ISSN 2178-2229

EVIDÊNCIAS DE VALIDADE BASEADAS NO CONTEÚDO DE UM INSTRUMENTO DE AVALIAÇÃO DA LITERACIA CIENTÍFICA[1]

VALIDITY EVIDENCE BASED ON CONTENT OF A SCIENTIFIC LITERACY ASSESSMENT INSTRUMENT

EVIDENCIAS DE VALIDEZ DE CONTENIDO DE UN INSTRUMENTO DE EVALUACIÓN DE ALFABETIZACIÓN CIENTÍFICA

Marcelo Alves Coppi
https://orcid.org/0000-0001-6734-7592

Isabel Fialho
https://orcid.org/0000-0002-1749-9077

Marília Cid
https://orcid.org/0000-0002-6009-0242

Resumo: O estudo tem como objetivo a recolha de evidências de validade baseadas no conteúdo de um instrumento piloto de avaliação da literacia científica. A recolha foi realizada em sete etapas: definição dos domínios cognitivos, definição do universo e da representatividade do conteúdo, elaboração da tabela de especificação, construção do instrumento, análise teórica dos itens e análise empírica dos itens. Foram selecionados 35 itens, que avaliam a compreensão, a análise e a avaliação das competências presentes nos principais documentos curriculares portugueses. O teste piloto foi aplicado a 176 alunos de oito escolas da região sul de Portugal. A análise empírica revelou a presença de 14 itens muito fáceis e sete muito difíceis, os quais devem ser revistos para adequar o instrumento às competências da população-alvo.
Palavras-chave: Literacia científica. Avaliação. Evidências de validade baseadas no conteúdo.

Abstract: The study aims to collect validity evidence based on the content of a pilot scientific literacy assessment instrument. The data collection was carried out in seven stages: definition of the cognitive domains, definition of the universe and the representation of the content, elaboration of the specification table, instrument construction, theoretical analysis of the items and empirical analysis of the items. Thirty-five items were selected that assess the understanding, analyzing, and evaluating competences present in the main Portuguese

1 Este trabalho é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito da Bolsa de Investigação com referência UI/BD/151034/2021.

curriculum documents. The pilot test was applied to 176 students from eight schools in southern Portugal. The empirical analysis revealed the presence of 14 very easy items and seven very difficult, which should be revised to adapt the instrument to the target population's abilities.
Keywords: Scientific literacy. Assessment. Validity evidence based on content.

Resumen: El estudio tiene como objetivo recopilar evidencias de validez basada en el contenido de un instrumento piloto de evaluación de la alfabetización científica. La recopilación se realizó en siete etapas: definición de los dominios cognitivos, del universo de contenido y de su representación, elaboración de la tabla de especificación, construcción del instrumento, análisis teórico y análisis empírico de los ítems. Se seleccionaron 35 ítems que evalúan la comprensión, análisis y evaluación de competencias de los principales documentos curriculares portugueses. El teste-piloto se aplicó a 176 estudiantes de ocho escuelas del sur de Portugal. El análisis reveló la presencia de 14 ítems muy fáciles y siete muy difíciles, que deben ser revisados para se adecuar a las habilidades de la población objetivo.
Palabras clave: Alfabetización científica. Evaluación. Evidencia de validez de contenido.

INTRODUÇÃO

A literacia científica é o principal objetivo do ensino de ciências e caracteriza-se pela capacidade de utilizar os conhecimentos científicos para lidar com os problemas do cotidiano e atuar como cidadão. A sua verdadeira procura deu-se no período pós Segunda Guerra Mundial, quando cientistas e governantes testemunharam a abrangência e a capacidade dos acervos científico e militar utilizados (MILLER, 1983). Desde então, muitos esforços vêm sendo realizados a fim de contribuir para a sua definição.

Em 1983, Miller propôs um conceito multidimensional para o termo. Segundo o autor, a literacia científica é definida por três dimensões: a compreensão do empreendimento da ciência, o conhecimento dos principais conteúdos da ciência e a consciência do impacto da ciência e da tecnologia na sociedade (MILLER, 1983).

Essa definição suscitou avanços importantes para a sua avaliação (Laugksch, 2000). Laugksch e Spargo (1996a) alegam que, após a publicação de Miller (1983), muitos estudos têm sido conduzidos com o objetivo de avaliar o conhecimento dos alunos em relação às três dimensões de literacia científica propostas pelo autor.

Contudo, Laugksch e Spargo (1996a) alegam que a maior parte dos instrumentos disponíveis na literatura apresenta validade inespecífica e que avaliam as três dimensões separadamente. Gormally, Brickman e Lutz (2012) reforçam esse fato, referindo que a maioria dos instrumentos desenvolvidos mede as três dimensões de forma individual. Dentre aqueles que avaliam as três dimensões propostas por Miller (1983), destaca-se o *Test of Basic Scientific Literacy* (TBSL), desenvolvido por Laugksch e Spargo (1996a). O TBSL é composto por 110 itens no formato "verdadeiro-falso-não sei", que avaliam ideias e atitudes importantes dos alunos no final do ensino secundário sobre a ciência.

Ademais, Fives et al. (2014) apontam que nenhum dos instrumentos até então desenvolvidos contempla as necessidades da literacia científica dos estudantes do 3º ciclo do

ensino básico. Em Portugal, este ciclo é marcado pela transição de uma única disciplina científica, Ciências Naturais, ministrada por um único professor (no 2.º ciclo), para duas disciplinas, Ciências Naturais e Físico-Química, as quais apresentam particularidades próprias e professores específicos. O 3º ciclo do ensino básico representa, também, o último ciclo em que há a obrigatoriedade de os alunos frequentarem disciplinas científicas.

Considerando a inexistência de instrumentos de avaliação que compreendam as três dimensões de literacia científica propostas por Miller (1983) e que sejam direcionadas aos alunos do 3.º ciclo do ensino básico português, desenvolveu-se um projeto de pesquisa no âmbito do doutoramento em curso, com o intuito de elaborar um instrumento de avaliação do nível de literacia científica dos estudantes portugueses no final do 3.º ciclo do ensino básico. Este projeto pretende contribuir com a investigação do nível de literacia científica dos estudantes ao fim deste ciclo, fornecendo indicadores capazes de auxiliar a monitorização do progresso da educação científica a níveis regional e nacional. Aspira-se que os resultados da aplicação do teste sejam utilizados pelos professores de uma forma orientadora, para a reformulação das suas aulas, planos de ensino e práticas em sala de aula, a fim de que os alunos completem o 3º ciclo do ensino básico como cidadãos cientificamente literatos.

CONTEXTUALIZAÇÃO TEÓRICA

Elaborar um instrumento de avaliação exige a utilização de procedimentos que garantam indicadores confiáveis (ALEXANDRE; COLUCI, 2011) e que evidenciem uma alta qualidade técnica. De acordo com especialistas da área da avaliação (DEPRESBITERIS; TAVARES, 2009; HALADYNA; RODRIGUEZ, 2013; POPHAM, 2017), a propriedade mais importante para a qualidade de um instrumento de avaliação é a validade. Russel e Airasian (2014, p. 26) respaldam essa ideia, argumentando que “a característica mais importante de uma boa avaliação é a sua habilidade de ajudar o professor a tomar as decisões adequadas. Essa característica é chamada de validade”.

Tradicionalmente, a validade é concebida como a capacidade de um instrumento de avaliação medir o que ele foi projetado para medir (GIPPS, 2003). Dentro dessa perspectiva, a literatura destaca três tipos de validade: de conteúdo, a qual se refere à relevância e à representatividade dos conteúdos que serão avaliados; de construto, que concerne à capacidade do teste em medir a competência que se está a avaliar, ou seja, o construto; e de critério, que pode ser concorrente ou preditiva, e está relacionada com a previsão de *performance* relativamente a algum critério externo (GIPPS, 2003).

Contudo, a literatura mais recente classifica a validade como um conceito unitário (AERA; APA; NCME, 2014; MILLER; LINN; GRONLUND, 2009; POPHAM, 2017), que reflete “o grau em que cada evidência e a teoria suportam a precisão das interpretações dos resultados dos testes para os usos propostos” (AERA; APA; NCME, 2014, p. 11, tradução

nossa). Nesta perspectiva, o que antes era entendido como tipos de validade passou a ser assumido como tipos de evidências de validade, as quais estão divididas em cinco categorias: evidências baseadas no conteúdo, evidências baseadas nos processos de resposta, evidências baseadas na estrutura interna, evidências baseadas na relação com outras variáveis e evidências baseadas nas consequências dos testes (AERA; APA; NCME, 2014).

As evidências de validade baseadas no conteúdo envolvem, além do próprio conteúdo, a elaboração, a redação e o formato dos itens e o processo de administração e de pontuação dos instrumentos de avaliação. De acordo com a versão atual dos *Standards for Educational and Psychological Testing*, os quais serão chamados de *Standards*, este tipo de evidência pode "incluir análises lógicas ou empíricas da adequação com a qual o conteúdo do teste representa o domínio do conteúdo e a relevância do domínio do conteúdo para a interpretação da pontuação do teste proposto" (AERA; APA; NCME, 2014, p. 14, tradução nossa). Além disso, as evidências baseadas no conteúdo podem resultar da análise de especialistas quanto à correspondência entre os itens do instrumento e o conteúdo selecionado (AERA; APA; NCME, 2014).

