no domínio da avaliação no Ensino Básico e Secundário, o conhecimento sobre domínios de complexidade cognitiva (*pe.* Taxonomia de Bloom, de Marzano, *Depth of Knowledge*) e o conhecimento de ferramentas de auxílio à construção de instrumentos de avaliação;

3. **Conhecimentos sobre a utilização de instrumentos de avaliação diversificados**: Em especial, instrumentos de avaliação de diagnóstico, formativa e sumativa. Importa também verificar os conhecimentos dos professores na construção de diferentes itens de avaliação e a inclusão dos alunos no processo de avaliação;

4. **Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação**: Procura-se, nesta dimensão, verificar os conhecimentos e capacidades dos professores em calcular medidas de localização e dispersão, bem como algumas propriedades psicométricas dos instrumentos de avaliação.

Consideramos igualmente relevante aferir os conhecimentos e competências na construção de instrumentos de registo de avaliação e de utilização do *feedback* em sala de aula.

O QALA (ver Anexo 3) encontra-se organizado em 4 partes distintas às quais daremos especial atenção nos subcapítulos seguintes.

3.4.1 Parte 1 - Dados Gerais

A Parte 1 do QALA é constituída por um conjunto de 9 itens que permitiram a caracterização da amostra. O conhecimento de algumas características da amostra era essencial, visto que alguns dos objetivos da investigação compreendem a verificação de relações entre algumas das variáveis recolhidas nesta parte do QALA com os resultados alcançados nas restantes partes[11]. Assim, foram solicitadas informações como:

1. Grupo(s) de Recrutamento: o respondente selecionou o(s) Grupo(s) de Recrutamento para o qual possui habilitação profissional e/ou própria. Aqui, não foram incluídos os grupos de recrutamento referentes à Educação Pré-Escolar (GR100) e Educação Especial (GR910, 920 e 930) visto estarem fora do âmbito da presente investigação;

2. Tipo de habilitação: O respondente assinalou se possui habilitação profissional ou habilitação própria;

3. Habilitações Literárias: O respondente selecionou uma de 7 opções referentes às suas habilitações literárias;

[11] Alguns estudos - por exemplo, Alkharusi (2011b) e Daniel e King (1998)- analisaram a relação de alguns indicadores como género, área disciplinar, experiência letiva e nível de ensino na literacia em avaliação, bem como nas perceções sobre competências em avaliação dos professores.

4. Subsistema de ensino em que leciona: O respondente selecionou entre as opções Público ou Particular e/ou Cooperativo;

5. Tipo de Vínculo: O respondente indicou a opção Contratado, Quadro ou Prestação de serviços;

6. Experiência Letiva: O respondente indicou o intervalo de tempo que corresponde à experiência letiva que possui. Os intervalos apresentados respeitam os ciclos de vida do professor definidos por Huberman (1995), nomeadamente:
   - Até aos 3 anos: Fase de Entrada;
   - Entre 4 e 6 anos: Fase da Estabilização;
   - Entre 7 e 25 anos: Fase da Diversificação;
   - Entre 26 e 35anos: Fase da Serenidade;
   - Mais de 35 anos: Fase de Desinvestimento.

7. Sexo;

8. Idade: O respondente indicou a opção correspondente ao intervalo da sua idade;

9. Realizou algum tipo de formação em avaliação após a sua formação inicial?: O respondente selecionou a opção sim ou não.

Para além da caracterização da amostra, a Parte 1 do QALA permite a recolha de indicadores que permitem verificar a existência de algum tipo de relação entre esses mesmos indicadores com os dados recolhidos das Partes 2 (Perceções sobre os conhecimentos e capacidades em avaliação), 3 (Conhecimentos em avaliação) e 4 (Cenários em contexto de avaliação).

3.4.2 Parte 2 - Perceções sobre conhecimentos e capacidades em avaliação

Com a Parte 2 do QALA pretendeu-se recolher informações sobre as perceções[12] que os professores tinham sobre os seus conhecimentos e capacidades em avaliação. Com os dados recolhidos, procurou-se:

- Analisar de que forma os professores do ensino básico e secundário percecionam os seus conhecimentos e capacidades em avaliação;
- Estabelecer relações entre as perceções dos professores e o seu desempenho nas partes 3 (Conhecimentos em avaliação) e 4 (Cenários em contexto de avaliação) do QALA;
- Estabelecer relações entre as perceções dos professores e algumas variáveis recolhidas na Parte 1 (Dados Gerais);

Assim, foram colocadas 20 questões do tipo *Likert*, variando de 1 (Discordo Totalmente) a 5 (Concordo Totalmente), e organizadas em torno dos 4 domínios considerados no presente estudo. Para cada item, cada professor escolheu a opção que melhor refletia a sua perceção face à afirmação colocada. A Tabela 11 apresenta os 20 itens colocados na parte 2 do QALA, com os respetivos domínios considerados. De salientar que este tipo de itens, embora de origem ordinal, assume propriedades muito próximas de uma escala de intervalos (Moreira, 2004), pelo que é comum tratar os dados obtidos como se tratando de uma escala métrica (Hill & Hill, 2002).

[12]Entenda-se aqui perceção como a atribuição de significado a estímulos sensoriais, a partir de estímulos passados. De acordo com Lent (2010) a perceção é a capacidade de interpretar as sensações, associando informações sensoriais a memórias, de modo a formar conceitos sobre o mundo, sobre nós mesmos e orientar o nosso comportamento.

Tabela 11: Relação Domínios/Itens da Parte 2 do QALA

Domínio	***Itens***
Conhecimento sobre objetivos e funções da avaliação	1. Considero que possuo sólidos conhecimentos sobre as diferentes funções e princípios da avaliação.
	2. Considero que possuo sólidos conhecimentos sobre as características e funções da avaliação de diagnóstico.
	3. Considero que possuo sólidos conhecimentos sobre as características e funções da avaliação formativa.
	4. Considero que possuo sólidos conhecimentos sobre as características e funções da avaliação sumativa.
	5. Considero que possuo sólidos conhecimentos sobre as diferenças entre avaliação referente a critério e avaliação referente a norma.
Conhecimento sobre currículo e sobre aquilo que é importante aprender e avaliar	6. Considero que possuo sólidos conhecimentos sobre os diferentes tipos de currículo (Oficial, Ensinado e Aprendido).
	7. Considero que possuo sólidos conhecimentos sobre as Aprendizagens Essenciais e o Perfil do Aluno à Saída da Escolaridade Obrigatória.
	8. Considero que possuo sólidos conhecimentos sobre a legislação em vigor relacionada com a avaliação das aprendizagens dos alunos.
	9. Considero que possuo sólidos conhecimentos sobre níveis de complexidade cognitiva (*p.e.* Taxonomia de Bloom, Taxonomia de Marzano, *Depth of Knowledge*) .
	10. Considero que possuo sólidos conhecimentos sobre a construção e utilização de ferramentas de auxílio à construção de instrumentos de avaliação
Conhecimento sobre a utilização de instrumentos de avaliação diversificados	11. Considero que possuo sólidas competências sobre como e quando fazer uso de instrumentos de avaliação de diagnóstico.
	12. Considero que possuo sólidas competências sobre como e quando fazer uso de instrumentos de avaliação formativa.
	13. Considero que possuo sólidas competências sobre como e quando fazer uso de instrumentos de avaliação sumativa.
	14. Considero que possuo sólidas competências sobre como construir diferentes tipos de itens de avaliação (p.e. escolha múltipla, verdadeiro e falso, questões de resposta aberta).
	15. Considero que possuo sólidas competências sobre como incluir os alunos no processo de avaliação.
Conhecimento sobre interpretação e utilização de informação recolhida no processo de avaliação	16. Considero que possuo sólidas competências sobre como calcular medidas de localização (p.e. média, moda e mediana) e dispersão (p.e. desvio-padrão) com a informação recolhida após a realização de um teste.
	17. Considero que possuo sólidas competências sobre como determinar algumas propriedades psicométricas dos instrumentos/itens de avaliação (p.e. índice de dificuldade, índice de discriminação).
	18. Considero que possuo sólidas competências sobre como construir instrumentos de registo da avaliação (p.e. Perfis de desempenho, listas de verificação).
	19. Considero que possuo sólidas competências sobre como e quando utilizar o *feedback*.
	20. Considero que possuo sólidas competências sobre como utilizar a informação recolhida de forma melhorar a minha prática pedagógica.

3.4.3 Parte 3 - Conhecimentos em Avaliação

A Parte 3 do QALA visa obter informações sobre o desempenho dos professores face aos seus conhecimentos em avaliação. Neste caso, foram colocados 40 itens de caráter dicotómico (Verdadeiro/Falso) organizados pelos quatro domínios considerados na presente investigação e considerando os itens colocados na Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação). Deste modo, foi possível comparar o desempenho dos professores nas várias questões colocadas na Parte 3, com informação recolhida na Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação). Procurou-se assim estabelecer a relação entre o desempenho e as perceções em avaliação. Desta forma, foi possível verificar, por exemplo, se os professores que têm uma melhor perceção sobre os seus conhecimentos e capacidades em avaliação alcançam melhores resultados na parte 'Conhecimentos em Avaliação' e vice-versa.

Nas tabelas 12a, 12b, 12c e 12d encontram-se elencados os vários itens colocados na Parte 3 com as respetivas respostas, assim como a relação de cada item com os itens colocados na Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação). Desta forma, procurámos estabelecer mais facilmente a relação entre os itens das duas partes do questionário, bem como com os respetivos domínios da literacia em avaliação considerados.

Tabela 12a: Relação entre itens da Parte 2 e 3 do QALA no domínio ’Conhecimentos sobre objetivos e funções da avaliação’

Item Parte 2	***Itens Parte 3***	***Solução***
1	De forma a garantir a princípio da justiça na avaliação, todos os alunos devem ser avaliados da mesma forma.	F
	No contexto português, a avaliação de natureza externa tem como principais finalidades a classificação e a certificação.	V
2	A avaliação de diagnóstico pode ocorrer em qualquer momento de um período escolar, até mesmo próximo do final do ano letivo.	V
	A avaliação de diagnóstico permite explorar ou identificar características dos alunos que sejam relevantes para a definição de estratégias de ensino e avaliação mais adequadas.	V
3	A avaliação formativa tem uma influência significativa nas aprendizagens dos alunos, contribuindo para a sua motivação e auto-estima.	V
	Uma das principais desvantagens da avaliação formativa é o facto de promover uma aprendizagem superficial dos conteúdos, ao invés de uma aprendizagem aprofundada dos mesmos.	F
4	A avaliação sumativa tem uma função predominantemente corretiva, visto permitir corrigir os erros cometidos ao longo do processo de ensino e aprendizagem.	F
	A avaliação sumativa incide sobre o produto da aprendizagem, sendo uma forma de recolha de informação que permite ao professor percecionar se os alunos atingiram os objetivos educacionais propostos.	V
5	Numa avaliação referente à norma, o objetivo principal é apreciar as aprendizagens efetivamente realizadas pelos alunos em relação às finalidades consideradas e aos objetivos definidos.	F
	A avaliação de referência a critério promove a competição entre alunos, enquanto que a avaliação referente à norma promove a competição do aluno consigo próprio.	F

Tabela 12b: Relação entre itens da Parte 2 e 3 do QALA no domínio ’Conhecimento sobre o currículo e sobre aquilo que é importante aprender e avaliar’

Item Parte 2	***Itens Parte 3***	***Solução***
6	A avaliação interna deverá incidir sobre o currículo oficial (ou formal), ao invés do currículo ensinado (ou real).	F
	A avaliação externa, em Portugal, tem por base o currículo oficial (ou formal).	V
7	A avaliação das aprendizagens dos alunos deverá ter como referência o programa curricular da respetiva disciplina, visto que as aprendizagens essenciais constituem-se apenas como um documento orientador.	F
	Uma das funções da avaliação deverá ser o de certificar as capacidades e atitudes no âmbito das competências inscritas no Perfil do Aluno à Saída do Ensino Obrigatório.	V
8	É exclusivamente a partir da avaliação formativa que se procede à verificação das condições de admissão aos exames nacionais.	F
	Segundo o quadro legal em vigor, a avaliação formativa constitui-se como a a principal modalidade de avaliação das aprendizagens.	V
9	Questões que procuram avaliar a capacidade de compreensão do aluno situam-se num nível de complexidade cognitivo elevado.	F
	Itens com elevado grau de dificuldade requerem uma maior capacidade em criar e avaliar situações com maior grau de complexidade.	V
10	As tabelas de especificações são ferramentas especialmente úteis para definir os objetivos que serão visados numa ficha de avaliação.	V
	A elaboração de matrizes de conteúdos, a serem alvo de avaliação através de exame nacional, são da exclusiva competência dos organismos centrais do Ministério da Educação.	V

Tabela 12c: Relação entre itens da Parte 2 e 3 do QALA no domínio ’Conhecimento sobre a utilização de instrumentos de avaliação diversificados’

Item Parte 2	***Itens Parte 3***	***Solução***
11	Os instrumentos de avaliação de diagnóstico devem possibilitar o despiste de variáveis exclusivamente cognitivas que possam interferir na aprendizagem dos alunos.	F
	A avaliação de diagnóstico poderá recorrer a testes de ensaio para verificar as condições dos alunos face às novas aprendizagens.	V
12	Com os portfólios os alunos têm mais possibilidades de mostrar o que sabem e são capazes de fazer, o que contribui para melhorar a sua auto-estima.	V
	O questionamento oral é um instrumento de avaliação predominantemente formativo, embora não permita uma regulação eficaz da aprendizagem.	F
13	Instrumentos de avaliação sumativa assentes em questões de verdeiro e falso, ou questões de escolha múltipla, apresentam uma maior fiabilidade, visto serem de correção objetiva.	V
	As provas de equivalência à frequência são instrumentos típicos de uma avaliação sobre o produto.	V
14	As questões do tipo completamento têm como desvantagem o facto de não permitirem a avaliação de um leque alargado de conteúdos.	F
	Itens de associação são indicados para testar a compreensão de conceitos, princípios e procedimentos.	V
15	O quadro legal em vigor privilegia o envolvimento dos alunos no processo de avaliação interna.	V
	A avaliação entre pares fomenta a autoconfiança dos alunos, mas põe em causa a qualidade das aprendizagens.	F

Tabela 12d: Relação entre itens da Parte 2 e 3 do QALA no domínio 'Conhecimento sobre interpretação e utilização da informação recolhida no processo de avaliação'

Item Parte 2	***Itens Parte 3***	***Solução***
16	Se os resultados de um instrumento de avaliação são heterogéneos, o instrumento tenderá a apresentar um desvio-padrão elevado.	V
	A classificação traduzida através de percentil, indica a percentagem de itens que o aluno respondeu corretamente.	F
17	Um instrumento de avaliação com índice de dificuldade de 0,5 permite uma maior diferenciação de resultados.	V
	A fiabilidade de um instrumento de avaliação garante a sua validade.	F
18	As listas de verificação são instrumentos especialmente úteis para registar a presença ou ausência de um comportamento ou procedimento desejado.	V
	As escalas de classificação são instrumentos especialmente indicados para registar a qualidade ou extensão de um comportamento.	V
19	O feedback pode ser considerado descritivo quando os alunos recebem uma classificação ou apreciação qualitativa com indicação das respostas corretas.	F
	As classificações, traduzidas por intermédio de uma nota, são a forma mais eficaz de comunicar os resultados da avaliação.	F
20	Após a comunicação dos resultados, da aplicação dos instrumentos de avaliação, o professor deverá prosseguir com a planificação da disciplina definida à priori.	F
	Um exemplo de boas práticas é a utilização da avaliação como forma de premiar os alunos que alcançam os melhores resultados escolares.	F

3.4.4 Parte 4 - Cenários em contexto de avaliação

A quarta, e última, parte do QALA contempla 20 questões de escolha múltipla organizadas em torno de cinco cenários hipotéticos em avaliação no contexto escolar. Em cada cenário são colocadas 4 questões, sendo que cada uma delas está alinhada com os vários domínios considerados, bem como com os itens constantes na Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA (ver Tabela 13). O objetivo da Parte 4 é, à semelhança da Parte 3 (Conhecimentos em avaliação), recolher informações sobre conhecimentos e capacidades em avaliação mas, neste caso, aplicado em situações hipotéticas em contexto escolar. Os dados recolhidos, permitiram aferir a literacia em avaliação dos professores e

comparar/relacionar com as perceções que os professores têm dos seus conhecimentos e capacidades em avaliação (Parte 2 do QALA).

Embora tratando-se de um instrumento desenvolvido especificamente no âmbito da presente investigação, a Parte 4 do QALA foi inspirada no *Assessment Literacy Inventory* (Mertler & Campbell, 2005), o qual também contempla questões de escolha múltipla organizadas em cenários hipotéticos em sala de aula.

Tabela 13: Correspondência entre itens da Parte 4 do QALA com os Domínios da Literacia em Avaliação e relação com os itens da Parte 2

Cenário	***Item Parte 4***	***Domínio em que se insere***	***Relação com item da Parte 2***
1	1.1.	Conhecimento sobre os objetivos e funções da avaliação	4
	1.2.	Conhecimento sobre o currículo e sobre aquilo que é importante aprender e avaliar	10
	1.3.	Conhecimento sobre a utilização de instrumentos de avaliação diversificados	13
	1.4.	Conhecimento sobre interpretação e utilização da informação recolhida	20
2	2.1.	Conhecimento sobre a utilização de instrumentos de avaliação diversificados	11
	2.2.	Conhecimento sobre os objetivos e funções da avaliação	2
	2.3.	Conhecimento sobre interpretação e utilização da informação recolhida	18
	2.4.	Conhecimento sobre o currículo e sobre aquilo que é importante aprender e avaliar	7
3	3.1.	Conhecimento sobre o currículo e sobre aquilo que é importante aprender e avaliar	6
	3.2.	Conhecimento sobre os objetivos e funções da avaliação	1
	3.3.	Conhecimento sobre a utilização de instrumentos de avaliação diversificados	15
	3.4.	Conhecimento sobre interpretação e utilização da informação recolhida	17
4	4.1.	Conhecimento sobre a utilização de instrumentos de avaliação diversificados	12
	4.2.	Conhecimento sobre os objetivos e funções da avaliação	3
	4.3.	Conhecimento sobre interpretação e utilização da informação recolhida	19
	4.4.	Conhecimento sobre o currículo e sobre aquilo que é importante aprender e avaliar	9
5	5.1.	Conhecimento sobre os objetivos e funções da avaliação	5
	5.2.	Conhecimento sobre o currículo e sobre aquilo que é importante aprender e avaliar	8
	5.3.	Conhecimento sobre a utilização de instrumentos de avaliação diversificados	14
	5.4	Conhecimento sobre interpretação e utilização da informação recolhida	16

3.5 Tratamento e Análise dos dados

3.5.1 Propriedades Psicométricas do QALA com recurso ao modelo Rasch

A análise das propriedades psicométricas e da validade de construto do QALA foi realizada com recurso ao Modelo Rasch, conjunto de técnicas estatísticas que se enquadram na Teoria de Resposta ao Item (TRI). O Modelo Rasch foi proposto pelo dinamarquês George Rasch, em 1960, e procurou resolver algumas das limitações reconhecidas à Teoria Clássica dos Testes (TCT) e, gradualmente, ganhou um importante campo de aplicação em psicologia e em educação (Maia, 2012).

O modelo Rasch é um modelo logístico de um parâmetro (a dificuldade). Neste modelo, considera-se que as respostas de um sujeito dependem da sua habilidade e da dificuldade dos itens que constituem o instrumento (Couto & Primi, 2011; Linacre & Wright, 2002). Para compreender o funcionamento do Modelo Rasch, vejamos a seguinte explicação:

> [...]consider a continuum of values that represent the construct of interest. Low levels of this continuum represent characteristics of the latent trait that frequently occur or are easily observed. Conversely, high levels of this continuum indicate characteristics that are more rarely observed or that are more difficult to achieve. A person's standing on the latent trait is denoted by θ [tetha], and it represents how far along the continuum that we expect the person to answer items correctly. The examinee will correctly answer all items below this point but incorrectly answer all items above this point. Also on this continuum are values of item difficulty, which represent how far along the continuum that we expect to obtain correct responses to the item. Item difficulty is denoted with the letter b. With both the person and item represented on this continuum, we would like to see a person answer an item correctly every time θ ≥ b and incorrectly otherwise. However, this relationship is deterministic and does not account for other possibilities during a test. It is possible for examinees with low values of θ to answer a difficult item correctly, and it is also possible for examinees with high values

> of θ to make a mistake and answer an easy item incorrectly (Meyer, 2014, p.83).

Dito de outra forma, a partir do modelo Rasch é possível estimar a habilidade dos respondentes em responder a um determinado item presente num determinado teste ou questionário. Este aspeto, assume-se de grande importância já que, desta forma, é possível analisar, numa mesma dimensão, a habilidade dos respondentes com as dificuldades dos itens e estabelecer, entre ambas, relações que nos permitem aferir as qualidades psicométricas dos testes ou questionários.

Nas últimas décadas, o Modelo Rasch tem sido amplamente utilizado para aferir as qualidades psicométricas e a validade de instrumentos, sejam eles constituídos por itens dicotómicos ou politómicos. A partir do Modelo Rasch é possível determinar vários parâmetros estatísticos que traduzem a qualidade psicométrica dos dados. No entanto, para a utilização do Modelo Rasch, deverão ser cumpridos os seguintes pressupostos:

- Unidimensionalidade: Um dos requisitos para a aplicação do Modelo Rasch é a unidimensionalidade dos dados (Linacre, 2002; Meyer, 2014). Assim, a verificação da unidimensionalidade é fundamental para aferir em que medida os itens que compõem o teste/questionário se relacionam com a variável latente em análise (Bond & Fox, 2015). Gomes e Borges (2009), a partir dos estudos de Tennant e Pallan (2006), identificaram três abordagens para a verificação da unidimensionalidade dos itens em contexto de modelo Rasch. A primeira recorre a uma análise prévia recorrendo a métodos da Teoria Clássica de Testes, como por exemplo a Análise Fatorial Exploratória ou a Análise de Componentes Principais, a segunda deriva da interpretação de que um modelo bem ajustado garante a unidimensionalidade de um instrumento e a terceira recorre a testes *post-hoc*, como por exemplo, a Análise de Componentes Principais dos

Resíduos[13] (ACPr). Na presente investigação, recorremos, à ACPr para uma primeira verificação da unidimensionalidade e, posteriormente, confirmámos através da análise ao ajuste dos dados ao modelo Rasch.

- Independência local: A independência local dos itens, tal como a unidimensionalidade, é um requisito necessário para a realização do modelo Rasch. Considera-se que existe independência local quando o desempenho de um respondente a um determinado item não afeta o desempenho a outros itens (Vieira, Ribeiro & Almeida, 2009), já que o seu desempenho está apenas dependente da sua habilidade.

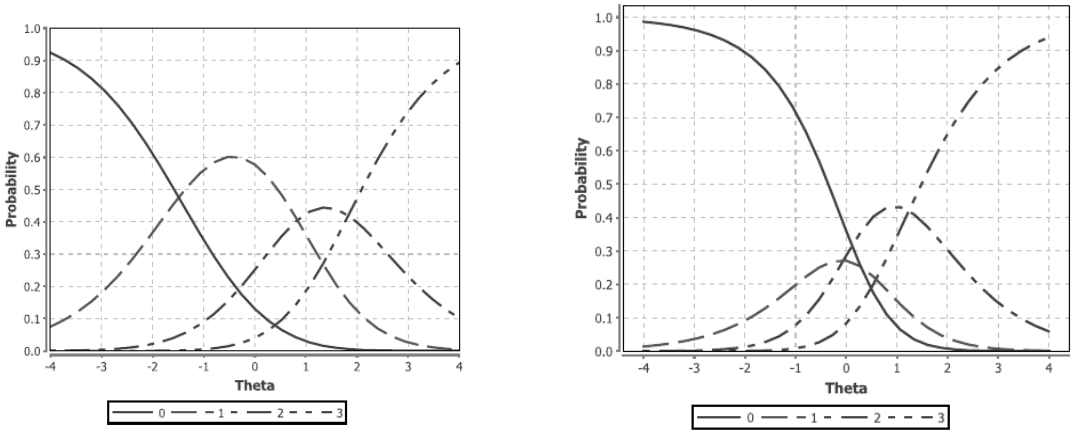
  Para aferir a independência local, analisou-se a tabela de correlações dos resíduos resultantes do Modelo Rasch. O valor crítico para identificar a independência local entre itens varia de autor para a autor. Autores como Chen e Thissen (1997) propõem o valor de 0,2 como sendo o valor de correlação entre resíduos a partir do qual os itens podem indicar dependência local. Já Nair, Moretin e Lincoln (2011), propõem o valor de 0,3 e Klooster, Taal e Laar (2008) utilizam como referência o valor de 0,5. No entanto, Linacre (2020), Lah e Tasir (2018) e González-de-Paz *et al.* (2015) consideram que existe dependência local entre itens quando os valores de correlação entre resíduos é superior a 0,7, uma vez que é a partir desse valor que o par de itens compartilha mais de metade da variância residual. Assim, no caso da presente investigação utilizou-se o valor de 0,7 como proposto por Linacre (2020). Espera-se que os valores de correlação de resíduos dos itens sejam inferiores a 0,7, caso contrário, considera-se que existe dependência local e apenas se utilizará um dos itens.

[13]São considerados resíduos as discrepâncias verificadas entre os valores preditos pelo modelo Rasch e os resultados efetivamente alcançados a partir das estatísticas de ajuste (Silva & Vendramini, 2006).

Cumpridos os pressupostos da Unidimensionalidade e da Independência local, o Modelo Rasch permite-nos aferir as qualidades psicométricas a partir dos seguintes parâmetros:

- Limiares de Categoria dos itens politómicos (apenas para a Parte 2 do QALA): Os limiares de categoria correspondem ao valor de habilidade em que uma pessoa tem igual probabilidade de selecionar duas opções de resposta adjacentes (Meyer, 2014; Robison *et al.*, 2019). A análise dos limiares de categoria implica a análise da Curva Característica do Item (CCI) dos vários itens de forma a verificar se as probabilidades de respostas estão organizadas em ordem ascendente e concordante com as categorias definidas (Robison *et al.*, 2019), indicando que os limites estão bem ordenados (ver Figura 9a). A desordem dos limiares de categorias (ver Figura 9b) pode ser indicador de categorias mal elaboradas ou, mais frequentemente, de excesso de opções de resposta. Quando a razão da desordem se deve ao excesso de opções, a forma mais comum de solucionar o problema passa pela junção de duas ou mais categorias de resposta (*p.e.* Discordo Totalmente e Discordo formam uma nova categoria designada de Discordo).

Figura 9: Exemplos de Curvas Características dos Itens (CCI) de itens politómicos (Fonte: Meyer, 2014)



(a) Limiares das categorias ordenadas (b) Limiares das categorias desordenadas

- Dificuldade dos itens: O modelo pressupõe que a probabilidade de uma determinada interação pessoa / item (em termos de classificação alta ou baixa) é determinada apenas pela dificuldade do item e pela habilidade da pessoa (Granger, 2007). O parâmetro dificuldade, usualmente representado pela letra β, é definido como o resultado no construto que se encontra associado a 50% de probabilidade de um item ser escolhido ou de receber uma resposta correta (Miguel, 2013). No caso dos itens politómicos a dificuldade é determinada pela proporção de respostas a uma categoria de escolha (Sartes & Souza-Formigoni, 2013). Assim, é necessário verificar em que medida os itens cobrem uma variada gama de dificuldades. Questionários/testes com um nível de dificuldade muito alto (*ceiling effect*) ou muito baixo (*floor effect*) colocariam em causa a utilidade dos mesmos como instrumentos de medida, pelo que é desejável contenham uma variedade de itens com níveis de dificuldade dispersos entre os fáceis (valores negativos na escala de *logits*[14] e os difíceis (valores positivos na escala de *logits*).

[14]Logits deriva de *log odds units*.

- Ajuste dos itens: As potencialidades da utilização do Modelo de Rasch só podem ser alcançadas caso os dados empíricos se ajustem ao modelo teórico (Prieto & Delgado, 2003). O ajuste do dados ao modelo é avaliado pela comparação entre a probabilidade teórica de acerto de cada pessoa a cada item com os valores observados. Assim, a presença de valores absurdos poderia colocar em causa os resultados alcançados, visto que careceriam de significado teórico (Prieto & Delgado, 2003). Segundo os mesmos autores, um modelo desajustado pode dever-se a múltiplos fatores, nomeadamente à multidimensionalidade dos dados, respostas dadas ao acaso, pouca cooperação ou motivação dos respondentes e instruções ou respostas pouco claras.

  Para avaliar o ajustamento dos dados ao modelo analisaram-se dois indicadores estatísticos de ajustamento, o *Weighted Mean Square* (WMS)[15] e o *Unweighted Mean Square* (UMS)[16]. Ambos os indicadores fornecem informações sobre as discrepâncias nas respostas, consoante o seu afastamento aos parâmetros estimados (Cadime *et. al*, 2017). Os valores de WMS são estimados dando um maior peso às pontuações de desempenho dos sujeitos mais próximos aos valores estimados (Brown & Bonsaksen, 2019). Já os valores de UMS são calculados sem qualquer ponderação, pelo que é um parâmetro mais sensível à influência das pontuações mais distantes (Brown & Bonsaksen, 2019). Esta particulariedade do cálculo do UMS leva a que muitos autores o descartem aquando da averiguação do ajuste dos itens ao modelo.

  Os valores de WMS e UMS que utilizaremos como referência foram propostos por Linacre (2002). Segundo o autor, itens que apresentem valores de ajuste próximos de 1,0 são explicados totalmente pelo modelo, ou seja, têm um ajuste

[15] Também designado por *Infit Mean Square*.
[16] Também designado por *Outfit Mean Square*.

perfeito. No entanto, o mesmo autor considera que os itens estão bem ajustados quando são produtivos para a medida, isto é, quando apresentam valores entre 0,5 e 1,5. Itens com valores entre 1,5 e 2,0 apresentam um desajuste moderado, mas não degradam as medidas (Linacre, 2002) pelo que podem ser mantidos. Já os valores acima de 2,0 apresentam um desajuste severo e degradam as medidas, pelo que devem ser revistos ou até mesmo descartados. Quanto aos valores inferiores a 0,5 são considerados improdutivos mas não degradam as medidas, pelo que podem ser mantidos.

