Validity evidence based on internal structure of a pilot scientific literacy assessment instrument

MARCELO COPPI[I]
ISABEL FIALHO[II]
MARÍLIA CID[III]
http://dx.doi.org/10.22347/2175-2753v16i51.4330

Abstract
This study presents the validity evidence based on internal structure gathered from the application of a pilot instrument to assess the scientific literacy of students at the end of the 3rd cycle of basic education. The instrument, composed of the subtests nature of science, impact of science and technology on society and content of science, was applied to 176 students in the 10th grade from eight Portuguese schools. Evidence was gathered using the Item Response Theory information function. The analysis revealed that the three subtests show evidence of validity based on internal structure that makes it possible to use the results for decision making. Nevertheless, it was found to be necessary to revise the items of the content of science subtest in order for them to better fit the characteristics of the sample.
Keywords: Assessment; Validity Evidence; Scientific literacy; Instruments; 3rd cycle of basic education.

Submetido em: 03/07/2023
Aprovado em: 27/05/2024

[I] Universidade de Évora (UE), Évora, Portugal; https://orcid.org/0000-0001-6734-7592; e-mail: mcoppi@uevora.pt.
[II] Universidade de Évora (UE), Évora, Portugal; https://orcid.org/0000-0002-1749-9077; e-mail: ifialho@uevora.pt.
[III] Universidade de Évora (UE), Évora, Portugal; https://orcid.org/0000-0002-6009-0242; e-mail: mcid@uevora.pt.

Evidencia de validez basada en la estructura interna de un instrumento piloto de evaluación de la alfabetización científica

Resumen
Este estudio presenta las pruebas de validez basadas en la estructura interna recogidas a partir de la aplicación de un instrumento piloto para evaluar la alfabetización científica de los alumnos del final del 3º ciclo de la enseñanza básica. El instrumento, compuesto por las subpruebas naturaleza de la ciencia, impacto de la ciencia y la tecnología en la sociedad y contenido de la ciencia, se aplicó a 176 alumnos de 10º curso de ocho escuelas portuguesas. Las pruebas se recogieron utilizando la función de información de la Teoría de Respuesta al Ítem. El análisis reveló que las tres subpruebas muestran evidencias de validez basadas en la estructura interna que permite utilizar los resultados para la toma de decisiones. No obstante, se constató la necesidad de revisar los ítems de la subprueba de contenido de ciencias para que se ajusten mejor a las características de la muestra.
Palabras clave: Evaluación; Pruebas de validez; Alfabetización científica; Instrumentos; 3.º ciclo de educación básica.

Evidência de validade baseada na estrutura interna de um instrumento piloto de avaliação de alfabetização científica

Resumo
Este estudo apresenta as evidências de validade baseadas na estrutura interna obtidas a partir da aplicação de um instrumento piloto para avaliar a literacia científica dos alunos no final do 3º ciclo do ensino básico. O instrumento, composto pelos subtestes natureza da ciência, impacto da ciência e da tecnologia na sociedade e conteúdo da ciência, foi aplicado a 176 alunos do 10º ano de oito escolas portuguesas. As evidências foram coletadas usando a função de informação da Teoria de Resposta ao Item. A análise revelou que os três subtestes apresentam evidências de validade baseadas na estrutura interna, o que possibilita o uso dos resultados para a tomada de decisões. Contudo, verificou-se que é necessário revisar os itens do subteste de conteúdo de ciência para que estes se ajustem melhor às características da amostra.
Palavras-chave: Avaliação; Evidências de validade; Literacia científica; Instrumentos; 3.º ciclo do ensino básico.

1 Introduction

Scientific literacy is considered to be one of the key skills for the 21st century and a fundamental aspect for the exercise of citizenship (World Economic Forum, 2015). Although it is still considered a polysemic term, due to the lack of consensus on its concept, scientific literacy can be defined as "the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen" (Organisation for Economic Co-Operation and Development, 2017, p. 22).

Introduced in the scientific field in the 1950s (Deboer, 2000; Laugksch, 2000), the term scientific literacy has been widely used as a slogan by educators, researchers and politicians, who use it in several contexts, in a broad and vague way, in order to describe a set of science-related skills (Deboer, 2000). However, for 30 years, scientific literacy presented itself as an uncertain and undefined term and only after the 1980s scientific literacy studies gained momentum in the scientific community.

One of the most influential studies of that decade was the article "Scientific literacy: a conceptual and empirical review", published in Daedalus - Journal of the American Academy of Arts and Sciences, by Miller (1983) (Laugksch, 2000; Laugksch; Spargo, 1996). In this paper, the author presents a multidimensional concept for scientific literacy, encompassing the dimensions of understanding the enterprise of science, knowledge of the main contents of science, and awareness of the impact of science and technology on society, and also proposes procedures for assessing scientific literacy.

Grounded on what Miller (1983) proposed, numerous studies have been conducted in countries such as the United States, Canada, Japan, and the European Community, and several instruments have been developed in order to assess the population's level of scientific literacy (Laugksch; Spargo, 1996). However, researchers claim that only a small amount of instruments assess the three dimensions together (Fives; Huebner; Birnbaum; Nicolich, 2014; Gormally; Brickman; Lutz, 2012; Laugksch; Spargo, 1996) and that many of them lack in providing the validity evidence necessary for the development of assessment instruments (Laugksch; Spargo, 1996).

Fives, Huebner, Birnbaum and Nicolich (2014) add that, until then, no instrument had been developed to assess the scientific literacy of students in the 3rd cycle of basic education. In the Portuguese education system, the transition from the 2nd to the 3rd cycle sets up the inclusion of a new scientific discipline, namely Physical-Chemistry, which is now part of the Physical and Natural Sciences area, jointly with the

natural sciences discipline, each one with its respective teachers, particularities and specificities.

Moreover, in Portugal, after the 3rd cycle of basic education, students are no longer required to take scientific-technological subjects, further emphasizing the importance of such subjects for students' education. In Secondary Education, the next three-year cycle (10th, 11th and 12th grades), it is only those students who opt for the Science and Technology course who will attend classes in subjects aimed at developing scientific literacy (Biology and Geology, Physics and Chemistry).

