Developing and validating an integrated three-tier multiple choice test for smartphone-based experiments in physics

Frank Angelo A. Pacala

University of San Carlos, Cebu City, Philippines

Abstract

The increasing use of smartphone-based physics experiments has become a theme of many physics education-related research and publications, including its actual implementation inside the classroom. Measuring this scientific process and teachers' and students' corresponding conceptual understanding is needed. This paper aims to create a well-designed, reliable, validated integrated instrument to measure the intervention's impact on the student's conceptual understanding and science process skills (SPS) of various physics topics. The reliability measure was internal consistency using Cronbach's alpha, and the validity measures were content and construct validity. The instrument was found reliable with an alpha point estimate value of 0.766, which is considered to be acceptably good. The content validity index revealed some items as inappropriate, so they were removed, and 30 items were appropriate. The exploratory factor analysis (EFA) revealed ten components that were measured in the instrument. These components were classified by their factor loads using orthogonal rotation in varimax. The manual analysis of these factor loads revealed that the EFA was measuring the teachers' cognition level based on the question's science process skills level. The most dominant scores were from remembering and lowly from evaluating level. This revealed the power of the instrument to integrate the SPS and conceptual understanding constructs into one instrument. The instrument is now validated; therefore, this paper suggests using this instrument to measure the level of integrated physics SPS and conceptual understanding during a smartphone-based physics experiment.

Keywords

physics education, smartphone physics experiment, three-tier multiple tests, instrument development, instrument validation, exploratory factor analysis

1. Introduction

Technology integration in education has revolutionized learning, particularly in physics. Smartphones have emerged as powerful tools for experiments, but the need for standardized assessment tools poses a challenge. To bridge this gap, this research was conducted to develop and validate robust evaluation instruments that adapt to the evolving educational technology landscape, enhancing the quality of physics education.

The traditional method of teaching physics involves conducting experiments in a lab using specialized equipment, which may not be easily accessible to all students. This results in unequal exposure to practical concepts, especially in resource-limited environments. Smartphone-based experiments offer a potential solution to this issue. However, the need for standardized assessment tools is a significant obstacle. Several studies (e.g., [1]; [2]) were conducted to measure the effect of smartphone-based experiments in Physics but could not produce standardized tests to measure this effect. The current evaluation methods may not fully capture smartphone-based experiments' unique features and outcomes, underscoring the need for tailored assessment instruments.

Moreover, the studies conducted by Cai [3] and Hochberg [4] explored the effect of various elements of smartphone experiments, like augmented reality, on the cognitive loads, self-efficacy, and conceptions of students' learning. They used validated research instruments to measure the different constructs of

DigiTransfEd 2024: 3rd Workshop on Digital Transformation of Education, co-located with the 19th International Conference on ICT in Education, Research, and Industrial Applications (ICTERI 2024, September 23-27, 2024, Lviv, Ukraine

frankpacala@gmail.com (F. A. A. Pacala)

https://scholar.google.com/citations?user=bvFlCogAAAAJ&hl=en&oi=ao (F. A. A. Pacala)

0000-0001-5774-0008 (F. A. A. Pacala)

students' learning. Also, Nikou and Economides [5] measured the impact of mobile-based experiments on the topic of electric currents. They successfully concluded that the instrument has an internal consistency of 0.85 on average. These sources used one-tier tests to measure the effect of smartphone experiments on student learning. This present research is focused on developing and validating a three-tiered multiple-choice test in the mechanics section of the physics curriculum.

Most of the test instruments in the literature measured a single construct like conceptual understanding, critical thinking, and science process skills (SPS). For instance, Tiruneh et al. [6] measured the critical thinking of students in electricity and magnetism, and He et al. [7] devised a test to diagnose students' conception of aqueous solutions in chemistry. This research created an integrated three-tier multiple-choice test (ITTMCT) measuring the students' conceptual understanding and SPS simultaneously. The approach discussed here aims to assess student's knowledge of physics principles theoretically and through their ability to design experiments, analyze data, and draw conclusions based on empirical evidence. This method combines conceptual understanding with science process skills, which allows students to gain a deeper understanding of the scientific method and its practical applications. Ultimately, this prepares them to face complex challenges in physics and related areas with greater confidence and competence.

Furthermore, multiple choice can be created in various ways. Single-tier and two-tier multiple-choice tests differ in assessing knowledge and understanding. In a single-tier test, each question is a standalone query with a fixed set of options, evaluating typically factual recall, conceptual understanding, or problem-solving skills. On the other hand, two-tier multiple-choice tests incorporate follow-up questions into each primary query, introducing an additional layer of complexity. The first tier resembles a traditional multiple-choice question, requiring selecting the correct answer from the provided options. However, the second tier prompts participants to justify their initial choice or reasoning, requiring them to apply critical thinking skills and provide a rationale for their selection. He et al. [7] argued that two-tier tests offer a more comprehensive assessment of understanding, probing factual knowledge and the thought processes and reasoning behind participants' responses. Hanson [8] noted that although they may take more time to administer and grade, two-tier tests offer deeper insights into participants' comprehension and analytical abilities. They are, therefore, valuable tools for assessing higher-order thinking skills and conceptual understanding.