- Mapa item-pessoa (Mapa de *Wright*): Os Mapas item-pessoa são representações gráficas da distribuição das habilidades dos respondentes e da dificuldade dos itens, ao longo do traço latente (Brown & Bonsaksen, 2019). Através da análise do mapa item-pessoa procura-se, por um lado, verificar se há uma boa dispersão tanto das habilidades dos respondentes como das dificuldades dos itens e, por outro, se as dificuldades dos itens têm a capacidade de cobrir os diferentes níveis de habilidade dos sujeitos. O cumprimento destes aspetos, segundo Franco *et al.* (2020), é indicador que o questionário tem um bom desempenho para medir o traço latente em análise, ou seja, é um bom indicador de validade de construto.

- Funcionamento Diferencial dos Itens (DIF): O DIF, segundo Bond e Fox (2015), procura identificar se os construtos estabelecem uma dificuldade consistente dos itens, independentemente do grupo de pessoas a que é aplicado. Desta forma, procura-se verificar se os itens funcionam de forma semelhante em respondentes de diferentes géneros, raças, grupos étnicos, religiões, ou outros (Brown & Bonsaksen, 2019). Conforme afirmam Fidalgo e Scalon (2012), um item funciona diferencialmente quando a probabilidade de sucesso no item é diferente entre os respondentes com o mesmo nível de habilidade, mas que pertencem a diferentes subgrupos da população determinada.

Para a verificação da existência de DIF foi utilizado o método de Cochran-Mantel-Haenszel. Os vários itens que compõem o QALA foram classificados consoante o seu grau de DIF. Itens classificados com A (dicotómicos) ou AA (politómicos) apresentam um baixo grau de DIF. Itens classificados com B ou BB sugerem um grau moderado de DIF. Já os itens classificados com C ou CC apresentam um elevado grau de DIF e deverão ser revistos ou retirados já que podem ser *uma clara ameaça à validade dos itens e do teste* (Fidalgo & Scalon, 2012, p.61).

- Índices de Fiabilidade e Separação dos Itens: Os índices de fiabilidade e separação dos itens referem-se, segundo Bond e Fox (2015), à capacidade do instrumento em definir uma hierarquia dos itens ao longo da variável (ou construto) medida. Dito de outra forma, estes índices revelam o quão bem os participantes separam os itens em diferentes níveis de dificuldade (Mofreh *et al.*, 2014). Assim, os índices de fiabilidade e separação dos itens assumem uma especial relevância já que são indicadores da validade de construto (Linacre, 2020).

  Para um instrumento ser útil, o valor do índice de separação dos itens deve ser superior a 1,0 (Green & Franton, 2002) e os valores de fiabilidade dos itens deverão ser superiores a 0,50 (Mohamad *et al.*, 2014). No entanto, quanto maiores forem os valores de ambos os indicadores, maior será também a confiança na replicabilidade do instrumento em outras amostras (Bond & Fox, 2015).

  A tabela 14 apresenta os critérios de qualidade dos dois indicadores segundo a proposta de Fisher (2007).

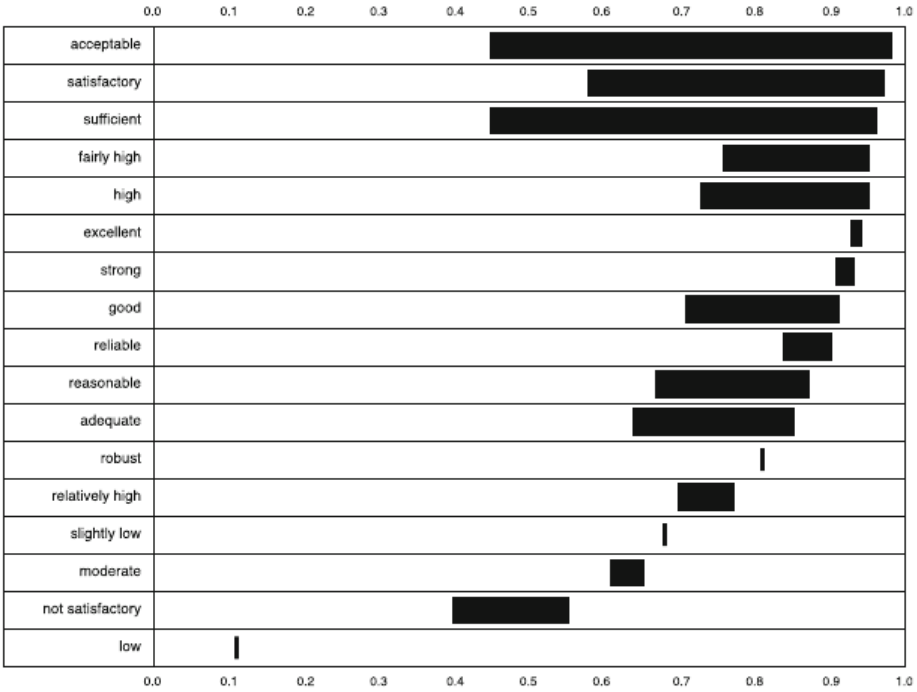
Tabela 14: Critérios de Qualidade dos Índices de Fiabilidade e Separação dos Itens (Adaptado de Fisher, 2007)

Critério	Baixo	Regular	Bom	Muito bom	Excelente
Fiabilidade dos Itens	<.67	.67 - .80	.81 - .90	.91 - .94	> .94
Separação dos Itens	<1,5	1,5 - 2	2 - 3	3 - 4	> 4

- Consistência interna: Para a análise da consistência interna das 3 partes do QALA, foi determinado o Coeficiente α (também designado de α de Cronbach). Os critérios para o nível de aceitação dos valores de α divergem de autor para autor. Taber (2017), elaborou um esquema (Figura 10) onde costam diferentes classificações de acordo com o valor do coeficiente α segundo a literatura.

Valores mais elevados do coeficiente α sugerem uma melhor consistência interna, logo uma maior confiabilidade, sendo que o valor mínimo aceitável e suficiente é de 0,45.

Figura 10: Critérios de classificação do Alfa de Cronbach (Fonte: Taber, 2017)



3.5.2 Análise estatística

Para além da análise com recurso ao modelo Rasch, para a aferição das qualidades psicométricas do QALA, o tratamento dos dados foi realizado com recurso a técnicas de análise descritiva e análise inferencial.

Na análise descritiva, recorremos a medidas de tendência central, como sejam a média, a moda e a mediana, bem como medidas de dispersão, em especial o desvio-padrão, o mínimo e o máximo. A análise descritiva foi realizada tendo em consideração, por um lado, os dados globais recolhidos pelo QALA e, por outro, algumas das variáveis de contexto recolhidas na Parte 1 (Dados Gerais).

Já a análise inferencial dos dados foi realizada a partir de um conjunto de estatísticas não-paramétricas. A utilização de técnicas não-paramétricas para a análise inferencial da dados deveu-se, como veremos adiante, ao facto de não ter sido garantido o pressuposto da distribuição normal, verificado a partir do teste de *Kolmogorov-Smirnov*. Assim, para a análise inferêncial recorremos ao teste *U* de Mann-Whitney para duas amostras independentes (alternativa ao teste *t*), ao teste *H* de Kruskal-Wallis para *k* amostras (alternativa ao *One-way* Anova) e ao coeficiente de correlação de Spearman para medir a intensidade das relações entre variáveis.

De salientar ainda que, para o tratamento dos dados, os *softwares* utilizados para estimar os diversos parâmetros foram o jMetrik (versão para *Linux*), desenvolvido por J. Patrick Meyer, o JASP (versão para *Linux*), desenvolvido pela Universidade de Amesterdão, e o GNU PSPP *Statistical Analysis Software* (versão para *Linux*). A utilização deste software, para além de ser *Open Source* e compatível com ambientes *Linux*, possibilita a exportação de gráficos e tabelas para LaTeX, ambiente em que a presente tese foi redigida.

Capítulo 4

Apresentação dos resultados

4.1 Qualidades Psicométricas do QALA

4.1.1 Unidimensionalidade

Conforme mencionado anteriormente, a análise da unidimensionalidade do QALA foi realizada a partir da Análise das Componentes Principais dos Resíduos (ACPr). A literatura existente indica que os dados revelam unidimensionalidade quando, a partir dos resultados da ACPr, a percentagem de variância explicada pelo primeiro fator é superior a 20% (Bond & Fox, 2015; Linacre, 2002) e o valor próprio (*eigenvalue*) do segundo fator apresenta valores absolutos inferiores a 3 (Brown & Bonsaksen, 2019; Bond & Fox, 2015).

A Tabela 15 revela que os dados das três partes do QALA apresentam valores concordantes com a unidimensionalidade (para mais detalhes ver Anexo 4), já que respeitam os critérios de percentagem de variância explicada do primeiro fator e o valor de *eigenvalue* do segundo fator.

Tabela 15: Resumo da ACPr realizada ao QALA

	Parte 2		Parte 3		Parte 4	
	Fator 1	Fator 2	Fator 1	Fator 2	Fator 1	Fator 2
Variância Explicada (%)	**31**	24	**32**	22	**27**	20
eigenvalue	3,18	**2,44**	3,54	**2,38**	2,03	**1,49**

Como mero exercício de confirmação do requisito da unidimensionalidade dos dados, recorremos também à Análise Fatorial Exploratória (AFE). No entanto, fizemo-lo apenas para a Parte 2 do QALA (Perceções sobre os conhecimentos e capacidades em avaliação). Esta opção deve-se ao facto de que a AFE, quando aplicada a dados dicotómicos, como são os casos da Parte 3 (Conhecimentos em Avaliação) e 4 (Cenários em contexto de avaliação), tende a produzir quase sempre muitos fatores, muitos deles considerados artificiais (Hattie, 1985; McDonald & Ahlawat, 1974), o que dificultaria a verificação da unidimensionalidade, pois tanto os resultados como a análise dos mesmos poderiam ser enviesados (Laros, 2012).

Assim, para a execução da AFE da Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação), foi necessário verificar, em primeiro lugar, a existência de correlações entre os itens. Para tal, recorremos ao teste de Keiser-Meyer-Olkin (KMO) e ao teste de esfericidade de Bartlett. Estes dois testes permitem aferir a qualidade das correlações entre variáveis de forma a prosseguir com a AFE (Pestana & Gageiro, 2003).

Tabela 16: Teste de KMO

	MSA
Overall MSA	0.902
P2.1	0.953
P2.2	0.927
P2.3	0.860
P2.4	0.812
P2.5	0.946
P2.6	0.944
P2.7	0.873
P2.8	0.909
P2.9	0.867
P2.10	0.927
P2.11	0.915
P2.12	0.898
P2.13	0.829
P2.14	0.893
P2.15	0.952
P2.16	0.896
P2.17	0.886
P2.18	0.933
P2.19	0.939
P2.20	0.907

Tabela 17: Teste de Esfericidade de Bartlett

χ^2	df	p
3289.682	190.000	$< .001$

Os resultados obtidos pela aplicação do teste de KMO (Tabela 16) revelam uma muito boa correlação entre os itens que constituem a Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA (overall MSA[17] $>$0,9), estando muito acima do patamar crítico de 0,6 (Pestana & Gageiro, 2003). Para além disso, o teste de Bartlett (Tabela 17) é estatisticamente significante (p$<$0,001), pelo que os dados podem ser considerados adequados à realização da AFE.

Dada a adequabilidade dos dados para a AFE, foi realizada a extração de fatores utilizando o método de *Principal Axis Fatoring* e como critério para a retenção de

[17] *Measure of Sampling Adequacy.*

fatores recorremos à Análise Paralela[18]. O recurso à análise paralela justifica-se pelo facto de diminuir a propabilidade de retenção equivocada de itens, por considerar o erro amostral e minimizar a influência do tamanho da amostra e das cargas fatoriais dos itens (Damásio, 2012; Nogueira, Seidl & Troccoli, 2016). Os resultados obtidos encontram-se resumidos nas Tabela 18 e 19.

Tabela 18: Fatores resultantes da aplicação da Análise Fatorial Exploratória

Item	Fator 1	Fator 2
P2.1	0.729	
P2.2	0.706	
P2.3	0.756	
P2.4	0.696	
P2.5	0.647	
P2.6	0.625	
P2.7	0.607	
P2.8	0.644	
P2.9	0.471	
P2.10	0.594	
P2.11	0.783	
P2.12	0.800	
P2.13	0.723	
P2.14	0.586	
P2.15	0.704	
P2.16	0.526	
P2.17	0.632	0.548
P2.18	0.665	0.448
P2.19	0.720	
P2.20	0.699	

Nota: São apenas apresentados os valores superiores a 0.4

[18]Segundo Freire e Motokane (2016), a análise paralela compara os *eigenvalues* obtidos empiricamente, com os valores médios de *eigenvalues* gerados a partir de matrizes hipotéticas (amostras de mesmo tamanho contendo dados aleatórios não correlacionados). Seguindo esse critério, devem ser extraídos apenas os fatores com *eigenvalues* empíricos superiores aos *eigenvalues* obtidos aleatoriamente.

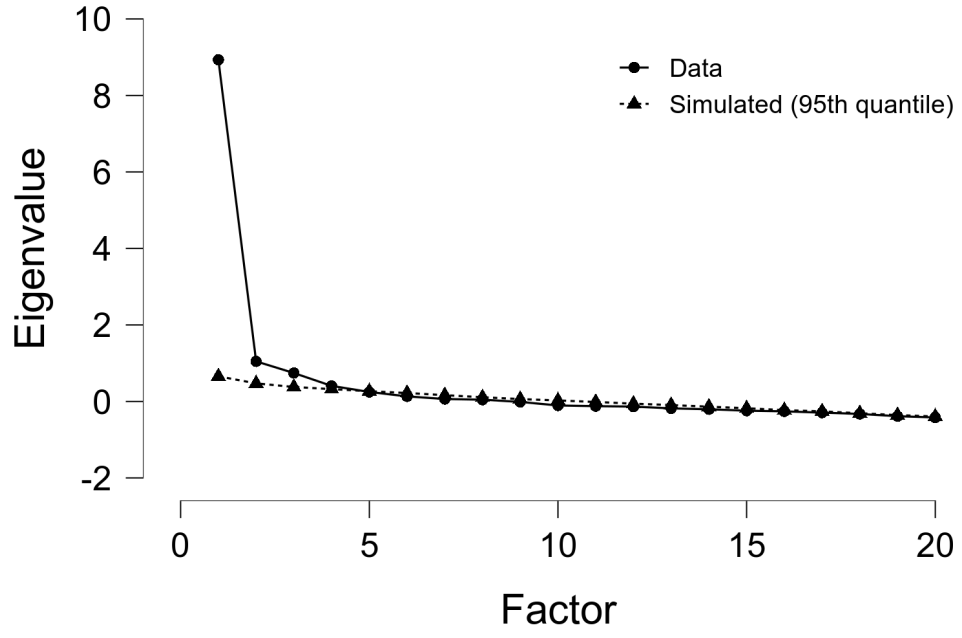
Tabela 19: Características dos fatores extraídos pela aplicação da Análise Fatorial Exploratória

	SumSq. Loadings	Proportion var.	Cumulative
Factor 1	8.991	0.450	0.450
Factor 2	1.200	0.060	0.510

Os resultados obtidos pela AFE parecem corroborar a tese da unidimensionalidade dos dados. Esta conclusão deriva do facto do primeiro fator extraído apresentar uma proporção de 45% da variância, ao passo que o segundo fator apresenta apenas um valor de 6%. Para além disso, o valor próprio (*eigenvalue*) estimado para o primeiro fator (8,991) é muito superior ao valor do segundo fator (1,200), o que vem confirmar a ideia de unidimensionalidade.

A própria análise ao *scree plot* (Figura 11), ou teste de Cattel, é mais uma evidência da unidimensionalidade já que é visível o destaque que o primeiro fator tem em relação aos demais.

Figura 11: *Scree plot* obtido após a AFE da Parte 2 do QALA



4.1.2 Independência Local

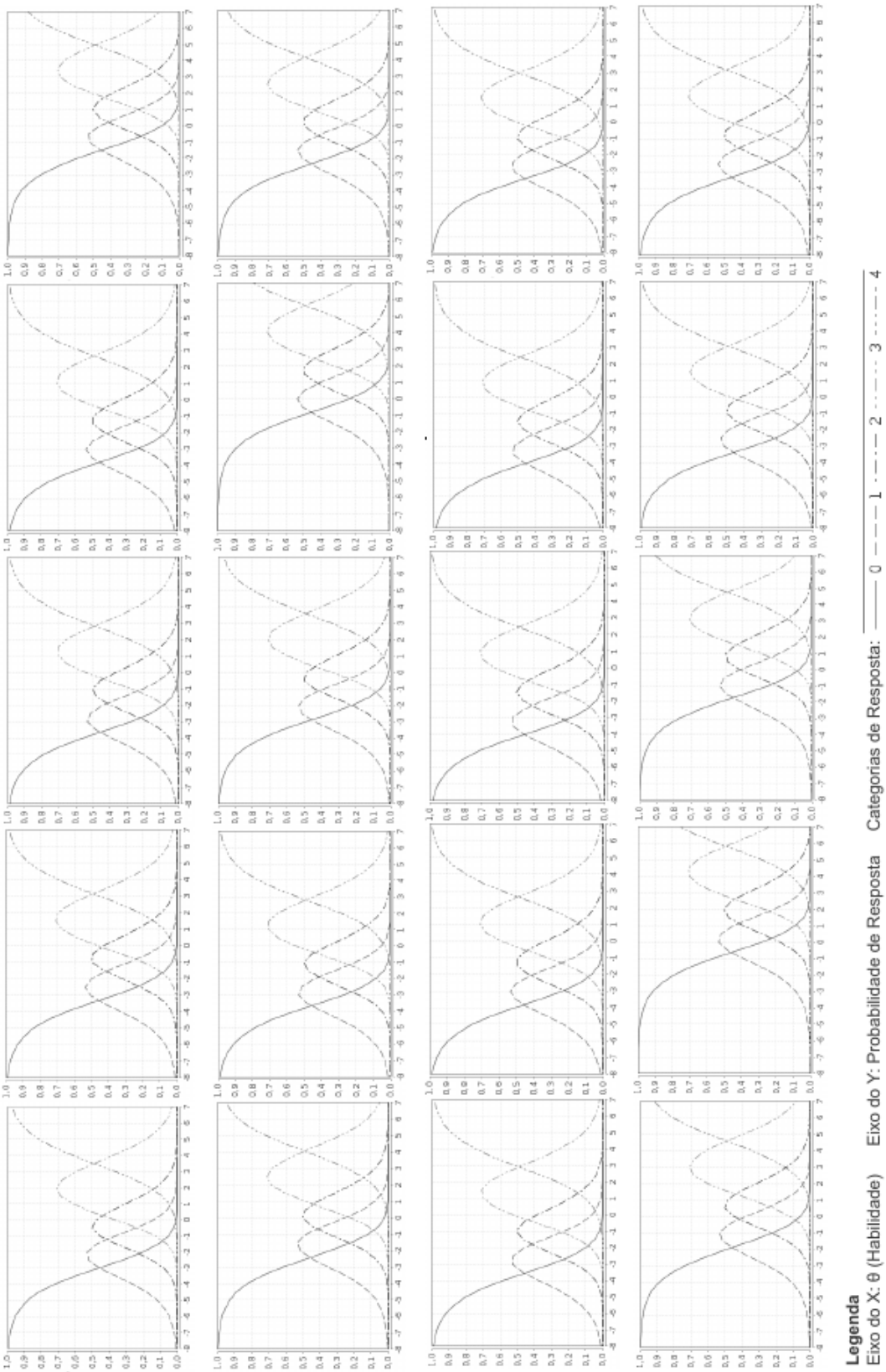
Os itens das Partes 2, 3 e 4 cumprem o requisito da independência local, já que as matrizes de correlações dos resíduos das 3 partes do QALA apresentam valores inferiores ao limiar de 0,7 (ver Anexo 5). Podemos assim concluir que as respostas dadas a cada item não estão dependentes das respostas dadas a outros itens.

4.1.3 Limiares de Categoria - Parte 2 do QALA

A análise aos limiares de categoria foi realizada apenas à parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA visto ser a única parte que é composta por itens politómicos (escala do tipo *Likert*). Pela análise às Curvas Características dos Itens que compõem a Parte 2 do QALA[19] (Figura 12), verifica-se que as probabilidades de resposta dos vários itens estão organizadas em ordem ascendente e concordante com as categorias definidas. Tal facto indica que as 5 categorias de resposta definidas (Discordo Totalmente(0), Discordo(1), Não Concordo Nem Discordo(2), Concordo(3) e Concordo Totalmente(4)) estão bem ajustadas, não havendo qualquer tipo de desordem nem necessidade de reclassificação.

[19]Mais detalhes no Anexo 6

Figura 12: Curvas Características dos Itens da Parte 2 do QALA



4.1.4 Dificuldade dos Itens

Os valores de dificuldade dos itens (β) do QALA (Tabela 20) variam entre -1,19 e 2,24 (na escala de *logit*) na Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação), entre -2,06 e 2,64 na Parte 3 (Conhecimentos em avaliação) e -3,04 e 2,95 na Parte 4 (Cenários em contexto de avaliação). Para além disso, verifica-se que a dificuldade média em cada uma das partes foi muito próxima de 0. Estes aspetos são reveladores de uma boa distribuição dos valores de β na escala *logit* e que o teste não é considerado nem fácil nem difícil, ou seja, não se verifica nem um *floor effect*, nem um *ceiling effect*.

Na Parte 2 os itens com os valores de β mais baixos foram os P2.14, P2.13 e P2.12 e os itens com os valores de β mais altos foram P2.17, P2.9 e P2.5. Na Parte 3, os itens mais fáceis foram os P3.3, P3.22 e P3.1 e os mais difíceis foram P3.37, P3.17 e P3.5. Por último, na Parte 4 os respondentes revelaram uma maior facilidade na resposta aos itens P4.3.1, P4.1.4 e P4.2.1 e uma maior dificuldade aos itens P4.2.3, P4.1.2 e P4.5.1.

Tabela 20: Dificuldades (em logit) dos itens do QALA

Parâmetro	Parte 2	Parte 3	Parte 4
Média	$\approx$ 0,0	$\approx$ 0,0	$\approx$ 0,0
DP	0,11	0,15	0,18
Máxima	2,24	2,64	2,95
Mínima	-1,19	-2,06	-3,04
Itens com β < -2	0	2	2
-2 < Itens com β < -1	2	5	4
-1 < Itens com β < 0	11	13	14
0 < Itens com β < 1	3	13	4
1 < Itens com β < 2	2	4	4
Itens com β > 2	2	3	2

4.1.5 Ajuste dos itens

Conforme referido anteriormente, o ajuste dos itens ao modelo é-nos dado a partir da análise de dois parâmetros, o WMS e o UMS. Para um modelo bem ajustado, os valores de WMS e UMS deverão estar enquadrados no intervalo entre 0,5 e 1,5 que, segundo Linacre (2002), são valores produtivos para a medida. Já itens com valores de WMS e UMS acima de 2,0 deverão ser retirados sob pena de haver distorção e degradação das medidas (Cadime *et. al*, 2017; Linacre, 2002).

Uma síntese dos resultados de ajuste dos itens do QALA pode ser visualizada na Tabela 21. De um modo geral, verifica-se que a maioria dos itens que compõem o QALA foi respondida de acordo com o modelo esperado, com itens bem ajustados e com média próxima de 1,0. Ao nível do parâmetro WMS (*Infit*), a Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) apresenta 19 (dos 20 itens), bem ajustados de acordo com o critério de Linacre (2002), tendo apenas 1 item (P2.16) com um desajuste moderado (WMS=1,70). Já todos os itens das Partes 3 e 4 encontram-se todos bem ajustados, ou seja, no intervalo de WMS entre 0,5 e 1,5. Relativamente ao UMS (*Outfit*), os valores médios das 3 partes encontram-se igualmente próximos de 1,0. Na Parte 2, dois dos itens apresentam um desajuste moderado (P2.9 e P2.16), estando os restantes bem ajustados. Na Parte 3 (Conhecimentos em avaliação), apenas 1 item (P3.5) apresenta um desajuste moderado, sendo inclusivé o item com maior valor de β, o que pode ser revelador que alguns respondentes tenham respondido ao item ao acaso. Já a parte 4 apresenta 1 item com um ajuste pouco produtivo, estando os restantes bem ajustados.

Nenhum dos itens do QALA apresenta valores de WMS e UMS superiores a 2,0, ou seja, nenhum item apresenta um desajuste severo (segundo Linacre, 2002) ao modelo esperado, pelo que todos os itens do QALA podem ser mantidos.

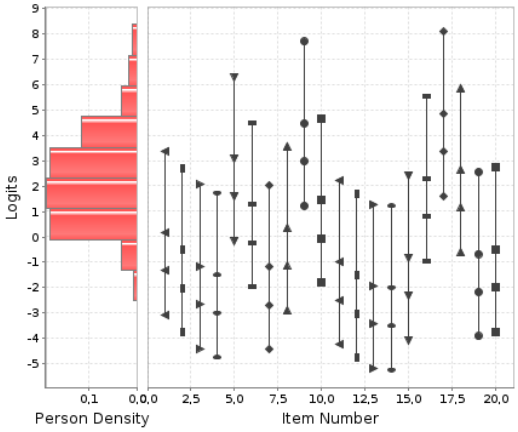
Tabela 21: Resumo dos Índices de Ajuste dos Itens do QALA

	Parte 2		Parte 3		Parte 4	
	WMS	UMS	WMS	UMS	WMS	UMS
Média	0,99	0,98	0,99	1,03	0,99	0,95
Desvio-Padrão	-0,2	-0,18	0,02	0,32	0,18	0,08
Máximo	1,70	1,74	1,14	1,64	1,15	1,43
Mínimo	0,62	0,60	0,83	0,75	0,82	0,45
Itens bem ajustados(0,5-1,5)	19	18	40	39	20	19
Itens com desajuste moderado(1,5-2,0)	1	2	-	1	-	-
Itens com desajuste severo(>2,0)	-	-	-	-	-	-
Itens pouco produtivos(<0,5)	-	-	-	-	-	1

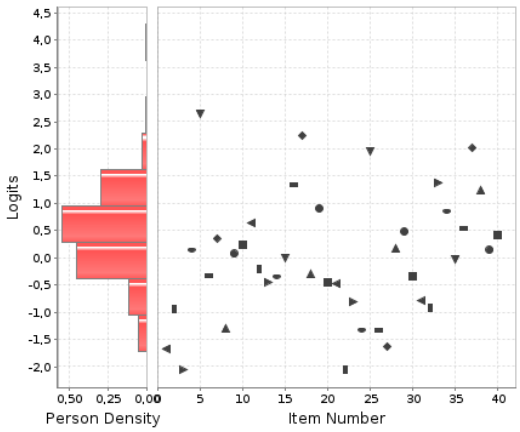
4.1.6 Mapas Item-Pessoa

Na Figura 13 estão presentes os Mapas Item-Pessoa das três partes do QALA. Os mapas de item-pessoa parecem confirmar o que foi referido anteriormente sobre a ausência de *floor* e *ceiling effect* dos itens. Verifica-se uma boa dispersão de habilidades (θ) das pessoas (à esquerda) e das dificuldades (β) dos itens (à direita) na escala de *logit*. Para além disso, a faixa das dificuldades dos itens sobrepõem-se adequadamente à faixa das habilidades das pessoas, pelo que podemos concluir que as três partes dos QALA apresentam um bom desempenho a medir os respetivos traços latentes.

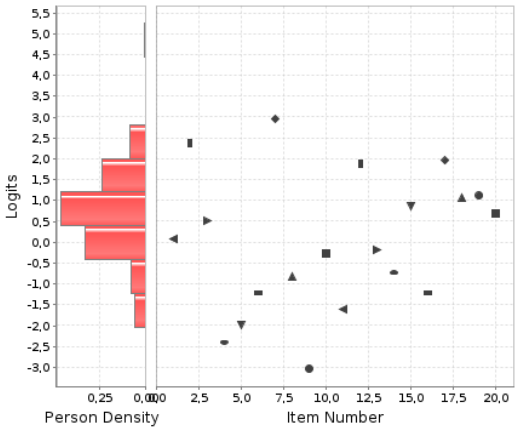
Figura 13: Mapas de Wright



(a) Parte 2



(b) Parte 3



(c) Parte 4

4.1.7 Funcionamento Diferencial dos Itens (DIF)

Para a verificação da existência de itens com DIF recorremos ao método Cochram-Mantel-Haenszel que classifica os itens quanto ao grau de DIF. Utilizaram-se como subgrupos em análise o susbsistema de ensino e o nível de ensino dos professores. Desta forma, verificou-se se os itens tinham um funcionamento semelhante entre professores do subsistema Público e Particular/Cooperativo, bem como entre os professores do 3°Ciclo e Secundário com

os professores do 1°e 2°Ciclos.

Tabela 22: Resumo dos resultados da análise de DIF

	Parte 2		Parte 3		Parte 4	
	S.E.*	N.E.**	S.E.*	N.E.**	S.E.*	N.E.**
Itens com DIF baixo (A/AA)	19	19	37	38	19	17
Itens com DIF moderado (B/BB)	1	1	3	2	1	3
Itens com DIF elevado (C/CC)	0	0	0	0	0	0

**Subsistema de Ensino **Nível de Ensino*

Da análise da Tabela 22 concluimos que não existem itens classificados com C ou CC (DIF elevado), ou seja, não há evidências que o QALA contenha itens que coloquem em causa a sua validade, pelo que nenhum item foi retirado.

4.1.8 Fiabilidade e Separação dos Itens

Os itens que compõem as 3 partes do QALA apresentam excelentes valores de fiabilidade e separação (Tabela 23), de acordo com os critérios definidos na Tabela 14 (ver página 119). Este aspeto é uma evidência importante da validade de construto e de replicabilidade.

