Revista Portuguesa de Educação

Qualidades Psicométricas do Questionário de Aferição da Literacia em Avaliação (QALA)

Luis Almeida[i]
Universidade Lusófona, Portugal

RESUMO

A literacia em avaliação é uma característica fundamental que todo o professor deve ter, uma vez que a avaliação dos alunos está numa estreita relação com todo o processo de ensino e aprendizagem. pelo que deverá ser considerada um elemento-chave na melhoria do ensino. O objetivo do presente artigo foi o de analisar as propriedades psicométricas de um instrumento desenvolvido para aferir a literacia em avaliação dos professores do ensino básico e secundário: o QALA (Questionário de Aferição da Literacia em Avaliação). A análise das propriedades psicométricas do QALA foi realizada a partir do Modelo Rasch, que se assume como um conjunto de técnicas estatísticas que se inserem na Teoria de Resposta ao Item (TRI). Os resultados evidenciam as boas qualidades psicométricas do QALA, pelo que o instrumento parece ser adequado para o objetivo a que se propõe, nomeadamente para aferir as perceções que os professores têm sobre os seus conhecimentos e capacidades em avaliação, por um lado, e a sua literacia em avaliação, por outro.

Palavras-chave: Literacia em Avaliação; Avaliação das Aprendizagens; QALA; Propriedades psicométricas; Modelo Rasch.

1. INTRODUÇÃO

A avaliação pedagógica é uma das mais importantes responsabilidades dos professores, assim como uma das tarefas nas quais os professores despendem mais tempo (Ramesal, 2011; Mertler, 2003). A avaliação pedagógica deve ser aqui entendida como o conjunto diversificado de processos que se desenvolvem nas salas de aula com o intuito de contribuir para o desenvolvimento das aprendizagens dos alunos (Fernandes, 2022). Neste contexto, a avaliação pedagógica integra tanto a avaliação formativa (ou avaliação para as aprendizagens) e a avaliação sumativa (ou avaliação das aprendizagens).

A avaliação formativa incide sobre o processo de aprendizagem, sendo uma forma de recolha de informação que permite ao professor perceccionar se os alunos atingiram os 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.

A avaliação formativa pode ser mesmo considerada 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)

Já a avaliação sumativa deverá corresponder a um balanço final, possibilitando uma visão de conjunto relativamente a um todo ao qual, até então, apenas se fizeram juízos parcelares (Ribeiro, 1993). A avaliação sumativa assume-se assim como um elemento complementar à avaliação formativa, na medida em que contribui para uma apreciação mais equilibrada do trabalho realizado pelos alunos.

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, 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 media 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 currículos e a dificuldade em encontrar desafios adequados para as necessidades dos alunos. No entanto, há um aspeto, a nosso ver, muito importante, que está diretamente relacionado com os objetivos deste artigo 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.

O conceito de literacia em avaliação foi primeiramente apresentado por Richard Stiggins (1991) como o conhecimento profundo das questões de avaliação. Do mesmo modo, Stiggins (1995) 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.

Na mesma linha de raciocínio, Popham (2011) refere que a literacia em avaliação trata do entendimento dos conceitos e procedimentos de avaliação fundamentais, suscetíveis 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).

Vários estudos de natureza quantitativa foram conduzidos com o objetivo de medir a literacia em avaliação dos professores, tanto em serviço como no decorrer da formação inicial, como sejam os exemplos do *Teacher Assessment Literacy Questionnaire* (TALQ), desenvolvido por Plake, Impara e Fager (1993), do *Assessment Literacy Inventory* (ALI), desenvolvido por Campbell et al. (2002) e mais tarde atualizado por Mertler e Campbell (2005), e do *Classroom Assessment Literacy Inventory* (CALI), desenvolvido por Mertler (2003). Estes questionários estão alinhados com os *Standards for Competence in Educational Assessment of Students* (American Federation of Teachers, National Council on Measurement in Education, National Education Association [AFT, NCME & NEA], 1990), pelo que são instrumentos que foram construídos considerando a realidade norte-americana. Assim, para a utilização destes instrumentos em outros contextos, seria necessária uma revisão profunda de forma a adequá-los às realidades de cada um desses mesmos contextos (Hailaya et al., 2014).

A investigação que se tem debruçado sobre as questões relacionadas com a literacia em avaliação revela 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 (Xu & Brown, 2016; Koh, 2011; DeLuca & Klinger, 2010) e, por outro, que os professores, seja no início 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 (Yamtim & Wongwanich, 2014; Koh, 2011; Volante & Fazio, 2007). 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.