Within this context, a research project was developed in the scope of an ongoing PhD project, whose objective is to develop and validate an instrument to assess the level of scientific literacy of Portuguese students at the end of the 3rd cycle of basic education. Given the relative lack of assessments directed to this cycle of education, it is intended that this instrument will provide indicators that can assist in monitoring the progress of students' scientific education, at regional and national levels (if applied on a large scale, in a measurement perspective).

It is also hoped that this scientific literacy assessment tool might be used by schools in a guiding way. Teachers, especially those who teach natural sciences and Physical-Chemistry, will also be able to use the data and information from the application of the instrument to redesign their lessons, teaching plans and classroom practice, aiming for students to complete, as far as possible, the 3rd cycle of basic education as scientifically literate citizens, able to deal with everyday scientific issues and questions, and prepared to deepen their scientific knowledge at the next level of education.

2 Theoretical framework

The development of an evaluation instrument requires the gathering of reliable indicators capable of demonstrating its high technical quality, enabling the use of its results for the proposed purposes. According to researchers in the field of assessment, validity is the main attribute for the quality of assessment instruments (Depresbiteris; Tavares, 2017; Haladyna; Rodriguez, 2013; Popham, 2018; Russel; Airasian, 2014). For Popham (2018), the most important assessment concept is validity. Russel and Airasian (2014, p. 26), in compliance with the same idea, recognize that "the single most important characteristic of good assessment is its ability to help the teacher make appropriate decisions. This characteristic is called validity".

Traditionally, the concept of validity refers to the ability of a test to measure what it was designed to measure (Gipps, 2003). Within this perspective, the literature highlights three types of validity: content validity, which competes to the relevance and representativeness of the contents that will be assessed; construct validity, which refers to the test's ability to assess the construct that is being measured; and criterion validity, which is related to the prediction of performance regarding some external criterion (Alexandre; Coluci, 2011; Gipps, 2003).

However, more contemporary literature recognizes validity as a unitary concept (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014; Miller; Linn; Gronlund, 2009; Popham, 2018), which represents "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014, p. 11). According to this perspective, what was previously understood as validity types gives way to validity evidence types, which are distributed into five categories, namely: evidence based on content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence based on consequences of tests (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014; Popham, 2018).

According to the Standards for Educational and Psychological Testing (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014, p. 21), validity evidence may indicate "the need for refining the definition of the construct, may suggest revisions in the test or other aspects of the testing process, and may indicate areas needing further study". However, the document claims that all five types of validity evidence are not always needed for all validation processes of assessment instruments, but that instead, "support is needed for the propositions underlying each interpretation for a specific use" (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014, p. 14).

Since the evidence of validity based on content of this instrument, which are the primary sources of validity evidence of an assessment instrument (Kane, 2013), has already been collected and published by Coppi, Fialho and Cid (2022, 2023), this

study aimed at collecting evidence of validity based on internal structure of the pilot instrument for the development of the final version of the instrument for assessing the scientific literacy of Portuguese students at the end of the 3rd cycle of basic education. This type of evidence is considered one of the fundamental forms for the analysis of the validation process of assessment instruments, since it "constitutes the direct way to verify the hypothesis of legitimacy of the behavioral representation of latent traits" (Pasquali, 2009a, p. 996).

3 Validity evidence based on internal structure

According to Braun (2016), validity evidence based on internal structure refers to the statistical analysis of items and their scores in order to ascertain the primary and, if any, secondary dimensions measured by a test. It is stated in the Standards (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014, 2014, p. 16) that analyses of the internal structure of an assessment instrument "can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based".

This document also clarifies that the types of analysis for the gathering of validity evidence based on internal structure, as well as for its interpretations, depend on the use for which the instruments are proposed (American Educational Research Association; American Psychological Association; National Council on Measurement in Education, 2014). Pasquali (2009a, p. 996) supports the idea, claiming that this gathering of evidence "can be worked from several angles: the analysis of the behavioral representation of the construct, the analysis by hypothesis, the IRT [Item Response Theory] information curve".

For the behavioral representation analysis of the construct, internal consistency and factor analysis are used. These allow to demonstrate the adequacy of the construct representation by the instruments (Pasquali, 2009b). According to the author (Pasquali, 2009b, p. 170), the internal consistency analysis comprises the calculation of the "correlation that exists between each item of the test and the rest of the items or the total (total score) of the items", while the "factor analysis has as its logic precisely to verify how many common constructs are necessary to explain the covariances (the intercorrelations) of the items" (Pasquali, 2009b, p. 173).

Analysis by hypothesis evaluates the relationship between an instrument's score and a particular external criterion. This analysis is based on the capacity of an assessment instrument "to be able to discriminate or predict a criterion external to itself; for example, to discriminate criterion groups that differ specifically on the trait that the test measures" (Pasquali, 2009b, p. 175).

The IRT information curve is used as a graphical representation of the information functions of the item and the test, able to reveal for which range of proficiency levels the instrument is particularly valid and for which ranges it is not (Pasquali, 2009b). The author also claims that the information function of the test can also be represented by the standard error of estimation, which represents the inverse of the information function and allows the analysis of the accuracy of an assessment instrument.

Braun (2016) includes multidimensional scaling, residual analysis, and differential item functioning analysis as strategies for gathering validity evidence based on internal structure. These analyses, as well as those mentioned previously, seek to "provide information about the psychometric adequacy of the scale and validity evidence based on internal structure" (Nunes; Noronha, 2011, p. 28).

Although behavioral representation analyses are most commonly found in the literature, the most recent validation process research has been using latent modeling, via IRT, as they are also concerned with the accuracy aspects of the assessment instruments (Mendonça Filho, 2017). The IRT information functions, as indices of the evaluation of accuracy, are classified as validity evidence based on internal structure, since they are able to assess the saliency of the main dimensions underlying an assessment, saliency which is related to the reliability of the internal consistency of the instrument (Braun, 2016).

4 Methodology

4.1 Sample

Participants in this study included 176 10th grade students from eight school groups, selected by convenience (Ghiglione; Matalon, 1992; Hill; Hill, 2005), from the southern region of continental Portugal, with an average age of 15.18 years (SD = 2.5). Of these, 87 individuals (49.43%) were female and 89 (50.57%) were male.

