

Development and validation of conceptual test instrument in Physics: A synthesis

Mary Allein Antoenette C. Bug-os^{1*}

¹University of Science and Technology of Southern Philippines, Philippines

*Corresponding author: allein.bugos@ustp.edu.ph

ARTICLE INFO

DOI:10.46223/HCMCOUJS.soci.en.13.2.2714.2023

Received: March 30th, 2023

Revised: May 29th, 2023

Accepted: June 07th, 2023

Keywords:

BSEd - Science curriculum; conceptual test; physics courses pre-service teachers

ABSTRACT

A well-developed physics conceptual test that can be a diagnostic, a misconception, or a summative test best suited to the science major will help educators determine the level of physics understanding of the prospective science teachers at public and private high schools. Hence, the purpose of this synthesis paper is to look into the key aspects of the process of the development and validation of conceptual test instruments in physics subjects for the undergraduate level. A literature review was used to determine the process of the development of conceptual tests. The process starts with preparation or content analysis, development, validation, and pilot testing. Statistical analysis, such as difficulty and discrimination indices and biserial coefficients used by the authors of various researchers, was also included. A stable, usable, and promising form test instrument can help instructors evaluate the effectiveness of the pedagogies and new curriculum.

1. Introduction

Conceptual understanding is very important, particularly for teacher education programs. It enables individuals to explain a phenomenon in different situations (Viennot, 2009), which only means it is far beyond knowing facts. One characteristic that promotes conceptual understanding is a task demanding the application and use of knowledge for solving a problem (Alonso-Tapia, 2006). Pre-service science teachers need to better understand science and its application because of their influence as future teachers on many students (Kaltakci-Gurel, 2023). On the other hand, a written examination is commonly conducted to determine conceptual understanding efficiently, combining lower-order thinking and higher-order thinking skill questions. A conceptual multiple-choice test differs from the traditional classroom multiple-choice tests used to grade the students. A well-developed conceptual tests result from thorough research in the field, and each distracter provides diagnostic information about common areas of difficulty and misconceptions of a certain subject (Engelhardt, 2009). As previous study revealed that pre-service science teachers still have misconceptions (Hartini, Liliyasi, Setiawan, & Ramalis, 2020) and difficulty in connecting from different disciplines (Winarno, Widodo, Rusdiana, Rochintaniawati, & Afifah, 2019), and their ideas prior to application affect their teaching activities (Takaoğlu, 2017).

New policies, standards, and guidelines for teacher education at the secondary level (CHED, 2017) have been implemented to cater to the enhanced basic education program (Department of Education, 2019; The Official Gazette, 2013), specifically to address the spiral progression approach. The new science program incorporates the different science disciplines in every grade level, starting in Grade 3 (Department of Education, 2019). The science concepts are learned each year at a higher degree in the succeeding years. Each grade has an enhanced and decongested

curriculum (Laureano, Espinosa, & Avilla, 2015). The new approach paves the way for a deeper understanding of the science core concepts and learning competencies to be mastered (Ely, 2019). In implementing the spiral approach, the prime movers are the teachers. Undeniably, the quality of teachers has a significant relationship with the mastery of the learning competencies of the students (Ely, 2019). With this, students of BSEd major in Science program must demonstrate a deep understanding of scientific concepts and principles in the different areas of science (CHED, 2017), such as chemistry, earth and space science, biology, and physics, to be able to effectively teach science after completing the program and passing the licensure examination.

Throughout the four-year program, students are bombarded with different physics courses, like mechanics, thermodynamics, and magnetism. Consequently, it is interesting to determine the conceptual understanding of the science majors in all the physics courses they took. A developed physics conceptual test that can be a diagnostic, a misconception, or a summative test (Engelhardt, 2009) best suited to the science major will help educators determine the level of physics understanding of the prospective science teachers at the junior high school and senior high school levels. Also, a validated multiple-choice conceptual test can be valuable in determining the effectiveness of the new BSEd - Science curriculum in conceptually teaching physics. Hence, constructing a high-quality multiple-choice test to assess the conceptual understanding of physics must also involve a series of processes that must be implemented.

2. Objective

This paper synthesizes the development and validation of conceptual test instruments in various physics subjects for the undergraduate level.

3. Methodology

This paper is a literature review. The literature search is done through online databases such as ERIC, EBSCO, and Google Scholar. Selected published articles were focused on developing and validating conceptual instruments, particularly in undergraduate physics subjects such as mechanics, thermodynamics, electricity and magnetism, and modern physics. These physics areas are covered in the new BSEd - Science program; hence the articles were chosen. The articles chosen were published from 2012 - 2021. The process of the development and validation of the instruments was synthesized based on the stages of developing and validating a conceptual test from preparation or content analysis, development, validation, and pilot testing. Statistical analysis conducted by the authors of various researchers was also included.

4. Result and discussion

Assessment of students' understanding is crucial, and one way to execute that is to construct a valid and reliable test instrument. Constructing a valid and reliable test involves various stages, such as preparation or content analysis, development, validation, and pilot testing (Morales, 2012; Rahmawati, Rustaman, Hamidah, & