Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Validation of the use of Bloom's revised taxonomy as a tool for the design of assessment tests.

Bartolomé Pizà-Mir[1]*

[1] University College Alberta Giménez – Comillas Pontifical University ; bpiza@comillas.edu
* Correspondence: bpiza@comillas.edu

Abstract: (1) Background: This study aims to validate the use of Bloom's revised taxonomy as an instrument for the design of assessment tests; (2) Methods: A validation has been carried out by external judges, as well as by teachers and students, validating the instrument by means of Aiken's V; (3) Results: Judges, teachers and students consider Bloom's revised taxonomy as an effective tool for the design of assessment tests; (4) Conclusions: Using Bloom's revised taxonomy as a model for designing assessment tests promotes learning.

Keywords: Bloom; Assessment; Validation

1. Introduction

The quality of assessment evidence is a major issue in education. The effectiveness of an education system can be measured in large part by the quality and accuracy of the tests used to assess student progress and achievement (Popham, 2008). In addition, assessment tests are also central to the work of teachers, as they allow them to measure the success of their teaching methods and to adapt accordingly (Shuell, 1986).

However, despite their importance, the quality of assessment tests is often questionable. A recent study showed that a large proportion of the tests currently used lack reliability and validity (Haladyna, 2004). This can have serious consequences for both teachers and students. Teachers may get misleading results about their students' performance and consequently implement ineffective teaching strategies (Popham, 2008). For their part, students may be evaluated unfairly and receive undeserved grades (Shuell, 1986).

In this context, Bloom's taxonomy has become a widely accepted theoretical framework in the educational community for categorising and classifying the different levels of skills and knowledge to be acquired in the teaching-learning process (Bloom, 1956). Bloom's taxonomy provides valuable guidance for teachers in designing tests that adequately address all relevant levels of skills and knowledge (Krathwohl, 2002).

The inclusion of students with diverse learning needs and backgrounds is a legal requirement and an ethical imperative in education. Various laws and policies, such as the Individuals with Disabilities Education Act (IDEA) in the United States and the United Nations Convention on the Rights of Persons with Disabilities, mandate that schools provide equal opportunities and accommodations for all students, regardless of their abilities or challenges (United Nations General Assembly, 2006).

Contrary to the claims of some skeptics, the revised version of Bloom's Taxonomy is not a "neuromyth" and is instead a valuable tool for improving education. A neuromyth is a false belief or misconception about the brain and how it works, which is often perpetuated by inadequate or misleading information (De Smedt et al., 2013). Examples of

© 2022 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

neuromyths include the idea that people only use 10% of their brain, or that certain learning styles are more effective for certain individuals.

The revised version of Bloom's Taxonomy, on the other hand, is based on extensive research and analysis of cognitive development and learning processes. This framework, proposed by Anderson and Krathwohl in 2001, builds on the original taxonomy proposed by Benjamin Bloom in 1956, and incorporates advances in cognitive psychology and educational research.

Given the abstract component that different academic disciplines (mathematics and science in particular) can have, in order to achieve the correct literacy of individuals, it is necessary to minimise anxiety (Hopko et al., 2003) and rejection of this type of discipline (Pérez-Martín, 2018), although not only in science, but in all types of academic knowledge. For this reason, curriculum design is fundamental not only in terms of content, but also in the way they are approached, with practical proposals and guidelines to guide students in their learning and assessment (Acevedo, 2004), taking into account the cognitive development of students (Sacristán, 2007; Razzouk, 2008). In terms of scientific knowledge, there are two similar concepts: literacy (acquisition of knowledge) and competence (use, applicability and transfer of content) (Cañal et al., 2012). Active learning methodologies such as Context-Based Learning (ABC) (Avargil et al., 2012; Sanmartí and Márquez, 2017), guided constructivist methodologies such as the 5E methodology (García-Grau et al., 2021) among others have been shown to be useful for student literacy. Educational taxonomies are used to organise the learning objectives and outcomes (Bakırcı and Erdemir, 2010) that are expected to be developed through planned classroom experiences. One of the most commonly used taxonomies is Bloom's revised taxonomy (Anderson et al., 2001) which establishes a hierarchy of categories, where performance in higher categories requires performance in lower categories. Methodologies based on taxonomies have three clear benefits:(1) It classifies learning objectives and therefore allows comparisons between different curricula and/or subjects. (2) It allows students to know precisely what is expected of each of the contents to be worked on. (3) It allows teachers to design learning sequences from the setting of objectives to their assessment (Amer, 2006).

Bloom's revised taxonomy and universal design for learning (UDL) have a close relationship in terms of addressing diversity in the classroom and learning disabilities. According to Bloom (1956), the taxonomy is a system that classifies learning objectives into six categories: knowledge, comprehension, application, analysis, synthesis, and evaluation. For its part, the SAD (Rose & Meyer, 2002) is a pedagogical approach that seeks to ensure that learning content and activities are accessible to all students, regardless of their abilities or disabilities.

The application of Bloom's revised taxonomy in the design of assessment tests ensures that they are relevant and pertinent to the learning objectives, and that they cover all categories of the taxonomy. In addition, the use of the SAD in teaching-learning sequencing ensures that content is presented clearly and concisely, and that strategies and resources are used that facilitate access to information for all students.

A study by Taylor and Marzano (2017) on the impact of SAD on the academic performance of students with intellectual disabilities found that the use of this approach significantly improved their grades compared to students who did not use it. In addition, another study by Smith (2013) that compared the performance of students with hearing impairment in a traditional class and in a class with a DUA approach found that students in the DUA class earned significantly higher grades.

Numerous countries have used the Revised Bloom Taxonomy (RBT) to evaluate their curriculum documents such as Singapore and Australia (Ang, 2019), Turkey (Seraceddin

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

et al., 2019; Elmas et al., 2020), Indonesia (Poluakan et al., 2019) or Finland (Elmas et al., 2020) or have used it for the analysis of diagnostic tests such as PISA (Vázquez-Alonso et al., 2018; Rosales et al., 2020).

The RBT categories (Table 1) provide a simple tool for analysis, design and comparison, because it forms a two-dimensional matrix in which in each cell there are identifying verbs in which the different assessment criteria can be located and included, as well as an estimate of the cognitive demand (LOTS: acronym for low-level thinking skills; and HOTS: acronym for high-level thinking skills).

		Cognitive dimension					
		LOTS (Low Cognitive Demand)			HOTS (High Cognitive Demand)		
		Remember	Understand	Apply	Analyse	Evaluate	Create
Type of knowledge	Factual	Enumerate, list	Summarise	Reply, answer	Select	Check	Generate
	Conceptual	Recognise, identify, locate	Classify, Explain	Provide, demonstrate	Differentiate, contrast, relate	Determine	Organise
	Procedural	Evoke	Clarify, clarify	Carry out	Integrate	Judging, criticising	Design
	Metacognitive	Identifying strategies and patterns	Predicting, interpreting	Extrapolate, use	Deconstructing	Reflect	Produce, elaborate

It is important not only to validate instruments that ensure their reliability, but also that when involving other educational agents, and especially those who will implement them (teachers) or participate in them (students), there is cohesion and agreement for the development of learning design strategies.

Student and teacher involvement in the design of assessment tests is a major issue in education. Collaboration between students and teachers in the creation of tests can help to ensure that they adequately reflect the content and skills taught in the classroom (Popham, 2008).

Firstly, student participation in test design can help ensure that tests are relevant and adapted to their needs and abilities. Students can provide their perspective and suggestions for improving the quality and relevance of tests (Nitko, 2007). For example, they can suggest to what extent the questions and tasks included in tests reflect the content and skills they have learned in the classroom (Bachman, 1990).

Second, teacher involvement in test design can help ensure that tests are appropriate and consistent with the educational goals and objectives of the school or curriculum. Teachers can bring their knowledge and experience in teaching to create tests that adequately address relevant content and skills (Worthen, Borg, & White, 2018). In addition, they can use their knowledge of the skills and knowledge students are expected to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

acquire in the classroom to design tests that accurately and fairly assess student progress (Messick, 1989).

Third, student and teacher participation in test design can contribute to improving the quality and reliability of tests. Collaboration between students and teachers can make it possible to detect and correct errors and deficiencies in assessment tests before they are implemented (Shuell, 1986). In this way, they can ensure that tests meet the necessary quality standards and provide accurate and reliable results (Nitko, 2007).

In conclusion, student and teacher involvement in the design of assessment tests is a beneficial practice to ensure that they are relevant, consistent and of high quality. Collaboration between students and teachers can help to improve the accuracy and reliability of assessment tests.

Collaboration between students and teachers in test design can also foster student engagement and motivation in the learning process. By participating in the design of the tests that will assess them, students may feel more involved and engaged in their own learning process (Shepard, 2000). In addition, student participation in test design can help develop important skills such as critical and analytical thinking (Nitko, 2007) which is one of the fundamental categories of Bloom's taxonomy.

Fourth, student and teacher participation in test design can contribute to improved communication and collaboration in the classroom. Collaboration in test design can be an opportunity for students and teachers to discuss and work together on the content and skills to be assessed (Shepard, 2000). This can help improve the quality of teaching and learning in the classroom and foster a culture of collaboration and communication (Worthen, Borg, & White, 2018).

In summary, student,and teacher participation in the design of assessment tests is a beneficial practice that can contribute to improving the relevance, accuracy, quality and reliability of tests. In addition, it can foster student engagement and motivation, develop important skills, and improve communication and collaboration in the classroom.

Instrument validation is a crucial process in science, as it ensures that the results obtained using instruments are accurate and reliable. Instrument validation refers to the assessment of the reliability and validity of a measure or tool used in scientific research. Reliability refers to the consistency and accuracy of the results obtained by using the instrument, while validity refers to the instrument's ability to measure what it is supposed to measure (Haladyna, 2004).

.2. Materials and Methods

This research is classified as an instrumental study aimed at the design and study of the psychometric properties of measurement instruments (Ato, López-García, & Benavente, 2013). The aim is to design an instrument to validly and reliably analyse (Corral, 2009) the planning process of assessment tests based on Bloom's revised taxonomy.

The sample was chosen deliberately and intentionally, selecting a group of experts who meet the suitability criteria (Rodríguez, Gil, & García, 1996). The sample selected to validate the analysis instrument consisted of a group of 35 expert teachers (judges), 42 non-expert teachers and 91 secondary school students. Those selected to form the sample of expert judges had to meet 75% of the inclusion criteria established by the researchers:

1. Hold a bachelor's or Graduate degree in your subject
2. Doctorate or postgraduate degree in didactics specific to their subject, without considering the didactic training required by law to teach.
3. 10 years of university, secondary and/or primary school teaching experience.
4. Have publications related to the area of general or specific didactics in database journals with quality indexes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

To ensure construct validity, three experts participated in the development of the instrument. These experts met the following inclusion criteria mentioned above. Consensus agreement was used in the process of constructing the categories of analysis (Anguera & Hernández-Mendo, 2013). The items or categories of the measurement instrument should be selected in a way that is tailored to the object of measurement (Thomas, Nelson, & Silverman, 2015).