A three-tier multiple-choice test has a unique feature called the certainty of response, which adds an extra layer of diagnostic assessment. In the first tier, participants answer a question, followed by justifying the second tier. The third tier prompts them to assess their confidence or certainty in their response. Laeli [9] emphasized that this feature provides valuable insight into their responses' reliability and accuracy, metacognitive awareness, and self-assessment skills. The third tier helps measure participants' knowledge and understanding while comprehensively evaluating their confidence levels [10]. By using this three-tier test in the diagnostic test, educators and evaluators gain a deeper understanding of the degree of certainty associated with participants' answers, ultimately contributing to informed decision-making in educational and diagnostic settings.

Therefore, developing and validating a three-tier multiple-choice test to measure the effect of smartphone-based experiments in mechanics is necessary. This research aims to create a well-designed, reliable, and validated instrument to measure the intervention's impact on the student's conceptual understanding of various physics topics. The research questions are as follows: What is the instrument's point estimate of Cronbach's alpha? How many factors this instrument can measure? What is the level of content validity of the instrument?

2. Methodology

2.1. Defining the constructs and formulating table of specifications

The first phase in developing the ITTMCT is to define and determine the constructs under which items will fall and formulate the table of specifications. For instance, if question 1 is under the construct of measuring or observing SPS. This paper used Vitti and Torres's [11] practical science process skills.

Table 1
The table of specifications of the instrument.

	Remembering	Understanding	Applying	Analyzing	Evaluating	Total items
Item No.	6,7,25,26,32, 5,24, 35	12,27,33,37, 2,13,21	22,23,24, 14,17,31	17,20, 1,16	4,9,19, 15, 29	30
SPS	Observing 6,7,25,26,32	Predicting (12,27,33,37)	Experimenting (22,23,24)	Communicating (17,20)	Controlling variable (4,9,19)	30
	Sorting and Classifying (5,24, 35)	Making inference (2,13,21)	Measuring (14,17,31)	Graphing (1,16)	Making inference (15,29)	

Their handbook provided several activities that can measure SPS, but this paper utilized their description of the process skills. These skills are observing, measuring, sorting and classifying, making inferences, predicting, experimenting, graphing, communicating, and controlling variables. All of these process skills were utilized.

Moreover, the test items followed Anderson and Krathwohl's [12] revised Bloom's taxonomy of learning cognition, which differentiates conceptual knowledge. In this taxonomy, the cognitive process categorizes cognitive skills into six levels that increase in complexity from Remembering to Creating. Unlike the original taxonomy, which only focused on the cognitive domain, this revision emphasizes the role of metacognition in the learning process. In this research, the cognition level started from remembering to evaluation.

The instrument followed the Philippines' Department of Education (DepEd) curriculum guide for the covered topics. However, the experiment is almost similar to CAIE practical exams, and it was made sure that the instrument would follow the Science Curriculum Guide of DepEd. This would ensure horizontalization of the skills and knowledge to be measured, plus the instrument is made to be localized.

Napal et al. [13] said science process skills could be seen as a progression or hierarchy, but they maintain that these skills are interconnected. Hence, both the science process skills and the taxonomy of cognition can be merged. Both of these constructs were combined to create the ITTMCT. The construct level in SPS is matched with the taxonomy of learning cognition. For instance, observing is matched with the remembering skill.

Each of the levels of cognition is given a total of five items. Observing, making inferences (understanding), experimenting, graphing, and controlling variables have three items each. The SPS of sorting and classifying, predicting, measuring, communicating, and making inferences (evaluating) is given two items each. There are 30 items in these instruments after the revision has been made.

2.2. Instrument development

The items were developed based on the principles of assessment design of Trends in International Mathematics and Science Study (TIMSS) and Cambridge Assessment International Education (CAIE). TIMSS consists of two main categories of questions. First is the multiple-choice format, where students are presented with response options, and second is the constructed response format, where students are expected to come up with their answers [14]. The CAIE also consists of multiple-choice tests, usually paper 1, and structured questions, generally papers 2 and 4.

The physics topics covered in this test are measuring acceleration due to gravity, magnetic fields, Doppler effect, and momentum and collision. These topics are commonly taken by students aged 16-18 under the Advanced Subsidiary (AS) and A-Level programs. These topics mostly come with practical activities under the Cambridge 9702 Curriculum. Although the topics do not represent the whole 9702 syllabus, these experiments are vital because they are core experiments.