4.2 Instrument

It was used the pilot version of a scientific literacy assessment instrument, which has already been submitted to an initial stage of the validation process, in which the Coppi, Fialho and Cid (2022, 2023) gathered and presented the evidences of validity based on content. This process was carried out in seven steps, as proposed by Pasquali (2009b): 1) definition of the cognitive domains, in which the cognitive or psychological processes that were intended to be assessed were determined; 2) definition of the universe of content, by delimiting the content into teaching units and sub-units; 3) definition of the representativeness of the content, in which the proportion of representation of each content in the instrument was established; 4) preparation of the specification table, which establishes the correspondence between the dimension of the content, the cognitive domains, and the number of items; 5) construction of the instrument, stage in which the format and configuration of the items' wording and the technical guidelines for their development were decided; 6) theoretical analysis of the items, with the participation of 10 experts in the areas of Education Sciences, Biology, Geology, Physics and Chemistry, who appreciated the representativeness, relevance and quality of the items in relation to the content areas and the objectives of the instrument; 7) and empirical analysis of the items, which consisted in the evaluation of psychometric characteristics of the items, more specifically, difficulty indices.

This pilot version of the instrument consists of 35 items in a "true-false-don't-know" format, grouped into three distinct subtests, namely: nature of science (NOS) subtest, consisting of six items; impact of science and technology on society (ISTS) subtest, also containing six items; and content of science (CS) subtest, composed by 23 items.

The set of items that make up the instrument includes the competencies present in the following Portuguese curricular documents of Physical and Natural Sciences of the 3rd cycle of basic education: Curricular Guidelines for the 3rd cycle of Basic Education - Physical and Natural Sciences (Galvão, 2001), Essential Learnings in Natural Sciences (Portugal, 2018a, 2018b, 2018c) and in Physical-Chemistry (Portugal, 2018d, 2018e, 2018f) and Students' Profile by the End of Compulsory Schooling (Martins, 2017). The items assess the cognitive domains of understanding, analysis and evaluation of everyday problems and phenomena that involve a set of skills from the disciplines of Natural Sciences and Physical-Chemistry for their resolution and/or explanation.

4.3 Proceedings

With the authorization of the Directorate-General for Innovation and Curricular Development (DGICD), Monitoring of Surveys in the School Environment, under the registration no. 0740900001, and of the directors and teachers of the participating educational institutions, the link to access the instrument on the LimeSurvey platform was made available to students through the teachers responsible for the application.

The assessment was conducted at the beginning of the 2020/2021 school year, in digital format, in the classroom and in the presence of the teachers. The average response time was 30 minutes.

4.4 Data analysis procedure

To gather evidence of validity based on internal structure was used the IRT technique, which fulfills this role through the information functions (Pasquali, 2009b). Through the application of IRT, in addition to the information functions of the item and subtests, the parameters of difficulty and discrimination of the items, the level of proficiency (θ) of the students and the Kernel density estimate were analyzed.

The two-parameter logistic model was used, since, when compared with the one- and three-parameter logistic models through the analysis of variance, it best fitted the data ($p < .05$). The IRT analyses were performed using RStudio software, version 3.6.0.

5 Results

The first step of the analysis consisted in identifying the psychometric parameters of the items of each subtest. The results showed that the average of the difficulty index (b) of the items of the subtests of the NS, the ISTS and the CS were -1.29 (SD = 1.73), -0.04 (SD = 1.93) and 1.03 (SD = 2.71), respectively. Regarding the discrimination index (a), average values of 1.20 (SD = 1.08), 0.89 (SD = 0.55), and 0.64 (SD = 0.34) were found for the respective subtests. Table 1 summarizes the values of the items of the instrument, by subtest. Items one to six belong to the NOS subtest, items seven to 12 to the ISTS subtest, and items 13 to 35 to the CS subtest.

Table 1 - Item difficulty and discrimination index, by subtest

Subtest									
NOS				**ISTS**			**CS**		
Item	**a**	**b**		**Item**	**a**	**b**	**Item**	**a**	**b**
1	0.87	0.81		7	0.46	-1.29	13	0.97	-0.96
2	0.26	-1.83		8	1.78	0.06	14	0.62	-0.52
3	1.16	-2.17		9	0.63	-1.45	15	0.52	0.28
4	0.51	-4.01		10	1.44	-0.47	16	0.49	1.58
5	1.08	-0.45		11	0.20	4.11	17	1.10	-2.01
6	3.29	-0.11		12	0.84	-1.23	18	0.74	-2.88
M	1.20	-1.29		*M*	0.89	-0.04	19	0.65	0.26
SD	1.08	1.73		*SD*	0.55	1.93	20	0.68	-0.41
							21	0.30	4.88
							22	0.36	0.72
							23	0.23	3.91
							24	0.19	5.45
							25	0.90	-0.76
							26	0.58	1.08
							27	0.17	9.66
							28	0.74	-0.56
							29	0.43	1.76
							30	0.43	-0.17
							31	0.29	2.52
							32	0.70	1.20
							33	1.67	-0.73
							34	0.99	-0.25
							35	0.86	-0.43
							M	0.64	1.03
							SD	0.34	2.71

Note. M = mean; SD = standard deviation; a = discrimination index; b = difficulty index.
Source: The authors (2023).

Next, the descriptive statistics of the θ level of the responding students were calculated. The data derived from this analysis revealed that the average θ of the students in the three subtests is located at the zero value, that is, at the center of the proficiency scale, as shown in Table 2.