To assess content validity, the technique of a panel of expert judges was used (Cabero & Barroso, 2013). The experts were asked to give a quantitative assessment on a Likert scale of 1 to 3 points, and at the same time they wanted to compare it with the rest of the teachers, as well as with the students.

The importance of having study subgroups within a population of external judges lies in the need to obtain accurate and reliable ratings of the instrument in question. Aiken's V is commonly used to assess the validity of an instrument by analyzing the ratings made by a group of expert judges. However, when working with a very large population of judges, it can be difficult to obtain accurate and consistent ratings (Worthen, Borg, & White, 2018).

It is important to note that not all subjects and their assessment are understood in the same way, so the responses of teachers (judges and non-judges) were divided into three categories, those in the sciences (including experimental sciences, health sciences, engineering and architecture), social sciences (including graduates in the social sciences, law or humanities) and a generic category including those who were considered not to fit into the above categories.

The use of external judges in the validation of instruments is particularly important in education. Assessment tests are a fundamental tool in teaching, as they allow teachers to measure the progress and performance of their students (Shuell, 1986). It is therefore crucial to ensure that the tests used are reliable and valid. The use of external judges in the validation of these tests can help to ensure their quality and accuracy (Bloom, 1956; Popham, 2008).

The creation of study subgroups within the judge population allows judges to be grouped according to their specialization's and areas of expertise. In this way, more accurate and consistent ratings can be obtained, as each subgroup will be composed of judges who are experts in a certain area (Nitko, 2007).

In addition, the creation of study subgroups can also facilitate discussion and the exchange of ideas among judges, which can contribute to improving the quality of the assessments made (Messick, 1989). By working in smaller groups, judges can establish more effective dialogue and develop greater group cohesion and collaboration (Bachman, 1990).

Materials

Statistical analysis

Data were compiled in Microsoft Excel 2016 for descriptive analysis, Aiken's V coefficient and its confidence intervals (Aiken, 1985), were calculated. This coefficient makes it possible to quantify the content validity or relevance of the item with respect to the opinion of a group n of expert judges. The algebraic equation modified by Penfield & Giacobbi (2004) was used to calculate Aiken's V coefficient.

$$V = \frac{\bar{X} - l}{k}$$

Figure 1. V for Aiken

$\bar{x}$ is the mean of the judges' ratings in the sample, l is the lowest possible rating, and k is the range of possible values of the Likert scale used.

Aiken's V coefficient was calculated with Microsoft Excel software at the minimum level of validity according to expert standards; this standard can be at a liberal level

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

(Cicchetti, 1994) of Vo>0.50, or at a more conservative level, such as V>0.70 or higher (Charter, 2003).

Aiken's V coefficient value and confidence intervals at the 95% and 99% levels were obtained using the score method (Wilson, 1927; Penfield & Giacobbi, 2004) according to the formulae below (Figure 2 and Figure 3):

$$L = \frac{2nkV + z^2 - z\sqrt{4nkV(1-V) + z^2}}{2(nk + z^2)}$$

Figure 2. Calculation of the lower limit of the confidence interval

$$U = \frac{2nkV + z^2 + z\sqrt{4nkV(1-V) + z^2}}{2(nk + z^2)}$$

Figure 3. Calculation of the lower limit of the confidence interval

Where L is the lower limit of the interval, U is the upper limit of the interval, Z is the standard normal distribution value, V is the coefficient of Aiken's V and n is the number of judges.

Aiken's V value was calculated using the initial formula proposed by (Aiken, 1985), applying the central limit theorem for large samples (m>25). The number of judges, teachers and students was 35, 42 and 91 respectively (n), the number of items 11 (m), with a response range of 3 according to a likert(c) scale.

The different subjects who participated in the surveys and data collection, before answering the questionnaire, observed how a possible evaluation test or the structuring of the teaching and learning sequences using Bloom's revised taxonomy would look like.

Level 1: Suspense/Inelegible

- Key verbs: identify, mention, describe

- Question types: open questions, closed questions.

- Questions: Can you mention the basic concepts of the topic? Can you point out the main characteristics of... ? Can you identify the key elements of...? ?

- Tips for writing questions: avoid questions that are too broad or with multiple options, give enough clues for the student to answer correctly, avoid questions that can be answered with a simple yes or no, focus on specific aspects of the topic in question.

Level 1: Pass/Sufficient

- Key verbs: identify, mention, describe

- Question types: questions that require reflection and analysis, questions that involve different perspectives.

- Questions: Can you compare the differences between...? Can you interpret the meaning of...? Can you relate... to...? ?

- Tips for writing questions: avoid questions that require extensive prior knowledge, provide examples that help illustrate the concept in question, establish a clear frame of reference for the evaluation, provide enough information for the student to conduct a thorough analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

At level 1, questions should be more reflective and require the student to analyze, evaluate or compare different aspects of the topic. Some indicators to consider at this level could be:

1. *The student's ability to analyze different components or aspects of the topic.*
2. *The student's ability to evaluate different perspectives or approaches to the topic.*
3. *The student's ability to compare different theories or concepts related to the topic.*
4. *The student's ability to use evidence or data to support their arguments and conclusions.*
5. *The student's ability to identify and discuss problems or limitations related to the topic.*

Level 2: Merit

- Key verbs: analyze, apply, develop, design

- Question types: questions that require the application of concepts to concrete situations, questions that involve the creation of original solutions.

- Questions: Can you apply... to a specific case? Can you design an experiment that demonstrates...? Can you justify your opinion about...? ?