Já as evidências baseadas nos processos de resposta referem-se aos domínios cognitivos requeridos pelos instrumentos de avaliação. De acordo com os *Standards* (AERA; APA; NCME, 2014, p. 15, tradução nossa), "algumas interpretações do construto envolvem suposições mais ou menos explícitas sobre os processos cognitivos aplicados pelos participantes do teste". Sendo assim, as análises teórica e empírica dos processos de resposta são capazes de gerar evidências a respeito da adequação entre o construto e a natureza da resposta. Os processos de resposta também estão relacionados com os sujeitos que analisarão as avaliações e, nesse caso, as evidências referem-se à consistência entre os critérios dos avaliadores e a interpretação e o uso pretendido dos resultados (AERA; APA; NCME, 2014).

No caso das evidências baseadas na estrutura interna, indicam o grau de relação entre os itens e os construtos sobre os quais as interpretações dos resultados serão fundamentadas (AERA; APA; NCME, 2014). Em outras palavras, essas evidências são capazes de revelar itens que avaliam construtos diferentes daqueles pretendidos, os quais podem interferir na validade da interpretação dos resultados e das decisões que serão tomadas a partir destes, como a aprovação ou reprovação dos alunos, o encaminhamento para especialistas, a atribuição de bolsas de estudo, entre outras decisões (POPHAM, 2017).

As evidências baseadas na relação com outras variáveis envolvem a relação entre os resultados de um determinado instrumento de avaliação com outros, provenientes de instrumentos de avaliação externos (AERA; APA; NCME, 2014). Esse tipo de evidência é necessário em casos em que a interpretação dos resultados requer a comparação dos construtos com outras variáveis, as quais incluem, por exemplo, a pontuação de uma avaliação externa que mede o mesmo construto (AERA; APA; NCME, 2014). De acordo com os *Standards*, as "evidências baseadas na relação com outras variáveis fornecem indicadores

sobre o grau em que essas relações são consistentes com o construto subjacente às interpretações da pontuação do teste propostas" (AERA; APA; NCME, 2014, p. 16, tradução nossa)

Por fim, as evidências de validade baseadas nas consequências da aplicação dos testes relacionam-se diretamente com a interpretação e o uso dos resultados. Assim, "o processo de validação envolve a recolha de evidências a fim de avaliar a adequação dessas interpretações propostas para os usos pretendidos" (AERA; APA; NCME, 2014, p. 19, tradução nossa). O objetivo da recolha desse tipo de evidência é evitar consequências não intencionais que, geralmente, são negativas (AERA; APA; NCME, 2014). Como, por exemplo, o fato de professores do ensino básico ou secundário passarem a focar as suas aulas nos conteúdos e nos construtos específicos exigidos em determinadas provas nacionais ou internacionais em detrimento de outros, devido às consequências para os próprios professores, para os alunos, para a escola ou para o sistema de ensino.

Nota-se, portanto, que a atual conceção de validade envolve o acúmulo de evidências relevantes, capazes de sustentar a interpretação e o uso dos resultados dos instrumentos de avaliação (AERA; APA; NCME, 2014; MESSICK, 1989; MILLER; LINN; GRONLUND, 2009; POPHAM, 2017). Desta forma, torna-se inadequado o debate sobre a validade de um instrumento de avaliação, já que são as interpretações e os usos dos seus resultados que podem ou não ser válidos, não o instrumento em si (KANE, 2013; POPHAM, 2017).

Brookhart (2000, p. 23, tradução nossa) corrobora a ideia, alegando que a "validade é uma característica de uma pontuação atribuída a um uso específico, não a característica de um teste ou de uma avaliação em si". Para Cizek (2016), a validade assegura o significado da pontuação de um instrumento de avaliação e infere sobre a possibilidade de utilização dos resultados para o propósito estabelecido. Por exemplo, para a reformulação da metodologia, a necessidade de aulas complementares ou o acesso ao Ensino Superior.

Considerando que as evidências de validade baseadas no conteúdo são as fontes primárias de evidências de um instrumento de avaliação (KANE, 2013)validation as formulated by Kane is fundamentally a simply-stated two-step enterprise: (1, este estudo teve como objetivo a recolha de evidências de validade baseadas no conteúdo para a elaboração de um instrumento piloto de avaliação da literacia científica dos alunos no final do 3º ciclo do ensino básico.

A PESQUISA

A recolha de evidências de validade baseadas no conteúdo seguiu as etapas propostas por Pasquali (2009b), a saber: 1) definição dos domínios cognitivos; 2) definição do universo do conteúdo; 3) definição da representatividade do conteúdo; 4) elaboração da tabela de especificação; 5) construção do instrumento; 6) análise teórica dos itens; e 7) análise empírica dos itens.

DEFINIÇÃO DOS DOMÍNIOS COGNITIVOS

Esta primeira etapa diz respeito à determinação dos processos cognitivos ou psicológicos que se pretende avaliar (PASQUALI, 2009a). De acordo com Anderson et al. (2001), processos cognitivos são estratégias pelas quais o conhecimento é adquirido, construído ou utilizado para a resolução de problemas.

Pasquali (2009b) ressalta a importância de se apoiar em alguma taxonomia clássica de objetivos educacionais e, com base nela, estabelecer os domínios cognitivos que se pretende avaliar. Desta forma, a taxonomia utilizada neste estudo foi a de Anderson et al. (2001), os quais, a pedido da Associação de Psicologia Americana, realizaram a revisão e a atualização da Taxonomia de Bloom apresentada em 1956 (FERRAZ; BELHOT, 2010).

Nesta nova versão, os domínios cognitivos são: lembrar, que significa "recuperar o conhecimento relevante da memória de longo prazo"; compreender, que pressupõe "construir significado a partir de mensagens instrucionais"; aplicar, que compreende "executar ou utilizar um procedimento em uma dada situação"; analisar, entendido como o processo de "dividir o material em partes e determinar como as partes se relacionam umas com as outras e com o todo"; avaliar, que expressa a capacidade de "fazer julgamentos com base em critérios e padrões"; e criar, que "implica reunir os elementos para formar um todo coerente e funcional" ou "reorganizar os elementos em um novo padrão ou estrutura" (ANDERSON et al., 2001, p. 67–68, tradução nossa).

A seleção dos domínios cognitivos teve em consideração a definição de literacia científica estipulada para o instrumento em desenvolvimento, a qual condiz com a compreensão do empreendimento científico e a utilização consciente dos conhecimentos científicos e tecnológicos para a resolução de problemas, a explicação de fenômenos naturais do cotidiano e para a participação ativa em debates de assuntos científicos que envolvem a sociedade, permitindo ao indivíduo atuar como cidadão. Tendo por base esta definição, os domínios cognitivos adotados para a elaboração dos itens foram: compreender, analisar e avaliar problemas e situações cotidianas que envolvem o conhecimento das Ciências Físicas e Naturais. De acordo com Russel e Airasian (2014), estes correspondem a domínios cognitivos de nível superior, uma vez que envolvem mais do que a simples memorização das informações.

DEFINIÇÃO DO UNIVERSO DO CONTEÚDO

A segunda etapa consistiu no estabelecimento do universo do conteúdo programático, posto que os itens do instrumento reproduzem apenas uma amostra representativa do conteúdo (PASQUALI, 2009a). Segundo o autor, este processo implica "delimitar o conteúdo em suas unidades e subunidades de ensino" (PASQUALI, 2009a, p. 190).

Considerando a realidade portuguesa, os conteúdos foram selecionados a partir dos principais documentos educacionais relacionados com a área das Ciências Físicas e Naturais então vigentes em Portugal, a saber: Orientações Curriculares de Ciências Físicas e

Naturais (OC), Aprendizagens Essenciais (AE) de Ciências Naturais e de Físico-Química e o Perfil dos Alunos à Saída da Escolaridade Obrigatória (PA). Tendo em vista que as OC, as AE e o PA correspondem a uma vasta quantidade de conteúdos e competências a serem adquiridos e alcançados pelos alunos ao longo do 3º ciclo do ensino básico, optou-se por utilizar como referência as diretrizes de literacia científica estipuladas pelos *Benchmarks for Science Literacy* (BFSL) (AAAS, 1993).

Orientações curriculares de Ciências Físicas e Naturais

As OC, publicadas em 2001 pelo Ministério da Educação português, apresentam as competências a serem desenvolvidas pela respectiva área e devem ser tomadas como referência para o trabalho em sala de aula (GALVÃO et al., 2001). Estas orientações curriculares estabelecem a literacia científica como a finalidade do ensino das ciências no 3º ciclo do ensino básico e qualificam-na como fundamental e indispensável para o exercício da cidadania (GALVÃO et al., 2001).

As competências específicas definidas para a área das Ciências Físicas e Naturais estabelecem-se nos domínios do conhecimento, do raciocínio, da comunicação e das atitudes, as quais são consideradas essenciais para o desenvolvimento da literacia científica (GALVÃO et al., 2001). Desta forma, apoiar-se nas OC permitiu ao instrumento reunir condições para avaliar o desenvolvimento das competências em literacia científica estabelecidas pelo Ministério da Educação de Portugal para os alunos no final do 3º ciclo.

Aprendizagens essenciais de Ciências Naturais e de Físico-Química

As AE de Ciências Naturais e de Físico-Química do 3º ciclo foram selecionadas por corresponderem a um conjunto de conhecimentos, capacidades e atitudes indispensáveis a serem adquiridas e desenvolvidas por todos os alunos nas disciplinas da área das Ciências Físicas e Naturais ao longo dos três anos deste ciclo (CETIC, 2014). Além disso, as AE estabelecem-se como um documento de orientação curricular, construídas com base em documentos curriculares existentes e constituem a base de referência, ou o denominador curricular comum, para a aprendizagem de todos os alunos (DGE, 2018).