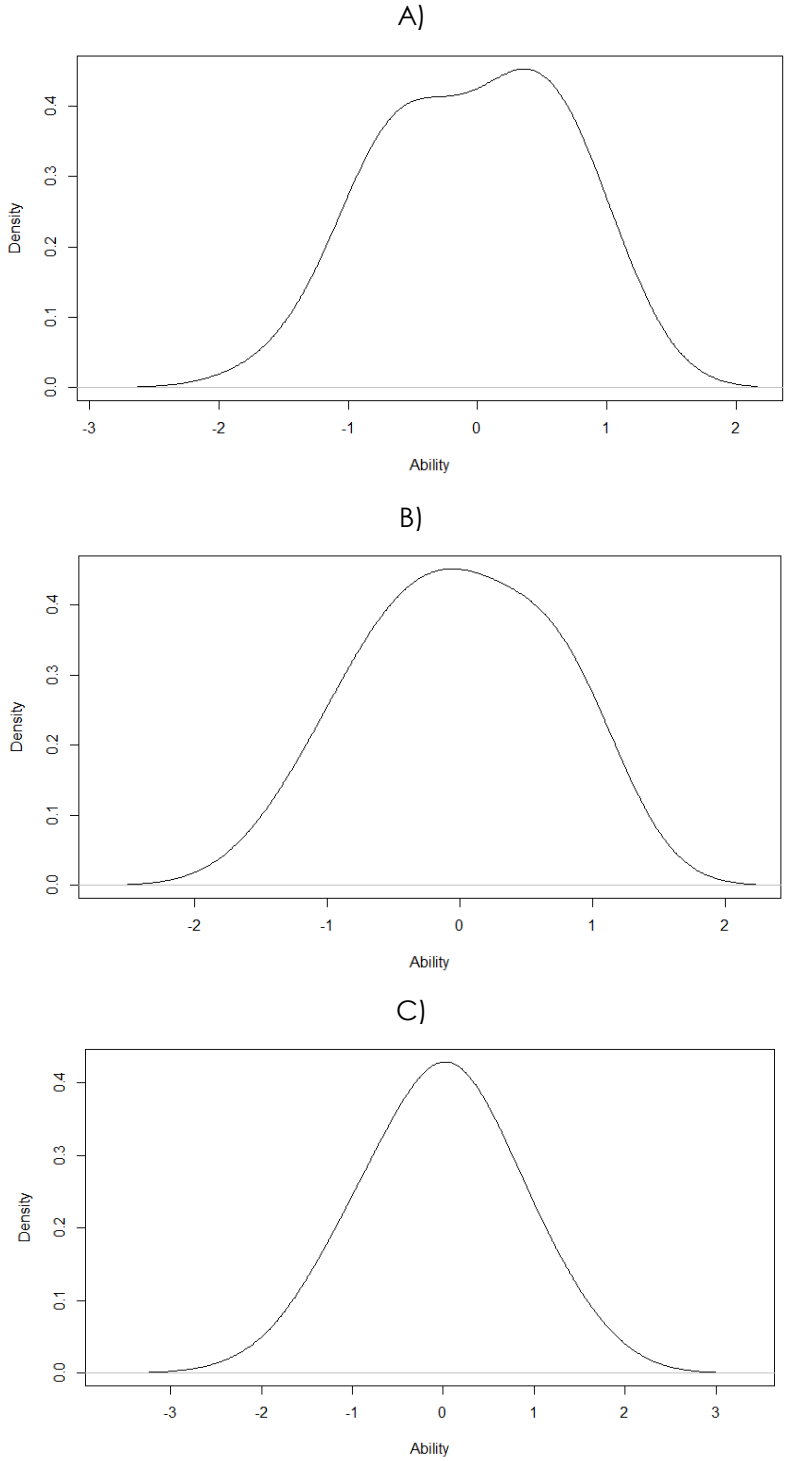
Table 2 - Average level of θ of the students by subtest

	θ		
	NOS	**ISTS**	**CS**
M	0.00	0.00	0.00
SD	0.78	0.72	0.79
Maximum	1.06	1.05	2.03
Minimum	-1.70	-1.48	-2.36

Note. θ = student proficiency; M = mean; SD = standard deviation.
Source: The authors (2023).

Additionally, by means of Kernel density estimation, student densities were calculated as a function of θ for each subtest. The results show that most students are concentrated between θ of -1 and 1, with the apex around zero, as presented in Figure 1.

Figure 1 - Kernel density estimation as a function of the proficiency of student respondents in each subtest. A) NOS subtest; B) IST subtest; C) CS subtest.

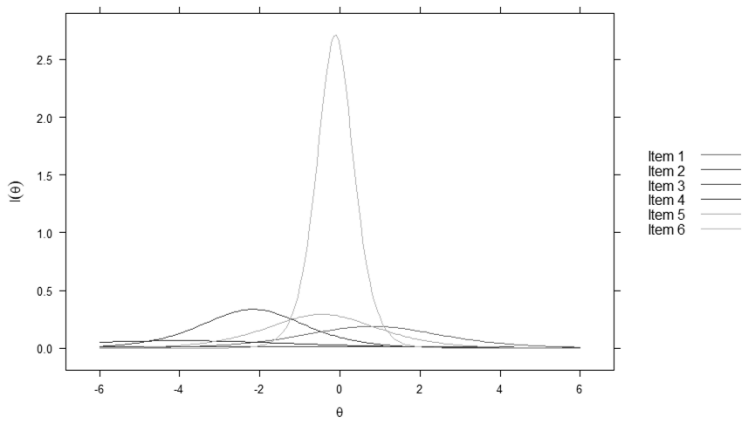


Source: The authors (2023).

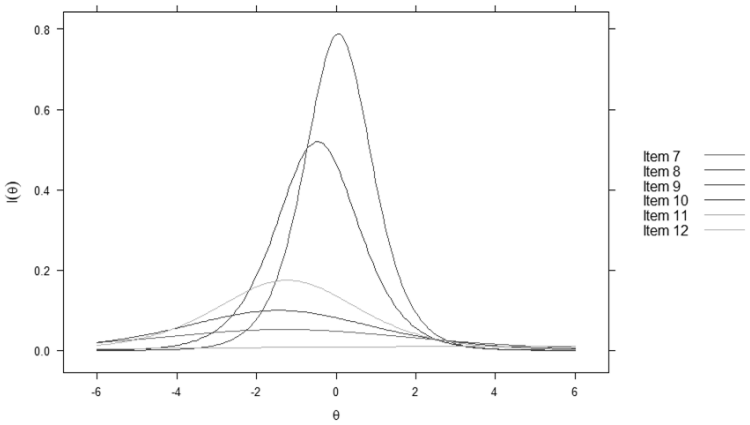
Finally, in order to analyze the data from the item and test information functions, we designed the item information curves (IIC) and the (sub)test information curves (TIC), shown in Figures 2 and 3, respectively. Along with the TIC of each subtest, the respective standard error of estimation is plotted.