Mertler (2003) sugere mesmo que os professores em formação raramente frequentam programas que lhes ensinem, por exemplo, 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 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 & Colby, 2014, p. 523).

De forma a avaliar a literacia em avaliação dos professores em Portugal, foi desenvolvido um instrumento o qual designámos por Questionário de Aferição da Literacia em Avaliação (QALA).

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) tendo, gradualmente, ganho 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). 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 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.

2. MÉTODOS

2.1. PARTICIPANTES

Foi utilizado um plano de amostra não-probabilístico, neste caso, inserido na amostragem por conveniência, tendo a recolha de dados ocorrido nos meses de junho e julho de 2020. Respondeu ao questionário um total de 253 professores do Ensino Básico e Ensino Secundário que lecionavam na região de Lisboa e Península de Setúbal (Portugal). Este território, inserido na Área Metropolitana de Lisboa, agrega um elevado número de escolas públicas e privadas/cooperativas nos vários níveis de ensino e, consequentemente, um elevado número de professores. A maioria dos respondentes foi do sexo feminino (79,4%) e o escalão etário mais representativo situou-se entre os 41 e 50 anos (37.9%).

2.2. INSTRUMENTO

O QALA é composto por quatro partes. A primeira parte corresponde à recolha de informações gerais dos respondentes. A segunda parte visa a recolha de informações sobre as perceções dos professores face aos seus conhecimentos e capacidades em avaliação, sendo constituída por 20 itens do tipo *Likert*, com

uma escala que varia de 1 (Discordo Totalmente) a 5 (Concordo Totalmente).

A terceira parte é composta por 40 itens de caráter dicotómico (Verdadeiro/Falso) e visa recolher informações sobre os conhecimentos que os professores têm em relação à avaliação em contexto de sala de aula. Já a quarta parte é constituída por 20 itens de escolha múltipla, e visa recolher informações sobre os conhecimentos dos professores em avaliação face a 5 cenários hipotéticos.

Os itens presentes na segunda, terceira e quarta partes estão organizados em torno de 4 (quatro) domínios da literacia em avaliação, inspirados na proposta de Abell e Siegel (2011), nomeadamente:

a) 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;
b) 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 (ex. Taxonomia de Bloom, de Marzano, *Depth of Knowledge* de N. L. Webb) e o conhecimento de ferramentas de auxílio à construção de instrumentos de avaliação;
c) Conhecimentos sobre construção e 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;
d) Conhecimentos sobre interpretação e utilização da informação recolhida no processo de avaliação: Procura-se, nesta dimensão, verificar os conhecimentos e competências 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.

2.3. ANÁLISE ESTATÍSTICA

Unidimensionalidade: Um dos requisitos para a aplicação do Modelo Rasch é a unidimensionalidade dos dados (Linacre, 2002; Meyer, 2014). Assim, a verificação deste pressuposto 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). 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).

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. Segundo, Linacre (2020), Lah e Tasir (2018) e González-de-Paz et al. (2015), existe dependência local entre itens quando os valores de correlação entre resíduos são superiores a 0,7, uma vez que é a partir desse valor que o par de itens compartilha mais de metade da variância residual.

Limiares de Categoria: Os limiares de categoria correspondem ao valor de habilidade em que uma pessoa tem igual probabilidade de selecionar duas opções de resposta adjacentes (Robison et al., 2019; Meyer, 2014). 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), indicador de que os limites estão bem ordenados. A desordem dos limiares de categorias pode ser indicadora de categorias mal elaboradas ou, mais frequentemente, de excesso de opções de resposta.

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). Assim, é necessário verificar em que medida os itens cobrem uma variada gama de dificuldades, desde os mais fáceis aos mais difíceis. 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 que contenham uma variedade de itens com níveis de dificuldade dispersos entre os fáceis (valores negativos na escala de *logits*) e os difíceis (valores positivos na escala de *logits*).

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 dos dados ao modelo é avaliado pela comparação entre a probabilidade teórica de acerto de cada pessoa a cada item e 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) e o *Unweighted Mean Square* (UMS). 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 particularidade 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 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 lado, 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). Na Tabela 1, apresentam-se os critérios de qualidade dos dois indicadores segundo a proposta de Fisher (2007).

Tabela 1

Critérios de Qualidade dos Índices de Fiabilidade e Separação dos Itens

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

Fonte: Adaptado de Fisher, 2007

Consistência interna: Para a análise da consistência interna das 3 partes do QALA, foi determinado o Coeficiente α.