These question structures benefit this three-tier test. The first part is the common multiple-choice

Table 2
Sample questions from the ITTMCT.

Tier 1	Tier 2	Tier 3
1. Explain the significance of waiting for 60 seconds after placing the phone near the solenoid in the experiment. What does this duration aim to achieve? A. To allow the conducting wire to cool down B. To synchronize with the voltage generator C. To stabilize the readings on the magnetometer D. To measure the length of the solenoid	Write a reason for your answer.	[ ] Sure [ ] Unsure [ ] I guessed my answer
2. When recording the time taken for the experiment using the Phyphox mobile application, what would be the most appropriate unit of measurement? A. Volts B. Seconds C. Tesla D. Meters	Describe the other units you did not choose.	[ ] Sure [ ] Unsure [ ] I guessed my answer
3. Explain how you measured the fractional uncertainty in magnetic field strength (B). A. dividing 0.2 seconds to the value of B B. multiplying 0.2 seconds by the value of B C. dividing the value of B to 0.2 seconds D. multiplying 0.02 seconds by the value of B	State the reasoning for your answer.	[ ] Sure [ ] Unsure [ ] I guessed my answer

question with a stem-option structure. The second tier is the structured/response question, which is a segment of papers 2 and 4 of CAIE and the constructed response of TIMSS. The addition is the certainty level of student response. The number of items in multiple choices and constructed responses is almost similar [15]. This paper followed this. The guiding principles of CAIE and TIMSS were followed as much as possible during this item development.

The third tier is the certainty level of the student's response; each question contains this level. This third tier asked the students to reveal the extent of their certainty in their answers. There are two options: sure, unsure, and guessed the answer.

Initially, five questions were constructed. Three colleagues with at least four years of teaching experience in the CAIE curriculum and experience dealing with TIMSS checked these five questions for consistency with the assessment design of the said two frameworks. Then, their comments and suggestions were incorporated, and the question-making process continued until the questions reached 37. In table 5, the questions were only 30 because this is after the disqualification of some items due to lower internal consistency.

2.3. Scoring guide

Each item is given a full five marks. Table 3 shows the full-scoring guide, which is borrowed from Pacala [16]. The provided rubric describes a methodical way to assess students' conceptual knowledge at different levels. Students who score at the lowest level, "No Understanding", give wholly inaccurate answers, demonstrating a lack of background knowledge or a conceptual misunderstanding. As they advance to "Alternative Conception", pupils could show signs of incomplete comprehension and false beliefs or explanations. "Partial Understanding with Alternative Conception" denotes an answer that contains both true and false information, frequently together with ambiguities or different interpretations. On the other hand, "Partial Understanding without Alternative Conception" denotes an answer that is partially accurate but lacks precision or clarity. When students reach the highest level, "Complete Understanding", they can demonstrate a thorough understanding of the idea and confidence in their

Table 3
The scoring guide for the integrated physics assessment borrowed from Pacala [16].

Level of Conceptual Understanding	Explanations	Score
Complete Understanding	The first tier is correct, the second tier is correct, and the third tier is I am sure of my answer.	5
Partial Understanding without Alternative Conception	The first tier is correct, the Second tier is correct, and the third tier is unsure of my answer.	4
Partial Understanding with Alternative Conception	The first tier is correct, the second tier is incorrect, and the third tier is I am sure of my answer.	3
Alternative Conception	The first tier is incorrect, the second tier is correct, and the third tier is I am sure of my answer. The first tier is incorrect, the second tier is correct, and the third, I am unsure of my answer. First tier is correct, second tier is incorrect, and third is I am not sure of my answer. The first tier is incorrect, the Second tier is incorrect, and the third tier is I am sure of my answer. The first tier is incorrect, the second tier is incorrect, and the third tier is I am not sure of my answer.	2
No Understanding	The first tier is correct, the second tier is correct, and the third tier is I completely guessed my answer. The first tier is correct, the second tier is incorrect, and the third tier is I completely guessed my answer. The first tier is incorrect, the second tier is correct, and third tier is I completely guessed my answer. The first tier is incorrect, the second tier is incorrect, and the third tier is I completely guessed my answer.	1

answers by providing entirely accurate solutions.

2.4. Pilot testing

The test instrument was piloted to a small circle of science teachers in the Division of Catbalogan City. They were contacted via Facebook Messenger and email to ask if they would participate (N = 8) in the pilot testing. Once they approved it, the researcher sent the test instrument using a Google form to their accounts in Messenger or email. At the end of this form is a statement about their comments/suggestions towards the instrument.

This pilot testing aims to evaluate the instrument's validity and reliability before the complete administration of the test. This pilot testing phase allows researchers to determine if the test items effectively measure the intended constructs or skills. Any issues with the test items can be identified and corrected at this stage to ensure that the final test is reliable and valid for assessing the desired outcomes.