- Question writing tips: avoid questions with obvious answers, provide enough detail so that the student can apply their knowledge effectively, provide enough context and detail so that the student can develop a creative and practical response.

At level 2, questions should require the student to apply, develop or design solutions or approaches in relation to the topic. Some indicators to consider at this level might include:

1. *The student's ability to apply the concepts and theories of the topic to concrete situations.*
2. *The student's ability to develop original or creative solutions in relation to the topic.*
3. *The student's ability to design and plan an approach or project in relation to the topic.*
4. *The student's ability to use specific tools or techniques to address the topic.*
5. *The student's ability to integrate different approaches or perspectives in their approach to the topic.*

Level 3: Outstanding

- Key verbs: evaluate, argue, justify.

- Typology of questions: questions that require critical evaluation of a topic, questions that involve defending a point of view.

- Questions: Can you create a new model that explains...? Can you research and develop a solution for...? Can you argue convincingly about...? ?

- Tips for writing questions: provide a realistic and challenging context, allow the student to use their creativity and critical thinking skills, provide a solid base of evidence and arguments so that the student can develop a strong and coherent response.

At level 3, questions should require the student to evaluate, argue and justify their point of view in relation to the topic. Some indicators to consider at this level might be:

1. *The student's ability to critically evaluate different approaches or perspectives in relation to the topic.*
2. *The student's ability to argue logically and coherently in relation to the topic.*
3. *The student's ability to justify his or her point of view using evidence or data.*
4. *The student's ability to discuss or refute arguments against his or her point of view.*
5. *The student's ability to integrate different theories or approaches in their critical evaluation of the topic.*

The items that respondents (judges, teachers and students) were asked about are shown below.

1. *A test/examination should contain questions of varying degrees of difficulty.*
2. *If a test/exam is graded by levels, learners should know in advance the type of questions for each level and have examples.*
3. *In a test designed by levels of attainment, the knowledge of one level should be fundamental to answering those of the next level and ensuring that the levels are well distributed.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

4. Imagine that in a test ordered by levels of achievement, the student answers the level 2 questions correctly, and not the level 1 questions (the easiest and most elementary ones). Would it make sense, in your opinion, that the result of that test would be a pass?

5. In the case of answering level 3 and 2 questions correctly, but not the first level, the grade should be passed.

6. Would you agree that the tests/exams of your subjects should follow this model if they had a guide/infographic and a small rubric?

7. Would you agree that the qualifications should be:

a. Ineligible

b. Level 1: Sufficient

c. Level 2: Merit

d. Level 3: Outstanding

8. If a test/exam were ordered in 2 or 3 levels of difficulty (e.g. pass, B and A), it would make no sense (if the exam is well designed) for a student to know how to answer the difficult ones and not the easy ones.

9. The questions in a test/exam do not need to have an associated numerical mark.

10. In a test/examination there should be compulsory questions that must be answered in order to be considered passed.

11. The marking of a test/exam should be qualitative (e.g. no pass, sufficient, so great/merit, excellent/outstanding) depending on the difficulty of the questions answered correctly.

3. Results

This scientific article discusses the use of an instrument for the design of assessment tests based on Bloom's taxonomy in schools. According to the author, Bloom's taxonomy is a theoretical framework widely accepted in the educational community to categorise and classify the different levels of skills and knowledge to be acquired in the teaching-learning process (Bloom, 1956). The results of the study demonstrate that Bloom's revised taxonomy is a useful tool for the analysis and development of assessment tests.

Aiken's V was used to analyze the assessment tests designed with Bloom's revised taxonomy. The results showed high agreement between the taxonomy and the tests, indicating that the taxonomy is an effective tool to guide the design of tests that adequately assess the content and skills taught in the classroom (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956).

The figure below (Figure 4) shows that there are no significant differences in the number of teachers in the two main categories, representing an unbiased sample.

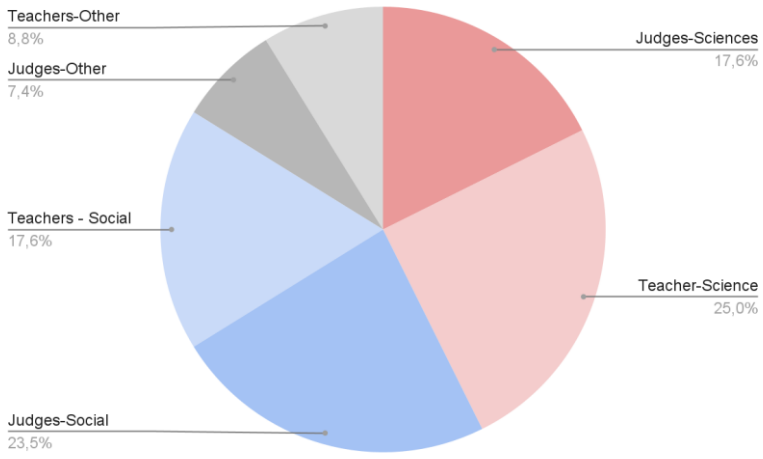

Figure 4. Distribution of teachers according to academic specialty

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

As can be seen in the following figure (Figure 5), although there are differences between the different expert judges and teachers according to their academic specialty, most of them meet the required Aiken V standards.