3. RESULTADOS

3.1. UNIDIMENSIONALIDADE

A Tabela 2 revela que os dados das três partes do QALA apresentam valores concordantes com a unidimensionalidade, já que respeitam os critérios de percentagem de variância explicada do primeiro fator e o valor de eigenvalue do segundo fator.

Tabela 2

Resumo da Análise de Componentes Principais dos Resíduos 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	3.28	2.44	3.54	2.38	2,03	1,49

3.2. INDEPENDÊNCIA LOCAL

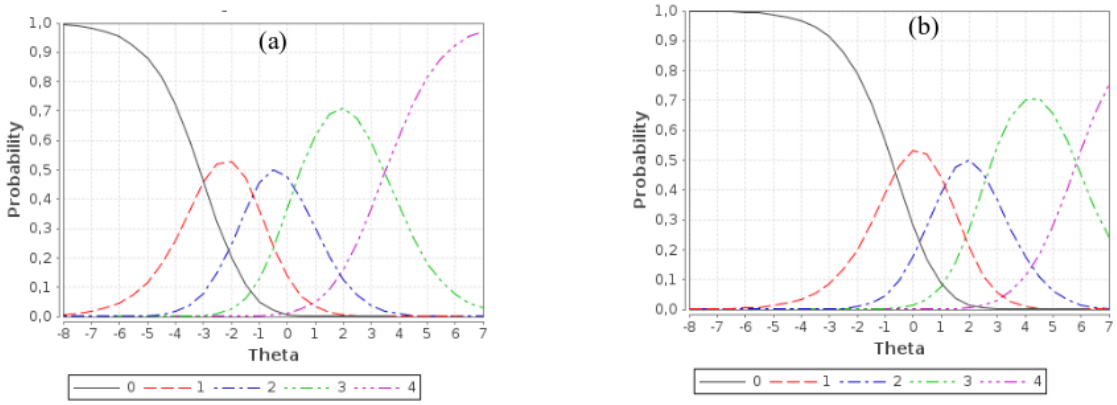
Os itens das Partes 2, 3 e 4 cumprem o requisito da independência local, já que as matrizes de correlações dos resíduos das 3 partes do QALA apresentam valores inferiores ao limiar de 0,7. Podemos assim concluir que as respostas dadas a cada item não estão dependentes das respostas dadas a outros itens.

3.3. LIMIARES DE CATEGORIA

A análise dos limiares de categoria foi realizada apenas à parte 2 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 (ver exemplos das figuras 1a e 1b), 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.

Figura 1

Exemplos das Curvas Características dos Itens 1(a) e 17(b) da Parte 2 do QALA



3.4. DIFICULDADE DOS ITENS

Os valores de dificuldade dos itens (β) do QALA (Tabela 3) variam entre -1,19 e 2,24 (na escala de *logit*) na Parte 2, entre -2,06 e 2,64 na Parte 3 e -3,04 e 2,95 na Parte 4. 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*.

Tabela 3

Dificuldades dos itens (em logits) dos itens QALA

	Parte 2	Parte 3	Parte 4
Média	≈ 0,0	≈ 0,0	≈ 0,0
Desvio-Padrão	0,11	0,15	0,18
Máxima	2,24	2,64	2,95
Mínimo	-1,19	-2,06	-3,04
Itens com $\beta < -2$	0	2	2
$-2 <$ Itens com $\beta < -1$	2	5	4
$-1 <$ Itens com $\beta < 0$	11	13	14
$0 <$ Itens com $\beta < 1$	3	13	4
$1 <$ Itens com $\beta < 2$	2	4	4
Itens com $\beta > 2$	2	3	2

3.5. AJUSTE DOS ITENS

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 (Linacre, 2002; Cadime et. al, 2017).

Uma síntese dos resultados de ajuste dos itens do QALA pode ser visualizada na Tabela 4. 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, a Parte 2 apresenta 19 (dos 20 itens), bem ajustados de acordo com o critério de Linacre (2002), tendo apenas 1 item 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, os valores médios das 3 partes encontram-se igualmente próximos de 1,0. Na Parte 2, dois dos itens apresentam um desajuste moderado, estando os restantes bem ajustados. Na Parte 3, apenas 1 item apresenta um desajuste moderado, sendo inclusive 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.