Based on this pilot testing, it is revealed that teachers left some items in tier two blank because they did not have an introductory statement that would guide them to the reason. This was mended by providing a command structure to the statements in tier 2. During this pilot testing, the Cronbach's Alpha was 0.62.

2.5. Instrument revision and final administration

The comments, suggestions, and insights from the pilot testing were incorporated into the updated instrument. The participants noted that the value of the constant (e.g., Planck's constant) should be present in the question's stem. They commended the standard length of the question and the separation

Table 4
Scoring guide for the expert.

Relevance	Clarity	Simplicity	Ambiguity
1 = not relevant	1 = not clear	1 = not simple	1 = doubtful
2 = item needed some revision	2 = item needed some revision	2 = item needed some revision	2 = item needed some revision
3 = relevant but need minor revision	3 = clear but need minor revision	3 = simple but need minor revision	3 = no doubt but need minor revision
4 = very relevant	4 = very clear	4 = very simple	4 = meaning is clear

of each paragraph to enhance readability.

The revised instrument was finally administered to the science teachers (N = 30) in one of the schools in the Catbalogan City Division. These teachers taught DepEd's science curriculum and had at least three years of teaching experience. The teachers were not participants in the pilot testing. During the day of testing, the tables and chairs were arranged following the CAIE standards as stated in the What to Say to Candidates document. For instance, they were told to fill out the instrument with their names, school names, testing date, and candidate numbers. The time given was one hour; however, many teachers finished 15-20 minutes before the hour. Once they were done, they were told that they could leave the testing room.

The teachers were sent a letter if they agreed to join this activity. Once they signed up, they were included in the pool of participants. The teachers were told that the data collected was for research purposes only. Their names, school names, and scores were not published on the Internet or social media to preserve their anonymity and confidentiality.

2.6. Data analysis

The ITTMCT results were subjected to validity and reliability testing. The validity measures were content validity and construct validity using exploratory factor analysis (EFA). The instrument's reliability was measured using the internal consistency value of Cronbach's alpha.

The content validity index (CVI) was used to measure content validity. The CVI is a way to assess the validity of a questionnaire or test. Polit et al. [17] argued that CVI determines how well the items in the instrument represent the content being measured. A panel of experts judges the relevance of each item using a 4-point scale. The higher the CVI score, the stronger the instrument's content validity, indicating that the items effectively measure the intended construct.

The instrument used was adapted from Waltz and Bussel [18]. This scale contains four sections: relevance, clarity, simplicity, and ambiguity. For each section, the expert rated the instrument from 1 to 4. As shown in table 4, the expert rated the instrument as 1 for not relevant and 4 for very appropriate under the relevance section.

The researcher calculated the average rating for each criterion to determine an item's CVI based on relevance, clarity, simplicity, and ambiguity. This was done by adding up the ratings given by all ten experts and then dividing by the total number of experts. Then, the overall CVI was taken from the average of these four sections.

Another measure to ensure the validity of the study's instrument is the EFA. It is a statistical method widely used in research to investigate the underlying structure of a set of variables and discover the relationships between them [19]. They added that this technique is beneficial in evaluating a measurement tool's validity by exploring the data's dimensionality and uncovering the hidden factors that may affect participants' responses. EFA offers valuable insights into the structure of the construct being measured, which can aid researchers in refining their measurement tools, creating theories, and guiding future research efforts.

This study has only 30 participants (N = 30). Some author recommends using a formula where the number of samples should be at least five times the number of variables to determine the sample size

Table 5
The Cronbach's Alpha per item.

Item	If Item Dropped Cronbach's Alpha	Mean	SD
Q1	0.770	2.800	0.887
Q2	0.761	2.700	0.837
Q3	0.751	3.067	0.868
Q4	0.757	2.800	0.887
Q5	0.746	2.700	0.988
Q6	0.760	2.600	0.621
Q7	0.745	2.967	0.964
Q9	0.784	2.867	0.900
Q12	0.756	2.233	0.679
Q13	0.755	2.800	0.887
Q14	0.751	3.067	0.868
Q15	0.768	2.733	0.944
Q16	0.769	2.733	0.828
Q17	0.771	2.733	0.785
Q19	0.773	2.533	0.629
Q20	0.770	2.100	0.845
Q21	0.761	2.467	0.571
Q22	0.763	2.667	0.606
Q23	0.758	2.433	0.504
Q24	0.746	2.600	0.932
Q25	0.760	2.600	0.621
Q26	0.753	2.733	0.740
Q27	0.761	2.567	0.679
Q29	0.772	2.567	0.568
Q31	0.763	2.700	0.702
Q32	0.758	2.367	0.765
Q33	0.758	2.500	0.731
Q34	0.758	2.433	0.504
Q35	0.746	2.600	0.932
Q37	0.756	2.233	0.679

[20]. This study's combination of SPS and conceptual understanding is only one variable. Hence, N = 30 is suitable for EFA.