Nesse sentido, ao se basear nas AE de Ciências Naturais e de Físico-Química do 3º ciclo, o instrumento compromete-se a avaliar os conhecimentos substanciais e fundamentais para o desenvolvimento da literacia científica dos alunos no final deste ciclo de ensino.

Perfil dos alunos à saída da escolaridade obrigatória

Em concordância com as OC, o PA afirma-se “como documento de referência para a organização de todo o sistema educativo, contribuindo para a convergência e a articulação

das decisões inerentes às várias dimensões do desenvolvimento curricular" (MARTINS et al., 2017, p. 8). Assim como nas OC e nas AE, o PA baseia-se em áreas de competências, entendidas como a combinação de conhecimentos, capacidades e atitudes, as quais favorecerão o desenvolvimento das diversas literacias, dentre elas, a científica (MARTINS et al., 2017).

As áreas de competências determinadas pelo PA são as seguintes: linguagens e textos, que remete para a utilização eficaz dos códigos para expressar e representar o conhecimento; informação e comunicação, que dize respeito à seleção, análise, produção e divulgação de experiências e de conhecimentos; raciocínio e resolução de problemas, relacionada com os processos lógicos de interpretação de fatos e a produção do conhecimento; pensamento crítico e pensamento criativo, a qual envolve as competências de observar, identificar, analisar e dar sentido à informação e à argumentação de ideias; relacionamento interpessoal, que compreende a interação com outros indivíduos em diferentes contextos sociais; desenvolvimento pessoal e autonomia, relacionada com a motivação, a autorregulação, a iniciativa e a tomada de decisão; bem-estar, saúde e ambiente, que leva em consideração a promoção da qualidade de vida; sensibilidade estética e artística, relacionada com os processos de experimentação e interpretação de diferentes realidades culturais; saber científico, técnico e tecnológico, que diz respeito à compreensão das consequências éticas, sociais, económicas e ecológicas da aplicação dos processos científicos e tecnológicos; e consciência do domínio do corpo, relacionada com a compreensão do corpo humano como um sistema integrado (MARTINS et al., 2017).

Utilizar o PA como referência para a elaboração do instrumento possibilitou que a definição das competências a serem avaliadas pelos itens estivesse em concordância com as áreas de competências essenciais a serem desenvolvidas nos alunos ao longo da escolaridade obrigatória.

Benchmarks for Science Literacy

Os BFSL correspondem ao segundo subproduto do *Project 2061*, o qual estabelece recomendações do que os alunos americanos deveriam saber sobre ciências, matemática e tecnologia no fim de cada ciclo de ensino (AAAS, 1993). Tais recomendações baseiam-se nas indicações de literacia científica propostas pelo primeiro subproduto do *Project 2061*, o *Science For All Americans* (SFAA) (AAAS, 1989).

A partir das recomendações dos BFSL, o projeto reformulou os objetivos de literacia científica propostos pelo SFAA, levando em consideração níveis intermediários de compreensão para os ciclos "K-2, 3-5, 6-8 e 9-12, correspondendo aproximadamente às idades de 5-7, 8-10, 11-13 e 14-17, respectivamente" (LAUGKSCH; SPARGO, 1996a, p. 57–58, tradução nossa). Para a construção do instrumento em questão foram utilizados os BFSL do 6º ao 8º ano, etapa correspondente ao 3º ciclo do ensino básico português.

A escolha pelas orientações dos BFSL (AAAS, 1993) deu-se pelo fato de este documento se basear no programa SFAA, o qual se fundamentou na pesquisa de Miller (1983). No seu estudo, o autor estabeleceu as três dimensões de literacia científica (LAUGKSCH, 2000) utilizadas na construção do instrumento que está sendo desenvolvido – natureza da ciência (NC), conteúdo da ciência (CC) e impacto da ciência e da tecnologia na sociedade (ICTS) – que também respaldaram diversas pesquisas nos Estados Unidos, no Reino Unido, na Comunidade Europeia, na China, no Canadá e no Japão (LAUGKSCH; SPARGO, 1996b).

Vale ressaltar que o único documento curricular da área das Ciências Físicas e Naturais até então vigente que não foi utilizado para a construção dos itens do instrumento foi o das Metas Curriculares de Ciências Naturais e de Físico-Química. A sua exclusão justifica-se por três motivos essenciais: não menciona a literacia científica como objetivo do ensino das ciências, baseia-se em objetivos e descritores e não em competências, como as AE, as OC e o PA (SERRA; GALVÃO, 2015), e porque o presente instrumento pretende avaliar os conhecimentos essenciais de literacia científica que um aluno no final do 9º ano deve apresentar a fim de dar sequência aos seus estudos e atuar como cidadão na sociedade, conhecimentos esses que estão presentes nas AE, não havendo necessidade de utilizá-lo.

Análise documental

A fim de identificar quais as competências das OC, das AE e do PA que deveriam ser selecionadas para compor o instrumento, foi realizada uma análise documental em três etapas. A primeira consistiu em elencar todas as competências relacionadas com a área das Ciências Físicas e Naturais de cada documento.

No caso do PA, por se tratar de um documento curricular de referência para todo o sistema educativo português, as suas áreas de competência não são específicas para a área das Ciências Físicas e Naturais e, por esse motivo, foram selecionadas três delas, a saber: raciocínio e resolução de problemas, a qual “diz respeito aos processos de encontrar respostas para uma nova situação, mobilizando o raciocínio com vista à tomada de decisão, à construção e uso de estratégias e à eventual formulação de novas questões” (MARTINS et al., 2017, p. 23); pensamento crítico e pensamento criativo, cujas competências “exigem o desenho de algoritmos e de cenários que considerem várias opções, assim como o estabelecimento de critérios de análise para tirar conclusões fundamentadas e proceder à avaliação de resultados” (MARTINS et al., 2017, p. 24); e o saber científico, técnico e tecnológico, que envolve a “mobilização da compreensão de fenómenos científicos e técnicos e da sua aplicação para dar resposta aos desejos e necessidades humanos, com consciência das consequências éticas, sociais, económicas e ecológicas” (MARTINS et al., 2017, p. 29).

Na segunda etapa, realizou-se o processo de comparação entre os documentos. O objetivo foi verificar a correspondência entre as competências das OC e das AE e os conhecimentos estabelecidos pelos BFSL, selecionando aquelas que apresentavam alguma correlação. Por exemplo, todas as competências que abordavam o tema "universo" nas OC e nas AE foram associadas aos conhecimentos que abordavam o tema "universo" nos BFSL. As competências que não apresentaram correlação entre os documentos foram eliminadas da análise e, consequentemente, não integraram os itens do instrumento piloto.

A terceira e última etapa teve por objetivo eliminar competências similares. O critério de seleção levou em consideração as competências que mais se assemelhavam às áreas de competências do PA. Esta etapa teve grande importância para a elaboração dos itens, pois a presença de competências semelhantes poderia gerar uma equivalência entre os itens, fazendo com que um item pudesse servir como pista para a resposta de outro, o que não é desejado para um instrumento de avaliação como este.

No total, foram selecionadas 60 competências, 10 das OC e 50 das AE, sendo 17 do 7º ano, 17 do 8º ano e 16 do 9º ano. Vale ressaltar que duas competências das OC e uma competência das AE do 8º ano foram utilizadas na elaboração de mais de um item. Isso ocorreu por se tratar de competências mais amplas do que as demais, nas quais foram encontradas mais de um correspondente nos BFSL.

DEFINIÇÃO DA REPRESENTATIVIDADE DO CONTEÚDO

Assim como na etapa anterior, a definição da representatividade do conteúdo, caracterizada pela proporção com que cada conteúdo deve ser representado no instrumento (PASQUALI, 2009a), foi estabelecida mediante a comparação dos documentos curriculares da área de Ciências Físicas e Naturais com os BFSL.

O número de itens por conteúdo foi determinado pela quantidade de competências similares entre as OC, as AE e o PA e os conhecimentos dos BFSL. Como o instrumento foi desenvolvido com base nas três dimensões propostas por Miller (1983) – natureza da ciência, conteúdo da ciência e impacto da ciência e da tecnologia na sociedade –, a representatividade do conteúdo também teve em conta estas dimensões, as quais foram transformadas em três subtestes. Nesse sentido, as 60 competências foram distribuídas da seguinte forma: 6 itens com competências da dimensão da NC, 7 itens com competências da dimensão do ICTS e 51 itens com competências da dimensão do CC. Como a dimensão do CC envolve competências mais específicas das disciplinas científicas do que as outras duas dimensões, a sua representatividade está detalhada na Tabela 1.

Tabela 1 – Representatividade do conteúdo da dimensão do conteúdo da ciência.

Conteúdo	N.º de itens	Conteúdo	N.º de itens	Conteúdo	N.º de itens
Alterações ambientais	3	Átomos e elementos químicos	3	Força, gravidade e movimento	2
Universo e sistema solar	3	Substâncias e Misturas	2	Ecologia	4
Geodinâmica interna	4	Reações químicas	2	Evolução	1
Geodinâmica externa	4	Energia	4	Células	2
Temperatura e mudanças de estado físico da matéria	3	Ondas	3	Fisiologia	11

Fonte: Elaboração dos autores.