Tabela 4

Resumo dos Índices de Ajuste dos Itens do QALA

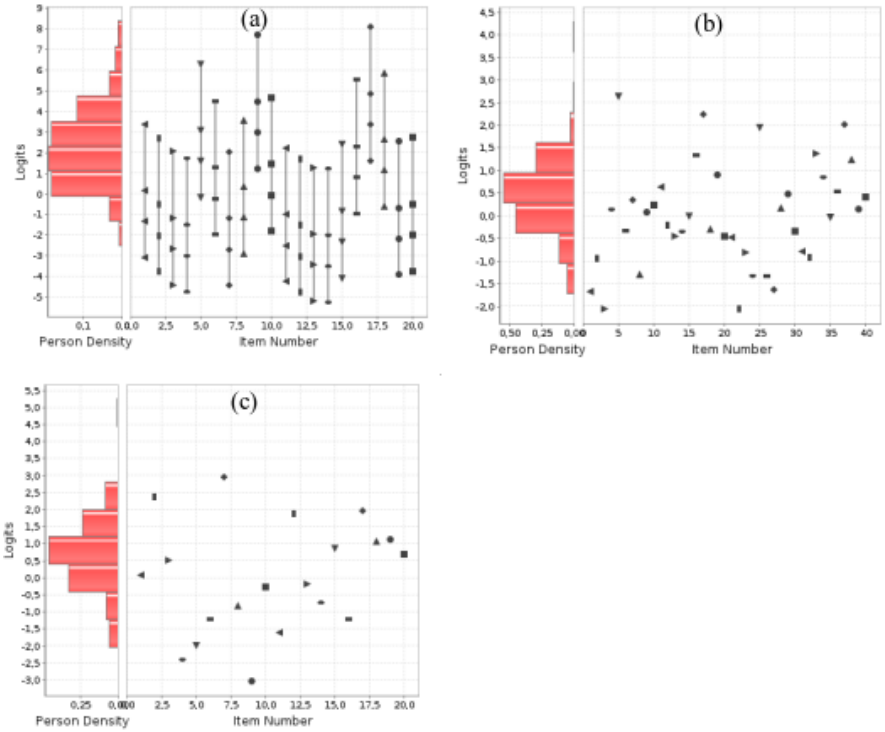
	Parte 2		Parte 3		Parte 4	
	WMS	UMS	WMS	UMS	WMS	UMS
Média	0.99	0.98	0.99	1.03	0.99	0.95
Desvio-padrão	-0.2	-0.18	0.02	0.32	0.18	0.08
Máximo	1.70	1.74	1.14	1.64	1.15	1.43
Mínimo	.62	.60	.83	.75	.82	.45
Itens bem ajustados (0,5-1,5)	19	18	40	39	20	19
Itens com desajuste moderado	1	2	0	1	0	0
Itens com desajuste severo (> 2,0)	0	0	0	0	0	0
Itens pouco produtivos (< 0,5)	0	0	0	0	0	1

3.6. MAPAS ITEM-PESSOA

Na Figura 2 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 direita) e das dificuldades (β) dos itens (à esquerda) na escala de logit. Para além disso, a faixa das dificuldades dos itens sobrepõe-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 2

Mapas Item-Pessoa das Partes 2(a) 3(b) e 4(c) do QALA



3.7. FUNCIONAMENTO DIFERENCIAL DOS ITENS

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 subsistema 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 3º Ciclo e Secundário e os professores do 1º e 2º Ciclos.

Pela análise à tabela 5, verifica-se a inexistência de itens com DIF elevado, pelo que os itens que constituem o QALA funcionam de forma semelhante, independentemente do subsistema e do nível de ensino em que os professores lecionam.

Tabela 5

Resumo dos resultados da análise de DIF do QALA

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

Nota: *Subsistema de Ensino; ** Nível de Ensino

3.8. FIABILIDADE E SEPARAÇÃO DOS ITENS

Os itens que compõem as 3 partes do QALA apresentam excelentes valores de fiabilidade e separação (Tabela 6), de acordo com os critérios definidos por Fisher (Tabela 1). Este aspeto é uma evidência importante da validade de construto e de replicabilidade do QALA.

Tabela 6

Índices de fiabilidade e separação dos itens

	Parte 2	Parte 3	Parte 4
Fiabilidade dos Itens	.988	.981	.987
Separação dos Itens	9.215	7.238	8.670

3.9. 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. 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 (Taber, 2017). 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. DISCUSSÃO E CONCLUSÕES

A aplicação do Modelo de Rasch permitiu avaliar as propriedades psicométricas do QALA à luz da Teoria de Resposta ao Item. 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, a partir das Curvas Características dos Itens, permitiu verificar que o sistema de 5 categorias utilizado 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 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 lado, 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 o valor mais alto (α =0,94) e a Parte 4 o valor mais baixo (α =0,59).