Figure 2 - Item information curve by subtest. A) NOS subtest; B) ISTS subtest; C) CS subtest

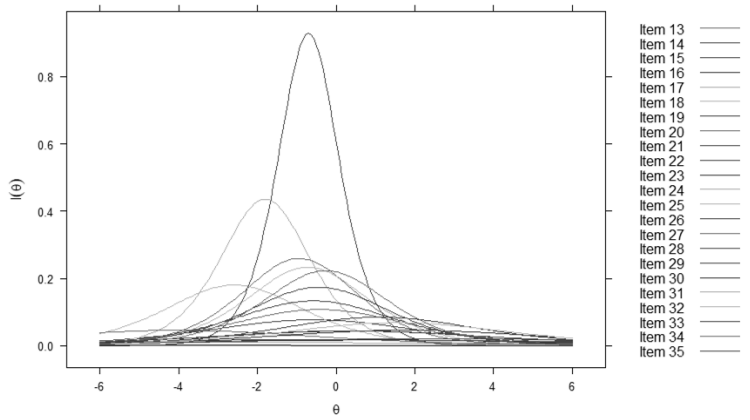
A)



B)



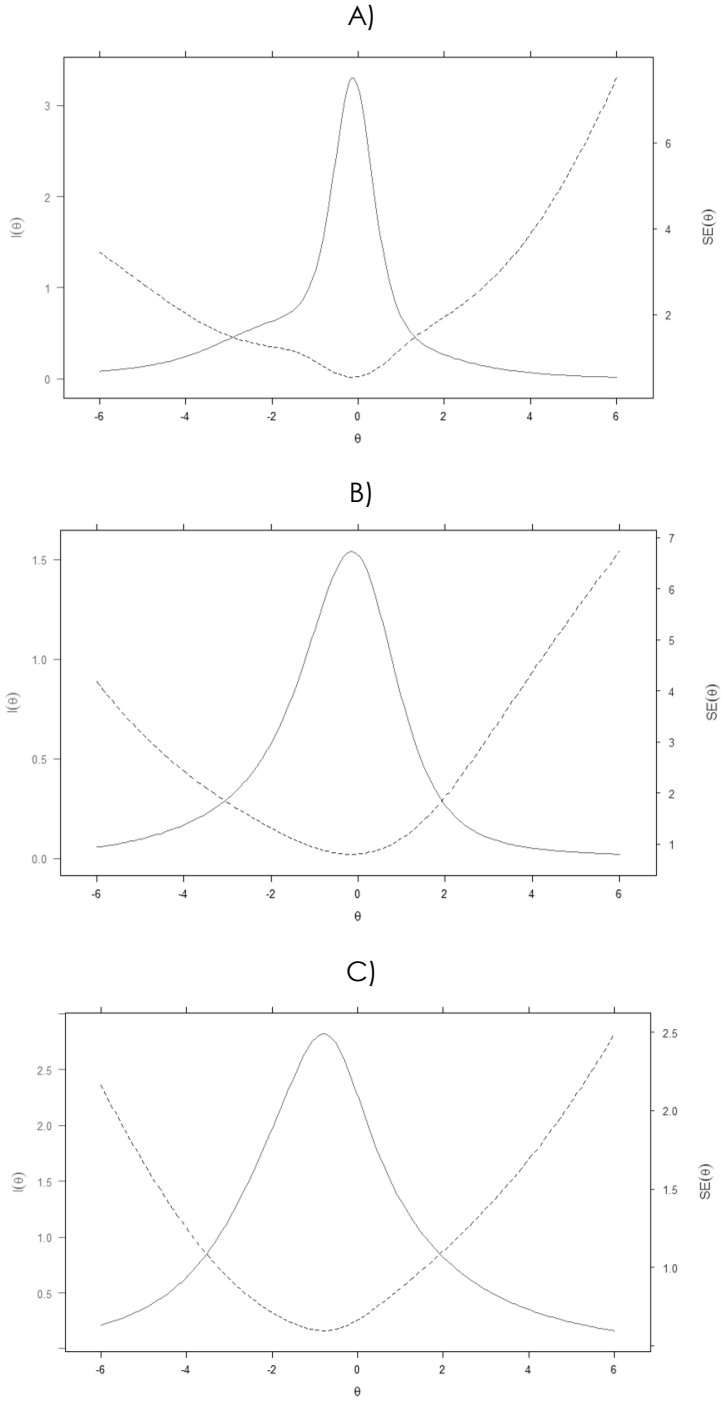
C)



Note. I(θ) = amount of subtest information; θ = proficiency.
Source: The authors (2023).

By observing Figure 2, it can be seen that the amount of information in most items of the three subtests is below the value of 0.5, with the exception of item 6, belonging to the NOS subtest, item 8, of the ISTS subtest, and item 33, of the CS subtest, which were the items with the highest amount of information.

Figure 3 - Test information curve and standard error of estimation by subtest. A) NOS subtest; B) ISTS subtest; C) CS subtest.



Note. I(θ) = amount of subtest information; θ = proficiency; SE(θ) = standard error of estimation; Continuous line (——) = I(θ); Dashed line (----) = SE(θ).
Source: The authors (2023).

It can be noted that the greatest amount of information from the NOS and ISTS subtests is produced in the θ range between -1 and 1, with a peak at θ of -0.06. In the case of the CS subtest, this range is located between -3 and 1 on the θ scale, with a peak at θ of -0.82.

Regarding the amount of total information available on each subtest, the IRT information function analysis indicated that the NS, ISTS, and CS subtests produce 7.11, 5.27, and 13.40 of information, respectively, as shown in Table 3. It could also be seen that the amount of information available in the range of θ between -3 and 3, a typical proficiency spectrum (Baker; Kim, 2017), was 6.11 for the NOS subtest, 4.54 for the ISTS subtest, and 10.09 for the CS subtest, as shown in Table 3.

Table 3 - Amount of information from each subtest

Subtest	Number of items	Total amount of information	Amount of information in the proficiency range of -3 to 3
NOS	6	7.11	6.11 (86.04%)
ISTS	6	5.27	4.54 (86.20%)
CS	23	13.40	10.09 (75.28%)

Source: The authors (2023).

6 Discussion

An educational assessment instrument, like any other test, must present validity evidence that allows its results to be used for the proposed uses. This section discusses the validity evidence based on internal structure collected by applying the pilot instrument of the ALCE, for which IRT was used.