Tabela 23: Índices de fiabilidade e separação dos itens

	Parte 2	Parte 3	Parte 4
Fiabilidade dos Itens	0,988	0,981	0,987
Separação dos Itens	9,215	7,238	8,670

4.1.9 Consistência Interna

Os valores do coeficiente α das Partes 2, 3 e 4 do QALA foram, respetivamente, de 0,94, 0,72 e 0,59 (conforme Anexo 7). Embora existam diferenças relevantes de consistência interna entre as 3 partes do QALA, os valores apresentados estão acima do limiar considerado como aceitável e suficiente (ver Figura 10, página 119). O valor mais baixo da Parte 4, embora sendo considerado satisfatório, pode ser explicado pelo baixo número de itens (20), no entanto, outros indicadores de consistência interna (como é o caso do ajuste dos itens) indicam que os itens funcionam corretamente.

4.1.10 Conclusões

A aplicação do Modelo de Rasch permitiu avaliar as propriedades psicométricas do QALA à luz da TRI. Para a utilização do Modelo Rasch foi necessário verificar se os dados cumpriam dois requisitos fundamentais, a unidimensionalidade e a independência local dos itens. A ACPr das 3 partes do QALA permitiu verificar que cada uma delas cumpria esse mesmo requisito, ou seja, medir apenas um traço latente. Já a análise das matrizes de correlação dos resíduos permitiu verificar que o pressuposto da independência local dos itens foi igualmente cumprido.

A análise dos limiares de categorias da Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação), a partir das CCI, permitiu verificar que o sistema de 5 categorias utilizado[20] apresenta boas qualidades psicométricas, já que os limiares de categoria estão organizados por ordem ascendente e concordante com o sistema adotado.

Os itens do QALA estão bem ajustados ao Modelo Rasch, havendo apenas um

[20] Discordo Totalmente, Discordo, Não Concordo Nem Discordo, Concordo e Concordo Totalmente

item que apresenta um valor de WMS que indica um desajuste moderado. A fiabilidade e separação dos itens é considerada excelente nas 3 partes do QALA, sendo um importante indicador de validade de construto e de replicabilidade.

Também os mapas item-pessoa parecem confirmar a validade de construto, já que dois aspetos foram verificados. Por um lado, há uma boa dispersão das habilidades dos respondentes e das dificuldades dos itens ao longo dos respetivos traços latentes. Por outro, é evidente uma especial sobreposição das dificuldades em relação às habilidades, pelo que se conclui que as 3 partes revelam um bom desempenho a medir os respetivos traços latentes.

Não foram encontrados valores de DIF que colocassem em causa a validade dos itens ou do QALA.

Finalmente, os valores de α de *Cronbach* são considerados satisfatórios, tendo a Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) os valor mais alto (α=0,94) e a Parte 4 (Cenários em contexto de avaliação) o valor mais baixo (α=0,59).

Com estes resultados, podemos concluir que as 3 partes do QALA apresentam boas propriedades psicométricas e evidenciam validade de construto. Assim, nos subcapítulos seguintes, iremos apresentar os resultados alcançados pela aplicação do QALA.

4.2 Análise Descritiva

Neste subcapítulo serão apresentados os resultados obtidos nas 3 partes do QALA considerando algumas estatísticas descritivas que, tal como o nome indica, permite

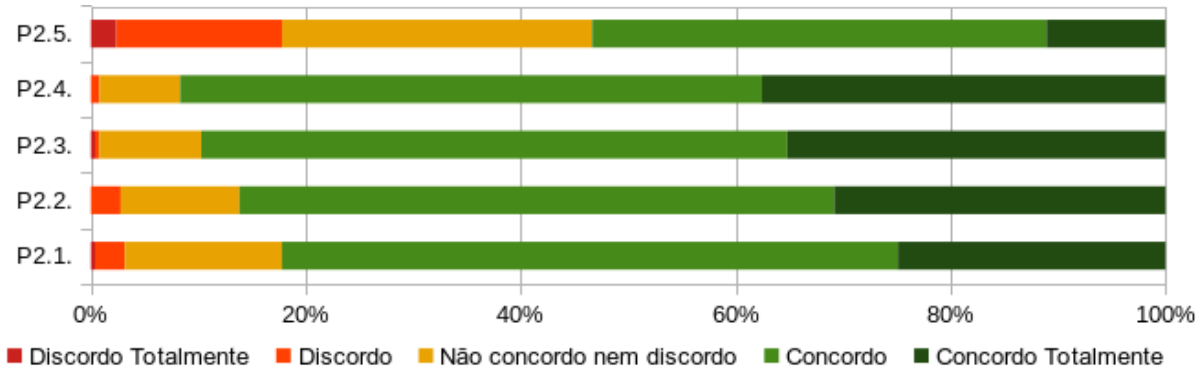
descrever os dados a partir de estatísticas como a média, a moda e o desvio-padrão (Pestana & Gageiro, 2003).

Para cada uma das partes que constituem o QALA, serão apresentados os resultados obtidos para cada um dos domínios que o constituem, bem como para a globalidade dos mesmos. Serão igualmente apresentados os resultados tendo em conta algumas variáveis de contexto que servirão de base para a análise inferencial, a realizar no próximo subcapítulo.

4.2.1 Perceções sobre conhecimentos e competências em avaliação

Para o domínio do 'Conhecimento sobre os objetivos e funções em avaliação' (P2D1) a distribuição das respostas obtidas nos 5 itens que o compõem encontra-se sistematizada na Figura 14.

Figura 14: Distribuição das respostas obtidas no domínio Conhecimentos sobre objetivos e funções da avaliação da Parte 2 do QALA

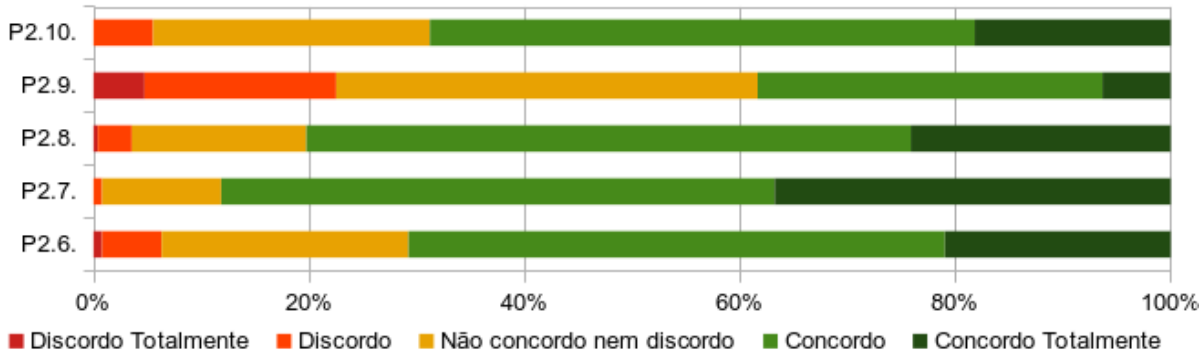


De acordo com os resultados obtidos neste domínio, verifica-se que a opção mais selecionada nos 5 itens foi Concordo (valor 4 na escala de tipo *Likert*). Em 4 dos 5 itens que constituem este domínio, a segunda opção mais selecionada foi o Concordo

Totalmente (valor 5 na escala de tipo *Likert*), sendo a exceção o item P2.5, o que revela uma menor confiança dos professores face aos seus conhecimentos em relação à distinção entre a avaliação criterial e normativa. No entanto, e de um modo geral, verifica-se que a amostra tem uma autoperceção, face ao domínio em análise, positiva ou muito positiva. De salientar ainda que a média alcançada na globalidade do domínio, ou seja, considerando o conjunto dos 5 itens, se situou nos 4,028 (de um máximo de 5), o que evidencia uma perceção positiva dos professores em relação aos seus conhecimentos sobre os objetivos e funções da avaliação.

Para o domínio do 'Conhecimentos sobre currículo e sobre o que é importante aprender e avaliar' (P2D2) a distribuição das respostas obtidas nos 5 itens que compõem o domínio encontra-se sistematizada na Figura 15.

Figura 15: Distribuição das respostas obtidas no domínio Conhecimentos sobre currículo e sobre o que é importante aprender e avaliar da Parte 2 do QALA



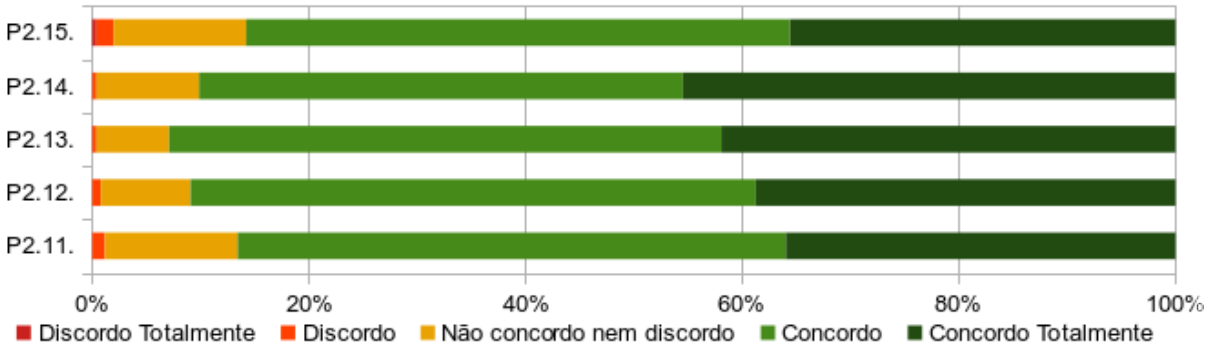
De acordo com os resultados obtidos neste domínio, verifica-se que a opção mais selecionada em 4 dos 5 itens foi Concordo (valor 4 na escala de tipo *Likert*), sendo que apenas o item P2.9[21] teve como opção mais selecionada o valor 3 (Não concordo nem discordo). A opção menos escolhida foi 1 (Discordo Totalmente), sendo que, em dois dos itens, não houve qualquer resposta com este valor. Já a média global verificada neste domínio situou-se nos 3,816 (de um máximo de 5). A

[21] Item P2.9.: Considero que possuo sólidos conhecimentos sobre níveis de complexidade cognitiva (p.e. Taxonomia de Bloom, Taxonomia de Marzano, Depth of Knowledge).

descida verificada neste domínio, em comparação com o domínio anterior, fica muito a dever-se aos modestos resultados alcançados no item P2.9., dado que apenas 38,3% dos respondentes selecionou a opção Concordo (32%) ou Concordo Totalmente (6,3%). Já na análise às dificuldades resultantes do modelo Rasch, este item foi identificado como tendo um dos mais altos índices de dificuldade.

Para o domínio do 'Conhecimentos sobre utilização de instrumentos de avaliação diversificados' (P2D3) a distribuição das respostas obtidas nos 5 itens que compõem o domínio encontra-se sistematizada na Figura 16.

Figura 16: Distribuição das respostas obtidas no domínio Conhecimentos sobre utilização de instrumentos de avaliação diversificados da Parte 2 do QALA



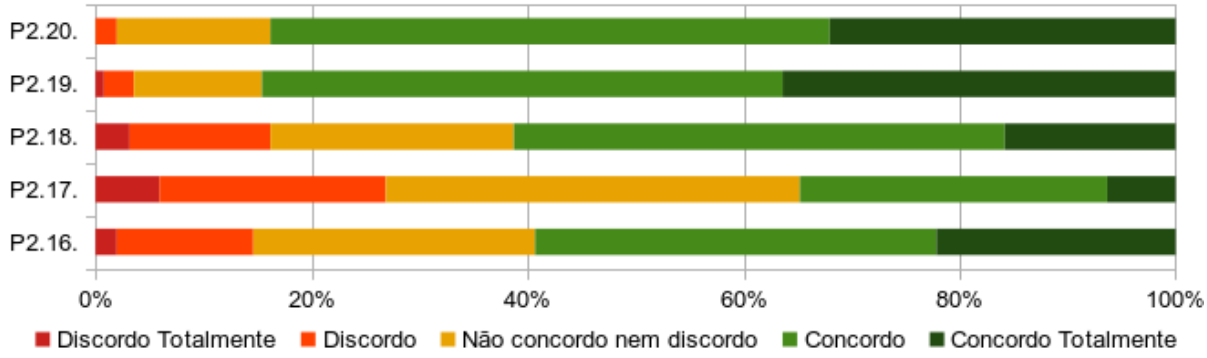
De acordo com os resultados obtidos neste domínio, verifica-se que a opção mais selecionada em 4 dos 5 itens foi Concordo (valor 4 na escala de tipo *Likert*), sendo que apenas o item P2.14[22] teve como opção mais selecionada o valor 5 (Concordo Totalmente). A opção menos escolhida foi 1 (Discordo Totalmente), sendo que em quatro dos itens, com exceção do P2.15[23], não houve qualquer resposta com este valor. Dos quatro domínios que constituem a Parte 2 do QALA, este foi o que apresentou a média global mais elevada com 4,277. Este aspeto parece evidenciar que os professores têm uma perceção muito positiva em relação aos seus

[22] Item P2.14: Considero que possuo sólidas competências sobre como construir diferentes tipos de itens de avaliação.
[23] Item P2.15: Considero que possuo sólidas competências sobre como incluir os alunos no processo de avaliação.

conhecimentos e capacidades sobre a utilização de instrumentos de avaliação diversificados.

As respostas ao domínio 'Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação' (P2D4), encontram-se sistematizadas na Figura 17.

Figura 17: Distribuição das respostas obtidas no domínio Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação da Parte 2 do QALA



Constata-se, a partir dos resultados obtidos no presente domínio, que em 4 dos 5 itens a opção mais selecionada foi a 4 (concordo). Só o item P2.17[24] tem como opção mais selecionada a 3 (Não concordo nem discordo). Mais uma vez, a opção menos selecionada foi a 1 (discordo totalmente), havendo mesmo itens sem qualquer resposta com este valor, como é o caso do item P2.20[25]. A média da globalidade deste domínio foi de 3,723 (em 5 possíveis), sendo este o valor mais baixo entre os quatro domínios considerados. Este facto é reflexo dos resultados modestos alcançados no item P2.17, já que apenas 34,8% dos respondentes selecionaram as opções Concordo (28,5%) e

[24]Item 2.17: Considero que possuo sólidas competências sobre como determinar algumas propriedades psicométricas dos instrumentos/itens de avaliação, p.e índice de dificuldade e índice de discriminação.

[25]Item P2.20: Considero que possuo sólidas competências sobre como utilizar a informação recolhida de forma a melhorar a minha prática pedagógica.

Concordo Totalmente (6,3%), mas também nos itens P2.16[26] e P2.18[27] que, conforme se pode verificar, apresentam um volume significativo de resposta nas categorias 1 (Discordo Totalmente) e 2 (Discordo).

Seguidamente, analisar-se-ão os resultados obtidos na Parte 2 do QALA e nos respetivos domínios (Tabela 24), tendo em consideração algumas das variáveis de contexto recolhidas na Parte 1 (Dados Gerais). Assim temos que:

- Os professores do sexo feminino apresentam médias ligeiramente superiores aos professores do sexo masculino nos 4 domínios considerados. Este aspeto reflete-se na média global da Parte 2 do QALA, já que os professores do sexo feminino alcançaram uma média de 3,994 e do sexo masculino 3,835.

- Tendo em consideração o subsistema de ensino onde os respondentes lecionam, verifica-se que, para os 4 domínios do QALA, os professores do Ensino Público apresentam resultados ligeiramente melhores que os professores do Ensino Particular e Cooperativo. A média global é de 3,992 para os professores do Ensino Público, ao passo que os professores do Ensino Particular e Cooperativo é de 3,856.

- Considerando o tipo de habilitação para a docência, foram apurados resultados contrários àqueles que seriam expectáveis. De facto, seria expectável que os professores com habilitação profissional apresentassem valores mais altos que os professores com habilitação própria, já que a formação inicial de professores possibilitou, à partida, a aquisição de conhecimentos na área da avaliação das aprendizagens. No entanto, os resultados apurados parecem evidenciar que os

[26] Item P2.16: Considero que possuo sólidas competências sobre como calcular medidas de localização, p.e. média, moda e mediana, e dispersão, p.e. desvio-padrão, com a informação recolhida após a realização de um teste.

[27] Item 2.18: Considero que possuo sólidas competências sobre como construir instrumentos de registo da avaliação, p.e. perfis de desempenho, listas de verificação.

professores com habilitação própria têm uma melhor autoperceção sobre os seus conhecimentos e capacidades do que os professores com habilitação profissional, já que apresentam uma média superior nos 4 domínios em análise. Há que salientar, contudo, que as diferenças são muito pequenas já que a média global dos professores com habilitação própria é de 4,049, enquanto que a média global dos professores com habilitação profissional é de 3,937.

- Os professores com vínculo estável (Quadro) apresentam médias superiores aos professores contratados, tanto nos domínios considerados como nos resultados globais. Este aspeto pode ser explicado pelo facto de, normalmente, os professores do quadro, sejam eles do subsistema público ou particular/cooperativo, possuirem uma maior experiência profissional pelo que, à partida, terão também mais conhecimentos e capacidades ao nível da avaliação e, tal facto reflete-se na sua autoperceção sobre esses mesmos conhecimentos e competências. Nos resultados globais, os professores do quadro apresentaram uma média de 3,994 e os professores contratados de 3,858.

- De um modo geral, e salvo algumas exceções, verifica-se que à medida que a idade dos respondentes aumenta, também os valores médios aumentam. Este aspeto, tal como no caso anterior, pode ser explicado pelo facto de, em princípio, quanto maior for a idade maior será também a experiência profissional. Os resultados globais são crescentes conforme a classe etária, com exceção dos professores com mais de 60 anos que apresentam uma média ligeiramente inferior à dos professores do escalão etário entre 51 e 60 anos. A diferença entre a média global mais baixa (professores entre os 26 e 30 anos) e a média global mais alta (professores entre os 51 e 60 anos) é de cerca de 0,3.

- Tal como no caso anterior, e também com algumas exceções, verifica-se que as

médias, tanto nos domínos como nos resultados globais, aumenta com a experiência letiva. Este aspeto parece evidenciar que quanto maior é a experiência profissional, maior será também a autoperceção sobre os conhecimentos e capacidades em avaliar. Os resultados globais variaram de 3,879, para os professores entre 4 e 6 anos de serviço, até aos 4,122, para os professores com mais de 35 anos de serviço.

- Analisando os resultados com base no nível de ensino dos professores respondentes, verifica-se que os professores do 3°Ciclo e do Secundário apresentam valores superiores aos demais professores em 3 dos 4 domínios, sendo a exceção o domínio ’Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação’ onde os professores do 2°Ciclo apresentaram um resultado ligeiramente superior. Os resultados globais apresentam pequenas diferenças entre os professores dos diferentes níveis de ensino. Ainda assim, parecem indicar que as perceções sobre os conhecimentos e capacidades em avaliação tendem a ser ligeiramenre melhores quanto maior é o seu nível de ensino, já que as médias globais mais baixas foram alcançadas pelos professores de 1°Ciclo, seguindo-se os professores de 2°Ciclo e, por fim, os professores de 3°Ciclo e secundário.

- Os resultados apurados parecem evidenciar que os professores de Línguas e de Ciências Sociais e Humanas têm uma melhor perceção dos seus conhecimentos e capacidades em avaliação, já que cada uma das áreas disciplinares apresenta médias superiores em dois domínios. Já os professores de Matemática e Ciências Experimentais apresentam valores mais baixos em 3 dos 4 domínios, sendo a exceção o domínio ’Conhecimentos sobre currículo e sobre o que é importante aprender e avaliar’. Nos resultados globais, a área disciplinar que apresentou médias mais elevadas foi a de Ciências Sociais e Humanas, seguindo-se a de

Línguas, de Expressões e, por fim, a de Matemática e Ciências Experimentais. Importa, contudo, ressalvar que as diferenças nos resultados globais entre as quatro áreas disciplinares são muito pequenas.

- Por último, verifica-se, para os 4 domínios, que os professores que apostaram na formação contínua em avaliação, apresentam valores substancialmente superiores aos dos professores que não frequentaram este tipo de formação. Na globalidade dos resultados, os professores que frequentaram ações de formação em avaliação tiveram uma média de 4,019, ao passo que os professores que não frequentaram tiveram uma média de 3,804.

Tabela 24: Síntese dos resultados obtidos na Parte 2 do QALA

	P2D1 $\bar{x}$	P2D2 $\bar{x}$	P2D3 $\bar{x}$	P2D4 $\bar{x}$	Global $\bar{x}$
Total da Amostra	4,028	3,816	4,277	3,723	3,961
Sexo					
F	4,062	3,839	4,314	3,759	3,994
M	3,896	3,727	4,135	3,581	3,835
Subsistema de Ensino					
Público	4,062	3,834	4,335	3,738	3,992
Particular e Cooperativo	3,938	3,749	4,084	3,655	3,856
Tipo de Habilitação					
Própria	4,126	3,926	4,341	3,804	4,049
Profissional	4,002	3,786	4,260	3,701	3,937
Vínculo					
Contratado	3,939	3,749	4,170	3,566	3,858
Quadro	4,056	3,838	4,309	3,771	3,994
Idade					
Entre 26 e 30 anos	3,800	3,667	3,967	3,633	3,767
Entre 31 e 40 anos	3,900	3,672	4,116	3,528	3,804
Entre 41 e 50 anos	3,971	3,792	4,248	3,785	3,949
Entre 51 e 60 anos	4,168	3,924	4,411	3,771	4,068
Mais de 60 anos	4,136	3,904	4,384	3,744	4,042
Experiência letiva					
Até 3 anos	3,800	3,950	3,975	3,850	3,894
Entre 4 e 6 anos	3,967	3,750	4,167	3,633	3,879
Entre 7 e 25 anos	3,966	3,743	4,218	3,682	3,902
Entre 26 e 35 anos	4,117	3,889	4,383	3,736	4,031
Mais de 35 anos	4,200	3,970	4,430	3,889	4,122
Nível de Ensino					
1°Ciclo	3,941	3,744	4,219	3,719	3,905
2°Ciclo	3,947	3,750	4,268	3,725	3,922
3°Ciclo e Secundário	4,077	3,861	4,300	3,724	3,991
Área disciplinar					
MCE*	3,941	3,764	4,184	3,724	3,903
CSH**	4,055	3,909	4,314	3,759	4,009
LING***	4,099	3,894	4,349	3,656	3,999
EXP****	4,005	3,705	4,238	3,762	3,927
Formação Contínua em Avaliação					
Sim	4,089	3,881	4,317	3,789	4,019
Não	3,865	3,638	4,171	3,541	3,804

*Matemática e Ciências Experimentais **Ciências Sociais e Humanas *** Línguas ****Expressões

4.2.2 Conhecimentos em Avaliação

Na Parte 3 do QALA foram aplicadas 40 questões de verdadeiro e falso organizadas em torno dos 4 domínios considerados. Cada afirmação admitia 3 possibilidades de resposta: verdadeiro (V), falso (F) e não sei (NS). A cada resposta correta foi atribuído 1 ponto e às restantes 0 pontos.

A síntese dos resultados obtidos no domínio 'Conhecimentos sobre os objetivos e funções da avaliação' encontra-se sistematizada na Tabela 25. As médias para os 10 itens que compõem este domínio variaram entre 0,119 (P3.5[28]) e 0,909 (P3.3[29] e P3.22[30]). As duas questões com as médias mais baixas (P3.5 e P3.25[31]) estão relacionadas com a avaliação criterial e normativa.

Também a questão P3.4[32] apresentou um resultado relativamente baixo (0,569). Neste item, a afirmação apresentava uma das principais características da avaliação formativa atribuída à avaliação sumativa.

A média global deste domínio foi de apenas 6,652, o que se traduz numa percentagem de acertos ligeiramente superior a 66%. Considerando a importância que este domínio tem no contexto da avaliação das aprendizagens, os resultados ficaram aquém do esperado. Efetivamente, para uma boa avaliação é fundamental que os professores saibam porque avaliam, como avaliam e quando devem avaliar. No entanto, os dados obtidos revelam a existência fragilidades neste domínio.

[28] Item P3.5: Numa avaliação referente à norma, o objetivo principal é apreciar as aprendizagens efetivamente realizadas pelo aluno, em relação às finalidades consideradas e aos objetivos definidos.

[29] Item P3.3: A avaliação formativa tem uma influência significativa nas aprendizagens dos alunos, contribuindo para a sua motivação e auto-estima.

[30] Item P3.22: A avaliação de diagnóstico permite explorar ou identificar características dos alunos que sejam relevantes para a definição de estratégias de ensino e avaliação mais adequadas.

[31] Item P3.25: A avaliação de referência a critério promove a competição entre alunos, enquanto que a avaliação referente a norma promove a competição do aluno consigo próprio.

[32] Item P3.4: A avaliação sumativa tem uma função predominantemente corretiva, visto permitir corrigir os erros cometidos ao longo do processo de ensino e aprendizagem.

Tabela 25: Síntese dos resultados obtidos no domínio Conhecimentos sobre os objetivos e funções da Avaliação (P3D1) da Parte 3 do QALA

	P3.1	P3.2	P3.3	P3.4	P3.5	P3.21	P3.22	P3.23	P3.24	P3.25	Global
Média	0.874	0.779	0.909	0.569	0.119	0.700	0.909	0.759	0.834	0.202	6.652
DP	0.333	0.416	0.288	0.496	0.324	0.459	0.288	0.429	0.373	0.402	1.430
Mín.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	2.000
Máx.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	10.000

Os resultados no domínio 'Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar' encontram-se sistematizados na Tabela 26. Conforme se pode verificar, as médias para os 10 itens neste domínio variaram entre 0,494 (P3.29[33]) e 0,870 (P3.27[34]). Dois grupos de questões sobressaem pelos seus baixos resultados. O primeiro grupo, composto pelos itens P3.9[35] (média=0,585) e P3.29 (média=0,494), corresponde a questões relacionadas com o conhecimento sobre domínios de complexidade cognitiva, como por exemplo a *Taxonomia de Bloom, a Taxonomia de Marzano e o Depth of Knowledge*.

O segundo grupo é composto pelos itens P3.10[36] (média=0,549) e P3.30[37] (média=0,672). Estes itens procuraram aferir o conhecimento que os professores tinham sobre os instrumentos de auxílio à construção de instrumentos de avaliação. A partir destes instrumentos, os professores podem selecionar os conteúdos e objetivos que serão alvo de avaliação. Dois exemplos destes instrumentos são as tabelas de especificações e as matrizes de conteúdo. Os resultados alcançados pela aplicação do QALA foram baixos, já que a média para o item P3.10 foi de 0,549 e

[33]Item 3.29: Questões que procuram avaliar a capacidade de compreensão do aluno situam-se num nível de complexidade cognitivo elevado.

[34]Item 3.27: Uma das funções da avaliação deverá ser o de certificar as capacidades e atitudes no âmbito das competências inscritas no Perfil do Aluno à Saída da Escolaridade Obrigatória.

[35]Item P3.9: Itens com elevado grau de dificuldade requerem uma maior capacidade em criar e avaliar situações com maior grau de complexidade.

[36]Item P3.10: As tabelas de especificações são ferramentas especialmente úteis para definir os objetivos que serão visados numa ficha de avaliação.

[37]Item P3.30: A elaboração de matrizes de conteúdos, a serem alvo de avaliação através de exame nacional, é da exclusiva competência dos organismos centrais do Ministério da Educação.

para o item P3.30 foi 0,672.

Tabela 26: Síntese dos resultados obtidos no domínio Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar (P3D2) da Parte 3 do QALA

	P3.6	P3.7	P3.8	P3.9	P3.10	P3.26	P3.27	P3.28	P3.29	P3.30	Global
Média	0.668	0.522	0.830	0.585	0.549	0.834	0.870	0.561	0.494	0.672	6.585
DP	0.472	0.501	0.376	0.494	0.499	0.373	0.337	0.497	0.501	0.470	1.840
Mín.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Máx.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	10.000

Os resultados obtidos no domínio 'Conhecimentos sobre a utilização de instrumentos de avaliação diversificados' encontram-se sistematizados na tabela 27. As médias para os 10 itens variaram entre 0,300 (P3.33) e 0,775 (P3.32).

Dada a importância da utilização diversificada de instrumentos de avaliação em sala de aula, esperar-se-ia que os professores possuissem um bom nível neste domínio. No entanto, os resultados parecem evidenciar o oposto já que a média global foi de apenas 5,917 (num máximo de 10), constituindo-se como o segundo domínio com média mais baixa. Há também a salientar o facto de em 3 itens (P3.11[38], P3.33[39] e P3.34[40]) a médias terem sido inferiores a 0,500, ou seja, com menos de 50% de acertos.

Tabela 27: Síntese dos resultados obtidos no domínio Conhecimentos sobre Utilização de instrumentos de avaliação diversificados (P3D3) da Parte 3 do QALA

	P3.11	P3.12	P3.13	P3.14	P3.15	P3.31	P3.32	P3.33	P3.34	P3.35	Global
Média	0.458	0.644	0.692	0.672	0.601	0.755	0.775	0.300	0.411	0.609	5.917
DP	0.499	0.480	0.463	0.470	0.491	0.431	0.419	0.459	0.493	0.489	1.957
Mín.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
Máx.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	10.000

Relativamente ao domínio 'Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação' os resultados encontram-se

[38]Item P3.11: Os instrumentos de avaliação de diagnóstico devem possibilitar o despiste de variáveis exclusivamente cognitivas que possam interferir na aprendizagem dos alunos.

[39]Item P3.33: Instrumentos de avaliação sumativa assentes em questões de verdeiro e falso, ou questões de escolha múltipla, apresentam uma maior fiabilidade, visto serem de correção objetiva

[40]Item P3.34: As questões do tipo completamento têm como desvantagem o facto de não permitirem a avaliação de um leque alargado de conteúdos.

sistematizados na tabela 28. Conforme se pode verificar, as médias dos 10 itens variaram entre os 0,162 (P3.17[41]) e os 0,692 (P3.20[42]). Podemos assim concluir a existência de grandes fragilidades neste domínio, dados os baixos valores alcançados nos vários itens, o que se refletiu numa média global muito baixa (cerca de 43% de acertos).

Neste domínio foram colocadas questões que procuraram aferir os conhecimentos dos professores em áreas como a determinação de medidas de localização e dispersão, das qualidades psicométricas dos instrumentos de avaliação, da utilização de instrumentos de registos de avaliação e da comunicação da avaliação do professor aos alunos (*feedback*) e da utilização da informação recolhida no processo de avaliação na transformação das práticas pedagógicas dos professores.