Em jeito de conclusão, os dados obtidos, a partir do Modelo Rasch, parecem evidenciar boas qualidades psicométricas do QALA. Assim, o QALA parece ser adequado para o objetivo a que se propõe, nomeadamente para aferir as perceções que os professores têm sobre os seus conhecimentos e capacidades em avaliação, por um lado, e a sua literacia em avaliação, por outro.

REFERÊNCIAS

Abell, S., & Siegel, M. (2011). Assessment Literacy: What science teachers need to know and be able to do. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), *The Professional Knowledge Base of Science Teaching*. Springer. https://doi.org/10.1007/978-90-481-3927-9_12

American Federation of Teachers, National Council on Measurement in Education, National Education Association (1990). *The Standards for Competence in the Educational Assessment of Students*. http://files.eric.ed.gov/fulltext /ED323186.pdf

Bond, T., & Fox, C. (2015). *Applying the Rasch Model - Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698

Brown, G. (2008). Assessment literacy training and teachers' conceptions of assessment. In C. M. Rubie-Davies, & C. Rawlinson (Eds.), *Challenging Thinking about Teaching and Learning* (pp. 269-285). Nova Science Publishers.

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. https://doi.org/10.1080/2331186X.2019.1571146

Cadime, I., Santos, S., Leal, T., Viana, F., Rodrigues, B., Cosme, M. C., & 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. https://doi.org/10.14417/ap.1234

Campbell, C., Murphy, J., & Holt, J. (2002, October). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers* [Paper presentation]. Annual Meeting of the Mid-Western Educational Research Association, Columbus, OH, USA.

Couto, G., & Primi, R. (2011). Teoria de Resposta ao Item (TRI): Conceitos elementares para itens dicotómicos. *Boletim de Psicologia, 61*(134), 1-15.

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. https://doi.org/10.1080/0969594X.2010.516643

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. https://doi.org/10.1080/08878730.2012.760024

Fernandes, D. (2022). *Avaliar e aprender numa cultura de inovação pedagógica*. Leya Educação.

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. https://doi.org/10.1590/S0102-79722012000100008

Fisher, W. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transaction, 21*(1), 1095.

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. https://dx.doi.org/10.3390%2Fijerph17093063

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. https://doi.org/10.1093/fampra/cmu073

Granger, C. (2007). Rasch analysis is important to understand and use for measurement. *Rasch Measurement Transaction, 21*(3), 1122-1123. https://www.rasch.org/rmt/rmt213d.htm

Green, K., & Franton, C. (2002, Nov. 14-17). *Survey development and validation with the Rasch Model.* [Paper presentation]. International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC, USA.

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*(3), 297-317. https://doi.org/10.1177%2F0004944114542984

Koh, K. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education, 22*(3), 255-276. https://www.tandfonline.com/doi/abs/10.1080/10476210.2011.593164

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. https://doi.org/10.1166/asl.2018.12452

Linacre, J. (2020). *A user's guide to Winstep Ministep Rasch-model computer programs*. 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*(4), 486-512. https://pubmed.ncbi.nlm.nih.gov/12486312/

Luckesi, C. (2005). *Avaliação da aprendizagem escolar: Estudos e proposições*. 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.

McGee, J., & Colby, S. (2014). Impact of an assessment course on teacher candidates' assessment literacy. *Action in Teacher Education, 36*(5-6), 522-532. https://doi.org/10.1080/01626620.2014.977753

Mertler, C. (2003, Oct. 15-18). *Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference?* [Paper presentation]. Annual Meeting of the Mid-Western Educational Research Association, Columbus, Ohio, USA.

Mertler, C. & Campbell, C. (2005, April 11-15). *Measuring teachers' knowledge and application of classroom assessment concepts: Development of the Assessment Literacy Inventory* [Paper presentation]. Annual Meeting of the American Educational Research Association, Montreal, Canada.

Meyer, J. (2014). *Applied measurement with jMetrik.* Routledge.