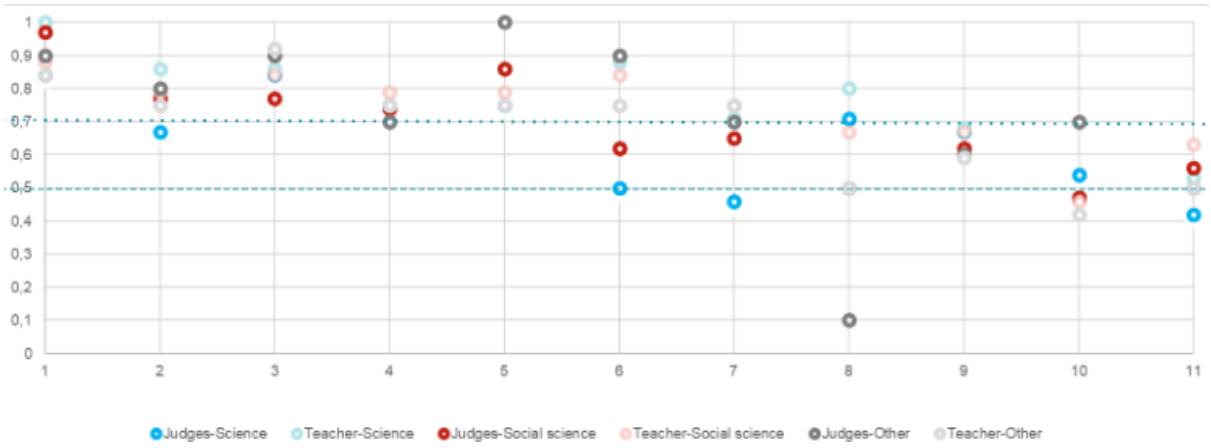

Figure 5. Representation of Aiken's V for each of the items in the teachers according to the different academic categories.

The following table (Table 2) shows the means and standard deviations of each of the items analyzed (according to the Likert scale), the corresponding Aiken's V value, as well as the values of the confidence intervals, in this case without differentiating between teachers (expert judges or not according to their academic specialty).

Table 2. Descriptive statistics and Aiken's V value

		Item	1	2	3	4	5	6	7	8	9	10	11
Judges		M±DE	2,77±0,6	2,49±0,7	2,63±0,69	2,43±0,74	2,63±0,55	2,23±0,91	2,11±0,93	2±0,91	2,29±0,79	2,09±0,85	1,97±0,92
		V	0,89	0,75	0,82	0,72	0,82	0,62	0,56	0,5	0,65	0,55	0,49
	95% CI	Inf	0,79	0,63	0,71	0,6	0,71	0,5	0,44	0,39	0,53	0,43	0,37
		Sup	0,94	0,83	0,89	0,81	0,89	0,72	0,66	0,61	0,75	0,65	0,6
	99% CI	Inf	0,75	0,6	0,67	0,56	0,67	0,46	0,41	0,35	0,49	0,4	0,34
		Sup	0,95	0,85	0,9	0,83	0,9	0,75	0,69	0,65	0,77	0,69	0,63
		Item	1	2	3	4	5	6	7	8	9	10	11
Teacher		M±DE	2,83±0,49	2,55±0,71	2,64±0,62	2,45±0,63	2,6±0,73	2,57±0,67	2,36±0,85	2,36±0,82	2,36±0,82	1,9±0,82	2,07±0,84
		V	0,92	0,78	0,82	0,73	0,8	0,79	0,68	0,68	0,68	0,45	0,54
	95% CI	Inf	0,83	0,67	0,72	0,61	0,69	0,68	0,57	0,57	0,57	0,34	0,42
		Sup	0,96	0,86	0,89	0,81	0,88	0,86	0,78	0,78	0,78	0,56	0,65

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

	99% CI	Inf	0,79	0,63	0,68	0,57	0,66	0,64	0,53	0,53	0,53	0,31	0,39
		Sup	0,97	0,88	0,91	0,84	0,89	0,88	0,8	0,8	0,8	0,6	0,68
		Item	1	2	3	4	5	6	7	8	9	10	11
Students		M±DE	2,6±0,63	2,63±0,63	2,4±0,7	2,69±0,63	2,86±0,44	1,99±0,77	1,96±0,89	1,9±0,87	1,46±0,69	1,57±0,85	2,3±0,82
		V	0,8	0,82	0,7	0,85	0,93	0,5	0,48	0,45	0,23	0,29	0,65
	95% CI	Inf	0,69	0,71	0,59	0,74	0,85	0,38	0,37	0,34	0,15	0,19	0,53
		Sup	0,88	0,89	0,79	0,91	0,97	0,61	0,59	0,56	0,34	0,4	0,75
	99% CI	Inf	0,66	0,67	0,55	0,71	0,81	0,35	0,34	0,31	0,13	0,17	0,5
		Sup	0,89	0,9	0,82	0,93	0,98	0,64	0,63	0,6	0,38	0,44	0,78

Considering the validity values for the Aiken V, all items are positively rated by teachers (judges or not) and students, with the exception of items 9 and 10.

It is worth noting that the first five items have an Aiken V greater than 0.7, which implies even greater reliability and consistency.

The following figure (Figure 6) shows the values of Aiken's V for the three populations studied, the expert judges, the teachers in general and the students with the liberal and acceptable reliability level, in yellow (>.05) and green (>0.7) respectively.