Dois itens sobressaem pelos seus muito baixos resultados, o P3.17 (média=0,162) e o P3.37[43] (média=0,194). Ambos os itens estão relacionados com a determinação das qualidades psicométricas dos instrumentos de avaliação pelo que os resultados alcançados parecem ser indicadores de grandes fragilidades neste domínio.

Tabela 28: Síntese dos resultados obtidos no domínio Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação (P3D4) da Parte 3 do QALA

	P3.16	P3.17	P3.18	P3.19	P3.20	P3.36	P3.37	P3.38	P3.39	P3.40	Global
Média	0.308	0.162	0.664	0.399	0.692	0.478	0.194	0.328	0.569	0.510	4.304
DP	0.463	0.369	0.473	0.491	0.463	0.501	0.396	0.470	0.496	0.501	1.786
Mín.	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Máx.	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	10.000

Por último, a tabela 29, sintetiza os resultados obtidos em cada um dos domínios da Parte 3 do QALA, bem como os resultados globais tendo em consideração as váriaveis

[41] Item P3.17: Um instrumento de avaliação com índice de dificuldade de 0,5 permite uma maior diferenciação de resultados.

[42] Item P3.20: Após a comunicação dos resultados da aplicação dos instrumentos de avaliação, o professor deverá prosseguir com a planificação da disciplina definida à priori.

[43] Item P3.37: A fiabilidade de um instrumento de avaliação garante a sua validade.

de contexto consideradas no presente estudo.

Tabela 29: Síntese dos resultados obtidos na Parte 3 do QALA

	P3D1 $\bar{x}$	P3D2 $\bar{x}$	P3D3 $\bar{x}$	P3D4 $\bar{x}$	Global $\bar{x}$
Total da Amostra	6,652	6,585	5,917	4,304	58,646
Sexo					
F	6,627	6,751	5,985	4,139	58,756
M	6,750	5,942	5,654	4,942	58,221
Subsistema de Ensino					
Público	6,686	6,796	6,037	4,335	59,634
Particular e Cooperativo	6,527	5,818	5,564	4,273	55,455
Tipo de Habilitação					
Própria	6,611	6,722	5,852	4,056	58,102
Profissional	6,663	6,548	5,935	4,372	58,794
Vínculo					
Contratado	6,356	6,373	5,678	3,949	55,890
Quadro	6,741	6,668	6,005	4,425	59,598
Idade					
Entre 26 e 30 anos	6,500	6,167	5,000	3,167	52,083
Entre 31 e 40 anos	6,560	6,040	5,820	4,080	56,250
Entre 41 e 50 anos	6,469	6,396	5,750	4,469	57,708
Entre 51 e 60 anos	7,013	7,066	6,276	4,368	61,809
Mais de 60 anos	6,480	7,040	5,880	4,200	59,000
Experiência letiva					
Até 3 anos	6,500	6,250	5,375	2,375	51,250
Entre 4 e 6 anos	6,750	6,667	6,333	4,417	60,417
Entre 7 e 25 anos	6,507	6,284	5,619	4,448	57,146
Entre 26 e 35 anos	6,972	7,056	6,486	4,375	62,222
Mais de 35 anos	6,519	6,889	5,852	3,926	57,963
Nível de Ensino					
1°Ciclo	6,750	6,250	5,594	4,016	56,523
2°Ciclo	6,438	6,388	5,850	4,150	57,063
3°Ciclo e Secundário	6,727	6,673	5,933	4,473	59,517
Área disciplinar					
MCE*	6,492	6,328	5,656	4,590	57,664
CSH**	7,023	7,000	6,455	4,932	63,523
LING***	6,797	6,861	6,228	3,911	59,494
EXP****	6,357	6,357	5,595	4,524	57,083
Formação Contínua em Avaliação					
Sim	6,659	6,746	6,097	4,519	60,054
Não	6,632	6,147	5,426	3,721	54,816

*Matemática e Ciências Experimentais **Ciências Sociais e Humanas *** Línguas ****Expressões

Na Parte 3 do QALA, a totalidade da amostra alcançou uma média de acertos na ordem dos 58,6%. O domínio que apresentou melhores resultados foi o ’Conhecimentos sobre os objetivos e funções da avaliação’, com uma média de acertos de cerca de 66,5%, e o que piores resultados alcançou foi o domínio ’Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação’ com uma média de acertos de apenas 43,04%.

Analisando os resultados considerando as variáveis de contexto, temos que:

- Os professores do sexo feminino alcançaram melhores resultados em dois domínios, tendo os professores do sexo masculino alcançado melhores resultados nos outros dois domínios considerados. Globalmente, os professores do sexo feminino obtiveram um resultado ligeiramente superior aos do sexo masculino, apresentando valores de 58,8% e 58,2% respetivamente;

- Os professores que lecionam no ensino público alcançaram médias superiores, aos professores do ensino particular e cooperativo, nos 4 domínios considerados. Consequentemente, os professores do ensino público obtiveram uma média global superior, em cerca de 4 pontos percentuais, que os professores do ensino particular e cooperativo;

- Os professores com habilitação profissional obtiveram melhores resultados em 3 dos 4 domínios considerados, sendo a exceção o domínio ”Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar”(P3D2). Os resultados globais não diferem muito entre professores com habilitação profissional e habilitação própria, já que os primeiros tiveram uma média superior em cerca de 0,7 pontos percentuais quando comparados com os segundos;

- Os professores pertencentes aos quadros alcançaram médias mais altas nos 4 domínios, quando comparados com os professores contratados. Tal facto teve reflexo nas médias globais, já que os professores dos quadros tiveram uma média global de 59,6%, ao passo que os professores contratados alcançaram apenas 55,9%;

- Os professores dos escalões etários mais altos apresentam melhores resultados que os restantes, embora se verifique um ligeiro decréscimo, em todos os domínios, entre os professores com idades entre os 51 e 60 anos e os professores com mais de 60 anos. Analisando as médias globais, verifica-se que, tendencialmente, as médias aumentam com o escalão etário, com a exceção do caso anteriormente identificado;

- Os professores com experiência entre os 26 e 35 anos apresentam médias superiores aos demais em 3 dos 4 domínios. Já os professores com experiência até 3 anos alcançaram as médias mais baixas nos 4 domínios. Os resultados globais mostram que as médias mais altas foram alcançadas pelos professores com experiência letiva entre os 26 e 35 anos, seguindo-se os professores com 4 a 6 anos de experiência. A média global mais baixa foi alcançada pelos professores com até 3 anos de experiência letiva. A diferença entre a média mais alta e a média mais baixa foi de cerca de 11 pontos percentuais;

- Os professores do 3°Ciclo do Ensino Básico e Secundário obtiveram melhores resultados em 3 dos 4 domínios. Já os professores do 1°Ciclo tiveram as médias mais baixas em 3 dos 4 domínios. Este aspeto reflete-se nos resultados globais, verificando-se um aumento das médias com o aumento do nível de ensino. A diferença entre a média dos professores do 1°Ciclo e dos professores do 3°Ciclo do Ensino Básico e Secundário foi de cerca de 3 pontos percentuais;

- Os professores de Ciências Sociais e Humanas obtiveram os melhores resultados em todos os domínios. Já os professores de Expressões obtiveram as médias mais baixas em 3 dos 4 domínios. Os resultados globais mostram que os professores de Ciênciais Sociais e Humanas obtiveram a média mais alta, seguindo-se os professores de Línguas, os professores de Matemática e Ciências Experimentais e, por fim, os professores de Expressões. A diferença entre a média mais baixa e a média mais alta foi de cerca de 6,5 pontos percentuais;

- Os professores que frequentaram ações de formação na área da avaliação alcançaram melhores resultados nos 4 domínios. Assim, também a média global foi mais elevada, sendo a diferença de cerca de 5,2 pontos percentuais em relação aos professores que não frequentaram formação contínua nesta área.

4.2.3 Cenários em contexto de avaliação

A Parte 4 do QALA consistiu na aplicação de 20 questões de escolha múltipla organizadas em torno dos 4 domínios considerados e de 5 cenários hipotéticos em contexto de sala de aula. Para cada questão foram apresentadas 4 possibilidades de resposta, sendo que a cada resposta correta foi atribuído 1 ponto e às restantes 0 pontos.

Em relação aos resultados obtidos no domínio Conhecimentos sobre os objetivos e funções da Avaliação (tabela 30), verificamos que a média se situou entre 0,241 (P4.5.1[44]) e 0,834 (P4.2.2[45]). Já a média global, para o domínio considerado, foi de 3,150, para um máximo possível de 5, o que se traduz numa percentagem de acertos

[44] Item relacionado com a distinção entre avaliação referente a critério e avaliação referente a norma.
[45] Item relacionado com os conhecimentos sobre as características e funções da avaliação de diagnóstico.

de 63%.

Neste domínio, e à semelhança das Partes 2 (Perceções sobre os conhecimentos e capacidades em avaliação) e 3 (Conhecimentos em avaliação), o item que apresentou a média mais baixa está relacionada com a avaliação criterial e normativa. Este aspeto parece confirmar que os professores têm grandes dificuldades em distinguir ambos os conceitos e/ou formas de avaliação. Os restantes itens apresentam resultados mais satisfatórios, situando-se acima dos 0,621, ou seja, acima dos 62% de acertos.

Tabela 30: Síntese dos resultados obtidos no domínio Conhecimentos sobre os objetivos e funções da Avaliação (P4D1) da Parte 4 do QALA

	P4.1.1	P4.2.2	P4.3.2	P4.4.2	P4.5.1	Global
Média	0.621	0.834	0.688	0.767	0.241	3.150
DP	0.486	0.373	0.464	0.424	0.429	1.084
Mín.	0.000	0.000	0.000	0.000	0.000	0.000
Máx.	1.000	1.000	1.000	1.000	1.000	5.000

Já no domínio 'Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar', a média dos resultados (Tabela 31) variou entre 0,178 (P4.1.2[46]) e 0,964 (P4.3.1[47]). A média global foi de 3,170, em 5 possíveis, o que representa 63,4% de acertos.

O item que alcançou a média mais baixa (P4.1.2) procurou avaliar o conhecimento dos professores em relação às ferramentas de auxílio à construção de instrumentos de avaliação. Já na Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA, o conjunto de itens que procurava avaliar este aspeto apresentou dos resultados mais baixos. Também o item relacionado com os

[46] Item relacionado com os conhecimentos sobre a construção e utilização de ferramentas de auxílio à construção de instrumentos de avaliação.

[47] Item relacionado com os conhecimentos sobre os vários tipos de currículo.

conhecimentos sobre a legislação em vigor relacionada com a avaliação das aprendizagens dos alunos (P4.5.2) apresentou um valor abaixo ao que seria desejável, já que a percentagem de acertos se fixou apenas nos 41,1%.

Tabela 31: Síntese dos resultados obtidos no domínio Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar (P4D2) da Parte 4 do QALA

	P4.1.2	P4.2.4	P4.3.1	P4.4.4	P4.5.2	Global
Média	0.178	0.783	0.964	0.834	0.411	3.170
DP	0.383	0.413	0.186	0.373	0.493	0.912
Mín.	0.000	0.000	0.000	0.000	0.000	0.000
Máx.	1.000	1.000	1.000	1.000	1.000	5.000

No domínio 'Conhecimentos sobre Utilização de instrumentos de avaliação diversificados' verifica-se, a partir da tabela 32, que as médias variaram entre 0,399 (P4.5.3[48]) e 0,909 (P4.2.1[49]). A média global foi de 3,387, num máximo de 5, o que corresponde a uma média de acertos de 67,74%.

No conjunto dos itens que compõem este domínio, o P4.5.3 foi o que apresentou uma média mais baixa. Este item enquadra-se no conjunto de itens que procurava aferir o conhecimento dos professores em relação à construção e utilização de diferentes itens de avaliação. Também na Parte 2 do QALA, um dos itens (P3.34[50]) que tinha o mesmo objetivo, apresentou um resultado abaixo dos 0,500, ou seja, mais de metade dos participantes não acertou na questão colocada. Este aspeto parece ser revelador de algumas fragilidades, por parte dos professores, no que diz respeito à construção e/ou utilização de itens de avaliação.

[48] Item relacionado com as competências na construção de diferentes itens de avaliação.

[49] Item relacionado com as competências sobre como e quando fazer uso de instrumentos de avaliação de diagnóstico.

[50] Item P3.34: As questões do tipo completamento têm como desvantagem o facto de não permitirem a avaliação de um leque alargado de conteúdos.

Tabela 32: Síntese dos resultados obtidos no domínio Conhecimentos sobre Utilização de instrumentos de avaliação diversificados (P4D3) da Parte 4 do QALA

	P4.1.3	P4.2.1	P4.3.3	P4.4.1	P4.5.3	Global
Média	0.530	0.909	0.877	0.672	0.399	3.387
DP	0.500	0.288	0.329	0.470	0.491	1.008
Mín.	0.000	0.000	0.000	0.000	0.000	1.000
Máx.	1.000	1.000	1.000	1.000	1.000	5.000

Por último, no domínio 'Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação' as médias variaram entre 0,115 (P4.2.3[51]) e 0,937 (P4.1.4[52]). A média global foi 2,245, em 5 possíveis, o que reflete uma percentagem de acertos de apenas 45%. Assim, este assume-se como o domínio que apresenta as médias globais mais baixas, à semelhança do que ocorreu nas Partes 2 e 3 do QALA.

Dos cinco itens considerados, quatro apresentam uma média de acertos inferior a 50%. Só o item P4.1.4 apresenta um resultado satisfatório com uma média de acertos de cerca de 94%. Fica uma vez mais evidente as fragilidades que os professores têm na interpretação e utilização da informação recolhida no processo de avaliação. Mais uma vez, e conforme foi referido aquando da análise aos resultados neste domínio na Parte 3 (Conhecimentos em avaliação) do QALA, estes dados vão ao encontro dos resultados alcançados nos trabalhos de Plake, Impara e Fager (1993), Alkharusi *et al*. (2012) e Yamtin e Wongwanich (2014).

[51] Item relacionado com as competências na construção de instrumentos para registo de avaliação.
[52] Item relacionado com a utilização da informação recolhida no processo de avaliação para a melhoria da prática pedagógica

Tabela 33: Síntese dos resultados obtidos no domínio Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação (P4D4) da Parte 4 do QALA

	P4.1.4	P4.2.3	P4.3.4	P4.4.3	P4.5.4	Global
Média	0.937	0.115	0.253	0.451	0.490	2.245
DP	0.244	0.319	0.436	0.499	0.501	0.969
Mín.	0.000	0.000	0.000	0.000	0.000	0.000
Máx.	1.000	1.000	1.000	1.000	1.000	5.000

Na tabela 34 podemos encontrar uma síntese dos resultados obtidos em cada um dos domínios da Parte 4 do QALA, bem como os resultados globais considerando as diversas variáveis de contexto em análise.

Tabela 34: Síntese dos resultados obtidos na Parte 4 do QALA

	P4D1 $\bar{x}$	P4D2 $\bar{x}$	P4D3 $\bar{x}$	P4D4 $\bar{x}$	Global $\bar{x}$
Total da Amostra	3,150	3,170	3,387	2,245	59,763
Sexo					
F	3,139	3,114	3,418	2,209	59,403
M	3,192	3,385	3,269	2,385	61,154
Subsistema de Ensino					
Público	3,173	3,162	3,403	2,230	59,843
Particular e Cooperativo	3,109	3,218	3,327	2,309	59,818
Tipo de Habilitação					
Própria	2,981	3,278	3,407	2,259	59,630
Profissional	3,196	3,141	3,382	2,241	59,799
Vínculo					
Contratado	2,898	3,119	3,407	2,085	57,542
Quadro	3,228	3,181	3,389	2,290	60,440
Idade					
Entre 26 e 30 anos	3,000	3,333	3,500	1,833	58,333
Entre 31 e 40 anos	3,120	3,140	3,420	2,080	58,800
Entre 41 e 50 anos	3,135	3,115	3,354	2,198	59,010
Entre 51 e 60 anos	3,184	3,237	3,368	2,408	60,987
Mais de 60 anos	3,200	3,200	3,480	2,360	61,200
Experiência letiva					
Até 3 anos	2,000	3,125	3,250	1,500	49,375
Entre 4 e 6 anos	3,500	3,250	3,500	2,250	62,500
Entre 7 e 25 anos	3,149	3,112	3,351	2,239	59,254
Entre 26 e 35 anos	3,278	3,264	3,444	2,361	61,736
Mais de 35 anos	3,000	3,185	3,407	2,185	58,889
Nível de Ensino					
1°Ciclo	2,781	2,984	3,219	1,969	54,766
2°Ciclo	3,100	3,000	3,400	2,100	58,000
3°Ciclo e Secundário	3,287	3,260	3,413	2,387	61,733
Área disciplinar					
MCE*	3,311	3,328	3,344	2,377	61,803
CSH**	3,182	3,227	3,386	2,500	61,477
LING***	3,139	3,063	3,450	2,203	59,304
EXP****	3,214	3,167	3,476	2,214	60,357
Formação Contínua em Avaliação					
Sim	3,168	3,141	3,515	2,297	59,730
Não	3,103	3,250	3,341	2,103	59,853

*Matemática e Ciências Experimentais **Ciências Sociais e Humanas *** Línguas ****Expressões

Na Parte 4 do QALA, a totalidade da amostra obteve uma média de acertos de cerca de 59,8%. Este valor está muito próximo do alcançado na Parte 3 (Conhecimentos em avaliação) que, recorde-se, foi de cerca de 58,6%. Ao contrário do que aconteceu na Parte 3, aqui o domínio que apresentou melhores resultados foi o 'Conhecimentos sobre a utilização de instrumentos de avaliação diversificados'(P4D3) com uma média de 3,387, num máximo de 5, correspondente a uma média de acertos de 67,74%. Já o domínio que apresenta os valores médios mais baixos é o 'Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação' (P4D4), algo que já se havia verificado tanto na Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) como na Parte 3 do QALA. Analisando os resultados obtidos na Parte 4 do QALA, e considerando as variáveis de contexto, temos que:

- Os professores do sexo masculino apresentam uma média global superior às professoras do sexo feminino, com uma diferença perto dos 1,8 pontos percentuais. Em 3 dos domínios, os professores do sexo masculino obtiveram uma média superior, sendo que as professoras do sexo feminino conseguiram uma média superior no domínio 'Conhecimentos sobre a utilização de instrumentos de avaliação diversificados';

- Mais uma vez, os professores que lecionam no ensino público alcançaram um resultado melhor que os professores do ensino particular e cooperativo. No entanto, a diferença existente é muito pouco expressiva, já que foi de apenas 0,025 pontos percentuais. Recorde-se que na Parte 3 do QALA a diferença foi muito superior, na ordem dos 4 pontos percentuais;

- Os professores com habilitação profissional apresentam um resultado global mais alto que os professores com habilitação própria. No entanto, e tal como no ponto anterior, a diferença é residual (0,169 pontos percentuais);

- Relativamente ao tipo de vínculo, verifica-se que os professores do quadro apresentam resultados globais superiores, em quase 3 pontos percentuais, aos professores contratados. Esta diferença, um pouco mais expressiva, já havia sido identificada tanto na Parte 2 (cerca de 4 pontos percentuais) como na Parte 3 (cerca de 3,7 pontos percentuais);

- Tal como na Parte 3, são os professores dos escalões etários superiores aqueles que apresentam médias globais mais elevadas. Aliás, verifica-se mesmo um crescimento da média global conforme aumenta o escalão etário. Parece assim existir uma possível relação entre as duas variáveis, algo que será verificado no próximo subcapítulo;

- Os professores com menos experiência letiva apresentam médias globais muito abaixo dos restantes, com uma média de acertos de apenas 49,4%. Mais uma vez, e à semelhança da Parte 3, os professores com experiência letiva entre os 4 e os 6 anos e entre os 26 e 35 anos são os que apresentam as médias globais mais altas. A diferença entre os valor médio mais baixo e o valor médio mais elevado é superior a 13 pontos percentuais;

- Os professores do 3°Ciclo e Secundário são os que alcançam médias superiores nos 4 domínios e, consequentemente, na média global. Por outro lado, os professores do 1°Ciclo são os que têm as médias mais baixas, tanto nos 4 domínios, como nos resultados globais. A diferença entre estes dois grupos de professores foi de quase 7 pontos percentuais;

- Relativamente à área disciplinar dos professores respondentes, verificamos que, neste caso, foram os professores de Matemática e Ciências Experimentais os que obtiveram uma média global superior, seguindo-se os professores de Ciências Sociais e Humanas, os professores de Expressões e, por fim, os

professores de Línguas. Nesta parte do QALA, os professores de Línguas apresentam as médias mais baixas em 3 dos 4 domínios, enquanto que os professores de Matemáticas e Ciências Experimentais apresentam médias mais altas em 2 domínios, à semelhança dos professores de Ciências Sociais e Humanas;

- Por último, e ao contrário do verificado nas Partes 2 e 3, os professores que não frequentaram ações de formação em avaliação apresentaram uma média ligeiramente superior aos professores que frequentaram tais formações. No entanto, há que ressalvar que a diferença é muito pouco expressiva, sendo de apenas 0,123 pontos percentuais.

4.3 Análise Inferencial

Neste subcapítulo iremos apresentar os resultados com base em algumas estatísticas indutivas. Salientamos, uma vez mais, que devido ao método de amostragem utilizado (amostragem por conveniência), os resultados obtidos neste subcapítulo devem ser interpretados com prudência. Serão testadas várias hipóteses de relação entre as diferentes variáveis consideradas na presente investigação. No entanto, antes de avançar com tal análise é imperativo decidir os conjuntos de técnicas a adotar, ou seja, se se enquadram nas técnicas paramétricas ou não-paramétricas.

Para a utilização de técnicas paramétricas devem ser satisfeitos alguns pressupostos, como sejam o caráter intervalar das medidas, a distribuição normal e a homogeneidade de variância (Hill & Hill, 2002). Quando tais pressupostos não são verificados, devem ser aplicadas um conjuntos de técnicas estatísticas classificadas

de não-paramétricas (Hill & Hill, 2002; Pestana & Gageiro, 2003).

Relativamente à natureza dos dados, e conforme verificado na página 103, embora de origem ordinal, o tipo de escala utilizado apresenta propriedades próximas de uma escala de intervalos, pelo que assumiremos que o caráter intervalar das medidas foi cumprido.

Para aferir a normalidade da distribuição dos dados, foi aplicado o teste de *Kolmogorov-Smirnov*[53] a cada uma das partes do QALA, bem como aos respetivos domínios. Para este teste, considera-se uma distribuição normal quando o valor de *p* não é significativo, ou seja, superior a 0.05 (Hill & Hill, 2002).

Conforme podemos constatar na tabela 35, só poderemos considerar uma distribuição normal na globalidade da Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) e na Parte 3 (Conhecimentos em avaliação) do QALA. No entanto, a distribuição normal não se verifica nem na Parte 4 (Cenários em contexto de avaliação), nem nos vários domínios considerados em cada uma das partes, visto que, em cada uma delas, o valor de *p* é inferior a 0.05.

[53]A opção de utilizar o teste de Kolmogorov-Smirnov é adequada, na medida que a nossa amostra é superior a 50. Caso o amostra fosse inferior a 50, recomendar-se-ia a utilização do teste *Shapiro-Wilk*.

Tabela 35: Teste de Normalidade (Kolmogorov-Smirnov)

	Z	*p*
PARTE 2	0.61	.854
P2D1	2.28	.000
P2D2	1.56	.010
P2D3	2.45	.000
P2D4	1.53	.012
PARTE 3	1.20	.096
P3D1	3.19	.000
P3D2	2.90	.000
P3D3	2.12	.000
P3D4	1.88	.001
PARTE 4	2.22	.000
P4D1	3.34	.000
P4D2	3.76	.000
P4D3	3.66	.000
P4D4	3.57	.000

Para a verificação da homogeneidade de variâncias foi utilizado o *Teste de Levene*. Neste teste, considera-se que a homogeneidade de variâncias existe quando o valor de *p* não é significativo, ou seja, superior a 0.05 (Hill & Hill, 2002).

Na tabela 36 constam os resultados obtidos pelo *Teste de Levene* às 3 partes do QALA, tendo em consideração duas variáveis de contexto, nomeadamente o sexo e a formação contínua em avaliação. Conforme se pode constatar, a homogeneidade de variâncias não é garantida, já que se verificam casos em que *p* é significativo.

Tabela 36: Teste de homogeneidade de variâncias de Levene para as variáveis Sexo e Formação Contínua em Avaliação

	Variável "Sexo"			Variável "Formação Contínua"		
	F	*df*	*p*	F	*df*	*p*
PARTE 2	1.18	1	.278	5.42	1	**.021**
P2D1	1.24	1	.266	3.15	1	.077
P2D2	.98	1	.323	6.89	1	**.009**
P2D3	1.93	1	.166	3.55	1	.061
P2D4	.12	1	.733	.35	1	.557
PARTE 3	2.86	1	.092	4.64	1	**.032**
P3D1	.56	1	.453	.44	1	.508
P3D2	1.48	1	.225	.89	1	.346
P3D3	6.64	1	**.011**	.16	1	.692
P3D4	.39	1	.533	2.02	1	.156
PARTE 4	.87	1	.352	.24	1	.622
P4D1	.89	1	.347	.20	1	.658
P4D2	8.25	1	**.004**	.29	1	.590
P4D3	2.02	1	.156	.22	1	.640
P4D4	.73	1	.394	5.20	1	**.023**

Visto não serem garantidos os pressupostos da normalidade e da homogeneidade de variâncias, descartaremos a possibilidade de utilizar técnicas paramétricas, pelo que se opta pela utilização de técnicas não-paramétricas para a análise indutiva dos dados. As técnicas não-paramétricas que utilizaremos na análise inferencial serão o Teste de Mann-Whitney, o Teste de Kruskal-Wallis e a Correlação de Spearman.

O teste de Mann-Whitney é a alternativa não-paramétrica ao teste *t* de *student* para duas amostras independentes. Segundo Pestana e Gageiro (2003), este teste permite verificar a igualdade de comportamentos de dois grupos de casos, ou a existência de diferenças entre duas condições experimentais.

A partir da aplicação do teste de Mann-Whitney, estabelecem-se duas hipóteses. A hipótese nula (H_0), quando as duas populações têm distribuições idênticas, ou seja, as diferenças entre as populações não são estatisticamente significativas, e a hipótese

alternativa (H_1), quando as duas populações não apresentam distribuições idênticas, ou seja, as diferenças entre as populações são estatisticamente significativas.

A hipótese nula não deverá ser rejeitada quando o valor de significância (*p*) é superior a 0.05. Quando o valor de *p* é inferior, ou igual, a 0.05 deveremos rejeitar a hipótese nula e assumir a hipótese alternativa.

Já o teste de Kruskal-Wallis é a alternativa não-paramétrica ao *One-way Anova* para *k* amostras. O teste de Kruskal-Wallis é utilizado para comparar as distribuições de três ou mais grupos em amostras independentes (Marôco, 2018).

A partir da aplicação do teste de Kruskal-Wallis, estabelecem-se igualmente duas hipóteses. A hipótese nula, quando as distribuições das variáveis consideradas são idênticas para as *k* populações, ou seja, as diferenças entre as *k* populações não são estatisticamente significativas, e a hipótese alternativa, quando pelo menos uma das *k* populações apresenta distribuições diferentes às demais, ou seja, as diferenças entre as *k* amostras são estatisticamente significativas.

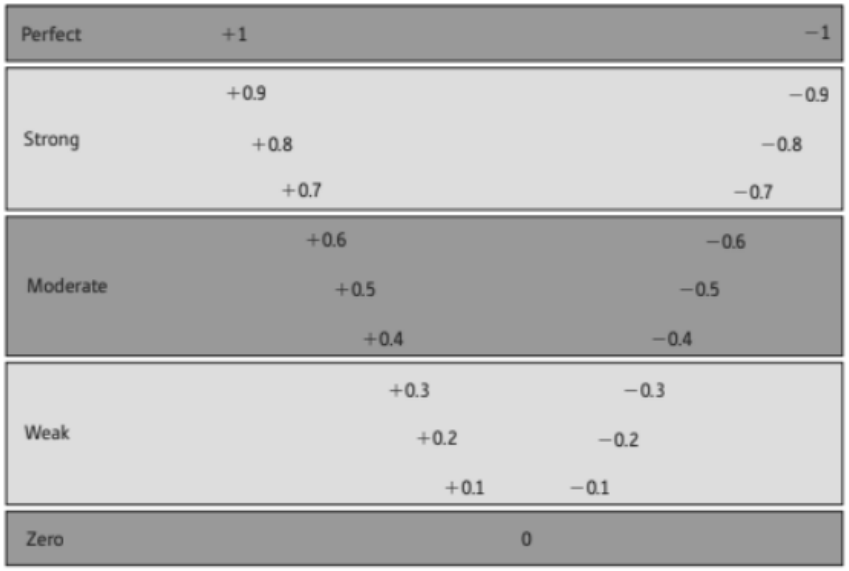
À semelhança do teste de Mann-Whitney, a hipótese nula não deverá ser rejeitada quando o valor de (*p*) é superior a 0.05. Quando o valor de *p* é inferior, ou igual, a 0.05 deveremos rejeitar a hipótese nula e assumir a hipótese alternativa.

O coeficiente de correlação de Spearman é uma estatística não-paramétrica que mede a intensidade das relações entre variáveis (Pestana & Gageiro, 2003). O coeficiente r_s de Spearman varia entre -1 e 1, sendo a correlação entre variáveis tanto maior quanto maior for a aproximação a estes dois extremos. Valores de r_s positivos indicam que ambas as variáveis variam no mesmo sentido. Já os valores negativos indicam que as variáveis variam em sentido contrário, ou seja, categorias mais altas numa variável estão associadas a categorias mais baixas na outra variável

(Pestana & Gageiro, 2003). Já um valor de zero indica a inexistência de correlação entre as duas variáveis em análise, ou seja, não existe associação linear entre elas.

A literatura sugere diferentes formas de classificar o grau de correlação entre duas variáveis por intermédio do r_s de Spearman. Desta forma, iremos utilizar a classificação proposta por Dancey e Reidy (2017), visto tratar-se de uma classificação mais fina utilizando 5 classes, como nos retrata a figura 18.