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. https://doi.org/10.5539/ies.v7n11p47

Mohamad, M., Sulaiman, N., Sern, L., & Salleh, K. (2015). Measuring the Validity and Reliability of Research Instruments. *Procedia - Social and Beahavioral Sciences, 204*, 164-171. https://doi.org/10.1016/j.sbspro.2015.08.129

Neto, A., & Aquino, J. (2009). A avaliação da aprendizagem como ato amoroso: o que o professor pratica?. *Educação em Revista, 25*(2), 223-240. https://doi.org/10.1590/S0102-46982009000200010

Plake, B., Impara, J., & Fager, J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10-12. https://doi.org/10.1111/j.1745-3992.1993.tb00548.x

Popham, W. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator, 46*(4), 265–73. https://doi.org/10.1080/08878730.2011.605048

Prieto, G., & Delgado, A. (2003). Análisis de un test mediant el modelo de Rasch. *Psicothema, 15*(1), 94-100. https://www.psicothema.com/pdf/1029.pdf

Ramesal, A. (2011). Primary and secondary teachers' conceptions of assessment: a qualitative study. *Teaching and Teacher Education, 27,* 472-482. https://doi.org/10.1016/j.tate.2010.09.017

Ribeiro, L. (1993). *Avaliação da aprendizagem.* Texto Editora.

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. https://doi.org/10.1186/s12874-019-0680-5

Santiago, P., Donaldson, G., Looney, A., & Nusche, D. (2012). *OECD reviews of Evaluation and Assessment in Education: Portugal 2012*. OECD Publishing.

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. https://doi.org/10.1590/S0104-40362016000300006

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.

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. https://doi.org/10.1007/s11165-016-9602-2

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. https://doi.org/10.2307/20466661

Xu, Y., & Brown, G. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education, 58*, 149-162. https://doi.org/10.1016/j.tate.2016.05.010

Yamtim, V., & Wongwanich, S. (2014). A study of classroom assessment literacy of primary school teachers. *Procedia: Social and Behavioral Sciences, 116*, 2998-3004. https://doi.org/10.1016/j.sbspro.2014.01.696

i Centro de Estudos Interdisciplinares em Educação e Desenvolvimento, Universidade Lusófona, Portugal.
https://orcid.org/0000-0001-9655-5876

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

Luis Almeida
Rua S. João, 924B, 2975-158 Quinta do Conde
lmp.almeida@sapo.pt

Recebido em 27 de março de 2021

Aceite para publicação em 05 de dezembro de 2022

Psychometric Properties of the Assessment Literacy Admeasurement Questionnaire (QALA)

ABSTRACT

Assessment literacy is a fundamental characteristic that every teacher should have, since student's assessment is closely related to the entire teaching and learning process, so it should be considered a key element in improving teaching. The aim of this article was to analyse the psychometric properties of an instrument developed to measure assessment literacy of teachers of basic and secondary education: the QALA (Assessment Literacy Admeasurement Questionnaire). The analysis of the psychometric properties of QALA was carried out using the Rasch Model, which is assumed to be a set of statistical techniques that are part of the Item Response Theory (IRT). The results show the good psychometric qualities of QALA, so the instrument seems to be suitable for its purpose, namely to measure the perceptions that teachers have about their knowledge and skills in assessment, on the one hand, and their assessment literacy, on the other.

Keywords: Assessment Literacy; Learning assessment; QALA; Psychometric properties; Rasch Model.

Cualidades Psicométricas del Cuestionario de Medición de la Competencia en Evaluación (QALA)

RESUMEN

La Competencia en Evaluación es una característica fundamental que todo docente debe tener, ya que la evaluación del alumno está íntimamente relacionada con todo el proceso de enseñanza y aprendizaje, por lo que debe considerarse un elemento clave para mejorar la docencia. El objetivo de este artículo fue analizar las propiedades psicométricas de un instrumento desarrollado para medir la competencia en evaluación de profesores de educación básica y secundaria: el QALA (Cuestionario de Medición de la Competencia en Evaluación). El análisis de las propiedades psicométricas de QALA se realizó mediante el Modelo de Rasch, que se supone que es un conjunto de técnicas estadísticas que se insertan en la Teoría de Respuesta al Ítem (TRI). Los resultados muestran las buenas cualidades psicométricas de QALA, por lo que el instrumento parece adecuado para el propósito que se propone, es decir, medir las percepciones que los docentes tienen sobre sus conocimientos y habilidades en evaluación, por un lado, y su competencia evaluativa, en el otro.

Palabras clave: Competencia en Evaluación; Evaluación del aprendizaje; QALA; Cualidades psicométricas; Modelo Rasch.