ELABORAÇÃO DA TABELA DE ESPECIFICAÇÃO

A tabela de especificação (Tabela 2) foi elaborada mediante a atribuição da correspondência entre as competências das OC, AE e do PA, indicadas na etapa da definição do universo do conteúdo, com os domínios cognitivos da Taxonomia atualizada de Bloom (ANDERSON et al., 2001). Dentre os 64 itens, 26 pertencem ao domínio cognitivo de avaliar, 21 de compreender e 17 de analisar.

Tabela 2 – Tabela de especificação por dimensão do conteúdo/domínio cognitivo.

Dimensão do conteúdo	Domínio cognitivo			Total
	Compreender	Analisar	Avaliar	
Natureza da ciência (NC)	4	1	1	6
Impacto da ciência e da tecnologia na sociedade (ICTS)	2	1	4	7
Conteúdo da ciência (CC)	15	15	21	51
Total	21	17	26	64

Fonte: Elaboração dos autores.

CONSTRUÇÃO DO INSTRUMENTO

A construção do instrumento relaciona-se, de fato, com a elaboração dos itens que o constituem (PASQUALI, 2009a). Este processo envolveu decisões sobre: a) o formato dos itens; b) as diretrizes técnicas de construção – *guide lines* – utilizadas; e c) a configuração dos enunciados dos itens.

Formato dos itens

Uma grande variedade de tipos de itens vem sendo utilizada em testes avaliativos: resposta livre (RL), escolha múltipla (EM), verdadeiro-falso (VF), múltiplo-verdadeiro-falso (MVF), respostas múltiplas (RM) entre outros (BERK, 1996). A fim de garantir que o conteúdo desejado possa ser avaliado no tempo disponível de uma aula, o formato de item escolhido para o instrumento foi o VF.

Itens de VF são "afirmações declarativas simples absolutamente verdadeiras ou falsas" (HALADYNA, 2018, p. 4, tradução nossa). Pelo fato de não apresentar opções alternativas, itens neste formato exigem que o aluno crie mentalmente um contraexemplo da afirmação e opte pela opção verdadeira ou falsa (HALADYNA, 2004). Desta forma, testes assim configurados podem gerar informações importantes a respeito da compreensão do conteúdo (GATES; HOYER, 1986).

Frisbie (1973) alega que muitos autores da área de avaliação aprovam o uso de itens de VF em testes de avaliação educacional elaborados por professores. Haladyna (2004) corrobora a ideia, alegando que testes de VF foram bem aceites para avaliações de sala de aula.

Além de serem relativamente fáceis e menos demorados para construir, os itens VF apresentam uma alta eficiência (EBEL, 1979; HALADYNA, 2004; RUSH; RANKIN; WHITE, 2016). Esta característica está associada ao fato de que um maior número de itens de VF pode ser respondido em um tempo limitado. De acordo com Frisbie e Becker (1991, p. 72, tradução nossa), "evidências consideráveis de pesquisas mostram que pelo menos 50% a mais de itens de verdadeiro-falso do que itens de escolha múltipla podem ser utilizados num determinado intervalo de tempo de teste".

Os autores alegam ainda que, além de permitirem avaliar um maior número de conteúdos, os testes no formato VF são capazes de avaliar os respondentes de forma mais reveladora, pois contam com mais itens para um mesmo tema. Desta forma, os "testes de verdadeiro-falso podem fornecer uma amostra muito mais ampla do conhecimento dos alunos sobre o assunto" (MAIHOFF; MEHRENS, 1985, p. 3, tradução nossa).

Para Ebel e Frisbie (1991), a justificação da utilização de itens de VF em avaliações educacionais pode ser resumida em quatro afirmações:

> 1. A essência do desempenho escolar é o domínio do conhecimento verbal útil.
> 2. Todo o conhecimento verbal pode ser expresso em proposições.
> 3. Uma proposição é simplesmente uma afirmação que pode ser entendida como verdadeira ou falsa.
> 4. A amplitude do domínio de um aluno sobre uma área particular do conhecimento é indicada pelo seu sucesso em julgar a veracidade ou a falsidade das proposições relacionadas com ela. (EBEL; FRISBIE, 1991, p. 133, tradução nossa)

Os autores consideram como conhecimento verbal qualquer conhecimento que possa ser expresso em frases, fórmulas ou símbolos.

No entanto, a utilização deste formato de item em testes educacionais é motivo de grandes debates na literatura. Ebel (1971) argumenta que muitos especialistas em avaliação consideram os itens de VF como inadequados. Frisbie e Becker (1991, p. 67, tradução nossa) reafirmam a ideia, alegando que "muitos educadores e pesquisadores que são especialistas em testes de desempenho tendem a considerar os itens de verdadeiro-falso como um dos formatos de item menos satisfatórios".

Em contrapartida, outros especialistas "reconhecem que as deficiências encontradas em itens de testes de verdadeiro-falso não são inerentes ao formato, mas estas frequentemente refletem uma escrita descuidada ou incompetente dos itens" (FRISBIE; BECKER, 1991, p. 67, tradução nossa). Além disso, "alguns veem neles virtudes especiais de eficiência e facilidade de preparação e advogam a sua utilização de forma mais ampla" (EBEL, 1971, p. 1, tradução nossa).

Talvez a crítica mais recorrente sobre itens de VF seja a do acerto por adivinhação. Tasdemir (2010) afirma que esta questão vem sendo discutida há muito tempo na literatura. Isso porque, devido ao seu formato, "os alunos têm 50% de chance de responder corretamente um item sem conhecimento do conteúdo" (RUSH; RANKIN; WHITE, 2016, p. 3, tradução nossa). Entretanto, Frisbie e Becker (1991, p. 74, tradução nossa) afirmam que "autores que consideram os efeitos da adivinhação em testes de verdadeiro-falso como uma das fraquezas primárias falharam completamente em considerar evidências racionais e empíricas".

Burton e Miller (1999, p. 399, tradução nossa) corroboram a ideia, afirmando que embora testes de VF possam ser afetados pela adivinhação, "a extensão real em que o acaso afeta as pontuações ainda é muito pouco assentida". Além disso, esta não é uma crítica exclusiva do formato VF. Tasdemir (2010, p. 259, tradução nossa)) lembra que "quando os testes de escolha múltipla começaram a ser amplamente utilizados, eles foram criticados porque os examinados podiam responder corretamente através da adivinhação".

De fato, é muito difícil, se não impossível, prevenir a adivinhação em testes, principalmente a adivinhação informada (CHANDRATILAKE; DAVIS; PONNAMPERUMA, 2011). No entanto, Haladyna (2004) defende que a adivinhação não é um fator muito importante em testes de VF. Segundo o autor, a base de uma escala de um teste VF é 50% e o topo é 100%, desta forma "ultrapassar 60% nestes testes quando o tamanho do teste é substancial é difícil para um adivinhador aleatório" (HALADYNA, 2004, p. 80, tradução nossa).

A fim de reduzir a adivinhação, optou-se por utilizar uma versão adaptada do formato de itens de VF, o formato de "verdadeiro-falso-não sei". Nessa configuração, acrescentou-se a opção "não sei", a qual deve ser assinalada caso o aluno não tenha conhecimento do conteúdo ou da competência solicitada pelo item. De acordo com Ebel e Frisbie (1991), ao

adicionar esta opção em questionários de pesquisas, diminui-se significativamente o número de respostas corretas obtidas apenas por meio de adivinhação.

Além disso, a inclusão da opção “não sei” também teve por objetivo a recolha de dados importantes para os professores a respeito do conhecimento dos alunos. Itens com grandes quantidades de resposta nesta opção são capazes de informar que determinado conteúdo ou competência não foi efetivamente assimilada pelos alunos ou, até mesmo, que não foi trabalhada em sala de aula, merecendo maior atenção dos professores na reformulação das suas aulas e estratégias de ensino.

Diretrizes técnicas

Uma vez estabelecido o formato dos itens, foi tido um grande cuidado no seu processo de elaboração, utilizando, para isso, duas diretrizes principais, a de Haladyna (2004) e a de Ebel e Frisbie (1991). Haladyna (2004) apresenta instruções detalhadas para a elaboração de itens objetivos, as quais estão organizadas em três categorias: diretrizes de conteúdo, orientações de estilo e formato e a elaboração do enunciado. Na primeira categoria, o autor defende que cada item deve refletir um único domínio cognitivo presente na tabela de especificação, basear-se em conhecimentos e evitar conteúdos triviais, utilizar exemplos para medir a aplicação do conhecimento, conter informação diferente dos outros itens e evitar generalizações, opiniões e expressões que enganem os alunos.

Na categoria de estilo e formato, o autor sugere que os itens sejam editados verticalmente; escritos de forma clara e com a correta gramática e pontuação; simplificados, para que as nomenclaturas e terminologias não interfiram no conhecimento do conteúdo do item; minimizados de tempo de leitura, evitando o uso de vocabulário excessivo; e revistos, para que possíveis falhas de elaboração sejam detectadas. E, na categoria da elaboração do enunciado, Haladyna (2004) recomenda que as indicações devem ser breves, claras e que a principal ideia do item esteja no enunciado e não na alternativa.

Já as propostas de Ebel e Frisbie (1991), são específicas para itens no formato VF. Nas suas diretrizes, os autores referem que itens neste formato devem: testar apenas uma ideia central; avaliar a compreensão e a explicação dos eventos e não apenas a simples memorização de fatos triviais; apresentar uma resposta correta que seja defensável cientificamente e que não seja óbvia para qualquer aluno, mas apenas para aqueles que realmente detenham determinado conhecimento; serem escritos de forma clara e concisa; e não apresentar a dupla negação.