Figura 18: Classificação do grau de correlação entre duas variáveis
(Fonte: Dancey e Reidy, 2017)



Outro parâmetro resultante da aplicação do coeficiente de correlação de Spearman é a significância que nos indica, para o caso do Universo, se essa correlação é igual ou diferente de zero. Assim, estabelecem-se as seguintes hipóteses:

- H_0: O coeficiente de correlação no Universo é igual a zero ($p > .05$);
- H_1: O coeficiente de correlação no Universo é, provavelmente, diferente de zero ($p < .05$).

4.3.1 Relação entre os resultados obtidos no QALA com a variável Sexo

Para avaliar a relação entre os resultados obtidos no QALA e o sexo dos respondentes, recorremos ao teste U de Mann-Whitney para amostras independentes. Desta forma, colocaram-se as seguintes hipóteses:

- Hipótese nula (H_0): As diferenças nas distribuições dos valores obtidos pelos professores do sexo feminino e masculino não são estatisticamente significativas;
- Hipótese alternativa (H_1): As diferenças nas distribuições dos valores obtidos pelos professores do sexo feminino e masculino são estatisticamente significativas;

O teste U realizado à globalidade da Parte 2 (Tabela 37) revela-nos que devemos rejeitar a hipótese nula, pois o valor de *p* é inferior a 0.05. Desta forma, consideramos a existência de diferenças estatisticamente significativas nas distribuições dos valores alcançados pelos respondentes do sexo feminino e masculino. Relembramos que, aquando da análise descritiva dos resultados do QALA, os professores do sexo feminino apresentaram médias globais mais elevadas que os professores do sexo masculino (ver página 143).

Tabela 37: Testes de Mann-Whitney realizados à Parte 2 do QALA para a variável Sexo

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 2	4228.50	5606.50	-2.12	**.034**
P2D1	4442.00	5820.00	-1.68	.092
P2D2	4598.50	5976.50	-1.34	.179
P2D3	4212.50	5590.50	-2.19	**.029**
P2D4	4430.50	5808.50	-1.70	.089

Ao analisarmos cada um dos domínios da Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA, verificamos que no caso do domínio ’Conhecimentos sobre utilização de instrumentos de avaliação diversificados’ existem também diferenças significativas entre ambos os sexos (*p*=.029), pelo que rejeitamos a hipótese nula. Nos restantes domínios, não rejeitamos a hipótese nula, pois embora os valores de *p* sejam relativamente baixos, encontram-se acima do limiar de significância de 0.05.

O teste U, realizado à globalidade da Parte 3 (Conhecimentos em avaliação) do QALA (tabela 38), revela-nos que não existem diferenças estatisticamente significativas entre os dois sexos (p=.936). Resultado semelhante foi verificado nos domínios ’Conhecimentos sobre os objetivos e funções da avaliação’ (*p*=.917) e ’Conhecimento sobre utilização de instrumentos de avaliação diversificados’ (*p*=.330). No entanto, tais diferenças parecem verificar-se nos domínios ’Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar’ (*p*=.005) e ’Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação’ (*p*=.005).

Tabela 38: Testes de Mann-Whitney realizados à Parte 3 do QALA para a variável Sexo

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 3	5188.50	6566.50	-0.08	.936
P3D1	5178.50	25479.50	-0.10	.917
P3D2	3938.50	5316.50	-2.78	**.005**
P3D3	4773.50	6151.50	-0.97	.330
P3D4	3930.00	24231.00	-2.79	**.005**

Analisando os resultados alcançados pela aplicação do teste U à Parte 4 (Cenários em contexto de avaliação) do QALA (tabela 39), concluimos que não devemos rejeitar a hipótese nula, nem na sua globalidade, nem nos 4 domínios que o

compõem já que, para todos os casos, o valor de p é superior a .05. Consideramos assim que não existem diferenças estatisticamente significativas nas distribuições dos valores de ambos os sexos na Parte 4 do QALA.

Tabela 39: Testes de Mann-Whitney realizados à Parte 4 do QALA para a variável Sexo

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 4	4989.00	25290.00	-0.51	.611
P4D1	5192.00	6570.00	-0.08	.940
P4D2	4453.00	24754.00	-1.76	.079
P4D3	4884.00	6262.00	-0.76	.446
P4D4	4925.00	25226.00	-0.68	.499

4.3.2 Relação entre os resultados obtidos no QALA com a variável Tipo de Habilitação

Para aferir a relação entre os resultados obtidos no QALA e a variável Tipo de Habilitação recorremos, à semelhança do caso anterior, ao teste U de Mann-Whitney. Foram estabelecidas as seguintes hipóteses:

- H_0: As diferenças nas distribuições dos valores obtidos pelos professores com habilitação própria e profissional não são estatisticamente significativas;
- H_1: As diferenças nas distribuições dos valores obtidos pelos professores com habilitação própria e profissional são estatisticamente significativas.

Ao analisar os resultados do teste U à Parte 2 (tabela 40), Parte 3 (tabela 41) e Parte 4 (tabela 42), bem como aos 4 domínios que compõem cada uma delas, concluimos que não devemos rejeitar a hipótese nula, pois nenhum dos testes apresenta um valor de *p* inferior a .05. Desta forma, parece não existir especial relação entre o tipo de habilitação dos professores e os resultados obtidos pela

aplicação do QALA. Tal resultado era expectável dadas as pequenas diferenças verificadas aquando da análise descritiva.

Tabela 40: Testes de Mann-Whitney realizados à Parte 2 do QALA para a variável Tipo de Habilitação

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 2	4609.00	24509.00	-1.60	.109
P2D1	4714.00	24614.00	-1.40	.163
P2D2	4613.00	24513.00	-1.60	.109
P2D3	4964.50	24864.50	-0.87	.385
P2D4	4991.00	24891.00	-0.80	.421

Tabela 41: Testes de Mann-Whitney realizados à Parte 3 do QALA para a variável Tipo de Habilitação

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 3	5215.50	6700.50	-0.33	.741
P3D1	5002.00	6487.00	-0.81	.421
P3D2	5044.00	24944.00	-0.70	.483
P3D3	5280.50	6765.50	-0.20	.844
P3D4	4945.00	6430.00	-0.91	.363

Tabela 42: Testes de Mann-Whitney realizados à Parte 4 do QALA para a variável Tipo de Habilitação

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 4	5264.50	6749.50	-0.23	.818
P4D1	4656.00	6141.00	-1.58	.115
P4D2	4911.50	24811.50	-1.03	.301
P4D3	5307.00	25207.00	-0.15	.885
P4D4	5225.00	25125.00	-0.33	.743

4.3.3 Relação entre os resultados obtidos no QALA com a variável Subsistema de Ensino

Na recolha de informações gerais do QALA, mais concretamente no ponto em que se solicita ao respondente para selecionar o subsistema de ensino em que leciona,

eram apresentadas 3 opções de resposta: 'Público', 'Particular e Cooperativo' e 'Ambos'. Conforme podemos constatar na tabela 4, o número de professores que responderam 'Ambos' foi baixo (n=7), pelo que decidimos exclui-los desta análise.

Assim, havendo apenas dois grupos independentes, optou-se, uma vez mais, pela utilização do teste U. Estabeleceram-se, assim, as seguintes hipóteses:

- H_0: As diferenças nas distribuições dos valores obtidos pelos professores que lecionam no ensino público e pelos professores que lecionam no ensino particular e cooperativo não são estatisticamente significativas;

- H_1: As diferenças nas distribuições dos valores obtidos pelos professores que lecionam no ensino público e pelos professores que lecionam no ensino particular e cooperativo são estatisticamente significativas;

Os resultados obtidos pela aplicação do teste U à Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA (tabela 43), mostra-nos que devemos rejeitar a hipótese nula para a globalidade da Parte 2 (p=.050), bem como para o domínio 'Conhecimentos sobre a utilização de instrumentos de avaliação diversificados' (p=.001). No entanto, para os restantes domínios, não devemos rejeitar a hipótese nula , ou seja, não há evidências de que existam diferenças significativas entre os resultados alcançados pelos professores do ensino público e pelos professores do ensino particular e cooperativo.

Tabela 43: Testes de Mann-Whitney realizados à Parte 2 do QALA para a variável Subsistema de Ensino

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 2	4340.00	5880.00	-1.96	**.050**
P2D1	4628.00	6168.50	-1.36	.175
P2D2	4744.50	6284.00	-1.10	.271
P2D3	3785.50	5325.50	-3.20	**.001**
P2D4	4730.50	6270.50	-1.13	.260

Já na Parte 3 (Conhecimentos em avaliação) do QALA(tabela 44), verificamos que deveremos rejeitar a hipótese nula para a globalidade da parte 3 (p=.033) e para um dos domínios, mais concretamente o 'Conhecimento sobre o currículo e sobre aquilo que é importante aprender e avaliar' (p=.002). Os restantes domínios da parte 3, apresentam valores de p superiores ao limiar de significância, pelo que não rejeitamos a hipótese nula.

Tabela 44: Testes de Mann-Whitney realizados à Parte 3 do QALA para a variável Subsistema de Ensino

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 3	4265.50	5805.50	-2.13	**.033**
P3D1	4910.00	6450.00	-0.76	.446
P3D2	3811.00	5351.00	-3.15	**.002**
P3D3	4456.00	5996.00	-1.73	.083
P3D4	5103.50	6643.50	-0.32	.745

Parece não existir relação entre os resultados alcançados na parte 4 (Cenários em contexto de avaliação) do QALA e a variável subsistema de ensino (tabela 45). Tanto na sua globalidade, como nos vários domínios que o compõem, não se verificam valores de p inferiores a .05, pelo que não rejeitamos a hipótese nula. Este resultado era expectável dadas as pequeníssimas diferenças verificadas aquando da análise descritiva.

Tabela 45: Testes de Mann-Whitney realizados à Parte 4 do QALA para a variável Subsistema de Ensino

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 4	5165.50	6705.50	-0.19	.850
P4D1	5054.50	6594.50	-0.45	.656
P4D2	5050.50	23386.50	-0.46	.642
P4D3	5079.50	6619.50	-0.39	.697
P4D4	5060.00	23396.00	-0.44	.633

4.3.4 Relação entre os resultados obtidos no QALA com a variável Nível de Ensino

Para a variável 'Nível de ensino', optámos por realizar dois testes estatísticos, o teste U de Mann-Whitney para duas amostras independentes e o teste H de Krukal-Wallis para *k* amostras independentes. A decisão de realizar os dois testes deve-se ao facto de vários professores terem habilitação para mais de um nível de ensino o que, à partida, colocaria em causa a independência das amostras. Assim, foi necessário codificar os dados da seguinte forma:

- Foi criada uma coluna para cada nível de ensino (1°ciclo[C1], 2°ciclo [C2] e 3°ciclo e secundário[C3S]). A cada respondente, foi atribuído o valor 1 ('Sim') para o nível (ou níveis) de ensino para o qual possui habilitação e o valor 0 ('Não') para o nível (ou níveis) de ensino para o qual o professor não detinha habilitação (ver exemplo na tabela 46);
- Foi criada uma coluna designada 'Ciclo' (conforme tabela 46) à qual se atribuiu um dos seguintes códigos para cada respondente:

  0 - Possui habilitação para mais de um nível de ensino;

  1 - Possui habilitação apenas para 1°Ciclo;

  2 - Possui habilitação apenas para 2°Ciclo;

3 - Possui habilitação apenas para 3°Ciclo e Secundário;

Tabela 46: Exemplo de codificação dos dados para a variável Nível de Ensino

Exemplo	GR1*	GR2*	C1	C2	C3S	Ciclo
A	110	210	1	1	0	0
B	230	-	0	1	0	2
C	320	330	0	0	1	3
D	240	600	0	1	1	0
E	110	-	1	0	0	1

*Grupo de Recrutamento

Tal codificação garantiu as amostras independentes para a realização dos testes acima referidos. Assim, para o teste U foram formuladas as seguintes hipóteses:

- H_0: As diferenças nas distribuições dos valores obtidos pelos professores que têm habilitação para um determinado nível de ensino e os professores que não têm habilitação para esse mesmo nível, não são estatisticamente significativas;

- H_1: As diferenças nas distribuições dos valores obtidos pelos professores que têm habilitação para um determinado nível de ensino e os professores que não têm habilitação para esse mesmo nível, são estatisticamente significativas;

Já para o teste de Kruskal-Wallis, foram ignorados os professores com habilitação a mais do que um nível de ensino. Este aspeto permitiu analisar as diferenças nos resultados dos professores com habilitação para os diferentes níveis de ensino. Neste caso, estabeleceram-se as seguintes hipóteses:

- H_0: As diferenças nas distribuições dos valores obtidos pelos professores dos diferentes níveis de ensino, não são estatisticamente significativas;

- H_1: As diferenças nas distribuições dos valores obtidos pelos professores dos diferentes níveis de ensino são estatisticamente significativas;

Dado este enquadramento inicial, analisemos os dados obtidos pela aplicação do teste U à parte 2 do QALA para a variável em questão. Na tabela 47 constatamos que não existem diferenças estatisticamente significativas entre os professores que têm habilitação para o 1°ciclo e para os professores que não têm essa habilitação. Esta conclusão aplica-se tanto para a globalidade da Parte 2 como para os domínios que o compõem. Conclusão semelhante poder ser tirada em relação aos professores que têm habilitação para lecionar no 2°ciclo e aqueles que não têm essa habilitação.

Já em relação ao professores que têm habilitação para lecionar no 3°ciclo e secundário, as conclusões são ligeiramente diferentes. Isto acontece pois dois dos domínios que compõem a Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA, nomeadamente os domínios 'Conhecimentos sobre os objetivos e funções da avaliação' e 'Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar', apresentam valores de *p* inferiores a .05, pelo que, nestes dois casos, rejeitamos hipótese nula. Assim, para os dois domínios identificados, parecem existir diferenças estatisticamente significativas consoante os professores possuam, ou não, habilitação para lecionar no 3°ciclo do ensino básico e ensino secundário.

Tabela 47: Testes de Mann-Whitney realizados à Parte 2 do QALA para a variável Nível de Ensino

Entre professores com e sem habilitação para o **1°Ciclo**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE2	5652.00	7732.00	-0.78	.434
P2D1	5255.00	7335.00	-1.58	.114
P2D2	5480.00	7560.00	-1.13	.259
P2D3	5883.50	7963.50	-0.33	.741
P2D4	6013.00	23968.00	-0.07	.945
Entre professores com e sem habilitação para o **2°Ciclo**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 2	6315.00	9555.00	-1.12	.263
P2D1	6202.00	9442.00	-1.34	.180
P2D2	6183.50	9423.50	-1.37	.171
P2D3	6712.50	9952.50	-0.39	.697
P2D4	6718.50	9958.50	-0.37	.709
Entre professores com e sem habilitação para o **3°Ciclo e Secundário**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 2	6864.50	12220.50	-1.51	.132
P2D1	6610.00	11966.00	-1.97	**.049**
P2D2	6607.50	11963.50	-1.97	**.049**
P2D3	7336.50	12692.50	-0.69	.491
P2D4	7502.50	12858.50	-0.39	.696

Ao analisarmos os resultados ao teste H de Kruskal-Wallis, realizado à parte 2 do QALA (tabela 48) e considerando apenas os professores que têm habilitação para 1 dos níveis de ensino, parece não existir diferenças significativas entre professores dos três níveis de ensino considerados. Isto aplica-se à globalidade da Parte 2 como para os respetivos domínios.

Tabela 48: Testes H de Kruskal-Wallis realizados à Parte 2 do QALA para a variável Nível de Ensino

	Nível de Ensino	N	Mean Rank	χ^2 (df)	p
Parte 2	Apenas 1°Ciclo	34	106.40	1.78 (2)	.410
	Apenas 2°Ciclo	42	94.95		
	Apenas 3°Ciclo e Secundário	135	109.34		
P2D1	Apenas 1°Ciclo	34	99.44	2.43 (2)	.297
	Apenas 2°Ciclo	42	95.88		
	Apenas 3°Ciclo e Secundário	135	110.80		
P2D2	Apenas 1°Ciclo	34	100.16	3.31 (2)	.191
	Apenas 2°Ciclo	42	93.12		
	Apenas 3°Ciclo e Secundário	135	111.48		
P2D3	Apenas 1°Ciclo	34	113.01	0.69 (2)	.706
	Apenas 2°Ciclo	42	101.60		
	Apenas 3°Ciclo e Secundário	135	105.60		
P2D4	Apenas 1°Ciclo	34	113.63	1.21 (2)	.547
	Apenas 2°Ciclo	42	98.36		
	Apenas 3°Ciclo e Secundário	135	106.46		

Relativamente à parte 3 (Conhecimentos em avaliação) do QALA, o teste U de Mann-Whitney (tabela 49) mostra-nos que não existem diferenças significativas nas distribuições dos valores, entre quem tem habilitação para um determinado nível de ensino e de quem não o tem, seja na globalidade da parte 3, seja nos 4 domínios que o constituem.

Tabela 49: Testes de Mann-Whitney realizados à Parte 3 do QALA para a variável Nível de Ensino

Entre professores com e sem habilitação para o **1°Ciclo**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE3	5339.50	7419.50	-1.40	.161
P3D1	5652.50	23607.50	-0.81	.418
P3D2	5331.00	7411.00	-1.44	.150
P3D3	5239.00	7319.00	-1.62	.105
P3D4	5383.00	7463.00	-1.33	.183
Entre professores com e sem habilitação para o **2°Ciclo**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 3	6260.50	9500.50	-1.22	.222
P3D1	6172.50	9412.50	-1.43	.153
P3D2	6412.50	9652.50	-0.95	.340
P3D3	6682.00	9922.00	-0.45	.656
P3D4	6492.50	9732.50	-0.80	.423
Entre professores com e sem habilitação para o **3°Ciclo e Secundário**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 3	7023.50	12379.50	-1.23	.219
P3D1	7239.50	12595.50	-0.88	.379
P3D2	7239.50	12595.50	-0.86	.388
P3D3	7618.00	12974.00	-0.19	.850
P3D4	6879.50	12235.50	-1.50	.134

O teste de Kruskal-Wallis, também aplicado à parte 3 (tabela 50), dá-nos uma perspetiva semelhante à anterior. Desta forma, podemos concluir que não existe especial relação entre os resultados obtidos na parte 3 do QALA e o nível de ensino para o qual os professores estão habilitados.

Tabela 50: Testes de Kruskal-Wallis realizados à Parte 3 do QALA para a variável Nível de Ensino

	Nível de Ensino	N	Mean Rank	χ^2 (df)	p
Parte 3	Apenas 1°Ciclo	34	90.72	2.71 (2)	.257
	Apenas 2°Ciclo	42	105.61		
	Apenas 3°Ciclo e Secundário	135	109.97		
P3D1	Apenas 1°Ciclo	34	106.43	3.16 (2)	.206
	Apenas 2°Ciclo	42	91.80		
	Apenas 3°Ciclo e Secundário	135	110.31		
P3D2	Apenas 1°Ciclo	34	94.26	1.56 (2)	.458
	Apenas 2°Ciclo	42	109.18		
	Apenas 3°Ciclo e Secundário	135	107.97		
P3D3	Apenas 1°Ciclo	34	93.59	2.95 (2)	.229
	Apenas 2°Ciclo	42	117.39		
	Apenas 3°Ciclo e Secundário	135	105.58		
P3D4	Apenas 1°Ciclo	34	85.68	5.26 (2)	.072
	Apenas 2°Ciclo	42	103.40		
	Apenas 3°Ciclo e Secundário	135	111.93		

Por fim, na Parte 4 (Cenários em contexto de avaliação) do QALA, o teste de Mann-Whitney apresenta-nos uma perspetiva diferente das anteriores. Conforme podemos constatar na Tabela 51, o nível de ensino dos professores parece ter especial relação com os resultados obtidos. Os valores de *p* indicam-nos diferenças significativas entre quem tem habilitação para o 1°ciclo e quem não tem, na globalidade da parte 4 (*p*=.002) e em 3 dos 4 domínios, sendo a execeção o domínio ’Conhecimentos sobre utilização de instrumentos de avaliação diversificados’ (*p*=.226).

Relativamente aos resultados entre quem tem habilitação para o 2°ciclo e quem não tem, verificamos que o valor de *p* é significativo no domínio ’Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar’ (*p*=.036). Nos restantes domínios, e na globalidade da parte 4, não rejeitamos a hipótese nula.

Já as diferenças nas médias entre professores com habilitação para o 3°ciclo do ensino básico e ensino secundário e sem, são estatisticamente significativas para a globalidade da parte 4 (*p*=.012) e para os domínios ’Conhecimentos sobre objetivos e

funções da avaliação' (*p*=.034) e 'Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação' (*p*=.009). De salientar que o domínio 'Conhecimentos sobre o curriculo e sobre aquilo que é importante aprender e avaliar' ficou muito próximo do limiar de significância estatística, já que o valor de *p*, resultante do teste de Mann-Whitney, foi de .051.

Tabela 51: Testes de Mann-Whitney realizados à Parte 4 do QALA para a variável Nível de Ensino

Entre professores com e sem habilitação para o **1°Ciclo**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE4	4478.50	6558.50	-3.13	**.002**
P4D1	4776.50	6856.50	-2.63	**.008**
P4D2	4898.50	6978.50	-2.43	**.015**
P4D3	5463.50	7543.50	-1.21	.226
P4D4	4849.00	6929.00	-2.50	**.012**
Entre professores com e sem habilitação para o **2°Ciclo**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 4	6176.50	9416.50	-1.39	.166
P4D1	6739.00	9979.00	-0.35	.726
P4D2	5856.00	9096.00	-2.10	**.036**
P4D3	6796.00	21847.00	-0.24	.810
P4D4	6087.50	9327.50	-1.62	.105
Entre professores com e sem habilitação para o **3°Ciclo e Secundário**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 4	6299.50	11655.50	-2.52	**.012**
P4D1	6568.50	11924.50	-2.12	**.034**
P4D2	6683.50	12039.50	-1.95	.051
P4D3	7656.50	13012.50	-0.13	.900
P4D4	6314.50	11670.50	-2.60	**.009**

O teste de Kruskal-Wallis parece confirmar uma especial relação entre o nível de ensino, para o qual os professores estão habilitados, e os resultados alcançados na parte 4.

Ao analisarmos a tabela 52 concluimos que devemos rejeitar a hipótese nula na globalidade da parte 4 (*p*=.007), bem como nos domínios 'Conhecimentos sobre os objetivos e funções da avaliação' (*p*=.010) e 'Conhecimentos sobre interpretação e

utilização da informação recolhida no processo de avaliação' (*p*=.023). Estes resultados mostram-nos que as diferenças nas distribuições dos valores alcançados pelos professores de diferentes níveis de ensino são estatisticamente relevantes. Recordamos que, aquando da análise à tabela 34, constatámos, por exemplo, que a diferença nas médias globais, entre professores do 1°ciclo e professores do 3°ciclo e secundário, foi próxima de 7 pontos percentuais.

Tabela 52: Testes de Kruskal-Wallis realizados à Parte 4 do QALA para a variável Nível de Ensino

	Nível de Ensino	N	Mean Rank	χ^2 (df)	p
Parte 4	Apenas 1°Ciclo	34	78.09	9.85 (2)	**.007**
	Apenas 2°Ciclo	42	102.33		
	Apenas 3°Ciclo e Secundário	135	114.17		
P4D1	Apenas 1°Ciclo	34	79.44	9.17 (2)	**.010**
	Apenas 2°Ciclo	42	104.27		
	Apenas 3°Ciclo e Secundário	135	113.23		
P4D2	Apenas 1°Ciclo	34	93.76	2.37 (2)	.305
	Apenas 2°Ciclo	42	102.80		
	Apenas 3°Ciclo e Secundário	135	110.08		
P4D3	Apenas 1°Ciclo	34	94.53	2.47 (2)	.291
	Apenas 2°Ciclo	42	115.60		
	Apenas 3°Ciclo e Secundário	135	105.90		
P4D4	Apenas 1°Ciclo	34	85.16	7.51 (2)	**.023**
	Apenas 2°Ciclo	42	98.21		
	Apenas 3°Ciclo e Secundário	135	113.67		

4.3.5 Relação entre os resultados obtidos no QALA com a variável Área Disciplinar

À semelhança do ponto anterior, a análise da relação entre as várias partes do QALA e a variável 'Área disciplinar', foi realizada através dos testes estatísticos U de Mann-Whitney e H de Kruskal-Wallis. Mais uma vez, vários professores tinham habilitação para mais do que uma área disciplinar, o que colocaria em causa a independência das amostras. Desta forma, os dados foram codificados da seguinte

forma:

- Foi criada uma coluna para cada área disciplinar (’Matemática e Ciências Experimentais’, ’Ciências Sociais e Humanas’, ’Línguas’ e ’Expressões’). A cada respondente, foi atribuído o valor 1 (’Sim’) para a área (ou áreas) disciplinar para a qual possui habilitação e o valor 0 (’Não’) para a área (ou áreas) disciplinar para o qual a professor não detinha habilitação;
- Foi criada uma coluna designada ’adisciplinar’ à qual se atribuiu um dos seguintes códigos para cada respondente:

  0 - Possui habilitação para o 1°ciclo ou a mais do que uma área disciplinar;

  1 - Possui habilitação apenas para a área disciplinar ”Matemática e Ciências Experimentais”;

  2 - Possui habilitação apenas para a área disciplinar ”Ciências Sociais e Humanas”;

  3 - Possui habilitação apenas para a área disciplinar ”Línguas”;

  4 - Possui habilitação apenas para a área disciplinar ”Expressões”;

Esta codificação dos dados garantiu a independência das amostras permitindo assim a realização dos testes estatísticos. Assim, para o teste U de Mann-Whitney foram definidas as seguintes hipóteses:

- H_0: As diferenças nas distribuições dos valores obtidos pelos professores que têm habilitação para uma determinada área disciplinar e os professores que não têm habilitação para essa mesma área disciplinar, não são estatisticamente significativas;

- H_1: As diferenças nas distribuições dos valores obtidos pelos professores que têm habilitação para uma determinada área disciplinar e os professores que não têm habilitação para essa mesma área disciplinar, são estatisticamente significativas;

Já para o teste de Kruskal-Wallis, foram ignorados os professores com habilitação para o 1°ciclo ou com habilitação a mais de uma área disciplinar. Este aspeto permitiu analisar as diferenças nos resultados dos professores com habilitação para diferentes áreas disciplinares. Neste caso, estabeleceram-se as seguintes hipóteses:

- H_0: As diferenças nas distribuições dos valores obtidos pelos professores das diferentes áreas disciplinares, não são estatisticamente significativas;

- H_1: As diferenças nas distribuições dos valores obtidos pelos professores das diferentes áreas disciplinares, são estatisticamente significativas;

Passemos então à análise dos resultados alcançados pelo teste U de Mann-Whitney à parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA. Ao analisar a tabela 53, constatamos que não devemos rejeitar a hipótese nula, independentemente de os professores possuirem, ou não, habilitação a uma determinada área disciplinar.

Tabela 53: Testes de Mann-Whitney realizados à Parte 2 do QALA para a variável Área Disciplinar

Entre professores com e sem habilitação para **Matemática e Ciências Experimentais**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE2	5482.00	7373.00	-0.75	.452
P2D1	5436.50	7327.50	-0.85	.395
P2D2	5451.50	7342.50	-0.82	.414
P2D3	5291.50	7182.50	-1.15	.250
P2D4	5850.50	24378.50	-0.01	.991
Entre professores com e sem habilitação para **Ciências Sociais e Humanas**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 2	4328.00	26273.00	-0.61	.540
P2D1	4386.00	26331.00	-0.49	.628
P2D2	4069.00	26014.50	-1.21	.228
P2D3	4542.00	26487.00	-0.13	.898
P2D4	4571.00	5561.00	-0.06	.951
Entre professores com e sem habilitação para **Línguas**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 2	6336.00	21561.00	-1.00	.319
P2D1	6251.50	21476.50	-1.16	.245
P2D2	6064.50	21289.50	-1.51	.131
P2D3	6270.50	21495.50	-1.13	.257
P2D4	6500.00	9660.00	-0.69	.487
Entre professores com e sem habilitação para **Expressões**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 2	4106.00	5009.00	-0.75	.453
P2D1	4250.00	5153.00	-0.42	.673
P2D2	3899.00	4802.00	-1.24	.216
P2D3	4144.00	5047.00	-0.67	.501
P2D4	4262.00	26628.00	-0.39	.695

Também o teste de Kruskal-Wallis realizado à parte 2 (Tabela 54) parece indicar que não existe especial relação entre os resultados alcançados e a área disciplinar dos professores.

Tabela 54: Testes de Kruskal-Wallis realizados à Parte 2 do QALA para a variável Área Disciplinar

	Nível de Ensino	N	Mean Rank	χ^2 (df)	p
Parte 2	Apenas MCE (1)	60	100.33	2.68 (3)	.443
	Apenas CSH (2)	38	113.03		
	Apenas Línguas	73	113.03		
	Apenas Expressões	41	97.85		
P2D1	Apenas MCE (1)	60	98.42	3.01 (3)	.391
	Apenas CSH (2)	38	112.09		
	Apenas Línguas	73	114.02		
	Apenas Expressões	41	99.76		
P2D2	Apenas MCE (1)	60	99.15	5.16 (3)	.161
	Apenas CSH (2)	38	116.55		
	Apenas Línguas	73	114.72		
	Apenas Expressões	41	93.30		
P2D3	Apenas MCE (1)	60	97.78	3.61 (3)	.307
	Apenas CSH (2)	38	110.95		
	Apenas Línguas	73	115.38		
	Apenas Expressões	41	99.32		
P2D4	Apenas MCE (1)	60	106.72	0.34 (3)	.953
	Apenas CSH (2)	38	107.80		
	Apenas Línguas	73	103.57		
	Apenas Expressões	41	110.20		

Nota: (1) Matemática e Ciências Experimentais; (2) Ciências Sociais e Humanas.

Na tabela 55, verificamos que o valor de *p* é estatisticamente significativo quando comparamos as diferenças nas distribuições dos valores obtidos pelos professores com habilitação na área das Ciências Sociais e Humanas e os professores que não têm habilitação nessa área, na globalidade da Parte 3 (*p*=.012) e no domínio ’Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação’ (*p*=.010). Também o facto de os professores terem ou não habilitação para a área das ’Línguas’ parece ter relação com os resultados alcançados no domínio ’Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação’ (*p*=.021). Nos casos da ’Matemática e Ciências Experimentais’ e ’Expressões’ não rejeitamos a hipótese nula, nem na parte 3 como um todo, nem em nenhum dos domínios que a constituem.