The data analysis for the EFA and Cronbach's Alpha were conducted using the JASP free software, while the descriptive statistics like mean and standard deviation were from Microsoft Excel.

3. Results and discussion

3.1. Reliability of the instrument

The instrument's reliability was measured by internal consistency using Cronbach's Alpha. A new instrument is said to have good internal consistency if the Alpha value is higher than 0.60 [21]. The values of the alpha per item are found in table 5.

The data were uploaded to Jasp Software. Reliability analysis and unidimensional reliability were chosen, and Cronbach's Alpha was selected. The study revealed the overall point estimate of alpha is 0.766 when items 8, 10, 11, 18, 28, 30, and 36 were removed.

The data in table 5 provides information regarding the participants' preliminary test scores. Most obtained a score of 2, meaning their concepts contain alternative conceptions. Very few got a score of 3, meaning they have a partial understanding with alternative conceptions.

Table 6
Distribution of Experts' Appraisal to the Instrument.

Score Range	Relevance	Clarity	Simplicity	Ambiguity
1.00-1.99	3	1	2	0
2.00-2.99	5	4	6	7
3.00-3.99	27	32	27	30
4.00	2	0	2	0

Table 7
The sample CVI computation for each item under the criteria of relevance.

Item No.	Relevant	CVI	Interpretation
1	9	0.90	Appropriate
2	7	0.70	Needs Revision
3	9	0.90	Appropriate
4	10	1.00	Appropriate
5	10	1.00	Appropriate
6	8	0.80	Appropriate
7	2	0.20	Eliminated
8	10	1.00	Appropriate
9	9	0.90	Appropriate
10	9	0.90	Appropriate

3.2. Content validity of the instrument

The ten experts who judged the instruments were teachers of the CAIE curriculum for at least three years. They are familiar with the CAIE content and how TIMSS questions are structured. They were given a printed copy of the instrument and the scoring guide for content validity. It took the judges four days to finish all the scoring and writing additional comments for the instrument.

The experts appraised most items with a score between 3.00 and 3.99, meaning minor revision was needed. Only a tiny number were rated perfect. On the other hand, some items were not relevant, not clear enough, not simple or very difficult, and too doubtful, while some needed significant revision. This means that most of the instrument's contents are relevant, clear, and simple, and the meaning is clear.

Zamanzadeh et al. [22] recommended that new instruments have 80 percent agreement or higher among experts when developing them. The CVI of each item should be considered to determine its appropriateness. An item is considered appropriate if the CVI is greater than 79 percent. However, it requires revision if it falls between 70 and 79 percent. The item is eliminated if the CVI is less than 70 percent. The relevant items mean that experts rated it by 3 or 4. This CVI was computed for all four criteria to carefully examine the items needing revision. The data in table 4 and table 5 also match and agree with one another. Both ideas in table 4 and table 6 were considered which item to eliminate or revise. When the item is lower than 2 in table 6 and is to be eliminated in its CVI, it is completely removed.

Table 7 shows the sample distribution of items and their CVI under the relevance section. The CVI of each item is calculated as the number of 3 or 4 ratings divided by the number of experts (N = 10). The researcher eliminated all items with an average rating of 1.00-1.99 and those with CVI below 0.50. The final number of items with appropriate and considerable content validity and reliability was 30. The items that needed revisions were revised based on the comments of the ten experts.

3.3. Construct validity of the instrument

The Kaiser–Meyer–Olkin (KMO) test and Bartlett sphericity test were conducted to ensure that the assumptions in this validity are met. The KMO value was 0.500. The (Measures of Sampling Adequacy) MSA value should equal or exceed 0.500 for consideration for further analysis [21]. The Bartlett

Table 8
The component characteristics of the EFA.

Components	Eigenvalue	Unrotated solution		Rotated solution		
		Proportion var.	Cumulative	SumSq. Loadings	Proportion var.	Cumulative
1	5.208	0.174	0.174	4.551	0.152	0.152
2	4.483	0.149	0.323	3.174	0.106	0.258
3	3.884	0.129	0.453	3.119	0.104	0.361
4	2.879	0.096	0.548	3.001	0.100	0.462
5	2.299	0.077	0.625	2.909	0.097	0.558
6	2.190	0.073	0.698	2.833	0.094	0.653
7	1.721	0.057	0.755	1.772	0.059	0.712
8	1.418	0.047	0.803	1.715	0.057	0.769
9	1.179	0.039	0.842	1.696	0.057	0.826

sphericity test yielded the Chi-Square test with a p-value of <0.001, lower than the significance value of 0.05. Both tests concluded that a factor analysis could be conducted for factor loading analysis.

Ten factors were derived from the EFA. Components with eigenvalues greater than 1.0 are separated into distinct components [23]. These components and their eigenvalues are found in table 8, corroborated by the scree plot in figure 1. Therefore, this research obtained ten components from the instrument's EFA.