Nota-se que as diretrizes propostas por Ebel e Frisbie (1991), embora tenham sido elaboradas especificamente para os itens de VF, se assemelham àquelas recomendadas por Haladyna (2004). A fim de que este instrumento não apresentasse, ou minimizasse ao máximo, falhas na elaboração, as diretrizes acima foram seguidas de forma rigorosa.

Configuração do item

Com o intuito de elaborar itens que avaliassem os domínios cognitivos definidos anteriormente, optou-se pela construção de itens interpretativos. Estes fornecem informações em forma de textos, gráficos, quadros, imagens ou tabelas, que servem de base para os alunos o responderem (RUSSEL; AIRASIAN, 2014). De acordo com os autores, para responder a este tipo de item, "os alunos têm de interpretar, compreender, analisar, aplicar ou sintetizar as informações apresentadas" (RUSSEL; AIRASIAN, 2014, p. 146) e, por esse motivo, são itens capazes de mobilizar domínios cognitivos superiores.

Para a configuração dos itens deste instrumento, optou-se por apresentar as informações em forma de texto. Consequentemente, todos os itens foram estruturados com uma ou mais frases destacadas em itálico – as quais descrevem uma situação, um caso, um fenómeno ou um evento –, seguidas de uma afirmação sem destaque, que deve ser analisada e respondida como verdadeira ou falsa. Nesta configuração, os alunos respondem aos itens levando em consideração uma introdução contendo informações verdadeiras e claramente identificadas, conforme o exemplo abaixo:

> *Numa experiência, o professor amarrou um balão bem esticado numa das extremidades abertas de uma lata e colocou uma pequena bola de esferovite sobre o balão. Em seguida, tocou um tambor perto da lata.* **É** correto afirmar que a vibração do toque do tambor originou uma onda que se propagou pelo ar até à lata, fazendo com que o balão vibrasse, movimentando a bola de esferovite.

Tendo em vista o estabelecimento do contexto português como fundamentação para a elaboração dos itens, teve-se o cuidado de elaborar cada item de forma que este fizesse sentido aos respondentes. Para isso, quando possível, as situações problema ou os casos criados foram formulados com exemplos de locais portugueses ou com fenómenos ocorridos em Portugal. Na impossibilidade de trazer para o item o contexto português, optou-se por criar situações adequadas ao ciclo a que se destina o instrumento. Em ambos os casos, foi dada uma grande atenção para evitar a condução do aluno para a resposta correta e para não gerar pistas.

A seguir, são apresentados dois exemplos de itens. O primeiro evidencia a utilização do contexto português e o segundo demonstra a adequação do item ao 3º ciclo do ensino básico:

> *Aproximadamente 60% da população portuguesa adulta está acima do peso e 24% é obesa. O sobrepeso e a obesidade aumentam o risco de doenças cardiovasculares, a principal causa de doença e morte em Portugal.* Neste cenário, a adoção da dieta mediterrânica torna-se uma estratégia de promoção da saúde, já que tem como objetivo diminuir a quantidade dos alimentos ingeridos, gerando um déficit calórico e evitando o sobrepeso.

> *Foi encontrado fora do nosso sistema solar um planeta muito semelhante à Terra em tamanho, composição química e distância à sua estrela. Este planeta apresenta uma atmosfera que bloqueia a entrada da maior parte da radiação solar e possui água apenas no estado sólido.* Devido às suas características, este planeta poderia abrigar vida na forma como a conhecemos.

Análise teórica dos itens

A análise teórica dos itens tem como objetivo a verificação da "representatividade dos itens em relação às áreas de conteúdo e à relevância dos objetivos a medir" (RAYMUNDO, 2009, p. 87). Considerando a inexistência de um teste específico para este tipo de análise, realizou-se uma abordagem qualitativa, seguida de uma quantitativa, conforme proposto por Alexandre e Coluci (2011).

Abordagem qualitativa

Na etapa qualitativa, foi constituído um painel de especialistas, que avaliou os seguintes aspetos: a correspondência entre o item e os documentos curriculares; a veracidade das afirmações; a presença de ambiguidades lógicas e científicas; a adequação da linguagem e do vocabulário para o público-alvo; e a relevância do item para a literacia científica. Esta etapa compreendeu duas fases: a seleção dos especialistas e a elaboração dos formulários de revisão dos itens.

Seleção dos especialistas

Tratando-se de uma análise subjetiva, a seleção dos especialistas deve ser bastante criteriosa e deve levar em consideração a qualificação profissional e o número de especialistas (ALEXANDRE; COLUCI, 2011). No que se refere à qualificação, os membros do painel foram selecionados de acordo com as seguintes características: área e nível de formação e área e nível de atuação. Levou-se em consideração as áreas das Ciências da Educação e das Ciências Físicas e Naturais, representadas pelos ramos específicos da biologia, da geologia, da física e da química.

Com relação à quantidade de especialistas necessários, Alexandre e Coluci (2011) afirmam não existir um consenso na literatura. No entanto, Rubio et al. (2003)argumentam que, em geral, são recomendados de três a vinte especialistas. De acordo com os autores, "a utilização de um maior número de especialistas pode fornecer mais informações sobre o instrumento" (RUBIO et al., 2003, p. 96, tradução nossa)

Considerando essas orientações, foram selecionados 10 especialistas. Destes, quatro são professores do ensino básico e/ou secundário e seis são professores universitários. Dos dez especialistas, sete são doutorados e três são mestres, como mostra a Tabela 3.

Tabela 3 – Painel de especialistas: área de atuação, nível de ensino.

Área de atuação	Ciências Naturais		Físico-Química		Ciências da educação		Biologia		Geologia		Física		Química	
Nível de ensino	M	D	M	D	M	D	M	D	M	D	M	D	M	D
N.º	1	1	2	-	-	2	-	1	-	1	-	1	-	1
Total	2		2		2		1		1		1		1	

Nota: M = mestre; D = doutorado.
Fonte: Elaboração dos autores.

Após a seleção, os especialistas foram contactados via e-mail, solicitando a sua participação no processo de análise teórica dos itens do instrumento em construção. Anexa ao e-mail foi enviada uma carta de apresentação contendo: o objetivo e a justificativa do estudo; a razão pela qual o especialista foi escolhido; a descrição do instrumento; e as instruções de como preencher o formulário de revisão dos itens, conforme sugerido por Rubio et al. (2003). Todos os especialistas escolhidos aceitaram integrar o painel.

Elaboração dos formulários de revisão

Levando em consideração a área de atuação dos especialistas, foram elaborados seis formulários de revisão dos itens do instrumento (Tabela 4).

Tabela 4 – Formulários de revisão dos itens do instrumento.

Área de atuação	Formulário a preencher	Dimensão da literacia científica	Total de itens
Ciências Naturais	A	NC, CC (ciências naturais) e ICTS	45
Físico-Química	B	NC, CC (físico-química) e ICTS	36
Biologia	C	NC, CC (biologia) e ICTS	37
Geologia	D	NC, CC (geologia) e ICTS	24
Física	E	NC, CC (física) e ICTS	26
Química	F	NC, CC (química) e ICTS	26

Fonte: Elaboração dos autores.

Os formulários foram divididos em duas partes. Na primeira, foi solicitado aos especialistas que preenchessem com as suas informações pessoais: o nome, a atividade profissional, a instituição onde a exerce e o domínio da atividade. A segunda parte contou com perguntas relacionadas com a análise dos itens, para as quais foram disponibilizados os seguintes dados: as OC ou as AE e os BFSL referentes a cada item, o item propriamente dito e a sua respectiva opção de resposta correta.

Para cada item, foram elaboradas cinco perguntas: 1) Há correspondência entre o item e as OC/AE/BFSL? 2) A afirmação é claramente verdadeira ou claramente falsa? 3)

O item apresenta ambiguidades lógicas e/ou científicas? 4) A linguagem, o vocabulário e/ou a estrutura da frase podem ser de difícil interpretação/entendimento para alunos do 9º ano? 5) Qual a relevância deste item para a literacia científica que foi definida para este instrumento?

Na pergunta 1, os especialistas assinalaram as opções “sim”, “não” ou “precisa de ajustes”. Nos casos em que a opção “precisa de ajustes” foi selecionada, os especialistas registaram comentários e sugestões para melhorar a adequação do item aos documentos curriculares.

As opções de resposta das perguntas 2, 3 e 4 foram “sim” e “não”. Sendo que, na pergunta 3, quando a opção “sim” foi assinalada, solicitou-se ao especialista que identificasse a ambiguidade. Já na pergunta 5, as opções de resposta foram “muito relevante”, “relevante” e “pouco relevante”.

Abordagem quantitativa

A abordagem quantitativa compreendeu a utilização de recursos estatísticos para quantificar o grau de concordância entre os especialistas. Alexandre e Coluci (2011) afirmam que as pesquisas têm dado destaque principalmente a dois métodos: a Percentagem de Concordância (PC) – PC = número de especialistas que concordaram totalmente com o item / número total de especialistas x 100 – e o Índice da Validade de Conteúdo (IVC) – IVC = número de respostas válidas / número total de respostas. Neste estudo, foi utilizado o método do IVC, o qual “mede a proporção ou percentagem de juízes que estão em concordância sobre determinados aspetos do instrumento e de seus itens” (ALEXANDRE; COLUCI, 2011, p. 3065).

Levando em consideração que as quatro primeiras perguntas estavam associadas às características técnicas dos itens, cujos resultados das análises dos especialistas auxiliaram na melhoria das suas respectivas qualidades, e que a quinta pergunta se referia à relevância dos itens para a literacia científica, o IVC foi calculado apenas para esta última.