Tabela 55: Testes de Mann-Whitney realizados à Parte 3 do QALA para a variável Área Disciplinar

Entre professores com e sem habilitação para **Matemática e Ciências Experimentais**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 3	5783.00	7674.00	-0.15	.883
P3D1	5417.00	7308.00	-0.91	.361
P3D2	5267.00	7158.00	-1.20	.229
P3D3	5522.50	7413.50	-0.68	.498
P3D4	5130.50	23658.50	-1.48	.140
Entre professores com e sem habilitação para **Ciências Sociais e Humanas**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 3	3487.00	25432.00	-2.52	**.012**
P3D1	3857.50	25802.50	-1.74	.082
P3D2	3954.50	25899.50	-1.48	.138
P3D3	3968.00	25913.00	-1.45	.148
P3D4	3483.00	25428.00	-2.56	**.010**
Entre professores com e sem habilitação para **Línguas**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 3	6604.50	21829.50	-0.50	.618
P3D1	6166.00	21391.00	-1.36	.175
P3D2	5935.50	21160.50	-1.77	.077
P3D3	6016.50	21241.50	-1.61	.108
P3D4	5646.50	8806.50	-2.31	**.021**
Entre professores com e sem habilitação para **Expressões**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 3	4000.00	4903.00	-1.00	.319
P3D1	3761.00	4664.00	-1.60	.109
P3D2	4043.50	4946.50	-0.91	.363
P3D3	3947.00	4850.00	-1.13	.258
P3D4	4109.00	26475.00	-0.75	.451

Quando analisamos as diferenças dos resultados alcançados, na parte 3 (Conhecimentos em avaliação) do QALA, entre os professores que apenas possuem habilitação a uma das áreas disciplinares (tabela 56), verificamos que são significativas apenas no domínio ’Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação’ (p=.010). Já na globalidade da Parte 3 e nos restantes domínios, verificamos que não existem diferenças significativas nas distribuições dos valores obtidos pelos professores das diferentes áreas disciplinares.

Tabela 56: Testes de Kruskal-Wallis realizados à Parte 3 do QALA para a variável Área Disciplinar

	Nível de Ensino	N	Mean Rank	χ^2 (df)	p
Parte 3	Apenas MCE (1)	60	103.05	5.40 (3)	.145
	Apenas CSH (2)	38	125.45		
	Apenas Línguas	73	106.19		
	Apenas Expressões	41	94.54		
P3D1	Apenas MCE (1)	60	99.84	6.10 (3)	.197
	Apenas CSH (2)	38	119.82		
	Apenas Línguas	73	113.21		
	Apenas Expressões	41	91.95		
P3D2	Apenas MCE (1)	60	96.76	5.16 (3)	.160
	Apenas CSH (2)	38	116.30		
	Apenas Línguas	73	115.04		
	Apenas Expressões	41	96.46		
P3D3	Apenas MCE (1)	60	101.12	3.17 (3)	.366
	Apenas CSH (2)	38	114.12		
	Apenas Línguas	73	112.95		
	Apenas Expressões	41	95.83		
P3D4	Apenas MCE (1)	60	112.61	11.38 (3)	**.010**
	Apenas CSH (2)	38	128.04		
	Apenas Línguas	73	89.29		
	Apenas Expressões	41	108.23		

Nota: (1) Matemática e Ciências Experimentais; (2) Ciências Sociais e Humanas.

Em relação à parte 4 (Cenários em contexto de avaliação), o teste U indica que não devemos rejeitar a hipótese nula, independentemente de os professores possuirem, ou não, habilitação a uma determinada área disciplinar (tabela 57). No entanto, salientamos que, no domínio ’Conhecimentos sobre os objetivos e funções da avaliação’, o valor de *p*, para os professores com habilitação para ’Matemática e Ciências Experimentais’ e professores sem essa habilitação, ficou muito próxima do limiar de significância (*p*=.052).

Tabela 57: Testes de Mann-Whitney realizados à Parte 4 do QALA para a variável Área Disciplinar

Entre professores com e sem habilitação para **Matemática e Ciências Experimentais**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 4	5054.00	23582.00	-1.63	.104
P4D1	4931.50	23459.50	-1.95	.052
P4D2	5312.50	23840.50	-1.17	.243
P4D3	5540.00	7431.00	-0.67	.506
P4D4	5113.00	23641.00	-1.57	.115
Entre professores com e sem habilitação para **Ciências Sociais e Humanas**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 4	4349.50	26294.50	-0.57	.570
P4D1	4515.00	5505.00	-0.20	.844
P4D2	4487.00	26432.00	-0.27	.788
P4D3	4460.50	5450.50	-0.33	.744
P4D4	3918.50	25863.50	-1.63	.104
Entre professores com e sem habilitação para **Línguas**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 4	6450.00	9610.00	-0.79	.426
P4D1	6531.50	9691.50	-0.66	.507
P4D2	6446.00	9606.00	-0.85	.397
P4D3	6372.00	21597.00	-0.97	.330
P4D4	6537.00	9697.00	-0.66	.511
Entre professores com e sem habilitação para **Expressões**				
	Mann-Whitney (U)	Wilcoxon(W)	Z	p
PARTE 4	4183.00	26549.00	-0.58	.563
P4D1	4222.50	26588.50	-0.50	.614
P4D2	4388.50	26754.50	-0.10	.916
P4D3	4160.00	26526.00	-0.66	.512
P4D4	4293.00	5196.00	-0.34	.737

Também o teste de Kruskal-Wallis, realizado à parte 4 (tabela 58), parece confirmar um baixo grau de relação entre os resultados e a variável 'Área disciplinar', não existindo, portanto, diferenças estatisticamente significativas.

Tabela 58: Testes de Kruskal-Wallis realizados à Parte 4 do QALA para a variável Área Disciplinar

	Nível de Ensino	N	Mean Rank	χ^2 (df)	p
Parte 4	Apenas MCE (1)	60	115.37	3.14 (3)	.371
	Apenas CSH (2)	38	107.28		
	Apenas Línguas	73	97.12		
	Apenas Expressões	41	109.51		
P4D1	Apenas MCE (1)	60	115.96	2.62 (3)	.454
	Apenas CSH (2)	38	104.22		
	Apenas Línguas	73	99.82		
	Apenas Expressões	41	106.66		
P4D2	Apenas MCE (1)	60	112.23	1.40 (3)	.705
	Apenas CSH (2)	38	109.13		
	Apenas Línguas	73	100.79		
	Apenas Expressões	41	105.84		
P4D3	Apenas MCE (1)	60	102.12	1.51 (3)	.680
	Apenas CSH (2)	38	100.08		
	Apenas Línguas	73	109.93		
	Apenas Expressões	41	112.76		
P4D4	Apenas MCE (1)	60	116.30	4.12 (3)	.249
	Apenas CSH (2)	38	112.51		
	Apenas Línguas	73	97.28		
	Apenas Expressões	41	103.00		

Nota: (1) Matemática e Ciências Experimentais; (2) Ciências Sociais e Humanas.

4.3.6 Relação entre os resultados obtidos no QALA com a variável Vínculo

A relação entre os resultados alcançados no QALA e a variável 'Vínculo' foi avaliada por intermédio do teste U de Mann-Whitney. Para essa mesma avaliação, foram formuladas as seguintes hipóteses:

- H_0: As diferenças nas distribuições dos valores obtidos pelos professores que pertencem aos quadros das escolas em que lecionam e pelos professores que são contratados, não são estatisticamente significativas;
- H_1: As diferenças nas distribuições dos valores obtidos pelos professores que

pertencem aos quadros das escolas em que lecionam e pelos professores que são contratados, são estatisticamente significativas.

O teste U aplicado à parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA (tabela 59), revela-nos que deveremos rejeitar a hipótese nula para a globalidade da Parte 2 (p=.049), bem como no domínio 'Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação' (p=.028). Nestes casos, as diferenças nas distribuições dos valores entre professores do quadro e professores contratados, são estatisticamente significativas. Nos restantes casos, o valor de p é superior a .05, pelo que não rejeitamos a hipótese nula.

Tabela 59: Testes de Mann-Whitney realizados à Parte 2 do QALA para a variável Vínculo

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 2	4731.50	6501.50	-1.96	**.049**
P2D1	5031.00	6801.00	-1.36	.172
P2D2	5207.00	6977.00	-1.00	.318
P2D3	4939.00	6709.00	-1.56	.118
P2D4	4618.50	6388.50	-2.20	**.028**

Relativamente aos resultados obtidos na parte 3 (Conhecimentos em avaliação) do QALA(tabela 60), o teste U mostra-nos que deveremos rejeitar a hipótese nula para a globalidade da parte 3, mas não nos domínios que o compõem.

Tabela 60: Testes de Mann-Whitney realizados à Parte 3 do QALA para a variável Vínculo

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 3	4670.00	6440.00	-2.09	**.036**
P3D1	5012.50	6782.50	-1.44	.150
P3D2	5015.00	6785.00	-1.41	.159
P3D3	5158.00	6928.00	-1.11	.268
P3D4	4791.50	6561.50	-1.87	.062

Já na parte 4 (Cenários em contexto de avaliação), o teste U (tabela 61) revela que apenas deveremos rejeitar a hipótese nula para um dos domínios, designadamente o 'Conhecimentos sobre os objetivos e funções da avaliação' (p=.018). Nos restantes casos não rejeitamos a hipótese nula.

Tabela 61: Testes de Mann-Whitney realizados à Parte 4 do QALA para a variável Vínculo

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 4	4794.00	6564.00	-1.85	.064
P4D1	4584.00	6354.00	-2.37	**.018**
P4D2	5498.50	7268.50	-0.43	.670
P4D3	5608.50	24329.50	-0.18	.856
P4D4	4847.00	6617.00	-1.82	.068

4.3.7 Relação entre os resultados obtidos no QALA com a variável Formação Contínua em Avaliação

Para a variável 'Formação Contínua em Avaliação', foi uma vez mais utilizado o teste de Mann-Whitney. As hipóteses aqui colocadadas foram:

- H_0: As diferenças nas distribuições dos valores obtidos pelos professores que realizaram formação contínua em avaliação e pelos professores que não

realizaram este tipo de formação, não são estatisticamente significativas;

- H_1: As diferenças nas distribuições dos valores obtidos pelos professores que realizaram formação contínua em avaliação e pelos professores que não realizaram este tipo de formação, são estatisticamente significativas.

O teste realizado à parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação), mostra-nos, para a globalidade da parte 2, bem como para 3 dos 4 domínios, que as diferenças são estatisticamente significativas para a variável em análise, ou seja, deveremos rejeitar a hipótese nula. A única exceção encontra-se no domínio 'Conhecimentos sobre a utilização de instrumentos de avaliação diversificados' (*p*=.054) que ultrapassa ligeiramente o limiar de significância estatística.

Tabela 62: Testes de Mann-Whitney realizados à Parte 2 do QALA para a variável Formação Contínua em Avaliação

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 2	4637.00	6983.00	-3.20	**.001**
P2D1	4865.50	7211.50	-2.79	**.005**
P2D2	4532.00	6878.00	-3.43	**.001**
P2D3	5310.00	7656.00	-1.93	.054
P2D4	4960.00	7306.00	-2.59	**.010**

Na parte 3 (Conhecimentos em avaliação) do QALA, o teste U (tabela 63) revela-nos, uma vez mais, especial relação entre a variável 'Formação Continua em Avaliação' e os resultados alcançados. Esta afirmação assenta no facto de o valor de *p* ser inferior a .05 na generalidade da parte 3 e em 3 dos domínios. No entanto, neste caso, a exceção foi o domínio 'Conhecimentos sobre os objetivos e funções da avaliação' que apresenta um valor de *p* de .874.

Tabela 63: Testes de Mann-Whitney realizados à Parte 3 do QALA para a variável Formação Contínua em Avaliação

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 3	4509.00	6855.00	-3.46	**.001**
P3D1	6211.00	8557.00	-0.16	.874
P3D2	4934.00	7280.00	-2.67	**.008**
P3D3	5076.00	7422.00	-2.38	**.017**
P3D4	4621.00	6967.00	-3.28	**.001**

Pela análise ao teste U da parte 4 (Cenários em contexto de avaliação) do QALA (tabela 64), concluímos que não devemos rejeitar a hipótese nula em nenhum dos casos. Assim, parece não existir especial relação entre os resultados obtidos nesta parte do QALA e a frequência, ou não, de formação contínua em avaliação.

Tabela 64: Testes de Mann-Whitney realizados à Parte 4 do QALA para a variável Formação Contínua em Avaliação

	Mann-Whitney (U)	Wilcoxon (W)	Z	p
PARTE 4	6238.50	8584.50	-0.10	.920
P4D1	6163.50	8509.50	-0.26	.797
P4D2	5859.50	23064.50	-0.89	.372
P4D3	5774.50	22979.50	-1.05	.295
P4D4	5540.00	7886.00	-1.53	.125

4.3.8 Relação entre os resultados obtidos no QALA com a variável Idade

Para a variável idade, e visto tratarem-se de 5 escalões etários, o teste utilizado foi o teste H de Kruskal-Wallis para *k* amostras independentes. Desta forma, foram estabelecidas as seguintes hipóteses:

- H_0: As diferenças nas distribuições dos valores obtidos pelos professores dos

diferentes escalões etários, não são estatisticamente significativas;

- H_1: As diferenças nas distribuições dos valores obtidos pelos professores dos diferentes escalões etários, são estatisticamente significativas.

O teste H realizado à parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA (tabela 65), revela-nos uma relação estatisticamente significativa para a globalidade da parte 2 (p=.039), mas também para o domínio 'Conhecimentos sobre a utilização de instrumentos de avaliação diversificados' (p=.046). Nestes casos, rejeitamos a hipótese nula, pois existem diferenças significativas nas distribuições dos valores obtidos em 1 ou mais escalões etários. Para os restantes domínios, não rejeitamos a hipótese nula.

Tabela 65: Testes de Kruskal-Wallis realizados à Parte 2 do QALA para a variável Idade

	Experiência	N	Mean Rank	χ^2	p
Parte 2	Entre 26 e 30 anos	6	92.00	10.07 (4)	**.039**
	Entre 31 e 40 anos	50	103.14		
	Entre 41 e 50 anos	96	127.57		
	Entre 51 e 60 anos	76	139.90		
	Mais de 60 anos	25	141.70		
P2D1	Entre 26 e 30 anos	6	97.17	8.91 (4)	.063
	Entre 31 e 40 anos	50	112.03		
	Entre 41 e 50 anos	96	120.02		
	Entre 51 e 60 anos	76	143.80		
	Mais de 60 anos	25	139.84		
P2D2	Entre 26 e 30 anos	6	111.00	7.92 (4)	.095
	Entre 31 e 40 anos	50	106.50		
	Entre 41 e 50 anos	96	124.39		
	Entre 51 e 60 anos	76	140.28		
	Mais de 60 anos	25	141.52		
P2D3	Entre 26 e 30 anos	6	88.25	9.69 (4)	**.046**
	Entre 31 e 40 anos	50	106.25		
	Entre 41 e 50 anos	96	125.56		
	Entre 51 e 60 anos	76	140.73		
	Mais de 60 anos	25	141.58		
P2D4	Entre 26 e 30 anos	6	108.17	6.91 (4)	.141
	Entre 31 e 40 anos	50	104.43		
	Entre 41 e 50 anos	96	135.38		
	Entre 51 e 60 anos	76	130.38		
	Mais de 60 anos	25	132.46		

Relativamente à parte 3 (Conhecimentos em avaliação), o teste H (tabela 66) indica-nos que existem diferenças significativas no domínio 'Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar' (p=.011). Nos restantes casos, não rejeitamos a hipótese nula. No entanto, salientamos que a globalidade da parte 3 e o domínio 'Conhecimentos sobre os objetivos e funções da avaliação' apresentam valores de p muito próximos do limiar de significância estatística.

Tabela 66: Testes de Kruskal-Wallis realizados à Parte 3 do QALA para a variável Idade

	Experiência	N	Mean Rank	χ^2	p
Parte 3	Entre 26 e 30 anos	6	89.00	8.55 (4)	.073
	Entre 31 e 40 anos	50	115.16		
	Entre 41 e 50 anos	96	122.11		
	Entre 51 e 60 anos	76	145.80		
	Mais de 60 anos	25	121.44		
P3D1	Entre 26 e 30 anos	6	112.67	9.02 (4)	.061
	Entre 31 e 40 anos	50	127.64		
	Entre 41 e 50 anos	96	116.84		
	Entre 51 e 60 anos	76	145.92		
	Mais de 60 anos	25	110.66		
P3D2	Entre 26 e 30 anos	6	116.08	13.01 (4)	**.011**
	Entre 31 e 40 anos	50	105.50		
	Entre 41 e 50 anos	96	118.77		
	Entre 51 e 60 anos	76	146.75		
	Mais de 60 anos	25	144.18		
P3D3	Entre 26 e 30 anos	6	91.83	4.06 (4)	.398
	Entre 31 e 40 anos	50	124.58		
	Entre 41 e 50 anos	96	122.41		
	Entre 51 e 60 anos	76	138.90		
	Mais de 60 anos	25	121.72		
P3D4	Entre 26 e 30 anos	6	78.58	4.38 (4)	.357
	Entre 31 e 40 anos	50	117.59		
	Entre 41 e 50 anos	96	131.65		
	Entre 51 e 60 anos	76	132.16		
	Mais de 60 anos	25	123.88		

Na parte 4 (Cenários em contexto de avaliação) do QALA, parece não existir especial relação entre a variável 'Idade' e os resultados alcançados, dado que os valores de *p*, seja para a globalidade da parte 4, seja para os domínios que o compõem, estão muito acima do valor de .050. Logo, para a parte 4, não rejeitamos a hipótese nula.

Tabela 67: Testes de Kruskal-Wallis realizados à Parte 4 do QALA para a variável Idade

	Experiência	N	Mean Rank	χ^2	p
Parte 4	Entre 26 e 30 anos	6	114.42	1.74 (4)	.784
	Entre 31 e 40 anos	50	119.51		
	Entre 41 e 50 anos	96	124.49		
	Entre 51 e 60 anos	76	133.45		
	Mais de 60 anos	25	135.02		
P4D1	Entre 26 e 30 anos	6	116.33	0.62 (4)	.961
	Entre 31 e 40 anos	50	121.57		
	Entre 41 e 50 anos	96	127.92		
	Entre 51 e 60 anos	76	130.17		
	Mais de 60 anos	25	127.24		
P4D2	Entre 26 e 30 anos	6	141.00	0.75 (4)	.946
	Entre 31 e 40 anos	50	122.89		
	Entre 41 e 50 anos	96	125.22		
	Entre 51 e 60 anos	76	130.92		
	Mais de 60 anos	25	126.76		
P4D3	Entre 26 e 30 anos	6	132.58	0.46 (4)	.977
	Entre 31 e 40 anos	50	130.69		
	Entre 41 e 50 anos	96	124.95		
	Entre 51 e 60 anos	76	125.02		
	Mais de 60 anos	25	132.16		
P4D4	Entre 26 e 30 anos	6	95.50	5.29 (4)	.258
	Entre 31 e 40 anos	50	115.47		
	Entre 41 e 50 anos	96	123.63		
	Entre 51 e 60 anos	76	137.91		
	Mais de 60 anos	25	137.40		

4.3.9 Relação entre os resultados obtidos no QALA com a variável Experiência

À semelhança do caso anterior, para a analisarmos a relação existente entre os resultados obtidos no QALA e a variável 'Experiência', foi utilizado o teste H de Kruskal-Wallis, dado que foram consideradas 5 categorias alinhadas com os ciclos de vida do professor propostos por Huberman (1995). Neste caso, foram definidas as seguintes hipóteses:

- H_0: As diferenças nas distribuições dos valores obtidos pelos professores, tendo em consideração a sua experiência letiva , não são estatisticamente significativas;

- H_1: As diferenças nas distribuições dos valores obtidos pelos professores, tendo em consideração a sua experiência letiva, são estatisticamente significativas.

Constatamos, pela análise à tabela 68, que parece não existir especial relação entre a variável em análise e os resulados alcançados na parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação), já que o teste H não devolveu nenhum valor de *p* inferior a .05. Assim, neste caso, não rejeitamos a hipótese nula.

Tabela 68: Testes de Kruskal-Wallis realizados à Parte 2 do QALA para a variável Experiência

	Experiência	N	Mean Rank	χ^2 (df)	p
Parte 2	Até 3 anos	8	122.19	5.48 (4)	.241
	Entre 4 e 6 anos	12	110.21		
	Entre 7 e 25 anos	134	119.54		
	Entre 26 e 35 anos	72	136.26		
	Mais de 35 anos	27	148.20		
P2D1	Até 3 anos	8	105.56	5.74 (4)	.219
	Entre 4 e 6 anos	12	121.00		
	Entre 7 e 25 anos	134	119.14		
	Entre 26 e 35 anos	72	138.37		
	Mais de 35 anos	27	144.70		
P2D2	Até 3 anos	8	151.56	6.60 (4)	.159
	Entre 4 e 6 anos	12	118.63		
	Entre 7 e 25 anos	134	117.32		
	Entre 26 e 35 anos	72	136.33		
	Mais de 35 anos	27	146.61		
P2D3	Até 3 anos	8	92.88	6.94 (4)	.139
	Entre 4 e 6 anos	12	114.54		
	Entre 7 e 25 anos	134	120.54		
	Entre 26 e 35 anos	72	136.91		
	Mais de 35 anos	27	148.30		
P2D4	Até 3 anos	8	138.06	2.19 (4)	.701
	Entre 4 e 6 anos	12	114.67		
	Entre 7 e 25 anos	134	123.07		
	Entre 26 e 35 anos	72	129.40		
	Mais de 35 anos	27	142.31		

Já o teste H realizado à parte 3 (Conhecimentos em avaliação) do QALA(tabela 69) parece dar-nos uma perspetiva diferente em dois dos domínios. Assim, nos casos dos domínios ’Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar’ (*p*=.041) e ’Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação’ (*p*=.022), parece existir uma relação estatisticamente significativa para a variável ’Experiência’. Nos restantes casos, não rejeitamos a hipótese nula.

Tabela 69: Testes de Kruskal-Wallis realizados à Parte 3 do QALA para a variável Experiência

	Experiência	N	Mean Rank	χ^2	p
Parte 3	Até 3 anos	8	91.00	7.74 (4)	.102
	Entre 4 e 6 anos	12	128.88		
	Entre 7 e 25 anos	134	119.76		
	Entre 26 e 35 anos	72	145.13		
	Mais de 35 anos	27	124.43		
P3D1	Até 3 anos	8	121.94	3.94 (4)	.414
	Entre 4 e 6 anos	12	126.75		
	Entre 7 e 25 anos	134	121.72		
	Entre 26 e 35 anos	72	140.75		
	Mais de 35 anos	27	118.13		
P3D2	Até 3 anos	8	114.31	9.96 (4)	**.041**
	Entre 4 e 6 anos	12	127.04		
	Entre 7 e 25 anos	134	114.81		
	Entre 26 e 35 anos	72	144.55		
	Mais de 35 anos	27	144.43		
P3D3	Até 3 anos	8	114.06	8.95 (4)	.062
	Entre 4 e 6 anos	12	140.38		
	Entre 7 e 25 anos	134	117.87		
	Entre 26 e 35 anos	72	147.12		
	Mais de 35 anos	27	116.57		
P3D4	Até 3 anos	8	46.69	11.46 (4)	**.022**
	Entre 4 e 6 anos	12	122.13		
	Entre 7 e 25 anos	134	132.50		
	Entre 26 e 35 anos	72	130.40		
	Mais de 35 anos	27	116.57		

Por último, na parte 4 (Cenários em contexto de avaliação) do QALA(tabela 70), a

variável em análise parece ter especial relação com o domínio 'Conhecimentos sobre os objetivos e funções da avaliação' (*p*=.009), pelo que, neste caso, rejeitamos a hipótese nula. Nos restantes casos, não rejeitamos a hipótese nula, muito embora a globalidade da parte 4 tenha um valor muito próximo do limiar de significância estatística.

Tabela 70: Testes de Kruskal-Wallis realizados à Parte 4 do QALA para a variável Experiência

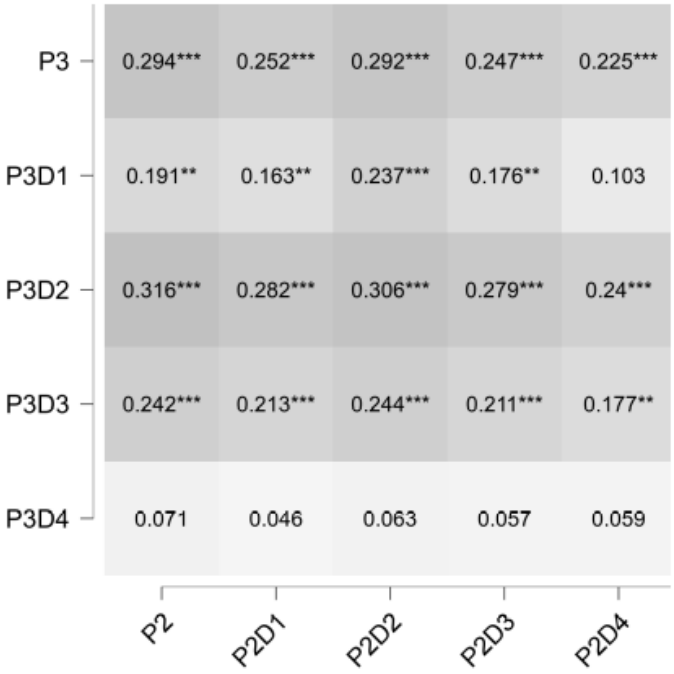
	Experiência	N	Mean Rank	χ^2	p
Parte 4	Até 3 anos	8	62.63	8.82 (4)	.066
	Entre 4 e 6 anos	12	130.71		
	Entre 7 e 25 anos	134	124.71		
	Entre 26 e 35 anos	72	139.72		
	Mais de 35 anos	27	121.87		
P4D1	Até 3 anos	8	45.31	13.51 (4)	**.009**
	Entre 4 e 6 anos	12	148.96		
	Entre 7 e 25 anos	134	127.76		
	Entre 26 e 35 anos	72	134.60		
	Mais de 35 anos	27	117.41		
P4D2	Até 3 anos	8	114.81	1.10 (4)	.895
	Entre 4 e 6 anos	12	128.71		
	Entre 7 e 25 anos	134	124.35		
	Entre 26 e 35 anos	72	133.40		
	Mais de 35 anos	27	125.93		
P4D3	Até 3 anos	8	114.25	0.77 (4)	.942
	Entre 4 e 6 anos	12	134.92		
	Entre 7 e 25 anos	134	124.74		
	Entre 26 e 35 anos	72	130.55		
	Mais de 35 anos	27	129.02		
P4D4	Até 3 anos	8	75.81	5.88 (4)	.208
	Entre 4 e 6 anos	12	112.33		
	Entre 7 e 25 anos	134	126.76		
	Entre 26 e 35 anos	72	135.05		
	Mais de 35 anos	27	128.39		

4.3.10 Análise Correlacional

A primeira análise correlacional será realizada às partes 2 e 3 do QALA. Com a utilização do r_s de Spearman procuraremos verificar se existe algum grau de correlação entre os resultados das 'Perceções sobre os conhecimentos e capacidades em avaliação' (Parte 2) dos professores e o seu desempenho na parte 'Conhecimentos em avaliação' (Parte 3).

A figura 19 apresenta um *heatmap* onde estão sistematizados os resultados obtidos pela aplicação do coeficiente de correlação de Spearman às Partes 2 e 3 do QALA, bem como aos respetivos domínios que os constituem (mais detalhe no Anexo 8). Os valores apresentados respeitam-se ao coeficiente (r_s) resultante da aplicação da correlação de Spearman, enquanto que a presença, ou ausência, dos caracteres (*) remetem-nos para o nível de significância estatística, conforme nota abaixo da figura.

Figura 19: *Heatmap* com os Coeficientes de Correlação de Spearman (r_S) entre a Parte 2 e a Parte 3 do QALA



Nota: * $p < .05$; ** $p < .01$; *** $p < .001$

A tabela 71 sistematiza algumas das conclusões que podemos retirar da análise à figura 19. De salientar que não analisaremos todas as correlações possíveis entre ambas as partes. Consideramos relevante analisarmos apenas as correlações existentes entre a globalidade das partes 2 e 3, que constituem o QALA, bem como os domínios que tenham correspondência entre ambas as partes (por exemplo P2D1 com P3D1).

Tabela 71: Sistematização dos resultados obtidos pela aplicação do Coeficiente de Correlação de Spearman (r_S) entre a Parte 2 e a Parte 3 do QALA

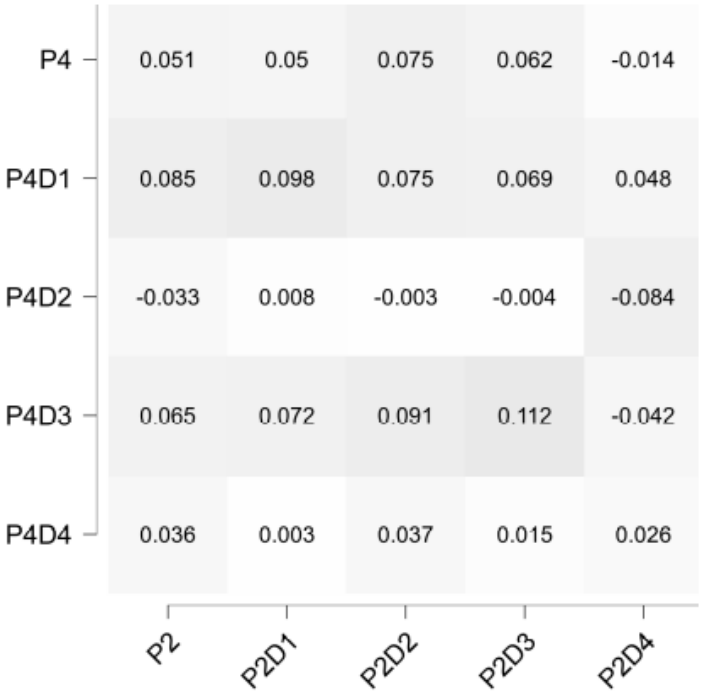
Variáveis	*Correlação na amostra*	*Teste de Significância*
Parte2 - Parte 3	As variáveis têm uma fraca correlação positiva (r_S=.294).	Rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é, provavelmente, diferente de zero ($p < .001$)
P2D1 - P3D1	As variáveis têm uma fraca correlação positiva (r_S=.163).	Rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é, provavelmente, diferente de zero ($p < .01$)

P2D2 - P3D2	As variáveis têm uma fraca correlação positiva (r_S=.306).	Rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é, provavelmente, diferente de zero (p < .001)
P2D3 - P3D3	As variáveis têm uma fraca correlação positiva (r_S=.211).	Rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é, provavelmente, diferente de zero (p < .001)
P2D4 - P3D4	As variáveis têm uma fraca correlação positiva (r_S=.059).	Não rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é igual a zero (p > .05)

Relativamente às partes 2 e 4 do QALA procuraremos verificar, à semelhança do caso anterior, se existe algum grau de correlação entre os resultados das ’Perceções sobre os conhecimentos e capacidades em avaliação’ (Parte 2) dos professores e o seu desempenho na parte ’Cenários em contexto de avaliação’ (Parte 4).