The table shows that component 1 has the greatest eigenvalue in both the unrotated and rotated solutions, which implies that it explains the highest amount of variance in the data. The portion of the overall variation in the data that each component can account for can help comprehend the significance of each component in capturing the fundamental structure of the data. When a component has a more significant proportion of variance, it indicates that it has a greater impact in explaining the variability observed in the dataset.

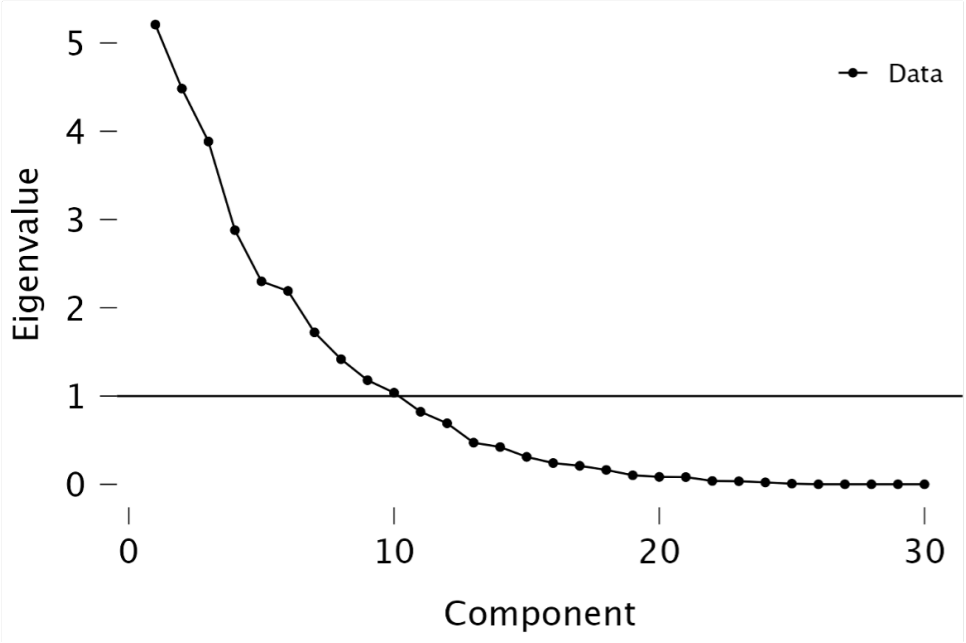

Figure 1: The Scree Plot of EFA shows ten components derived from JASP software.

The ten identified components account for 87.6% of the total variance. This statement suggests that the integrated assessment tool evaluates a primary element or aspect known as the ability, as stated in the work of Hambleton, Swaminathan, and Rogers [24]. In this study, the ability is known as the

Table 9
The factor loadings per item.

Item	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	Uniqueness
Q6	0.936										0.048
Q25	0.936										0.048
Q26	0.878										0.103
Q7	0.848										0.125
Q32	0.622										0.264
Q12		0.889									0.106
Q37		0.889									0.106
Q27		0.805									0.211
Q33		0.467					-0.644				0.110
Q24			0.956								0.023
Q35			0.956								0.023
Q5			0.923								0.062
Q2				0.939							0.084
Q13				0.894							0.073
Q21				0.830							0.134
Q23					0.941						0.023
Q34					0.941						0.023
Q22					0.603						0.158
Q4					0.455				0.496		0.113
Q3						0.945					0.067
Q14						0.945					0.067
Q31						0.573					0.378
Q17							0.630				0.163
Q20							0.538				0.119
Q1								0.853			0.165
Q16								0.661			0.091
Q9									0.873		0.086
Q19									-0.495		0.155
Q15										0.793	0.249
Q29										0.591	0.325

Note: Applied rotation method is varimax.

combination of SPS and conceptual understanding.

Factor loadings indicate how strongly individual items are related to the underlying components and in which direction [25]. This analysis used a varimax rotation to make the factor structure easier to understand. The displayed loadings in table 9 show the loads beyond 0.45. The missing loads are below 0.45 and were not shown.

The factor loadings show how the individual items are associated with each principal component. Items such as Q6 and Q25 have high loadings on PC1, which suggests they are strongly connected to this component. On the other hand, Q24 and Q35 have high loadings on PC3, indicating their association with a different underlying dimension. In contrast, Q33 and Q37 have low loadings across all principal components, which suggests weaker associations with the identified factors.

Interpreting the factor loadings can help understand the latent with. High loadings on the same principal component likely measurement principal component are likely measuring a common underlying construct. For instance, Q6, Q25, Q26, and Q7 all have strong loadings on PC1, suggesting they contribute to measuring a specific dimension or trait together. Conversely, items with high loadings on different principal components may represent distinct constructs or dimensions.