From a set of techniques that IRT has, the IICs were initially analyzed. These allow the analysis of the amount of psychometric information that an item contains at all points along the continuum of θ that it represents (Pasquali, 2009b). Although rarely performed, prioritizing the function for the test (Baker; Kim, 2017), the item information function

> is a powerful instrument for item analysis, making it possible to know not only how much information an item accumulates at a given value of θ, but also at what value of θ the item has the most information (Couto; Primi, 2011).

Reeve and Fayers (2005) and Baker and Kim (2017) discuss the influence of item discrimination and difficulty parameters for their respective information functions. According to these authors, the higher the item's discrimination index, the greater the

amount of information it brings to θ, since "higher discrimination means the item can better differentiate among individuals who lie near the thereshold value" (Reeve; Fayers, 2005, p. 60). In the case of the difficulty index, it determines the location of the information curve on the horizontal axis of θ (Baker; Kim, 2017; Reeve; Fayers, 2005).

Comparing the results presented in Table 1 and Figure 2, it can be seen that the three items that revealed the most information - items 6, 8 and 33 - had the highest discriminative power: 3.29, 1.78 and 1.67, respectively, confirming the findings of Baker and Kim (2017) and Reeve and Fayers (2005).

However, Klein (2013, p. 43) states that "items with difficulty parameter 'b' close to the student's proficiency provide more information". From this perspective, the amount of information in the items is not only influenced by discrimination, but also by the difficulty of the items, aiding the analysis of the results. These results showed that, in general, most of the items with low amount of information also showed lower discriminative power and difficulty index more distant from the average θ of the students, of zero.

As for the location of the IICs, a characteristic determined by the difficulty index of the items, it can be noted that, in general, these are located between θ of -2 and 2 on the proficiency scale. Thus, it might be stated that the items of the three subtests best assess the constructs corresponding to θ in this range (Reeve; Fayers, 2005). Furthermore, although most items did not present high levels of information, their information apexes are distributed along the continuum of the proficiency scale and therefore become informative for the purposes of the instrument (Edelen; Reeve, 2007).

As referred to in the literature (Baker; Kim, 2017; Klein, 2013; Pasquali, 2009b; Reeve; Fayers, 2005), the information functions of items can be summed, originating the test information function. This is represented graphically by the TIC and is used to evaluate the performance of assessment instruments, ensuring that they provide "adequate precision across the entire range of interest as well as maximizing precision along critical segments of the construct continuum" (Edelen; Reeve, 2007, p. 6).

Analyzing the TICs of each subtest, it was found that, in accordance with the proposed by Pasquali (2009b), the intervals of θ for which the instrument's results are particularly valid are located between -1 and 1, in the NOS and ISTS subtests, and between -3 and 1 in the CS subtest. Consequently, the same intervals comprise the range of θ in which the standard error of estimation is smaller, since this refers to "the

inverse of the square root of the information function" (Edelen; Reeve, 2007, p. 6) and its graphical function is inversely proportional to the IRT information curve (Baker; Kim, 2017; Pasquali, 2009b). In this sense, it is assumed that the intervals of θ from -1 to 1, in the NOS and ISTS subtests, and -3 to 1, in the CS subtest, are those in which the three subtests are most accurate.

It can also be ascertained that the accuracy of the subtests is not the same across the scale. In all three subtests, accuracy decreases as the scale approaches the extremes of the TIC, since the error curve exceeds the information curve, making these intervals produce more information error than legitimate information (Andrade; Laros; Gouveia, 2010; Cuartas, 2020).

According to Baker and Kim (2017), the optimal information function should be horizontal over the widest possible area, making the largest amount of information include as much θ as possible. However, the authors claim that such an event is very unlikely to occur. Therefore, Baker and Kim (2017, p. 134) argue that "the test information function should be rounded in appearance over the ability range of most interest". In this instrument's sense, this range corresponds to the average θ of the student respondents, which has been shown to be zero for all three subtests.

By comparing the TICs of the three subtests with the respective mean θ values of the student respondents and with the results of the Kernel density estimation analysis, it is possible to infer that the NOS and ISTS subtests showed the highest accuracy and the best ability to assess the student respondents. This is because the apex of the information and density curves conform to the average θ presented by the students, zero. For the CS subtest, there was a slight difference between the results of the three analyses, as the highest amount of information is in the range of -3 and 1, with a peak at θ of -0.82, while the average θ of the respondents is equal to zero and the highest density of the students is located around θ zero. However, although the items of the CS subtest prove to be moderately less accurate, they are still able to generate valid information by means of assessing students' scientific literacy.

This slightly divergence found between the results referring to the CS subtest seems to be associated with the difficulty and discrimination parameters of the items that compose it. According to Baker and Kim (2017), the difficulty of the items should be located around the midpoint of the θ range of interest and the discrimination should be as broad as possible. However, the results regarding these item parameters revealed that the CS subtest had the lowest percentage of items whose difficulty index

was close to θ zero (26%, compared to 33% of the NOS and ISTS subtests), taking into account the range between -0.5 and 0.5, in addition to the lowest average item discrimination (M = 0.64; SD = 0.34), when compared to the other two subtests (M = 1.20; SD = 1.08 and M = 0.89; SD = 0.55, respectively).

Baker and Kim (2017, p. 134) also clarify that "items whose values of the item difficulty parameters are within the ability range of interest should have larger values of the item discrimination parameters than items whose values of the item difficulty parameters are outside this range". This condition was not satisfied, since the highest discrimination values were evidenced in items whose difficulty parameter is located farther from the θ of interest. These results corroborate the results of the study by Coppi, Fialho and Cid (2022, 2023) and demonstrate the need to revise the items of the CS subtest in order to try and reduce the difficulty level of items whose index was too elevated.

Regarding the amount of information available in each subtest, it is observed that the CS subtest presented the highest amount of total information (13.40), followed respectively by the NOS (7.11) and ISTS (5.27) subtests. According to Baker and Kim (2017) and Klein (2013), the level of the test's information function depends on the number of items and the average discrimination value of the test items. Comparing these data with the number of items and the average discrimination values of each subtest, it is possible to notice that the results are in accordance with the proposed by the authors, since the subtest of the CS is the one with the largest number of items (23) and that the subtest of the NS, although it consists of the same number of items as the subtest of the ISTS (6), has items whose average discrimination is higher than the latter's (M = 1.20; SD = 1.08).