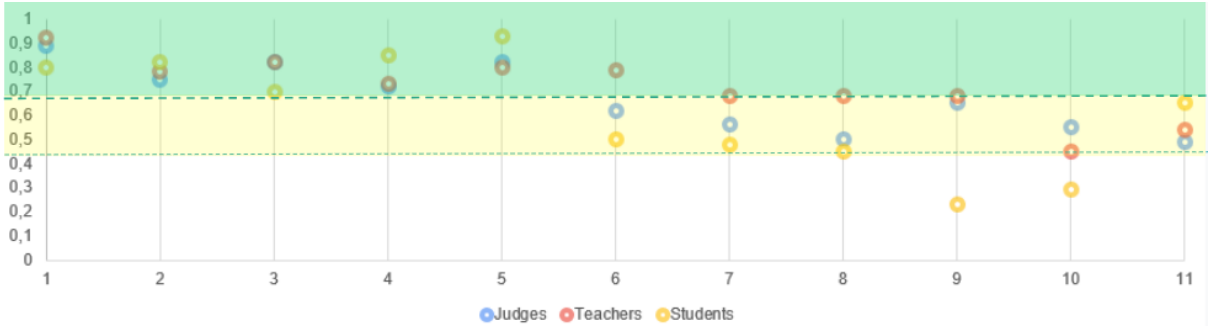

Figure 6. Representation of Aiken's V for each of the items in the expert, teacher and student judges.

Items 9 and 10 refer exclusively to the assessment of one question of an assessment test, and whether there should be compulsory questions to be answered, so item 10 agrees with items 4 and 5, both with an Aiken V value above 0.7 for all three populations, which shows the degree of agreement between the different parties involved in the teaching-learning process as well as in its assessment.

The instrument in question consists of a set of verbs that facilitate the creation of assessment tests (the chosen verb being the central core of the test questions that teachers ask) that cover all areas of Bloom's taxonomy in a balanced way. In this way, teachers can

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

design tests that adequately assess their students' progress and performance in all areas of knowledge (Krathwohl, 2002).

The author presents empirical evidence demonstrating the effectiveness of the instrument in terms of its ability to improve the quality and reliability of assessment tests. In a pilot study conducted (Pizà-Mir, 2021) in a school, the scores obtained by students on tests designed using the instrument were compared with the scores obtained on tests designed in the traditional way. The results showed a significant improvement in the quality and reliability of tests designed with the instrument (Haladyna, 2004). For example, a study by Wang and Hannafin (2005) found that the use of Bloom's Taxonomy in instruction improved students' achievement on standardized tests, particularly in the higher levels of the taxonomy, such as analysis and evaluation. Another study by Kalyuga, Ayres, Chandler, and Sweller (2003) found that the use of Bloom's Taxonomy in the design of exams was more effective for promoting learning than traditional exams that only tested lower levels of the taxonomy.

In addition, many teachers and students report that the use of Bloom's Taxonomy in instruction helps to increase their engagement and motivation. For example, a survey of teachers by Chen and Hsu (2008) found that the majority of respondents believed that the use of Bloom's Taxonomy improved students' interest and participation in the classroom. Similarly, a survey of students by Chen and Chang (2010) found that the majority of respondents believed that the use of Bloom's Taxonomy helped them to understand the material better and to apply their knowledge to real-world situations.

In addition, the instrument can help teachers develop more effective teaching, as it allows them to design tests that adequately address the different levels of Bloom's taxonomy and thus ensure that their students are acquiring all the necessary skills and knowledge (Popham, 2008; Shuell 1986) and the alignment of exam design with Bloom's Taxonomy is closely related to Universal Design for Learning (UDL), which is a framework for designing curriculum and instruction that is flexible and adaptable to the diverse needs of learners. UDL emphasizes the importance of providing multiple means of representation, expression, and engagement, in order to support learners with different abilities and interests (Universal Design for Learning Guidelines, 2018). By incorporating these principles into exam design, teachers can foster a more inclusive and equitable learning environment.

As mentioned above, the most basic and necessary contents to reach the higher stages of taxonomy and achieve true competence are not consolidated, which is in line with the thesis of Thamraksa (2005), Willingham (2009) or Brown, Roediger and McDaniel (2014) and which Ausubel (1978) already pointed out in the seventies.

The results of curriculum analyses from different countries show that there is an expectation of reaching higher stages of Bloom's revised taxonomy, but the need for a factual and conceptual learning base is evident, as already stated by authors such as Willingham (2009) or Ausubel (1978).

The assessment with 4 levels of achievement, using Bloom's revised taxonomy, is a useful tool for addressing diversity in the classroom as it allows teachers to set clear expectations and objectively assess student performance based on different levels of achievement. This allows teachers to tailor their teaching to the individual needs and abilities of each student, providing a fair and equitable assessment (Marzano, Pickering, & Pollock, 2001). In addition, also facilitates communication between teachers and students about assessment criteria, allowing students to better understand expectations and work toward achieving concrete goals (Brookhart, 2010).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

Overall, the use of Bloom's Taxonomy in exam design is a powerful tool for promoting learning and diversity inclusion. As stated by Bloom (1956), *"the purpose of education is to provide opportunities for students to develop to the fullest their potential for intellectual, emotional, physical, and social growth"* (p. 22). By considering the different levels of learning objectives, teachers can create exams that are challenging and engaging for all students and support the development of their knowledge and skills. This approach is also closely related to UDL and its potential for enhancing the educational experience of learners.

In the light of the results of the analyses of the curricular texts as well as the design of tests according to Bloom's revised taxonomy, the theses of Agarwal et al. (2012), McDaniel et al. (2013), McDermott et al. (2014) or Roediger et al. (2011) on how the design of tests and assessment instruments favour the learning of higher processes and promote the learning of the knowledge they are intended to assess are also confirmed (Agarwal, 2019; Butler et al., 2017; Jensen et al., 2014).

4. Conclusions

In summary, Bloom's revised taxonomy is a good tool for the design of assessment tests, as well as for the design of teaching-learning sequences and their subsequent assessment, whether by means of a written test or otherwise, as can be seen in the studies described in the previous sections.

Therefore, the use of the established levels, according to the category of verbs of the RBT presented, is validated as an instrument for the design of assessment tests, since they conform different psychological processes linked to the learning of the different subjects, and at the same time show the consensus between teachers (experts or not) and students.