Deste modo, as opções de resposta da pergunta 5 foram transformadas em uma pontuação de 0 a 1, conforme se descreve: às pontuações de 0, 0.5 e 1 foram atribuídas as opções “pouco relevante”, “relevante” e “muito relevante”, respectivamente. Apenas a opção “muito relevante” foi considerada como resposta válida para o cálculo do IVC. A fim de compor o instrumento piloto apenas com os itens considerados essenciais, foram excluídos aqueles que apresentaram IVC menor do que 0.8, representando 29 dos 64 itens.

Um desses itens pertencia à dimensão do ICTS e os outros 28 à dimensão do CC. Consequentemente, o instrumento piloto foi constituído por 35 itens, seis da dimensão da NC, seis da dimensão do ICTS e 23 da dimensão do CC, conforme a tabela de especificação atualizada do instrumento piloto (Tabela 5).

Tabela 5 – Tabela de especificação do instrumento piloto por dimensão do conteúdo/domínio cognitivo.

Dimensão do conteúdo	Domínio cognitivo			Total
	Compreender	Analisar	Avaliar	
Natureza da ciência (NC)	4	1	1	6
Impacto da ciência e da tecnologia na sociedade (ICTS)	1	1	4	6
Conteúdo da ciência (CC)	5	9	9	23
Total	10	11	14	35

Fonte: Elaboração dos autores.

Vale ressaltar que o instrumento está dividido em três subtestes, os quais correspondem às três dimensões do conteúdo referidas anteriormente. Assim, os 35 itens estão distribuídos da seguinte forma: seis itens pertencem ao subteste da NC, seis ao subteste do ICTS e 23 ao subteste do CC. Tal distribuição é necessária para a realização do processo de categorização do nível de literacia científica dos inquiridos, o qual exigirá um número de acertos mínimo nos itens de cada subteste.

ANÁLISE EMPÍRICA DOS ITENS

Teste piloto

A análise empírica foi realizada com os dados da aplicação do teste piloto. Este foi respondido por 176 alunos do 10º ano de oito agrupamentos de escolas/escolas não agrupadas da região sul de Portugal continental, no início do ano letivo de 2020/2021. Desses, 87 indivíduos são do sexo feminino e 89 do sexo masculino. A idade média dos alunos foi de 15.2 anos (*DP* = 2.5).

Em virtude das condições decorrentes da quarentena devida à Covid-19, os alunos responderam ao instrumento no formato *online*, através do *software LimeSurvey*, em sala de aula e na presença dos professores. Estes receberam as instruções de como aplicar o instrumento e ficaram encarregados de transmiti-las aos alunos, principalmente no que concerne a utilização da opção de resposta “não sei”, uma vez que o pesquisador responsável pelo estudo não pôde aceder aos espaços escolares durante o período da pandemia. O tempo máximo de resposta foi de 50 minutos.

A análise

De acordo com Pasquali (2009b), a análise empírica consiste na avaliação de um conjunto de características dos itens, a qual indicará se estes avaliam de forma adequada o que se propõem medir. O autor esclarece que, dentre essas características, a psicometria analisa, tradicionalmente, os índices de dificuldade e de discriminação dos itens e o índice de acerto ao acaso (PASQUALI, 2009a).

Considerando que o parâmetro do acerto ao acaso foi minimizado pela inclusão da opção de resposta “não sei”, essa característica não foi analisada. O índice de discriminação dos itens também não foi analisado, pois trata-se de um parâmetro relacionado com a capacidade que um item tem de “diferenciar sujeitos com magnitudes diferentes de traço do qual o item constitui a representação comportamental” (PASQUALI, 2009a, p. 139). Ou seja, este é um parâmetro importante para os testes de larga escala, já que é capaz de descrever quão bem um item pode diferenciar os respondentes que dominam dos que não dominam a competência requerida pelo item. Contudo, esta não é uma característica fundamental para avaliações de sala de aula, como o instrumento que está a ser testado, cujo propósito é fornecer informações aos professores, para que estes reflitam sobre o desempenho dos alunos quanto às competências que estão sendo avaliadas (SMITH, 2003).

Desta forma, a análise empírica dos itens foi realizada com base no parâmetro de dificuldade dos itens, concebido e tratado por meio do modelo logístico de dois parâmetros da Teoria de Resposta ao Item (TRI), modelo que melhor se adequou aos dados ($p < 0.5$). Além disso, a fim de obter dados sobre a adequação dos itens às habilidades dos alunos, foi analisado, também pela TRI, o valor da habilidade, ou traço latente (θ), dos respondentes para cada um dos subtestes.

A escolha pela utilização da TRI para a análise empírica dos dados justifica-se pelo fato de esta teoria considerar os itens individualmente, sem que os *scores* totais do teste influenciem diretamente na análise (ARAUJO; ANDRADE; BORTOLOTTI, 2009; BAKER, 2001)some examples of applications are given, and some recent development of the method are summarised. Design. Secondary analysis of data obtained by cross-sectional survey methods, including self-report and observation. Methods. Data from the Edinburgh Feeding Evaluation in Dementia scale and the Townsend Functional Ability Scale were analysed using the Mokken scaling procedure within the ‘R’ statistical package. Specifically, invariant item ordering (the extent to which the order of the items in terms of difficulty was the same for all respondents whatever their total scale score. Assim, “as conclusões não dependem exclusivamente do teste ou questionário, mas de cada item que o compõe” (ARAUJO; ANDRADE; BORTOLOTTI, 2009, p. 1002).

Nesta lógica, a TRI tem interesse específico em cada um dos itens “e quer saber qual é a probabilidade e quais são os fatores que afetam esta probabilidade de cada item individualmente ser acertado ou errado” (PASQUALI, 2009b, p. 993). Para Araujo et al. (2009, p. 1002), a TRI é capaz de fornecer modelos matemáticos que representam as habilidades dos inquiridos, permitindo a representação da “relação entre a probabilidade de um indivíduo dar uma certa resposta a um item, seu traço latente e características (parâmetros) dos itens, na área de conhecimento em estudo”.

Considerando que o parâmetro do acerto ao acaso foi minimizado pela inclusão da opção de resposta “não sei”, essa característica não foi analisada. O índice de discriminação dos itens também não foi analisado, pois trata-se de um parâmetro relacionado com a

capacidade que um item tem de “diferenciar sujeitos com magnitudes diferentes de traço do qual o item constitui a representação comportamental” (PASQUALI, 2009a, p. 139). Ou seja, este é um parâmetro importante para os testes de larga escala, já que é capaz de descrever quão bem um item pode diferenciar os respondentes que dominam dos que não dominam a competência requerida pelo item. Contudo, esta não é uma característica fundamental para avaliações de sala de aula, como o instrumento que está a ser testado, cujo propósito é fornecer informações aos professores, para que estes reflitam sobre o desempenho dos alunos quanto às competências que estão sendo avaliadas (SMITH, 2003).

Desta forma, a análise empírica dos itens foi realizada com base no parâmetro de dificuldade dos itens, concebido e tratado por meio do modelo logístico de dois parâmetros da Teoria de Resposta ao Item (TRI), modelo que melhor se adequou aos dados ($p < 0.5$). Além disso, a fim de obter dados sobre a adequação dos itens às habilidades dos alunos, foi analisado, também pela TRI, o valor da habilidade, ou traço latente (θ), dos respondentes para cada um dos subtestes.

A escolha pela utilização da TRI para a análise empírica dos dados justifica-se pelo fato de esta teoria considerar os itens individualmente, sem que os scores totais do teste influenciem diretamente na análise (ARAUJO; ANDRADE; BORTOLOTTI, 2009; BAKER, 2001). Assim, “as conclusões não dependem exclusivamente do teste ou questionário, mas de cada item que o compõe” (ARAUJO; ANDRADE; BORTOLOTTI, 2009, p. 1002).

Nesta lógica, a TRI tem interesse específico em cada um dos itens “e quer saber qual é a probabilidade e quais são os fatores que afetam esta probabilidade de cada item individualmente ser acertado ou errado” (PASQUALI, 2009b, p. 993). Para Araujo et al. (2009, p. 1002), a TRI é capaz de fornecer modelos matemáticos que representam as habilidades dos inquiridos, permitindo a representação da “relação entre a probabilidade de um indivíduo dar uma certa resposta a um item, seu traço latente e características (parâmetros) dos itens, na área de conhecimento em estudo”.

Consequentemente, a TRI revela-se como um método eficiente para a análise dos processos quantitativos de avaliação educacional, permitindo, entre outros fatores, a elaboração de escalas de habilidades precisas (ANDRADE; TAVARES; VALLE, 2000). Baker (2001, p. 84, tradução nossa) respalda essa ideia, alegando que, na perspectiva da TRI, “o principal objetivo de administrar um teste a um examinado é localizá-lo na escala de habilidades”.

No modelo logístico de dois parâmetros da TRI, o parâmetro de dificuldade é definido como “o ponto na escala de habilidade na qual a probabilidade de resposta correta ao item é de .5” (BAKER, 2001, p. 2, tradução nossa). Pasquali (2009b, p. 122) acrescenta que a TRI define a dificuldade do item “em termos do traço latente, do teta (θ), dizendo que esta dificuldade é diretamente proporcional ao nível ou tamanho de teta necessário para que um dado item possa ser acertado”.