Furthermore, analyzing the amount of information in the typical range of θ for normal tests (Baker; Kim, 2017), from θ between -3 to 3, it is observed that the subtests of the NS, the ISTS, and the CS are able to provide more than 75% of the total information of the test (86.04%, 86.20%, and 75.28%, respectively). These results corroborate the previous evidence, showing a slight disparity on the part of the CS subtest, again highlighting the need for a review of the items that compose this subtest.

7 Final considerations

Aiming to gather validity evidence based on internal structure of a pilot instrument to assess the scientific literacy of students at the end of the 3rd cycle of basic education, which is under development, this study used the IRT technique for the analysis of the information functions of the items and subtests, the difficulty and discrimination parameters of the items, the proficiency of the respondents and the estimation of Kernel density.

In light of the foregoing, it was found that the three subtests present evidence of validity based on internal structure. However, it is evident that the items of the CS subtest need to be revised, since small discrepancies were found when compared to the results of the item analyses of the other two subtests. Nevertheless, based on the number of items and the average discrimination value of the items, the CS subtest had the highest amount of total information, a factor that reiterates the presentation of validity evidence based on internal structure.

Moreover, considering that ideally items whose difficulty index is around the average θ of the students and whose respective discrimination parameters are higher than items with difficulties far from the average θ of the students are desired, we recommend revising the items of the three subtests, so that they can, in the application of the final test, reach these values. This would increase the precision of the instrument, decrease its standard error of estimation, and, consequently, strengthen the presentation of validity evidence based on internal structure of the instrument.

The study presented the limitation of the number of student respondents (n = 176), which can directly interfere in the reported item parameters and, consequently, in the analyzed information functions and in the validity of the evidence presented. The results suggest that the final instrument ought to be applied to a wider and more diverse sample, including students from schools spread over most of the national territory, thus obtaining more robust and, preferably, generalizable results for the Portuguese population.

Funding

This work was funded by national funds through FCT – Foundation for Science and 800 Technology – IP, under the scope of the PhD Research Grant with reference UI/BD/151034/2021 and DOI https://doi.org/10.54499/UI/BD/151034/2021.

References

ALEXANDRE, N. M. C.; COLUCI, M. Z. O. Validade de conteúdo nos processos de construção e adaptação de instrumentos de medidas. *Ciência & Saúde Coletiva*, Rio de Janeiro, v. 16, n. 7, p. 3061–3068, 2011. DOI: https://doi.org/10.1590/S1413-81232011000800006. Available in: https://www.scielo.br/j/csc/a/5vBh8PmW5g4Nqxz3r999vrn/#. Access in: 20 oct. 2022.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. *Standards for educational and psychological testing*. Washington, DC: AERA, 2014.

ANDRADE, J. M.; LAROS, J. A.; GOUVEIA, V. V. O uso da teoria de resposta ao item em avaliações educacionais: diretrizes para pesquisadores. *Avaliação Psicológica*, Porto Alegre, v. 9, n. 3, p. 421–435, 2010.

BAKER, F. B.; KIM, S.-H. *The basics of item response theory*: statistics for social and behavioral sciences. Pittsburgh: Springer, 2017.

BRAUN, H. I. *Meeting the challenges to measurement in an era of accountability*. New York: Routledge, 2016.

COPPI, M. A.; FIALHO, I.; CID, M. Evidências de validade baseadas no conteúdo de um instrumento de avaliação da literacia científica. *Cadernos de Pesquisa*, São Luís, v. 29, n. 2, p. 99–127, 2022. DOI: https://doi.org/10.18764/2178-2229v29n2.2022.27. Available in: https://periodicoseletronicos.ufma.br/index.php/cadernosdepesquisa/article/view/17211. Access in: 18 oct. 2022.

COPPI, M.; FIALHO, I.; CID, M. Developing a scientific literacy assessment instrument for portuguese 3rd cycle students. *Education Sciences*, [*S. l.*], v. 13, n. 9, p. 941, 2023. DOI: https://doi.org/10.3390/educsci13090941. Available in: https://www.mdpi.com/2227-7102/13/9/941. Access in: 16 apr. 2023.

COUTO, G.; PRIMI, R. Teoria de resposta ao item (TRI): Conceitos elementares dos modelos para itens dicotômicos. *Boletim de Psicologia*, São Paulo, v. 61, n. 134, p. 1–15, 2011.

CUARTAS, J. Improving the measurement of children's mental health problems in Colombia with item response theory. *Revista Colombiana de Psicologia*, Bogotá, v. 29, n. 1, p. 87–103, 2020. DOI: https://doi.org/10.15446/.v29n1.77214. Available in: https://revistas.unal.edu.co/index.php/psicologia/article/view/77214. Access in: 07 mar. 2023.

DEBOER, G. E. Scientific literacy: another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching*, New York, v. 37, n. 6, p. 582–601, 2000. DOI: https://doi.org/10.1002/1098-2736(200008)37:6<582::AID-TEA5>3.0.CO;2-L. Available in: https://onlinelibrary.wiley.com/doi/10.1002/1098-2736(200008)37:6%3C582::AID-TEA5%3E3.0.CO;2-L. Access in: 07 mar. 2023.

DEPRESBITERIS, L.; TAVARES, M. R. *Diversificar é preciso...*: instrumentos e técnicas de avaliação de aprendizagem. São Paulo: Senac São Paulo, 2017.

EDELEN, M. O.; REEVE, B. B. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, Dordrecht, v. 16, supl. 1, p. 5–18, 2007. DOI: 10.1007/s11136-007-9198-0. Available in: https://link.springer.com/article/10.1007/s11136-007-9198-0. Access in: 20 oct. 2022.

FIVES, H.; HUEBNER, W.; BIRNBAUM, A. S.; NICOLICH, M. Developing a measure of scientific literacy for middle school students. *Science Education*, New York, v. 98, n. 4, p. 549–580, 2014. Doi: https://doi.org/10.1002/sce.21115. Available in: https://onlinelibrary.wiley.com/doi/abs/10.1002/sce.21115. Access in: 16 apr. 2023.