The factors were grouped according to their level of cognition. For the researcher, the EFA factor loads were categorized according to the level of cognition. Since the participants were answering the instrument with increasing levels of cognitive ability based on the revised Bloom's taxonomy, the degree of item connection was seen. These results were similar to the findings of Sadhu and Laksono

Table 10
Factor groupings and identification.

Components	Items	Factor
PC1	6,7,25,26,32	Remembering
PC2	12,27,33,37	Understanding
PC3	5, 24,35	Applying
PC4	2,13,21	Analyzing
PC5	22, 23, 34	Evaluating
PC6	14,17,31	Remembering
PC7	17,20	Applying
PC8	1,16	Analyzing
PC9	4,9,19	Understanding
PC10	15,29	Evaluating

[25], who found that their integral assessment instrument in chemistry can measure nine factors. They added that quantitative skills are the most dominant.

It was observed that items grouped in one principal component (PC) tend to come from similar items. For instance, questions 15 and 29 were identified as evaluation questions based on the EFA, but even before the analysis, these two questions were also classified as evaluation questions. The most dominant principal component is the PC1, which contributes to eight items, followed by the PC2 with seven items. This means that the participants' most dominant scores were from remembering questions.

Furthermore, these findings show that the instrument can measure the teachers' combined conceptual understanding and SPS. Thus, it is worth noting that this instrument is an integrated assessment for physics. This paper has found that integrating SPS and conceptual understanding in an instrument is feasible.

3.4. Limitations of the test

Administering a three-tier test in a classroom environment may pose time constraints compared to traditional assessments, which could be challenging due to the packed curriculum. Interpreting the results of a three-tier test can be complex and may require advanced statistical methods to differentiate between genuine comprehension and surface-level knowledge. Teachers may require additional training to effectively utilize and understand the outcomes of these tests, which could be a hurdle in some educational settings.

The second tier of the test depends on students' self-evaluation of their confidence, but students' self-perception may not always align with their actual understanding, leading to biased outcomes. High confidence does not necessarily indicate accuracy, especially when students are unaware of their misunderstandings ([9]).

Additionally, the integration of smartphone-based experiments assumes that all students have access to smartphones or similar technology, which may not be the case. Disparities in the type of smartphones or the availability of specific apps could impact how students engage in related experiments, potentially affecting the fairness and reliability of the assessment.

4. Conclusion

Based on the instrument's analysis results, the integrated three-tier multiple-choice test this paper developed is now considered valid and reliable. The instrument has good internal consistency based on Cronbach's Alpha. The revised and retained items were relevant, clear, and simple, and their meanings were clear.

The instrument passed the KMO and Barlett's sphericity test for EFA. The EFA results found that the instrument measured the teachers' combined conceptual understanding and science process skills. However, the most dominant level is the remembering level, while the evaluating level is the lowest.

Therefore, this paper suggests using this instrument to measure the level of integrated physics SPS and conceptual understanding during a smartphone-based physics experiment.

This paper recommends evaluating the level of teachers' SPS and conceptual understanding using a three-tier test to determine the alternative conceptions the teachers have and devise interventions. This intervention can be considered during a continuing professional development program. The use of this instrument with students is also feasible. Moreover, this instrument can be enhanced if the item's difficulty index and face validity are measured.

5. Acknowledgments

I would like to thank the teachers who participated in the instrument's pilot testing and final administration. I also appreciate the experts who thoroughly scrutinized the instrument. Thanks also to Dr. Reston of USC for the comments and suggestions for the instrument and this paper.

References

[1] F. S. Arista, H. Kuswanto, Virtual Physics Laboratory Application Based on the Android Smartphone to Improve Learning Independence and Conceptual Understanding, International Journal of Instruction 11 (2018) 1–16. doi:10.12973/iji.2018.1111a.

[2] A. Ismail, I. Festiana, T. Hartini, Y. Yusal, A. Malik, Enhancing students' conceptual understanding of electricity using learning media-based augmented reality, in: Journal of Physics: Conference Series, volume 1157, IOP Publishing, 2019, p. 032049. doi:10.1088/1742-6596/1157/3/032049.

[3] S. Cai, C. Liu, T. Wang, E. Liu, J.-C. Liang, Effects of learning physics using Augmented Reality on students' self-efficacy and conceptions of learning, British Journal of Educational Technology 52 (2021) 235–251. doi:10.1111/bjet.13020.

[4] K. Hochberg, S. Becker, M. Louis, P. Klein, J. Kuhn, Using smartphones as experimental tools—a follow-up: cognitive effects by video analysis and reduction of cognitive load by multiple representations, Journal of Science Education and Technology 29 (2020) 303–317. doi:10.1007/s10956-020-09816-w.