Nesse sentido, o índice de dificuldade do item representa o nível de habilidade necessário para que a probabilidade de um indivíduo responder corretamente ao item seja de 50% (ARAUJO; ANDRADE; BORTOLOTTI, 2009). A TRI nomeia este parâmetro como "b" ou "limiar" – thereshold, location –, "porque ele é definido pela perpendicular, sobre a abscissa, da curva característica do item (CCI) no momento da inflexão, isto é, no ponto onde ocorre a probabilidade de 50% de acertar e 50% de errar o item" (PASQUALI, 2009a, p. 122). Sendo assim, a dificuldade do item é um indicador de localização, o qual descreve a sua posição ao longo da escala de habilidades representada na Curva Característica do Item (CCI) (BAKER, 2001; PASQUALI, 2009a).

Levando em consideração a logística deste instrumento piloto, os resultados serão reportados por subteste. Como parâmetro do índice de dificuldade dos itens, foi utilizada a categorização proposta por Baker (2001), que distingue 5 níveis de dificuldade: "muito fácil", "fácil", "médio", "difícil" e "muito difícil". Contudo, Baker (2001) não estipula as faixas de valores de cada nível, por isso, foram usadas aquelas propostas por Ferreira (2018), conforme demonstra a Tabela 6.

Tabela 6 – Faixa de valores dos níveis de dificuldade dos itens.

Nível	Faixa de valores
Muito fácil	< -1.28
Fácil	-1.27 a -0.52
Médio	-0.51 a 0.51
Difícil	0.52 a 1.27
Muito difícil	> 1.28

Fonte: Elaboração dos autores.

Resultados da análise empírica dos itens

Os valores das médias do índice de dificuldade dos itens e da habilidade dos alunos para cada subteste são apresentados na Tabela 7. Observa-se que a habilidade média dos alunos para os três subtestes é de 0.00 e que o nível médio de dificuldade dos itens dos subtestes da NC, do ICTS e do CC se revelou muito fácil, médio e difícil, respectivamente.

Tabela 7 – Média do índice de dificuldade e das habilidades dos alunos por subteste.

Subteste	b (*DP*)	θ (*DP*)
Natureza da ciência (NC)	-1.29 (1.58)	0.00 (0.78)
Impacto da ciência e da tecnologia na sociedade (ICTS)	-0.04 (1.93)	0.00 (0.72)
Conteúdo da ciência (CC)	1.03 (2.71)	0.00 (0.79)

Nota. b = índice de dificuldade; θ = habilidade ou traço latente; *DP* = desvio padrão.

Fonte: Elaboração dos autores.

Considerando que o parâmetro da dificuldade do item, na TRI, representa o nível de habilidade necessário para que o respondente possa respondê-lo corretamente (PASQUALI, 2009a) e comparando os valores das habilidades dos alunos com os índices de dificuldade dos itens dos três subtestes, nota-se que, em geral, os itens dos subtestes da NC e do ICTS estão adequados à habilidade dos alunos. O mesmo não é constatado nos itens do subteste do CC, cujas habilidades para lhes responder são superiores àquelas apresentadas pelos alunos, mostrando-se inadequados para este grupo de alunos.

Analisando os itens individualmente (Tabela 8 e Gráfico1), observa-se uma quantidade considerável de itens muito fáceis nos subtestes da NC (três = 50%), do ICTS (dois = 33.3%) e do CC (11 = 43%). Nota-se, também, a presença de um item (17%) do subteste do ICTS e sete itens (30%) do CC categorizados no nível "muito difícil".

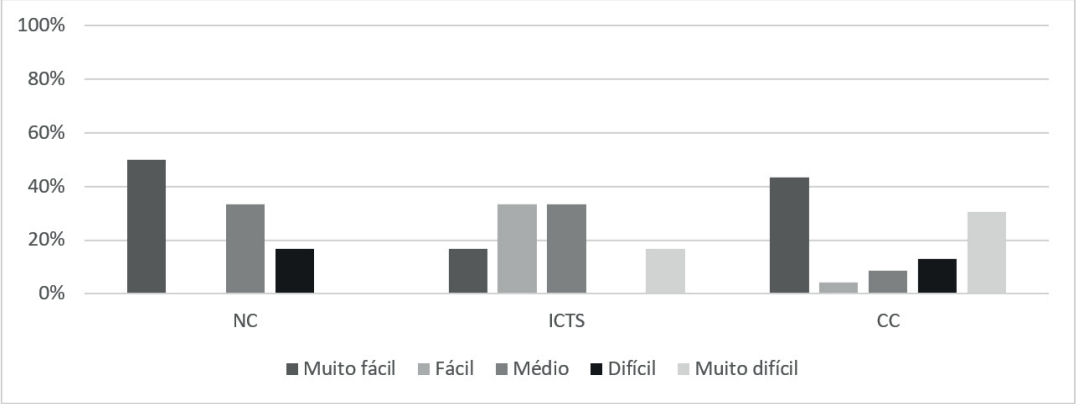
Tabela 8 – Índices e níveis de dificuldade dos itens.

NC			CC		
Item	Dificuldade (b)	Nível	Item	Dificuldade (b)	Nível
1	0.81	D	*17*	-2.01	MF
2	-1.83	MF	*18*	-2.88	MF
3	-2.17	MF	*19*	0.27	M
4	-4.01	MF	*20*	-0.41	MF
5	-0.45	M	*21*	4.88	MD
6	-0.11	M	*22*	0.72	D
			23	3.91w	MD
	ICTS		*24*	5.45	MD
7	-1.29	MF	*25*	-0.76	MF
8	0.06	M	*26*	1.08	D
9	-1.45	MF	*27*	9.66	MD
10	-0.47	M	*28*	-0.56	MF
11	4.11	MD	*29*	1.77	MD
12	-1.23	F	*30*	-0.17	MF
			31	2.52	MD
	CC		*32*	1.20	D
13	-0.96	F	*33*	-0.73	MF
14	-0.52	MF	*34*	-0.25	MF
15	0.28	MF	*35*	-0.43	MF
16	1.58	MD			

Nota. MF = muito fácil; F = fácil; M = médio; D = difícil; MD = muito difícil.

Fonte: Elaboração dos autores.

Gráfico 1 – Percentagem de itens de cada subteste pr nível de dificuldade.



Fonte: Elaboração dos autores.

Este fato indica a necessidade de revisão destes itens, principalmente daqueles que foram categorizados no nível “muito difícil”. Dentre os sete itens do subteste do CC categorizados neste nível, um pertence à área das ciências naturais (item 16), dois à área da química (itens 21 e 23), dois à área da física (itens 24 e 27) e dois à área da biologia (itens 29 e 31). Este nível de dificuldade evidenciado pode estar relacionado com diversos fatores, entre eles a ambiguidade, a clareza das informações, o vocabulário, os termos específicos de cada área, o domínio cognitivo exigido pelo item e as competências dos alunos.

Por conseguinte, aconselha-se o encaminhamento destes itens, assim como o item do subteste do ICTS, novamente ao painel de especialistas, para que este forneça novos pareceres mediante os resultados obtidos no teste piloto, contribuindo para o aperfeiçoamento destes itens e, consequentemente, do instrumento como um todo. Além disso, faz-se necessário aplicar o instrumento a um número maior de indivíduos e de diferentes regiões do país, para que seja possível inferir se a dificuldade dos itens está relacionada com problemas técnicos do instrumento ou com a ausência de competências dos alunos para responder corretamente àquelas requeridas pelos itens.

No que se refere aos itens muitos fáceis, este resultado indica que a maior parte dos inquiridos possui as competências necessárias para respondê-los. Este fato fornece um conjunto de informações importantes sobre o desenvolvimento dessas competências nas disciplinas científicas do 3º ciclo do ensino básico das escolas avaliadas, objetivo deste instrumento. Contudo, torna-se indispensável a revisão deste conjunto de itens, para que sejam identificadas, caso existam, possíveis falhas de construção que levem os inquiridos a responder corretamente aos itens, como, por exemplo, a obviedade da resposta ou a memorização de fatos triviais.

CONSIDERAÇÕES FINAIS

Com o intuito de recolher evidências de validade baseadas no conteúdo para a elaboração do instrumento piloto de avaliação da literacia científica dos alunos no final do 3º

ciclo do ensino básico, no âmbito de um doutoramento em curso, este estudo abordou os seguintes aspetos: a definição dos domínios cognitivos; a definição do universo do conteúdo; a definição da representatividade do conteúdo; a elaboração da tabela de especificação; a construção do instrumento; a análise teórica dos itens; e a análise empírica dos itens.

Como resultado, foram elaborados um total de 64 itens, no formato “verdadeiro-falso-não sei”, que avaliam os domínios cognitivos de compreender, analisar e avaliar problemas e situações cotidianas que envolvem as competências presentes nos principais documentos curriculares portugueses das Ciências Físicas e Naturais. Dentre estes, foram selecionados 35 itens para compor a primeira versão do instrumento em desenvolvimento, a qual foi aplicada em oito agrupamentos de escolas/escolas não agrupadas da região sul de Portugal continental no início do ano letivo de 2020/2021.

A análise empírica dos itens evidenciou que, devido aos altos níveis de dificuldade, sete itens não são compatíveis com as habilidades dos alunos avaliados para determinadas competências, uma vez que foram categorizados como muito difíceis. A fim de que o instrumento apresente evidências de validade baseadas no conteúdo e possibilite o uso dos resultados da avaliação da literacia científica dos alunos do 3º ciclo do ensino básico para a melhoria da qualidade do ensino das ciências neste ciclo, estes itens devem ser reencaminhados aos especialistas para que sejam revistos ou, se necessário, eliminados, para que o nível de habilidade requerido pelos itens do instrumento esteja de acordo com o dos alunos respondentes. No caso dos itens categorizados como muito fáceis, recomenda-se a revisão, e possível eliminação, caso sejam encontrados problemas relacionados com a qualidade técnica.