O *heatmap*, presente na figura 20, esquematiza os resultados obtidos pela aplicação do coeficiente de correlação de Spearman às Partes 2 e 4 do QALA, bem como aos respetivos domínios que os constituem.

Figura 20: *Heatmap* com os Coeficientes de Correlação de Spearman (r_S) entre a Parte 2 e a Parte 4 do QALA



Nota: * $p < .05$; ** $p < .01$; *** $p < .001$

A tabela 72 sistematiza algumas das conclusões que podemos retirar da análise à figura 20. Tal como na análise anterior, não analisaremos todas as correlações possíveis entre ambas as partes. Consideramos relevante analisarmos apenas as correlações existentes entre a globalidade das partes 2 e 4, que constituem o QALA, bem como os domínios que tenham correspondência entre ambas as partes (por exemplo P2D1 com P4D1).

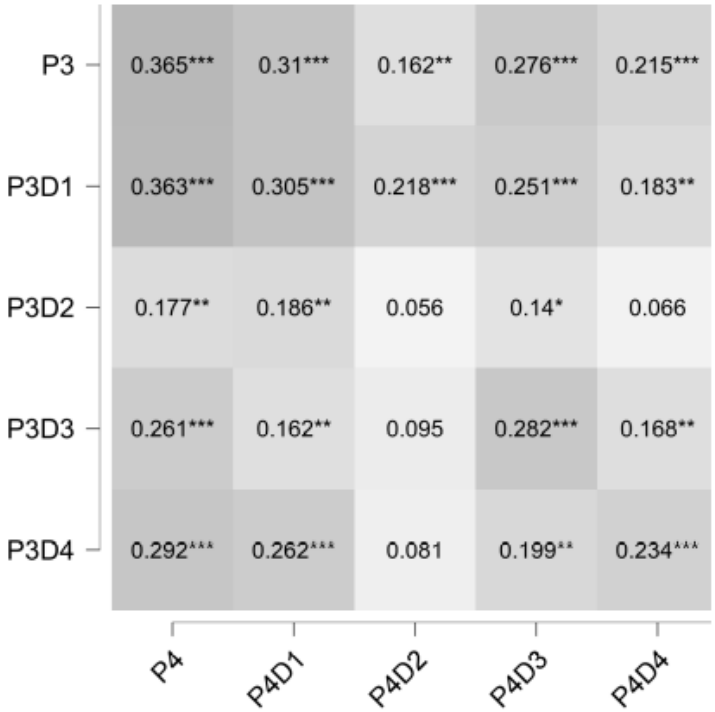
Tabela 72: Sistematização dos resultados obtidos pela aplicação do Coeficiente de Correlação de Spearman (r_S) entre a Parte 2 e a Parte 4 do QALA

Variáveis	*Correlação na amostra*	*Teste de Significância*
Parte2 - Parte 4	As variáveis têm uma fraca correlação positiva (r_S=.051).	Não rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é igual a zero ($p > .05$)
P2D1 - P4D1	As variáveis têm uma fraca correlação positiva (r_S=.098).	Não rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é igual a zero ($p > .05$)

P2D2 - P4D2	As variáveis têm uma fraca correlação negativa (r_S=-.003).	Não rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é igual a zero (p > .05)
P2D3 - P4D3	As variáveis têm uma fraca correlação positiva (r_S=.112).	Não rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é igual a zero (p > .05)
P2D4 - P4D4	As variáveis têm uma fraca correlação positiva (r_S=.026).	Não rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é igual a zero (p > .05)

Por último, procuraremos verificar a existência de correlações entre o desempenho dos professores nas partes ’Conhecimentos em avaliação’ (Parte 3) e ’Cenários em contexto de avaliação’ (Parte 4). O *heatmap*, presente na figura 21, esquematiza os resultados obtidos, pela aplicação do coeficiente de correlação de Spearman, às partes 3 e 4.

Figura 21: Heatmap com os Coeficientes de Correlação de Spearman (r_S) entre a Parte 3 e a Parte 4 do QALA



Nota: * p < .05; ** p < .01; *** p < .001

A tabela 73 sistematiza algumas das conclusões que podemos retirar da análise à

figura 21. Tal como até aqui, não analisaremos todas as correlações possíveis entre ambas as partes. Consideramos relevante analisarmos apenas as correlações existentes entre a globalidade das partes 3 e 4, que constituem o QALA, bem como os domínios que tenham correspondência entre ambas as partes (por exemplo P3D1 com P4D1).

Tabela 73: Sistematização dos resultados obtidos pela aplicação do Coeficiente de Correlação de Spearman (r_s) entre a Parte 3 e a Parte 4 do QALA

Variáveis	*Correlação na amostra*	*Teste de Significância*
Parte3 - Parte 4	As variáveis têm uma fraca correlação positiva (r_s=.365).	Rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é, provavelmente, diferente de zero (p < .001)
P3D1 - P4D1	As variáveis têm uma fraca correlação positiva (r_s=.305).	Rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é, provavelmente, diferente de zero (p < .001)
P3D2 - P4D2	As variáveis têm uma fraca correlação positiva (r_s=.056).	Não rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é igual a zero (p > .05)
P3D3 - P4D3	As variáveis têm uma fraca correlação positiva (r_s=.282).	Rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é, provavelmente, diferente de zero (p < .001)
P3D4 - P4D4	As variáveis têm uma fraca correlação positiva (r_s=.234).	Rejeitamos a hipótese nula. O Coeficiente de correlação no Universo é, provavelmente, diferente de zero (p < .001)

Conclusões

Neste derradeiro capítulo iremos proceder à discussão dos resultados obtidos no capítulo anterior, tendo em consideração os objetivos delineados para a presente investigação e alguns dos estudos realizados neste campo, procurando, sempre que possível, estabelecer pontos em comum.

Após a discussão dos resultados, procuraremos identificar as limitações do presente estudo e projetar desenvolvimentos futuros que esta investigação poderá proporcionar.

Discussão dos resultados

A presente investigação assentou em dois grandes objetivos gerais. O primeiro objetivo geral era o de analisar as perceções que os professores tinham sobre os seus conhecimentos e capacidades em avaliação e o segundo era o de aferir a sua literacia em avaliação. A partir destes dois grandes objetivos derivaram um conjunto de objetivos específicos que relembramos de seguida:

- Analisar as perceções sobre os conhecimentos e capacidades em avaliação dos professores em quatro domínios, nomeadamente: (a) Conhecimentos sobre os objetivos e funções da avaliação; (b) Conhecimentos sobre o currículo e sobre

aquilo que é importante aprender e avaliar; (c) Conhecimentos sobre a utilização de instrumentos de avaliação diversificados; (d) Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação;

- Aferir a literacia em avaliação dos professores nos quatro domínios identificados na alínea anterior;

- Analisar a relação entre as perceções sobre os conhecimentos e capacidades em avaliação com algumas variáveis de contexto, nomeadamente sexo, idade, subsistema de ensino, tipo de habilitação, vínculo, experiência letiva, nível de ensino, área disciplinar e formação contínua em avaliação;

- Analisar a relação entre a literacia em avaliação e as variáveis de contexto identificadas na alínea anterior;

- Analisar a eventual a relação entre a literacia em avaliação e as perceções sobre os conhecimentos e capacidades em avaliação.

Para a persecução dos objetivos delineados, foi necessária a construção de um questionário para a recolha de dados, o qual designámos de Questionário de Aferição da Literacia em Avaliação (QALA). As qualidades psicométricas das 3 partes que constituem o QALA (4 se considerarmos a Parte 1 - Dados Gerais) foram avaliadas com recurso ao Modelo Rasch. Os resultados obtidos parecem evidenciar as boas qualidades psicométricas do QALA bem como validade de construto.

A análise aos resultados alcançados na Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA, permite-nos afirmar que os professores têm uma autoperceção satisfatória sobre os seus conhecimentos e capacidades em avaliação. Os domínios para os quais os professores alcançaram os melhores resultados foram os domínios ’Conhecimentos sobre os objetivos e funções

da avaliação' e 'Conhecimentos sobre a utilização de instrumentos de avaliação diversificados'. Já os domínios 'Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar' e 'Conhecimentos sobre a interpretação e utilização da informação recolhida no processo de avaliação' tiveram os resultados mais baixos.

Embora os resultados sejam satisfatórios, tanto na globalidade da Parte 2, como nos domínios que a constituem, os dados mostram-nos que os professores apresentam algumas limitações face a alguns aspetos relacionados com a avaliação. Um dos itens que apresentou valores mais baixos na Parte 2 do QALA foi o P2.5, relacionado com a distinção entre avaliação criterial e normativa. Ao analisarmos os itens correspondentes nas Partes 3 (Conhecimentos em avaliação) e 4 (Cenários em contexto de avaliação), verificámos que estes apresentavam valores igualmente baixos, o que parece ser revelador de grandes lacunas por parte dos professores. Os conceitos de avaliação normativa e avaliação criterial são fundamentais no contexto da avaliação das aprendizagens, uma vez que estão intimamente relacionados com duas das principais funções da avaliação. A avaliação normativa tem a função de classificar e seriar, ou seja, o desempenho do aluno é comparado aos demais pelo que ele ocupa uma determinada posição numa "hierarquia". Já na avaliação criterial o desempenho do aluno é analisado em relação a um conjunto de objetivos e metas definidas. Assim, não existe propriamente uma competição/ hierarquização dos alunos, uma vez que todos têm a mesma possibilidade de progredir.

Outro item da Parte 2 que apresentou valores mais modestos, quando comparados com os demais, foi o P2.9 relacionado com os níveis de complexidade cognitiva, sistematizados, por exemplo, na *Taxonomia de Bloom, Taxonomia de Marzano e Depth of Knowledge*. Este conhecimento reveste-se de especial importância, pois permite, como referem Ferraz e Belhot (2010):

> a definição clara e estruturada dos objetivos instrucionais, considerando a aquisição de conhecimento e de competências adequados ao perfil profissional a ser formado, direcionará o processo de ensino para a escolha adequada de estratégias, métodos, delimitação do conteúdo específico, instrumentos de avaliação e, consequentemente, para uma aprendizagem efetiva e duradoura (p.422).

Assim, o conhecimento sobre os domínios de complexidade cognitiva auxilia no planeamento, na organização e no controlo dos objetivos de aprendizagem. Na parte 3 do QALA, os itens P3.9 e P3.29 estão diretamente relacionados com estes aspetos. Em ambos os itens, a taxa de acertos foi relativamente baixa. Já na parte 4, o item P4.4.4, relacionado com estas questões, teve uma taxa de acertos superior. No entanto, pelos resultados verificados, especialmente na Parte 2 e na Parte 3, poder-se-á concluir que os professores apresentam algumas fragilidades neste domínio.

Destaque ainda para o item P2.17, relacionado com os conhecimentos sobre a determinação das propriedades psicométricas dos instrumentos de avaliação, que apresentou um número significativo de respostas Discordo Totalmente e Discordo. Já nas partes 3 e 4, os itens correspondentes apresentaram médias muito baixas, o que se traduz numa franca limitação dos professores neste domínio. É a partir das qualidades psicométricas dos instrumentos de avaliação que poderemos garantir, entre outros, a validade e a fiabilidade dos mesmos que, como vimos nos subcapítulos 1.5 e 1.6.5, são fundamentais para garantir a qualidade da informação que é recolhida no processo de avaliação.

Num estudo levado a cabo por Alkharusi (2011b), o autor desenvolveu e aplicou um questionário designado de *Self Perceived Assessment Skills Scale* que procurava aferir as perceções dos professores em relação às suas capacidades de avaliação. Nesse estudo, o autor procurou verificar a relação do género, da área curricular, da

experiência letiva e da formação em avaliação com os resultados alcançados no referido questionário.

Analisando os resultados alcançados por Alkharusi (2011b) e comparando-os com os resultados verificados na Parte 2 (Perceções sobre os conhecimentos e capacidades em avaliação) do QALA, encontramos alguns aspetos em comum. Tendo em consideração o sexo dos professores, ambos os estudos apontam que os professores do sexo feminino têm uma melhor perceção dos seus conhecimentos e capacidades em avaliação do que os professores do sexo masculino. Tal como verificado no QALA, também no estudo de Alkharusi (2011b) os professores do sexo feminino alcançaram melhores resultados em todos os domínios.

Considerando a área curricular dos professores, Alkharusi (2011b) comparou os resultados entre professores de Inglês, professores de Ciências e professores de Artes. Os resultados alcançados pelo autor, apontam que os professores de Ciências apresentavam melhores resultados em todos os domínios, quando comparados com os professores de Inglês e Artes. No caso do QALA, verificámos que, embora os resultados fossem muito próximos, foram os professores de Ciências Sociais e Humanas os que apresentaram melhores resultados, seguindo-se os professores de Línguas, os professores de Expressões e, contrariamente ao estudo de Alkharusi (2011b), os professores de Matemática e Ciências Experimentais apresentaram os resultados mais baixos.

À semelhança do verificado pela aplicação do QALA, Alkharusi constatou que as médias eram crescentes com a experiência letiva. O autor comparou os resultados de professores até 5 anos de experiência, entre 6 e 10 anos de experiência e com mais de 10 anos de experiência. Em todos os domínios, foram os professores com mais de 10 anos de experiência aqueles que apresentaram melhores resultados e os professores

com até 5 anos de experiência os que apresentaram os resultados mais baixos. No caso do QALA, também foram os professores com mais experiência que apresentaram os valores mais altos em quase todos os domínios. A única exceção foi no domínio 'Conhecimentos sobre o currículo e sobre aquilo que é importante aprender e avaliar' onde foram o segundo grupo a apresentar melhores resultados.

Outro aspeto interessante do estudo de Alkharusi foi o facto de ele ter procedido a uma análise dos resultados, tendo em consideração o facto de os professores terem, ou não, frequentado formação específica em avaliação. O autor verificou, à semelhança do nosso estudo, que os professores que haviam frequentado formação específica em avaliação apresentavam resultados substancialmente superiores aos professores que não frequentaram tal formação.

Os resultados alcançados nas Partes 3 (Conhecimentos em avaliação) e 4 (Cenários em contexto de avaliação) do QALA revelaram igualmente aspetos interessantes e em linha com a investigação na área da literacia em avaliação.

A percentagem média de acertos alcançada na globalidade das Partes 3 e 4 foi francamente baixa, tendo em consideração a importância que a tarefa de avaliar tem no processo de ensino e aprendizagem. No estudo realizado por Plake, Impara e Fager (1993), a média de acertos foi de 66,3%. No caso de Mertler (2003), a média de acertos foi de 63%. No artigo publicado por Breziat e Coleman (2015), a média não foi além dos 52%. Por último, no caso do estudo de Alkharusi *et al*. (2012) a percentagem de acertos foi bastante mais baixa, não chegando sequer aos 40%. Os baixos valores alcançados no QALA parecem confirmar a ideia de que os professores apresentam baixos índices de literacia em avaliação, o que se poderá revelar em lacunas importantes na capacidade dos professores em avaliar os alunos de forma precisa e adequada (Daniel e King, 1998; Koh, 2011; Volante e Fazio, 2007; Yamtim e

Wongwanich, 2014). Recordando Malone (2013), um baixo nível de literacia em avaliação põe em causa tanto a avaliação dos alunos, como todo o processo de ensino e aprendizagem.

Na Parte 3 do QALA, o domínio que apresentou a média mais alta foi o ’Conhecimento sobre os objetivos e funções da avaliação’, enquanto que na Parte 4 foi o domínio ’Conhecimentos sobre a utilização de instrumentos de avaliação diversificados’ o que obteve melhores resultados. Embora os domínios do QALA sejam diferentes dos que organizam instrumentos como o TALQ, o ALI, o CALI e o MLQ, existem alguns paralelismos que podem ser realizados. No caso dos resultados alcançados na Parte 4 do QALA, podemos estabelecer alguns pontos em comum com os resultados alcançados por Plake, Impara e Fager (1993) e Breziat e Coleman (2015). Nestes dois estudos, o domínio do TALQ ’Escolher métodos de avaliação apropriados’ alcançou o segundo melhor e o melhor resultado respetivamente. Também no estudo levado a cabo por Mertler (2003), com recurso ao CALI, foi o domínio ’Escolher métodos de avaliação apropriados’ o que apresentou a média mais elevada.

Já o domínio que apresentou a média mais baixa foi o ’Conhecimentos sobre a interpretação e utilização da informação recolhida no processo de avaliação’, tanto na Parte 3 como na Parte 4 do QALA. No caso dos estudos conduzidos por Plake, Impara e Fager (1993) e Breziat e Coleman (2015), o domínio que apresentou o pior resultado foi o de ’Comunicar os resultados da avaliação’. A comunicação dos resultados da avaliação, em especial do feedback, é um aspeto que está presente no domínio ’Conhecimentos sobre a interpretação e utilização da informação recolhida no processo de avaliação’, pelo que concluímos que há alguma convergência nos resultados. Também Daniel e King (1998) identificaram grandes fragilidades, por parte dos professores, em interpretar os resultados obtidos pelos processos de

avaliação, assim como baixos conhecimentos sobre as qualidades psicométricas dos instrumentos de avaliação, em especial da validade e fiabilidade. Mais uma vez, estas conclusões estão em linha com os resultados alcançados nas Partes 3 e 4 do QALA.

A literatura sugere também alguns aspetos interessantes, relacionados com as variáveis de contexto, verificados nos resultados obtidos pela aplicação das partes 3 e 4 do QALA. Breziat e Coleman (2015) verificaram que os índices de literacia em avaliação eram tanto maiores quanto mais alto era o nível de ensino dos professores. No caso de Breziat e Coleman (2015), os melhores resultados foram alcançados pelos professores do *Secondary*, seguindo-se os professores do *Elementary* e, por fim, os professores do *Early Childhood*. Na presente investigação, verificámos que, tanto na globalidade da Parte 3 como da Parte 4, os índices mais altos de literacia em avaliação foram alcançados pelos professores do 3°Ciclo e Secundário, seguindo-se os professores do 2º Ciclo e, por último, os professores do 1º Ciclo. Já Plake, Impara e Fager (1993) verificaram que os valores de literacia em avaliação eram mais elevados nos professores que frequentaram formação em avaliação, comparativamente aos professores que não haviam frequentado tais formações. Este aspeto é igualmente verificado no nosso estudo, em especial nas Partes 2 (Perceções sobre os conhecimentos e capacidades em avaliação) e 3 (Conhecimentos em avaliação).

De salientar, em último lugar, o facto de se terem verificado correlações positivas entre as Partes 2 e 3 do QALA e entre as Partes 3 e 4. Embora os índices de correlação sejam baixos elas são estatisticamente significativas, com exceção de alguns domínios que constituem as diferentes partes do QALA. Assim, verificamos que parece existir uma tendência para que valores mais altos de literacia em avaliação ocorram em professores com uma melhor perceção sobre os seus

conhecimentos e capacidades em avaliação. Bem como professores com mais conhecimentos em avaliação conseguem tomar melhores decisões em contexto de sala de aula.

Limitações do estudo

As limitações à presente investigação podem ser organizadas em duas categorias, de ordem externa e de ordem interna. As limitações de ordem externa não estão diretamente relacionadas com a presente investigação mas, de algum modo, condicionaram-na. Já as limitações de ordem interna advêm diretamente da presente investigação. Numa perspetiva de análise *SWOT*[54], podemos afirmar que as limitações de ordem externa constituem-se como as 'ameaças', enquanto que as limitações de ordem interna se assumem como os pontos fracos da presente investigação. Desta forma, debruçar-nos-emos em primeiro lugar sobre os aspetos que consideramos serem limitações de ordem externa e, posteriormente, sobre as limitações de ordem interna.

Uma das principais limitações externas à presente investigação foi a escassez de estudos de natureza quantitativa relacionados com a literacia em avaliação. Com efeito, embora existam vários estudos relacionados com a literacia em avaliação, são poucos os que seguem uma abordagem quantitativa, o que acaba por limitar um pouco a triangulação dos resultados.

Outro aspeto que consideramos relevante, é o facto de grande parte dos estudos quantitativos em literacia em avaliação estarem assentes em questionários desenvolvidos nos anos 90 e na viragem do milénio, o que, a nosso ver, é algo

[54] Strengths, Weaknesses, Opportunities, and Threats

limitador, dados os avanços que têm ocorrido na área da avaliação. Para além disso, muitos desses mesmos questionários adequam-se à realidade norte-americana, pelo que a sua utilização, fora desse contexto, seja difícil, conforme demonstraram alguns estudos como os de Hailaya *et al.* (2014). Este aspeto levou à necessidade de desenvolver um instrumento novo que pudesse ser aplicado no contexto português. Embora consideremos este aspeto positivo, a verdade é que dificultou a discussão de resultados, visto não existirem estudos realizados com este instrumento, não havendo, portanto, um termo de comparação.

Uma outra limitação de ordem externa que, de certa forma, afetou a presente investigação foi a situação pandémica. Desde que foram delineados os objetivos da investigação, foi sempre nossa intenção que a recolha dos dados ocorresse de forma presencial, em ambiente escolar, com o intuito de diminuir os efeitos da desejabilidade social. Desta forma, foi necessário adaptar todo o processo de recolha de dados para ambiente à distância e reforçar, junto dos professores, a questão do anonimato para evitar tentações na busca de respostas às questões colocadas.

Analisando agora as limitações de ordem interna, temos o facto de a amostra utilizada ser não probabilística por conveniência o que, segundo Hill e Hill (2002), tem como desvantagem o facto dos resultados e as conclusões só se aplicarem à amostra, não podendo ser extrapolados, com confiança, para o universo.

Outra limitação de ordem interna que reconhecemos à presente investigação, é o facto de estar circunscrita no espaço, mais concretamente a Zona Pedagógica de Lisboa e Península de Setúbal. Embora reconheçamos a importância de alargar o âmbito geográfico da presente investigação, considerámos, nesta fase, que faria mais sentido, começarmos por um território de menores dimensões e, posteriormente, alargá-lo a outras regiões do país.

Ainda assim, e considerando todas as limitações identificadas, consideramos que os resultados alcançados não foram comprometidos e os objetivos estipulados foram globalmente alcançados.

Desenvolvimentos futuros

A nosso ver, o presente estudo não se assume como um ponto de chegada, mas sim como um ponto de partida para projetos de investigação futura. Assim, como desenvolvimentos futuros, esperamos:

- Alargar o âmbito territorial desta investigação até conseguir uma cobertura nacional que permita caracterizar a literacia em avaliação dos professores do ensino básico e secundário portugueses;
- Realizar outros estudos com recurso ao QALA de forma a, por um lado, melhorá-lo e, por outro, conferir-lhe melhores qualidades psicométricas;
- Complementar a investigação realizada com estudos de natureza qualitativa;
- Desevolver um estudo com professores em formação, de forma a analisar e compreender o contributo da formação inicial de professores no desenvolvimento da literacia em avaliação;
- Comparar as diferenças entre susbsistema Universitário e Politécnico no que à formação em avaliação dos futuros professores diz respeito.

Por último, gostariamos que este estudo fosse um ponto de partida para uma reflexão aprofundada e alargada sobre as questões da profissionalidade docente, em especial com as questões relacionadas com a avaliação das aprendizagens.

Conforme se pode constatar ao longo deste documento, as limitações e fragilidades dos professores quanto à tarefa de avaliar não é uma realidade exclusiva portuguesa. Este facto deriva, entre outros aspetos, de uma certa negligenciação, por parte das instituições de ensino superior, em relação às questões relacionadas com a avaliação na formação inicial de professores, pelo que urge uma revisão dos programas curriculares de forma a dar um maior enfâse a estes aspetos.

Bibliografia

Abell, S., & Siegel, M. (2011). Assessment Literacy: What science teachers need to know and be able to do. In D. Corrigan et al. (Eds.), *The Professional Knowledge Base of Science Teaching*. New York: Springer.

Afonso, A. (1998). *Políticas Educativas e Avaliação Educacional*. Braga: Edição do Centro de Estudos em Educação e Psicologia.

Afonso, A. (2011). *Concepções e práticas de avaliação de professores de Ciências da Natureza do 2ºCiclo do Ensino Básico: Um olhar dirigido para os testes de avaliação* (Dissertação de Mestrado). Disponível no Repositório Institucional do Instituto Politécnico de Bragança em http://hdl.handle.net/10198/6158.

Albuquerque, T., & Oliveira, E. (2008). *Avaliação da Educação e da Aprendizagem*. Curitiba: IESDE Brasil S.A.

Alexandre, N., & Coluci, M. (2011). Validade de conteúdo nos processos de construção e adaptação de instrumentos de medida. *Ciência & Saúde Coletiva*, 16(7), 3061-3068.

Aliaga, M., & Gunderson, B. (2002). *Interactive Statistics*. London: Sage Publications.

Alkharusi, H. (2011a). Psychometric properties of the teachers' assessment literacy:

Questionaire for preservice teachers in Oman. *Procedia – Social and Behavioral Sciences*, 29, 1614-1624.

Alkharusi, H. (2011b). Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teacher experience and in-service assessment training. *Journal of Turkish Science Education*, 8(2), 39-48.

Alkharusi, H., Aldhafri, S., Alnabhani, H., & Alkalbani, M. (2012). Educational Assessment Attitudes, Competence, Knowledge and Practices: An Exploratory study of Muscat Teachers in the Sultanate of Oman. *Journal of Education and Learning*, 1 (2), 217-232.

Almiro, P.(2017). Uma nota sobre a desejabilidade social e o enviesamento de respostas. *Avaliação Psicológica*, 16(3).

Alves, M. (2004). *Currículo e Avaliação: Uma perspectiva integrada*. Porto: Porto Editora.

American Federation of Teachers, National Council on Measurement in Education, National Education Association (1990). The Standards for Competence in the Educational Assessment of Students. Acedido em 29 de Março de 2017 em http://files.eric.ed.gov/fulltext /ED323186.pdf.

Amigues, R., & Zerbato-Poudou, M.(1996). *Les practiques scolaire d'apprentissage et d'évaluation*. Paris: Dunot.

Andrade, H.(2019). A critical review of research on student self-assessment. *Frontiers in Educations*, 4(87), 1-13.

Andrade, H., Brookhart, S. (2016). The role of classroom assessment in supportive self-regulated learning. In Dany Laveault and Linda Allal (Eds.), *Assessment*

for Learning: Meeting the challenge of implementation, Londres: Springer.

Angoff, W. (1988). Validity: An evolving concept. In H. Wainer and H. I. Braun (Eds.), *Test validity*, New Jersey: Lawrence Erlbaum Associates.

Babbie, E. (2003). *Métodos de Pesquisa de Survey*. Belo Horizonte: Editora UFMG.

Bayer, S., Klieme, E., & Jude, N. (2016). Assessment and Evaluation in Educational Contexts. In Susan Kuger, Eckhard Klieme, Nina Jude and David Kaplan (Eds.), *Assessing contexts of learning: An International Perspective*, Switzerland: Springer.

Bell, B., & Cowie, B. (2002). *Formative assessment and science education*. New York: Kluwer Academic Publishers.

Bennet, R. (2011). Formative assessment: A critical review.*Assessment in Education: Principles, Policy & Practice*, 18 (1), 5-25.

Berry, R. (2008). *Assessment for Learning*. Hong Kong: Hong Kong University Press.

Bessa, N. (2007). Validade - o conceito, a pesquisa, os problemas de provas geradas pelo computador. *Estudos em Avaliação Educacional*, 18 (37), 115-156.

Biggs, J. (1996). Assessing learning quality: reconciling institutional, staff and educational demands.*Assessment & evaluation*, 12 (1), 5-16.

Black, P., & William, B. (1998a). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5 (1), 7-74.

Black, P., & William, B. (1998b). Inside the Black Box. *Phi Delta Kappan*, 80 (2), 139-148.

Black, P., Harrison, C., Lee, C., Marshall, B.. & Wiliam, D. (2003) *Assessment for*

learning: Putting it into practice. Maidenhead: Open University Press.

Black, P., Harrison, C., Marshall, B., & William, D. (2004). Working inside de Black Box - Assessment for learning in the classroom. *Phi Delta Kappan*, 86 (1), 8-21.

Bloom, B., Hastings, J., & Madaus, G. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bond, T., & Fox, C.(2015). *Applying the Rasch Model - Fundamental Measurement in the Human Science*. New York: Routledge.

Breziat, T., & Coleman, B.(2015). Classroom assessment literacy: Evaluating pre-service teachers. *The Research*, 27 (1), 25-30.

Brookhart, S. (2011). Educational Assessment Knowledge and Skills for Teachers. *Educational Measurement: Issues and Practice*, 30 (1), 3-12.

Brown, G. (2008). Assessment literacy training and teachers' conceptions of assessment. In C.M. Rubic-Davis and C. Rawlinson (Eds.), *Challenging Thinking about Teaching and Learning*, New York: Nova Science Publishers.

Brown, T. (2015). *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford Press.