There is a strong consensus among teachers and students that the revised version of Bloom's Taxonomy is effective for promoting learning. This consensus is supported by a wealth of experimental and empirical evidence, as well as the positive experiences and perceptions of teachers and students who have used this framework in their instruction.

In conclusion, the scientific article presents solid evidence suggesting that the use of an instrument for the design of assessment tests based on Bloom's taxonomy is a useful and effective tool in the context of teaching. Its implementation in schools is recommended to improve the quality and reliability of assessment tests and to promote more effective teaching

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Acevedo, J. A. (2004). "Reflexiones sobre las finalidades de la enseñanza de las ciencias: Science education for citizenship". Eureka Journal on Science Education and Outreach, 1(1), 3-16.
2. Agarwal, P. K., Bain, P. M. and Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. Educational Psychology Review, 24, 437- 448. http://dx.doi.org/10.1007/s10648-012-9210-2
3. Agarwal, P. K. (2019). Retrieval Practice & Bloom's Taxonomy: Do Students Need Fact Knowledge Before Higher Order Learning?. Journal of Educational Psychology, 111 (2), 189-209. http://dx.doi.org/10.1037/edu0000282
4. Aiken, L. R. (1985). Three coeffients for analyzing the reability and validity of ratings. Educational and Psychological Measurement, 45(1), 131-143. doi:10.1177/0013164485451012
5. Amer, A. (2006). Reflections on bloom's revised taxonomy. Electronic Journal of Research in Educational Psychology, 4(1): 213-230.
6. Anderson, L. W. and Krathwohl, D. R. (2001). A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy. New York. Longman Publishing.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

7. Ang, G. X. (2019). Comparing science learning outcomes from Singapore and New South Wales, Australia, using Revised Bloom's Taxonomy. National Institute of Education, Nanyang Technological University.
8. Anguera, M. T., & Hernández-Mendo, A. (2013). Observational methodology in sport sciences. E-Handball Com, 9(3), 135-160.
9. Arı, A. and Gökler, Z.S. (2012). Evaluation of elementary science and technology course gains and SBS questions according to the new Bloom taxonomy. Nigde: X. National Science and Mathematics Education Congress.
10. Ato, M., López-García, J. J., & Benavente, A. (2013). A classification system for research designs in psychology. Anales de Psicología, 29(3). doi:10.6018/analesps.29.3.178511
11. Avargil, S., Herscovitz, O. and Dori, Y. J. (2012). Teaching thinking skills in context-based learning: Teachers' challenges and assessment knowledge. Journal of Science Education and Technology, 21, 207-225.
12. Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford, UK: Oxford University Press.
13. Bakırcı, H. and Erdemir, N. (2010). Physics teacher candidates' mechanical issues according to Bloom's taxonomy levels. Cukurova University Journal of the Faculty of Education, 38(3): 81-91.
14. -Brookhart, S. (2010). How to create and use rubrics for formative assessment and grading. ASCD.
15. Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain. New York, NY: David McKay Company.
16. Brown, P. C., Roediger, H. L. and McDaniel, M. A. (2014). Make it stick: The science of successful learning. Cambridge: Harvard University Press. http://dx.doi.org/10.4159/9780674419377
17. Butler, A. C., Black-Maier, A. C., Raley, N. D. and Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. Journal of Experimental Psychology: Applied, 23, 433-446. http://dx.doi.org/10.1037/xap0000142
18. Cabero, J., & Barroso, J. (2013). The use of expert judgement for ICT evaluation: the expert competence coefficient. Bordón, 65(2), 25-38. doi:10.13042/brp.2013.65202
19. Cañal, P. (2012). How to assess scientific competence? Investigación en la Escuela, 78, 5-17.
20. Cañas, A. M. and Nieda, J. (2013) A way of working on scientific competence in the classroom. Revista Digital de Educación y Formación del Profesorado, 10, 35-47.
21. Cicchetti, D. V. (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assesments, 6, 284-290.
22. Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. Journal of General Psychology, 130(3), 290-304.
23. Chen, G.-D., & Chang, H.-C. (2010). An empirical study of the effects of using Bloom's taxonomy in college English instruction. Educational Technology & Society, 13(3), 63-73.
24. Chen, G.-D., & Hsu, Y.-S. (2008). A survey of the effects of using Bloom's taxonomy in college English instruction. Educational Technology & Society, 11(4), 50-61
25. Corral, Y. (2009). Validity and reliability of research instruments for data collection. Journal of Educational Sciences, 19(33), 228-247.
26. De Smedt, B., Kirschner, P. A., & Hulshof, C. D. (2013). Neuromyths in education: Prevalence and predictors of misconceptions among teachers. Frontiers in Psychology, 4, 234. doi: 10.3389/fpsyg.2013.00234
27. Elmas, R., Rusek, M., Lindell, A.,Nieminen, P., Kasapoğlu, K., & Bílek, M. (2020). The intellectual demands of the intended chemistry curriculum in Czechia, Finland, and Turkey: A comparative analysis based on the revised Bloom's taxonomy. Chemistry Education Research and Practice, 21(3), 839-851. https://doi.org/10.1039/d0rp00058b
28. García-Grau, F., Valls, C. and Ruiz-Martín, H. (2021). The long-term effects of introducing the 5E model of instruction on students' conceptual learning, International Journal of Science Education. https://doi.org/10.1080/09500693.2021.1918354
29. Haladyna, T. M. (2004). Developing and validating multiple-choice test items. Mahwah, NJ: Lawrence Erlbaum Associates.
30. Hopko, D.R., Mahadevan, R., Bare, R.L., and Hunt, M. (2003). The Abbreviated Math Anxiety Scale (AMAS). Construction, validity and reliability. Assessment, 10, 178-182
31. Jensen, J. L., McDaniel, M. A., Woodard, S. M. & Kummer, T. A. (2014). Teaching to the test... or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. Educational Psychology Review, 26, 307-329. http://dx.doi.org/10.1007/s10648-013-9248-9
32. Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. Educational Psychologist, 38(1), 23-31
33. Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview, Theory into Practice, 41(4), 212-264.
34. Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). Classroom instruction that works: Research-based strategies for increasing student achievement. ASCD.
35. McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B. & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. Applied Cognitive Psychology, 27, 360-372. http://dx.doi.org/10.1002/acp.2914
36. McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. Journal of Experimental Psychology: Applied, 20, 3-21. http://dx.doi.org/10.1037/xap0000004
37. Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York, NY: Macmillan.
38. Nitko, A. J. (2007). Educational assessment of students (4th ed.). Upper Saddle River, NJ: Pearson.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 December 2022 doi:10.20944/preprints202212.0123.v2