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

GHIGLIONE, R.; MATALON, B. *O inquérito*: teoria e prática. Oeiras: Celta Editora, 1992.

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, Bethesda, v. 11, n. 4, p. 364–377, 2012. DOI: https://doi.org/10.1187/cbe.12-03-0026. Available in: https://www.lifescied.org/doi/10.1187/cbe.12-03-0026. Access in: 16 apr. 2023.

HALADYNA, T. M.; RODRIGUEZ, M. C. *Developing and validating test items*. New York: Routledge, 2013.

HILL, M. M.; HILL, A. *Investigação por questionário*. Lisboa: Edições Sílabo, 2005.

KANE, M. T. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, Washington, v. 50, n. 1, p. 1–73, 2013. DOI: https://doi.org/10.1111/jedm.12000. Available in: https://onlinelibrary.wiley.com/doi/abs/10.1111/JEDM.12000. Access in: 09 nov. 2022.

KLEIN, R. Alguns aspectos da teoria de resposta ao item relativos à estimação das proficiências. *Ensaio*: aval. Pol. Públ. Educ., Rio de Janeiro, v. 21, n. 78, p. 35–56, 2013. DOI: https://doi.org/10.1590/S0104-40362013005000003. Available in: https://www.scielo.br/j/ensaio/a/FxGm4KDdQ56hF8JMbwxjzJJ/abstract/?lang=pt#. Access in: 09 nov. 2022.

LAUGKSCH, R. C. Scientific literacy: a conceptual overview. *Science Education*, New York, v. 84, n. 1, p. 71–94, 2000. DOI: https://doi.org/10.1002/(SICI)1098-237X(200001)84:1<71::AID-SCE6>3.0.CO;2-C. Available in: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1098-237X(200001)84:1%3C71::AID-SCE6%3E3.0.CO;2-C. Access in: 18 oct. 2022.

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*, Bristol, v. 5, n. 4, p. 331–359, 1996. DOI: 10.1088/0963-6625/5/4/003. Available in: https://journals.sagepub.com/doi/10.1088/0963-6625/5/4/003. Access in: 18 oct. 2022.

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

MENDONÇA FILHO, E. J. *Evidências de validade relacionadas à estrutura interna da escala cognitiva do inventário dimensional de avaliação do desenvolvimento infantil*. 2017. 197 f. Orientador: Denise Ruschel Bandeira. Dissertação (Mestrado em Psicologia) - Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil, 2017.

MILLER, J. D. Scientific literacy: a conceptual and empirical review. *Daedalus*, Cambridge, 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, 2009.

NUNES, M. F. O.; NORONHA, A. P. P. Escala de autoeficácia para atividades ocupacionais: estudo da estrutura interna e precisão. *Avaliação Psicológica*, Porto Alegre, v. 10, n. 1, p. 25–40, 2011.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. *PISA 2015 assessment and analytical framework*: science, reading, mathematic, financial literacy and collaborative problem solving, revised edition. Paris: OECD Publishing, 2017.

PASQUALI, L. Psicometria. *Revista da Escola de Enfermagem da USP*, São Paulo, v. 43, n. spe, p. 992–999, 2009a. DOI: https://doi.org/10.1590/S0080-62342009000500002. Available in: https://www.scielo.br/j/reeusp/a/Bbp7hnp8TNmBCWhc7vjbXgm/. Access in: 18 oct. 2022.

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

POPHAM, W. J. *Classroom assessment*: what teachers need to know. 8. ed. Los Angeles: Pearson, 2018.

PORTUGAL. Ministério da Educação. Direção-Geral da Educação. *Aprendizagens essenciais*: articulação com o perfil dos alunos: ciências naturais**:** 7º ano, 3º ciclo do ensino básico. Lisboa, 2018a. Available in: http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/3_ciclo/ciencias_naturais_3c_7a_ff.pdf. Access in: 16 apr. 2023.

PORTUGAL. Ministério da Educação. Direção-Geral da Educação. *Aprendizagens essenciais*: articulação com o perfil dos alunos: ciências naturais**:** 8º ano, 3º ciclo do ensino básico. Lisboa, 2018b. Available in:

http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/3_ciclo/ciencias_naturais_3c_8a_ff.pdf. Access in: 16 apr. 2023.

PORTUGAL. Ministério da Educação. Direção-Geral da Educação. *Aprendizagens essenciais*: articulação com o perfil dos alunos: ciências naturais: 9º ano, 3º ciclo do ensino básico. Lisboa, 2018c. Available in: http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/3_ciclo/ciencias_naturais_3c_9a_ff.pdf. Access in: 16 apr. 2023.

PORTUGAL. Ministério da Educação. Direção-Geral da Educação. *Aprendizagens essenciais*: articulação com o perfil dos alunos: físico-química: 7º ano, 3º ciclo do ensino básico. Lisboa, 2018d. Available in: http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/3_ciclo/fisico-quimica_3c_7a_ff.pdf. Access in: 16 apr. 2023.

PORTUGAL. Ministério da Educação. Direção-Geral da Educação. *Aprendizagens essenciais*: articulação com o perfil dos alunos: físico-química: 8º ano. 3º ciclo do ensino básico. Lisboa, 2018e. Available in: http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/3_ciclo/fisico-quimica_3c_8a_ff.pdf. Access in: 16 apr. 2023.

PORTUGAL. Ministério da Educação. Direção-Geral da Educação. *Aprendizagens essenciais*: articulação com o perfil dos alunos: físico-química: 9º ano, 3º ciclo do ensino básico. Lisboa, 2018f. Available in: http://www.dge.mec.pt/sites/default/files/Curriculo/Aprendizagens_Essenciais/3_ciclo/fisico-quimica_3c_9a.pdf. Access in: 16 apr. 2023.

REEVE, B. B.; FAYERS, P. Applying item response theory modelling for evaluating questionnaire item and scale properties. *In*: FAYERS, P. M.; HAYS, R. D. (ed.). *Assessing quality of life in clinical trials*: methods and practice. 2. ed. Oxford: Oxford University Press, 2005. p. 55–73.

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

WORLD ECONOMIC FORUM. *New vision for education*: unlocking the potential of technology. Geneva: World Economic Forum, 2015.