Com o propósito de recolher informações mais abrangentes e indicadores mais compatíveis com aqueles da população para a qual o instrumento está a ser elaborado, sugere-se que, a partir da revisão dos itens, um novo teste piloto seja realizado, alargando a quantidade de respondentes e de regiões avaliadas.

REFERÊNCIAS

AAAS. **Project 2061:** science for all americans. Washington, DC: Oxford University Press, 1989.

AAAS. **Project 2061:** benchmarks for science literacy. Washington, DC: Oxford University Press, 1993.

AERA; APA; NCME. **Standards for educational and psychological testing**. Washington, DC: American Educational Research Association, 2014.

ALEXANDRE, N. M. C.; COLUCI, M. Z. O. Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. **Ciência & Saúde Coletiva**, v. 16, n. 7, p. 3061–3068, 2011.

ANDERSON, L. W. et al. **A taxonomy for learning, teaching and assessing:** a revison of Bloom's taxonomy of educational objectives. New York: Addison Wesley Longman, 2001.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. DA C. **Teoria da resposta ao item:** conceitos e aplicações. São Paulo: ABE - Associação Brasileira de Estatística, 2000.

ARAUJO, E. A. C.; ANDRADE, D. F.; BORTOLOTTI, S. L. V. Teoria da resposta ao item. **Revista da Escola de Enfermagem USP**, v. 3, n. especial, p. 1000–1008, 2009.

BAKER, F. B. **The basics of item response theory**. Washington, DC: ERIC, 2001.

BERK, R. A. A consumer's guide to multiple choice item formats that measure complex cognitive outcomes. In: National evaluation systems (Eds.). **From policy to practice**. Amherst, MA: Pearson Publishing, 1996. p. 101–127.

BROOKHART, S. M. **The art and science of classroom assessment:** the missing part of pedagogy. Washington, DC: The George Washington University, Graduate School of Education and Human Development, 1999.

BURTON, R. F.; MILLER, D. J. Statistical modelling of multiple/choice and true/false tests: ways of considering, and of reducing, the uncertainties attributable to guessing. **Assessment & Evaluation in Higher Education**, v. 24, n. 4, p. 399–411, 1999.

CETIC. **Pesquisa sobre o uso das tecnologias da informação e comunicação nas escolas brasileiras – TIC Educação 2013**. São Paulo: Comitê Gestor da Internet no Brasil, 2014.

CHANDRATILAKE, M.; DAVIS, M.; PONNAMPERUMA, G. Assessment of medical knowledge: The pros and cons of using true/false multiple choice questions. **Medical Education**, v. 24, n. 4, p. 225–228, 2011.

CIZEK, G. J. Validating test score meaning and defending test score use: different aims, different methods. **Assessment in Education: Principles, Policy and Practice**, v. 23, n. 2, p. 212–225, 2016.

DEPRESBITERIS, L.; TAVARES, M. R. **Diversificar é preciso...:**Instrumentos e técnicas de avaliação de aprendizagem. São Paulo: Senac São Paulo, 2009.

DGE. **Aprendizagens Essenciais**. Disponível em: <https://www.dge.mec.pt/aprendizagens-essenciais-0>. Acesso em: 18 out. 2019.

EBEL, R. L. **The comparative eftectiveness of true-false and multiple choice achievement test items**. American Educational Rsearch Association Annual Meeting. **Anais**...New York: 1971

EBEL, R. L. **Essentials of educational measurement**. 3. ed. Englewood Cliffs: Prentice Hail International, Inc., 1979.

EBEL, R. L.; FRISBIE, D. A. **Essentials of educational measurement**. 5. ed. Englewood Cliffs: Prentice Hail International, Inc., 1991.

FERRAZ, A. P. C. M.; BELHOT, R. V. Taxonomia de Bloom: revisão teórica e apresentação das adequações do instrumento para definição de objetivos instrucionais. **Gest. Prod.**, v. 17, n. 2, p. 421–431, 2010.

FERREIRA, E. A. **Teoria de tesposta ao item – TRI:** análise de algumas questões do ENEM: habilidades 24 a 30. Dissertação de Mestrado, Universidade Federal da grande Dourados, Mato Grosso do Sul, Brasil, 2018.

FIVES, H. et al. Developing a measure of scientific literacy for middle school students. **Science Education**, v. 98, n. 4, p. 549–580, 2014.

FRISBIE, D. A. Multiple choice versus true-false: a comparison of reliabilities and current validities. **Journal of Educational Measurement**, v. 10, n. 4, p. 297–304, 1973.

FRISBIE, D. A.; BECKER, D. F. An analysis of textbook advice about true-false tests. **Applied Measurement in Education**, v. 4, n. 1, p. 67–83, 1991.

GALVÃO, C. et al. **Ciências físicas e naturais - orientações curriculares para o 3º ciclo do ensino básico**. Lisboa: Ministério da Educação, 2001.

GATES, F. R.; HOYER, W. D. Measuring miscomprehension: a comparison of alternate formats. In: LUTZ, R. J. (Ed.). **Advances in Consumer Research**. Provo, UT: Association for Consumer Research, 1986. p. 143–146.

GIPPS, C. V. **Beyond testing:** towards a theory of educational assessment. Washington, DC: The Falmer Press, 2003.

GORMALLY, C.; BRICKMAN, P.; LUTZ, M. Developing a test of scientific literacy skills (TOSLS): measuring undergraduates' evaluation of scientific information and arguments. **CBE Life Sciences Education**, v. 11, n. 4, p. 364–377, 2012.

HALADYNA, T. M. **Developing and validating multiple-choice test items**. 3. ed. London: Lawrence Erlbaum Associates, 2004.

HALADYNA, T. M. Developing test items for course examinations. **IDEA**, v. 70, p. 1–16, 2018.

HALADYNA, T. M.; RODRIGUEZ, M. C. **Developing and validating test items**. New York: Taylor & Francis Group, 2013.

KANE, M. T. Validating the interpretations and uses of test scores. **Journal of Educational Measurement**, v. 50, n. 1, p. 1–73, 2013.

LAUGKSCH, R. C. Scientific literacy: a conceptual overview. **Science Education**, v. 84, n. 1, p. 71–94, 2000.

LAUGKSCH, R. C.; SPARGO, P. E. Construction of a paper-and-pencil test of basic scientific literacy based on selected literacy goals recommended by the american association for the advancement of science. **Public Understanding of Science**, v. 5, n. 4, p. 331–359, 1996a.

LAUGKSCH, R. C.; SPARGO, P. E. Development of a pool of scientific literacy test-items based on selected AAAS literacy goals. **Science Education**, v. 80, n. 2, p. 121–143, 1996b.

MAIHOFF, N. A.; MEHRENS, W. A. **A comparison of alternate-choice and true-false item forms used in classroom examinations**. Annual Researchers Meeting of the National Council on Measurement in Evaluation. **Anais**...Illinois: 1985.

MARTINS, G. O. et al. **Perfil dos alunos à saída da escolaridade obrigatória**. Lisboa: Ministério da Educação e Ciência - DGE, 2017.

MESSICK, S. Validity. In: LINN, R. L. (Ed.). **Educational measurement**. 3. ed. New York: Macmillan, 1989. p. 3–209.

MILLER, J. D. Scientific literacy: a conceptual and empirical review. **Daedalus**, v. 112, n. 2, p. 29–48, 1983.

MILLER, M. D.; LINN, R. L.; GRONLUND, N. E. **Measurement and assessment in teaching**. 10. ed. New Jersey: Pearson Education, Inc., 2009.

PASQUALI, L. **Psicometria teoria dos testes na psicologia e na educação**. 4. ed. Petrópolis: Vozes, 2009a.

PASQUALI, L. Psicometria. **Revista da Escola de Enfermagem da USP**, v. 43, n. spe, p. 992–999, 2009b.

POPHAM, W. J. **Classroom assessment:** what teachers need to know. 8. ed. Los Angeles: Pearson, 2017.

RAYMUNDO, V. P. Construção e validação de instrumentos um desafio para a psicolinguística. **Letras de Hoje**, v. 44, n. 3, p. 86–93, 2009.

RUBIO, D. M. G. et al. Objectifyng content validity: Conducting a content validity study in social work research. **Social Work Research**, v. 27, n. 2, p. 94–104, 2003.

RUSH, B. R.; RANKIN, D. C.; WHITE, B. J. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. **BMC Medical Education**, v. 16, n. 250, p. 1–10, 2016.

RUSSEL, M. K.; AIRASIAN, P. W. **Avaliação em sala de aula:** conceitos e aplicações. 7. ed. Porto Alegre: AMGH, 2014.

SERRA, P.; GALVÃO, C. Evolução do currículo de ciências em Portugal: será Bloom incontornável? **Interações**, v. 11, n. 39, p. 255–271, 2015.

SMITH, J. K. Reconsidering reliability in classroom assessment and grading. **Educational Measurement: Issues and Practice**, v. 22, n. 4, p. 26–33, 2003.

TASDEMIR, M. A comparison of multiple-choice tests and true-false tests used in evaluating student progress. **Journal of Instnjctional Psychology**, v. 37, n. 3, p. 258–267, 2010.