39. Penfield, R. D., & Giacobbi, P. J. (2004). Applying a Score Confidence Interval to Aiken's Item Content-Relevance Index. Measurement in Physical Education and Exercise Science, 8(4), 213-225. doi:10.1207/s15327841mpee0804_3
40. Pérez-Martín, J. M. (2018). A journey through scientific literacy in Spain. Revista de Didácticas Específicas, 18, 144-166.
41. Poluakan, C., Tilaar, A. F., Tuerah, P. and Mondolang, A. (2019) "Implementation of the revised bloom taxonomy in assessment of physics learning," in Proceedings of the 1st International Conference on Education, Science and Technology (ICES-Tech), Gorontalo, Indonesia.
42. Popham, W. J. (2008). Classroom assessment: What teachers need to know (6th ed.). Boston, MA: Pearson.
43. Razzouk, N.Y., (2008). Analysis in teaching with cases: A revised to bloom's taxonomy of learning objectives. College Teaching Methods & Styles Journal, 4(81), 49-56.
44. Richardson, F. C., & Suinn, R. M. (1972). The Mathematics Anxiety Rating Scale: Psychometric data. Journal of Counseling Psychology, 19(6), 551-554. https://doi.org/10.1037/h0033456
45. Rodríguez, G., Gil, J., & García, E. (1996). Qualitative research methods. Aljibe.
46. Roediger, H. L., Agarwal, P. K., McDaniel, M. A. & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. Journal of Experimental Psychology: Applied, 17, 382-395. http://dx.doi.org/10.1037/a0026252
47. Rosales Sánchez, E. M., Rodríguez Ortega, P. G., Romero Ariza, M. (2020). Knowledge, cognitive demand and contexts in the assessment of scientific literacy in PISA. Eureka Journal of Science Education and Outreach 17(2), 2302. http://dx.doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2020.v17.i2.2302
48. Rose, D. H., & Meyer, A. (2002). Teaching every student in the digital age: Universal design for learning. Alexandria, VA: Association for Supervision and Curriculum Development.
49. Sacristán, J.G. (2007). El currículum: una reflexión sobre la práctica. 9th edition. Madrid: Ediciones Morata.
50. Sanmartí, N. and Márquez, C. (2017). Project-based science learning: from context to action. Apex. Journal of Science Education, 1(1), 3-16.
51. Smith, J. (2013). Comparing academic achievement of deaf students in traditional and universal design for learning classrooms. American Annals of the Deaf, 158(1), 5-12.
52. Seraceddin L. Z. and Kızılaslan, A. (2019). Analysis of 10th Chemistry Curriculum According to Revised Bloom Taxonomy. Journal of Education and e-Learning Research, 6(2), 88-95. http://dx.doi.org/10.20448/journal.509.2019.62.88.95
53. Shepard, L. A. (2000). The role of assessment in a learning culture. Educational Researcher, 29(7), 4-14.
54. Shuell, T. J. (1986). Cognitive conceptions of learning. Review of Educational Research, 56(4), 411-436.
55. Taylor, J., & Marzano, R. (2017). The impact of universal design for learning on academic performance of students with intellectual disabilities. Journal of Special Education, 51(1), 17-25.
56. Tanık, N. and Saraçoğlu, S. (2011). Examination of the written questions of science and technology lessons according to the renewed bloom taxonomy. Journal of Science, 4(4), 235-246.
57. Thamraksa, C. (2005). Metacognition: a key to success for EFL learners. Bangkok University Academic Review, 4 (1), 95-99.
58. Thomas, J. R., Nelson, J. K., & Silverman, S. J. (2015). Research methods in physical activity. Human kinetics.
59. United Nations General Assembly (2006). Convention on the Rights of Persons with Disabilities. Retrieved from https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html.
60. Universal Design for Learning (UDL) Guidelines (version 2.0). (2018). National Center on Universal Design for Learning. Retrieved from http://udlguidelines.cast.org/.
61. Vázquez-Alonso, A. and Manassero M. A. (2018). Epistemic knowledge in the assessment of scientific competence in PISA 2015. Revista de Educación, 380, 103-128.
62. Wang, F., & Hannafin, M. J. (2005). The effects of using Bloom's taxonomy on student achievement in web-based instruction. Journal of Educational Computing Research, 32(3), 279-297.
63. Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association, 22, 209-212.
64. Worthen, B. R., Borg, W. R., & White, J. B. (2018). Educational assessment (10th ed.). Boston, MA: Pearson.