Brown, T., & Bonsaksen, T. (2019). An examination of the structural validity of the Physical Self-Description Questionnaire-Short Form (PSDQ–S) using the Rasch Measurement Model. *Cogent Education*, 6 (1), 1-28.

Brualdi, A. (1998). Classroom questions. *Practical Assessment, Research & Evaluation*, 6 (6), 1-3.

Cadime, I., Santos, S., Leal, T., Viana, F., Rodrigues, B., Cosme, M., & Ribeiro, I.

(2017). Compreensão de textos: diferenças em função da modalidade de apresentação da tarefa, tipo de texto e tipo de pergunta. *Análise Psicológica*, 3 (35), 351-366.

Campbell, C., Murphy, J., & Holt, J. (2002). Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Cardinet, J. (1983). *Des instruments d'évaluation pour chaque fonction*. Neuchatel: IRDP.

Chen, W., & Thissen, D.(2019). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, 22 (3), 265-289.

Child, D. (2006). *The essentials of factor analysis*. New York: Continuum.

Cochram, K., King, R., & DeRuiter, J. (1991). Pedagogical Content Knowledge: A tentative model for teacher preparation. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Illinois.

Cortez, P., & Veiga, H.(2019). Intenção empreendedora na universidade. *Ciências Psicológicas*, 13 (1), 134-149.

Cosme, A., Ferreira, D., Sousa, A., Lima, L., & Barros, M. (2020). *Avaliação das aprendizagens: Propostas e Estratégias de Ação*. Porto: Porto Editora.

Costelo, A., & Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendation for getting the most from your analysis. *Practical*

Assessment, Research & Evaluation, 10 (7), 1-9.

Couto, G., & Primi, R. (2011). Teoria de Resposta ao Item (TRI): Conceitos elementares para itens dicotómicos. *Boletim de Psicologia*, 61 (134), 1-15.

Cowie, B.(2013). Assessment in the science classroom: priorities, practices and prospects. In James H. Macmillan (Eds.), *Research on Classroom Assessment*, London: SAGE.

Creemers, B., Kyriakides, L., & Sammons, P. (2010). *Methodological advances in educational effectiveness research*. New York: Routledge.

Creswell, J. (2012). *Educational research: planning, conducting and evaluating quantitative and qualitative research*. Boston: Pearson.

Damásio, B.(2012). Uso da análise fatorial exploratória em psicologia. *Avaliação Psicológica*, 11 (2), 213-228.

Dancey, C., & Reidy, J. (2017).*Statistics without maths for psychology*. Harlow: Pearson.

Daniel, L., & King, D. (1998). Knowledge and Use of Testing and Measurement Literacy of Elementary and Secondary Teachers. *The Journal of Educational Research*, 91 (6), 331-344.

DeBlassie(1974).*Measuring and evaluating pupil progress*. New York: MSS Information Corporation.

DeLuca, C., & Klinger, D. (2010). Assessment literacy development: identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy and Practice*, 17 (4), 419-438.

DeLuca, C., Chavez, T., Bellara, A., & Cao, C. (2013). Pedagogies for preservice assessment education: Supporting teacher candidates' assessment literacy development. *The Teacher Educator*, 48 (2), 128-142.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2015). Teacher assessment literacy: a review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, 28 (3), 251-272.

Demo, P. (2008). *Avaliação Quantitativa*. São Paulo: Autores Associados.

Doel, M., Sawdon, C., & Morrison, D. (2002). *Learning, practice and assessment: signposting the portfolio*. London: Jessica Kingsley Publishers.

Edmonds, W., & Kennedy, T. (2017). *An applied guide to research designs: Quantitative, qualitative and mixed methods*. London: Sage Publications.

Engelsen, K., & Smith, K. (2014). Assessment Literacy. In Claire Wyatt- Smith, Valentina Klenowski e Peta Colbert (Eds.), *Designing Assessment for Quality for Learning*, New York: Springer.

Eyal, L. (2012). Digital Assessment Literacy – the Core Role of the Teacher in a Digital Environment. *Educational Technology and Society*, 15 (2), 37-49.

Falchikov, N. (2004). Involving students in assessment. *Psychology Learning and Teaching*, 3(2), 102-108.

Falchikov, N. (2005). *Improving assessment through student involvement*. New York: RoutledgeFalmer.

Fautley, M., & Savage, J. (2008). *Assessment for learning and teaching in secondary schools*. Exeter: Learning Matters.

Fernandes, D. (2004). *Avaliação das aprendizagens: Uma Agenda, Muitos Desafios*. Lisboa: Texto Editores.

Fernandes, D. (2005). *Avaliação das Aprendizagens: Desafios às Teorias, Práticas e Políticas*. Lisboa: Texto Editores.

Fernandes, D. (2013). Avaliação em Educação: Uma discussão de algumas questões críticas e desafios a enfrentar nos próximos anos. *Ensaio: Avaliação e Políticas Públicas em Educação*, 21 (78), 11-34.

Ferraz, A., & Belhot, R. (2010). Taxonomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais. *Gestão & Produção*, 17(2), 421-431.

Ferreira, C. (2007). *A avaliação no quotidiano da sala de aula*. Porto: Porto Editora.

Ferreira, C. (2018). Instrumentos de avaliação para a melhoria do ensino e da aprendizagem. *Revista Eletrónica de Educação e Psicologia*, 8, 12-17.

Ferreira, M. (2013). A ética da investigação em ciências sociais. *Revista Brasileira de Ciência Política*, 11, 169-191.

Fidalgo, A., & Scalon, J. (2012). Uso dos métodos Mantel-Haenszel para a detecção do funcionamento diferencial dos itens e software relacionado. *Psicologia: Reflexão e Crítica*, 25 (1), 66-68.

Figari, G. (1996). *Avaliar: Que referencial?*. Porto: Porto Editora.

Finder, M. (2004). *Educating America: How Ralph W. Tyler taught America to teach*. Connecticut: Praeger Publishers.

Fisher, D., & Frey, N. (2007). *Checking for understanding: formative assessment*

techniques for your classroom. Alexandria (EUA): ASCD.

Fisher, W. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transaction*, 21 (1), 1095.

Fives, H., & DiDonato-Barnes, N. (2013). Classroom Test Construction: The power of a table of specifications. *Practical Assessment, Research & Evaluation*, 18 (3), 1-7.

Fonseca, J., Carvalho, C., Conboy, J., Salema, H., Valente, M.; *et al.* (2015). Feedback na prática letiva: uma oficina de formação de professores. *Revista Portuguesa de Educação*, 28(1), 177-199.

Franco, M., Anguita, L., Sanz, I., & Hidalgo, P. (2020). Development and Psychometric Properties of the Pressure Injury Prevention Knowledge Questionnaire in Spanish Nurses. *International Journal of Environmental Research and Public Health*, 17(2), 1-16.

Frey, N., & Fisher, D. (2011). *The formative assessment action plan: practical steps to more sucessful teaching and learning*. Alexandria (EUA): ASCD.

Fulton, J., Kuit, J., Sanders, G., & Smith, P. (2013). *The Professional Doctorate: A Practical Guide*. New York: Palgrave Macmillan.

Furlan, M.(2007). *Avaliação da aprendizagem escolar: Covergências e desafios*. São Paulo: Annablume Editora.

Gareis, C., & Grant, L. (2015). *Teacher-made assessments: How to connect curriculum, instruction and student learning*. Londres: Routledge.

Gipps, C. (1994). *Beyond Testing: Towards a theory of educational assessment*. Londres: The Falmer Press.

Gipps, C., & Stobart, G. (2009). Fairness in Assessment. In Claire Wyatt-Smith and Joy Cumming (Eds.), *Educational Assessment in the 21st Century: Connecting Theory and Practice*, New York: Springer.

González-de-Paz, L., Kostov, B., López-Pina, J., Solans-Julian, P., Navarro-Rubio, M., & Sisó-Almirall, A. (2018). A Rasch analysis of patients' opinions on primary health care professionals' ethic behaviour with respect to communication issues. *Family Practice*, 32(2), 237-243.

Gomes, C., & Borges, D.(2009). Propriedades psicométricas do conjunto de testes de habilidade visuo espacial. *Psico*, 14(1), 134-149.

Gorard, S. (2001). *Quantitative methods in educational research: the role of numbers made easy*. London: Continuum.

Gotheiner, D., & Siegel, M. (2012). Experienced Middle School Science Teachers' Assessment Literacy: Investigating Knowledge of students conceptions in Genetics and ways to shape instruction. *Journal of Science Teacher Education*, 23, 531-557.

Granger, C. (2007). Rasch analysis is important to understand an use for measurement. *Rasch Measurement Transaction*, 21(3), 1122-1123.

Green, K., & Franton, C. (2002). Survey development and validation with the Rasch Model. International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC.

Guba, E., & Lincoln, Y. (1989). *Fourth Generation Evaluation*. California: Sage Publications.

Guilford, J., & Frutcher, B. (1978). *Fundamental Statistics in Psychology and Education*.

Auckland: McGraw-Hill International.

Gullickson, A. (1984). Teacher perspectives of their instructional use of tests. *The Journal of Educational Research*, 77(4), 244-248.

Gullickson, A. (1985).Student evaluation techniques and their relationship to grade and curriculum. *The Journal of Educational Research*, 79(2), 96-100.

Gustafsson, J. (2010). Longitudinal designs. In B. Creemers, L. Kyriakides and P. Sammins (eds), *Methodological advances in educational effectiveness research*. New York: Routledge.

Hadji, C. (2001). *Avaliação Desmistificada*. Porto Alegre: Artmed.

Hailaya, W., Alagumalai, S., & Ben. F. (2014). Examining the utility of Assessment Literacy Inventory and its portability to education systems in the Asia Pacific region. *Australian Journal of Education*, 58 (297), 297-317.

Hair, J., Black, W., Babin, B., Anderson, R., & Tathan, R. (2006). *Multivariate Data Analysis.* New Jersey: Pearson Prentice Hall.

Hamon, A., & Mesbah, M. (2002). Questionnaire Reliability under the Rasch Model. In M. Mesbah, B. Cole, M. Lee (Eds.), *Statistical methods for quality of life studies*, Boston: Kluwer Academic.

Harlen, W. (2007). *Assessment of Learning.* Londres: Sage Publications.

Harlen, W., & Crick, R. (2002). A systematic review of the impact of summative and tests on students' motivation for learning. In *Research Evidence in Education Library*, Issue 1, Londres: EPPI-Centre, Social Sciece Research Unit, Institute of Education.

Harrington, D. (2009). *Confirmatory Factor Analysis*. New York: Oxford University Press.

Hattie, J.(1985). Methodology Review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9 (2), 139-164.

Hattie, J. (1999). *Influences on student learning*. Inaugural professorial address, University of Auckland, New Zealand. Extraído (a 24 de novembro de 2018) de: https://cdn.auckland.ac.nz/assets/education/hattie/docs/influences-on-student-learning.pdf

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77 (1), 81-112.

Henderson, E. (1978). *The evaluation of in-service teacher training*. Londrea: Croom Helm ltd.

Hill, M., & Hill, A. (2002). *Investigação por Questionário.* Lisboa: Edições Sílabo.

Huberman, M. (2015).Professional careers and professional development: Some intersections. In T. Guskey and M. Huberman (Eds.), *Professional development in Education*, Nova Iorque: Teachers College Press.

Irons, A. (2008). *Enhancing learning through formative assessment and feedback*. Nova Iorque: Routledge.

Jeong, H. (2013). Defining assessment literacy: Is it different for language testers and non-language testers?. *Language Testing*, 30 (3), 345-362.

Jiang, H. (2015). *Learning to Teach with Assessment: A Student Teaching Experience in China*. New York: Springer.

Jorro, A. (2000). *L'enseignant et l'évaluation*. Bruxelas: Éditions De Boeck Université.

Kerlinger, F. (1979). *Behavioral Research - a conceptual approach*. New York: Holt, Rinehart and Winston.

Khadjeh, B., & Amir, R. (2015). Importance of teachers's assessment literacy. *International Journal of English Language Education*, 1 (3), 139-143.

Klooster, P., Taal, E., & Laar, M. (2008). Rasch analysis of the Dutch health assessment questionnaire disability index and the health assessment questionnaire II in patients with rheumatoid arthritis. *Arthritis & Rheumatism*, 59 (12), 1721-1728.

Koh, K. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education*, 22 (3), 255-276.

Kronowitz, E. (2004). *Your first year of teaching and beyond*. Boston: Pearson.

Kubiszy, T., & Borich, G. (1996). *Educational testing and measurement: Classroom application and practice*. New York: HarperCollins.

Lah, N., & Tasir, Z. (2018). Measuring Reliability and Validity of Questionnaire on Online Social Presence: A Rasch Model Analysis. *Advanced Science Letters*, 24(11), 7900-7903.

Lam, R.(2018). *Portfolio assessment for the teaching and learning of writing*. Singapura: Springer.

Landsheere, G.(1976). *Avaliação contínua e exames: Noções de docimologia*. Coimbra: Almedina.

Laros, J. (2012). O uso da Análise Fatorial: Algumas diretrizes para pesquisadores. In L. Pasquali (Ed.), *Análise Fatorial para pesquisadores*, Brasilia: LabPAM

Saber e Tecnologia.

Lent, R. (2010). *Cem Bilhões de Neurônios? Conceitos Fundamentais de Neurociência*. São Paulo: Atheneu.

Lian, L., Yew, W., & Meng, C. (2014). Enhancing Malaysian Teachers' Assessment Literacy. *International Education Studies*, 7 (10), 74-81.

Linacre, J. (2020). *A user's guide to Winstep Ministep Rasch-model computer programs*. Chicago: Winsteps.com.

Linacre, J. (2002). Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement*, 3 (1), 85-106.

Linacre, J., & Wright, B. (2002). Understanding Rasch measurement: Construction of measures from many-facets. *Journal of Applied Measurement*, 3 (2), 486-512.

Lomax, R. (1996). On becoming assessment literate: An initial look at preservice teachers' beliefs and practices. *The Teacher Educator*, 31 (4), 292-303.

Lopes, J., Silva, H. (2010). *O professor faz a diferença*. Lisboa: Lidel - Edições Técnicas.

Lucea, J. (2005). *La Evaluación formativa como instrumento de aprendizaje en Educación Física*. Barcelona: INDE Publicaciones.

Luckesi, C. (2005). *Avaliação da aprendizagem escolar: Estudos e proposições*. São Paulo: Cortez Editora.

Maia, L. (2012). El Modelo de Rasch aplicado a las Ciencias Psicológicas. *Revista Eletrónica de Psicologia, Educação e Saúde*, 2 (1), 1-34.

Malone, M. (2013). The esssentials of assessment literacy: Contrasts between testers

and users. *Language Testing*, 30 (3), 329-344.

Marinho, P., Fernandes, P., & Leite, C. (2014). Avaliação da aprendizagem: da pluralidade de enuncialão à dualidade de conceções. *Acta Scientaum - Education*, 36 (1), 153-164.

Marôco, J. (2018). *Análise Estatística com o SPSS Statistics*. Pêro Pinheiro: ReportNumber.

Martins, G. (2006). *Sobre confiabilidade e validade. RBGN*, 8 (20), 1-12.

Mathison, S. (Ed.) (2005). *Encyclopedia of evaluation*. California: Sage Publications.

McDonald, R., & Ahlawat, K.(1974). Difficulty factor in binary data. *British Journal of Mathematical and Statistical Psychology*, 27 (1), 82-99.

McGee, J., & Colby, S. (2014). Impact of an assessment course on teacher candidates' assessment literacy. *Action in Teacher Education*, 36 (5-6), 522-532.

Mertens, D. (2010). *Rersearch and evaluation in education and psychology: integrating diversity with quantitative, qualitative and mixed methods*. London: Sage Publications.

Mertler, C. (2003). *Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference?* Paper presented at the meeting of the Mid-Western Educational Research Association, Columbus, Ohio.

Mertler, C. (2004). Teachers' assessment literacy: Does classroom experience make a difference?. *American Secondary Education*, 33 (1), 49-64.

Mertler, C., & Campbell, C. (2005). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the Assessment Literacy*

Inventory. Paper presented at the meeting of the American Educational Research Association, Montreal, Canada.

Meyer, J. (2014). *Applied measurement with jMetrik*. New York: Routledge.

Miguel, J. (2013). *Teoria de Resposta ao Item: Representação e utilidade do modelo logístico de traço latente na psicometria actual* (Tese de Doutoramento). Disponível no Repositório Institucional da Universidade de Coimbra em http://hdl.handle.net/10316/24834.

Mofreh, S., Ghafar, M., Omar, A., Mosaku, M., & Ma'ruf, A. (2014). Psychometric Properties on Lecturers' Beliefs on Teaching Function: Rasch Model Analysis. *International Education Studies*, 7(11), 47-55.

Mohamad, M., Sulaiman, N., Sern, L., & Salleh, K.(2014). Measuring the Validity and Reliability of Research Instruments, artigo apresentado no 4th World Congress on Technical and Vocational Education and Training, Malásia (pp. 164-171). Procedia - Social and Beahavioral Sciences.

Monteiro, S., & Pissaia, L. (2018). Anedotário como ferramenta facilitadora no processo de avaliação escolar. *Revista Signos*, 39 (2), 104-114.

Monteiro, V., & Fragoso, R. (2005). *Avaliação entre pares*. Atas do VIII Congresso Galaico-Português de Psicologia, Instituto de Educação e Psicologia da Universidade do Minho, Braga.

Moon, J. (2004). *A Handbook of Reflective and Experiental Learning - Theory and Practice*. New York: RoutledgeFalmer.

Moreira, J. (2004). *Questionários: Teoria e Prática*. Coimbra: Almedina.

Muijs, D. (2004). *Doing quantitative research in education*. London: Sage Publications.

Nair, R., Moretin, B., & Lincoln, N. (2011).Rasch analysis of the Nottingham extended activities of daily living scale. *Journal of Rehabilitation Medicine*, 43 (10), 944-950.

Nérici, I. (1983). *Introdução à didática geral: Dinâmica da escola*. Rio de Janeiro: Editora Científica.

Neto, A., & Aquino, J. (2009). A avaliação da aprendizage como ato amoroso: o que o professor pratica?. *Educação em Revista*, 25 (2), 223-240.

Neuman, W. (2007). *Basics of Social Research: Qualitative and quantitative approaches*. Boston: Pearson Education.

Neves, A., & Ferreira, A. (2015). *Avaliar é preciso? Guia prático de avaliação para professores e formadores*. Lisboa: Guerra e Paz.

Newfields, T. (2006). Teacher development and assessment literacy. Authentic Communication: Proceedings of the 5th Annual JALT Pan-SIG Conference., 48-73.

Nezvalová, D. (2010). *Assessing science for understanding*. Palacký University: Olomouc.

Nogueira, G., Seidl, E., & Troccoli, B. (2016). Análise Fatorial Exploratória do Questionário de Perceção de Doenças versão breve (Brief IPQ). *Psicologia: Teoria e Pesquisa*, 32 (1), 161-168.

Nezvalová, D. (2010). *Assessing Science for understanding*. Olomouc: Palacký University.

Ogan-Bekiroglu, F., & Suzuk, E. (2014). Pre-service teachers' assessment literacy and its implementation into practice. *The Curriculum Journal*, 25 (3), 344-371.

OECD (2005). *Formative Assessment: Improving learning in secondary classroom*. OECD Publishing.

OECD (2011). *OECD Reviews of Evaluation and Assessment in Education: Norway 2011*. OECD Publishing.

Ozan, C., & Kincal, R.(2014). The effects of formative assessment on academic achievement, attitudes toward the lesson , and sel-regulation skills. *Educational Sciences: Theory and Practice*, 18 (1), 85-118.

Paterno, J. (2001). Measuring success: A glossary of assessment terms. In Building cathedrals: Compassion for the 21st century. Disponível em http://www.angel-fire.com/wa2/buiding cathedrals/measuringsuccess.html.

Perrenoud, P. (1996). *La construcción del éxito y del fracaso escolar*. Madrid: Ediciones Morata.

Perrenoud, P. (2001). Les trois fonctions de l'évaluation dans une scolarité organisée en cycles. *Éducateur*, 2, 19-25.

Pestana, M., & Gageiro, J. (2003). *Análise de dados para ciências sociais - A complementaridade do SPSS*. Lisboa: Edições Sílabo.

Pinhal, M. (2000). *Projecto Falar, Avaliação em Língua Portuguesa* [Página Web]. Disponível de https://area.dge.mec.pt/gramatica/lourdespinhal.htm.

Pinto, J., & Santos, L. (2006). *Modelos de Avaliação das Aprendizagens*. Lisboa: Universidade Aberta.

Plake, B., & Impara, J. (1992). *Teacher competencies questionnaire description*. Lincoln: University of Nebraska.

Plake, B., Impara, J., & Fager, J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, 12(4), 10-12.

Popham, W. (1995). *Classroom assessment: what teachers need to know*. Boston: Allyn-Bacon.

Popham, W. (2006). Needed: A dose of assessment literacy. *Educational Leadership*, 63, 84-85.

Popham, W. (2009). Assessment Literacy for Teachers: Faddish or Fundamental?. *Theory into Practice*, 48, 4-11.

Popham, W. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, 46 (4), 265–273.

Popham, W. (2018). *Assessment literacy for educators in a hurry*. Alexandria (Virginia): ASCD.

Prieto, G., & Delgado, A. (2003). Análisis de un test mediant em modelo de Rash. *Psicothema*, 15 (1), 94-100.

Proença, M. (1989). *Didáctica da História*. Lisboa: Universidade Aberta.

Quilter, S., & Gallini, J. (2000). Teacher assessment literacy and attitudes. *The Teacher Educator*, 36 (2), 115-131.

Race, P., Brown, S., & Smith, B. (2005). *500 tips on assessment*. Londres: RoutledgeFalmer.

Race, P. (2009). *Designing assessment to improve physical sciences learning*. Hull: Higher Education Academy.

Ramesal, A. (2011). Primary and secondary teachers' conceptions of assessment: a qualitative study. *Teaching and Teacher Education*, 27, 472-482.

Rampazzo, S. (2011). Instrumentos de avaliação: Reflexões e possibilidades de uso no processo de ensino e aprendizagem. In Secretaria de Estado da Educação do Paraná, *O professor PDE e os desafios da escola pública paranaense*, Volume II, Londrina: Produção Didático-Pedagógica.

Ribeiro, L. (1990). *Avaliação da aprendizagem*. Lisboa: Texto Editora.

Ribeiro, A., & Ribeiro, L. (1990). *Planificação e Avaliação do Ensino-Aprendizagem*. Lisboa: Universidade Aberta.

Ribeiro, L. (1993). *Avaliação da aprendizagem*. Lisboa: Texto Editora.

Rice, M. (2011). Examining my assessment literacy instruction practices with teacher candidates. *Brock Education*, 21 (1), 87-97.

Robison, M., Johnson, A., Walton, D., & MacDermid, J.(2019). A comparison of the polytomous Rasch analysis output of RUMM2030 and R (ltm/eRm/TAM/lordif). *BMC Medical Research Methodology*, 19 (1), 1-12.

Rogier, D. (2014). Assessment Literacy: Building a Base for Better Teaching and Learning. *English Teaching Forum*, 3, 2-13.

Rolheiser, C., Bower, B., & Stevahn, L.(2000). *The portfolio organizer: succeeding with portfolios in your classroom*. Alexandria: ASCD.

Ruiz-Primo, M., & Furtak, E. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching*, 44(1), 57-84.

Ruiz-Primo, M., Solano-Flores, G., & Li, M. (2014).Formative assessment as a process of interaction through language. In Claire Wyatt-Smith, Valentina Klenowski e Peta Colbert (Eds.), *Designing Assessment for Quality for Learning*, New York: Springer.

Sadler, D. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119-144.

Salmon-Cox, L. (1981). Teachers and standardized tests: What's really happening? *Phi Delta Kappan*, 62(9), 631-634.

Sant'Anna, I.(1995). *Por que avaliar? Como Avaliar? Critérios e Instrumentos*. Rio de Janeiro: Vozes.

Santiago, P., Donaldson, G., Looney, A., & Nusche, D.(2012). *OECD Reviews of Evaluation and Assessment in Education: Portugal 2012*. OECD Publishing.

Santos, P. (2016). Avaliação escolar para além da classificação: Perspectivas, desafios e apontamentos. *Saberes docentes em ação*, 2 (1), 15-27.

Santos, L. (2016). A articulação entre avaliação somativa e a formativa na prática pedagógica: uma impossibilidade ou um desafio?. *Ensaio-Avaliação e Políticas Públicas*, 24 (94), 637-669.

Santos, L., & Pinto, J. (2018). Ensino de conteúdos escolares: A avaliação como fator estruturante. In F. Veiga(coord.), *O Ensino como fator de envolvimento numa escola para todos*, Lisboa: Climepsi Editores.

Sartes, L., & Souza-Formigoni (2013). Avanços na Psicometria: da Teoria Clássica de Testes à Teoria de Resposta ao Item. *Psicologia: Reflexão e Crítica*, 26 (2), 241-250.

Scales, P. (1993). How teachers and education deans rate the quality of teacher preparation for the middle grades. *Journal of Teacher Education*, 44(5), 378-383.

Schafer, W., & Lissitz, R. (1987). Measurement training for school personnel: Recommendations and reality. *Journal of Teacher Education*, 38(3), 57-63.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagné e M. Scriven (Eds.), *Perspectives of curriculum evaluation*, Chicago: Rand McNally.

Shores, E., & Grace, C. (2001). *Manual de Portfólio: um guia passo a passo para o professor*. Porto Alegre: Artmed.

Shulman, L. (1986). Those who understand: Knowledge Growth in Teaching. *Educational Reseacher*, 15 (2), 4-14.

Siegel, M., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education*, 22, 371-391.

Silva, M., & Vendramini (2006). Evidências de validade de uma escala de autoconceito académico em estatística. *Educação Matemática Pesquisa*, 8 (1), 177-196.

Silva, M., & Souza, N. (2007). Portfólio: limites e possibilidades em uma avaliação formativa. *EDUCERE*, 7, 1295-1307.

Silva, H., & Lopes, J. (2015). O questionamento eficaz na sala de aula: Procedimentos e estratégias. *Revista Eletrónica de Educação e Psicologia*, 5, 1-17.

Sisto, F., Rueda, F., & Bartholomeu, D. (2006). Estudo sobre a unidimensionalidade do Teste Matrizes Progressivas Coloridas de Raven. *Psicologia: Reflexão & Crítica*, 19(1), 66-73.

Slaney, K. (2015).”I’m not that kind of Psychologist”: A caso for Methodological Pragmatism in Theoretical Inquiries into Psychological Science Practices. In Jack Martin, Jeff Sugarman and Kathleen Slaney (Eds.), *Theoretical and Philosophical Psychology: Methods, Approaches and New Directions for Social Sciences*, West Sussex: Wiley Blackwell.

Sohlberg, P., Czaplicka, M., Lindblad, S., Houtsonen, J., Müller, J., Morgan, M., et al. (2007). *Professional expertise under restructuring: comparative studies of education and health care: the survey study*. ProfKnow, EU sixth framework.

Stefanou, C., & Parker, J. (2003). Effects of classroom assessment on student motivation in fifth-grade science. *The Journal of Educational Research*, 96 (3), 152-162.

Stiggins, R., & Bridgeford, N. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22(4), 271-286.

Stiggins, R. (1991). Assessment literacy. *Phi Delta Kappan*, 72, 534-539.

Stiggins, R. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, 77 (3), 238-246.

Stiggins, R. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18, 23–27.

Stiggins, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, 83(10), 758-765.

Stufflebeam, D, Madaus, G., & Kallaghan, T. (2000). *Evaluation Models: Viewpoints on educational and human services evaluations*. Massachhussets: Klewer Academic Publishers.

Südkamp, A., Kaiser, J., & Moller, J. (2014). Teachers' judgements of students' academic achievement. In Sabine Krolak-Schwerdt et al.(Eds.), *Teachers' Professional Development*, Roterdão:Sense Publishers.

Taber, K. (2017). The use of Cronbach's alpha when developing and reporting research instruments in Science Education. *Research in Science Education*, 48, 1273-1296.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21-36.

Tennant, A., & Pallant, J. (2006). Unidimensionality matters. *Rasch Measurement Transactions*, 20, 1048-1051.

Valadares, J., & Graça, M. (1998). *Avaliando para melhorar a aprendizagem*. Venda Nova: Plátano Editora.

Vieira, S. (2009). *Como elaborar questionários*. São Paulo: Editora Atlas.

Vieira, M., Ribeiro, R.,& Almeida, L. (2009). As potencialidades da Teoria de Resposta ao Item na validade dos testes: Aplicação a uma prova de dependência-independência de campo. *Análise Psicológica*, 27(4), 455-462.

Volante, L., & Fazio, X. (2007). Exploring Teacher candidates' assessment literacy: implications for teacher education. *Canadian Journal of Education*, 30 (3), 749-770.

Wiliam, D., & Black, P. (1996). Meanings and Consequences: A basis for distinguishing formative and summative functions of assessment?. *British Educational Research Journal*, 22 (5), 537-548.

Willis, J, Adie, L., & Klenowski, V. (2013). Conceptualising teachers' assessment

literacies in an era of curriculum and assessment reform. *The Australian Association for Research in Education*, 40, 241-256.

Wolf, K. (1993). From informal to informed assessment: Recognizing the role of classroom teacher. *Journal of Reading*, 36 (7), 518-523.

Wragg, E. (2001). *Assessment and learning in the secondary school*. London: Routledge.

Xu, Y., & Brown, G. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149-162.

Xu, Y., & Brown, G. (2017). University English Teacher Assessment Literacy: A survey-test report from China. *Papers in Language Testing and Assessment*, 6(1), 133-158.

Yamtim, V., & Wongwanich, S. (2014). A study of classroom assessment literacy of primary school teachers. *Procedia – Social and Behavioral Sciences*, 116, 2998-3004.

Legislação

Decreto-Lei nº 55/2018, de 6 de julho. Diário da República, nº 129/18 - 1ª Série. Lisboa: Ministério da Educação.

Portaria nº 223-A/2018, de 3 de agosto. Diário da República, nº 149/18 - 1ª Série. Lisboa: Ministério da Educação.

Portaria nº 226-A/2018, de 7 de agosto. Diário da República, nº 151/18 - 1ª Série. Lisboa: Ministério da Educação.

Anexos