[5] S. A. Nikou, A. A. Economides, Mobile-Based micro-Learning and Assessment: Impact on learning performance and motivation of high school students, Journal of Computer Assisted Learning 34 (2018) 269–278. doi:10.1111/jcal.12240.

[6] D. T. Tiruneh, M. De Cock, A. G. Weldeslassie, J. Elen, R. Janssen, Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism, International Journal of Science and Mathematics Education 15 (2017) 663–682. doi:10.1007/s10763-016-9723-0.

[7] P. He, C. Zheng, T. Li, Upper Secondary School Students' Conceptions of Chemical Equilibrium in Aqueous Solutions: Development and Validation of a Two-Tier Diagnostic Instrument, Journal of Baltic Science Education 21 (2022) 428–444. doi:10.33225/jbse/22.21.428.

[8] R. Hanson, The impact of two-tier instruments on undergraduate chemistry teacher trainees: An illuminative assessment, International Journal for Infonomics 12 (2019) 9. doi:10.20533/iji.1742.4712.2019.0198.

[9] C. M. H. Laeli, et al., The 3 Tiers Multiple-Choice Diagnostic Test for Primary Students' Science Misconception, Pegem Journal of Education and Instruction 13 (2023) 103–111. doi:10.47750/pegegog.13.02.13.

[10] S. Türkoguz, Investigation of Three-Tier Diagnostic and Multiple Choice Tests on Chemistry Concepts with Response Change Behaviour, International Education Studies 13 (2020) 10–22. doi:10.5539/ies.v13n9p10.

[11] Debbye Vitti" and "Angie Torres, Practicing Science Process Skills at Home, 2006. URL: https://www.studocu.com/en-za/document/stadio/teaching-natural-sciences/teaching-scientific-process-skills/88305872.

[12] B. S. Bloom, A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives, Longman, 2010.
[13] M. Napal, A. M. Mendióroz-Lacambra, A. Peñalva, Sustainability teaching tools in the digital age, Sustainability 12 (2020) 3366. doi:10.3390/su12083366.
[14] I. Mullis, M. Martin, K. Cotter, V. Centurino, TIMSS 2019 Item Writing Guidelines, 2019. URL: https://timssandpirls.bc.edu/timss2019/methods/pdf/T19-item-writing-guidelines.pdf.
[15] M. O. Martin, I. V. Mullis, M. Hooper, Methods and procedures in TIMSS Advanced 2015, Boston College, TIMSS & PIRLS International Study, 2016. URL: http://timss.bc.edu/publications/timss/2015-a-methods.html.
[16] F. A. A. Pacala, Development and Validation of Three-Tier Multiple Choice Test for Conceptual Understanding in Momentum and Collision, International Journal of Multidisciplinary Approach & Studies 5 (2018) 1–7. URL: http://ijmas.com/upcomingissue/01.02.2018.pdf.
[17] D. F. Polit, C. T. Beck, S. V. Owen, Is the CVI an acceptable indicator of content validity? Appraisal and recommendations, Research in nursing & health 30 (2007) 459–467. doi:10.1002/nur.20199.
[18] C. F. Waltz, B. R. Bausell, Nursing research: design statistics and computer analysis, Philadelphia : F.A. Davis Co., 1981. URL: https://archive.org/details/nursingresearchd0000walt/page/n3/mode/2up.
[19] H. W. Marsh, B. Muthén, T. Asparouhov, O. Lüdtke, A. Robitzsch, A. J. Morin, U. Trautwein, Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching, Structural equation modeling: A multidisciplinary journal 16 (2009) 439–476. doi:10.1080/10705510903008220.
[20] Y. Chua, Research methods and statistics book 4: Univariate and multivariate tests, Shah Alam, Malaysia: McGraw-Hill Education, 2009.
[21] N. W. D. Ayuni, I. G. A. M. K. K. Sari, Analysis of factors that influencing the interest of Bali State Polytechnic's students in entrepreneurship, in: Journal of Physics: Conference Series, volume 953, IOP Publishing, 2018, p. 012071. doi:10.1088/1742-6596/953/1/012071.
[22] V. Zamanzadeh, A. Ghahramanian, M. Rassouli, A. Abbaszadeh, H. Alavi-Majd, A.-R. Nikanfar, Design and implementation content validity study: development of an instrument for measuring patient-centered communication, Journal of caring sciences 4 (2015) 165. doi:10.15171/jcs.2015.017.
[23] Z. Awang, A handbook on structural equation modelling using AMOS, Universiti Technologi MARA Press, Malaysia, 2012.
[24] R. Hambleton, H. Swaminathan, R. Jane, Fundamentals of item response theory, Sage Publications, 1991.
[25] S. Sadhu, E. W. Laksono, Development and Validation of an Integrated Assessment for Measuring Critical Thinking and Chemical Literacy in Chemical Equilibrium, International Journal of Instruction 11 (2018) 557–572. doi:10.12973/iji.2018.11338a.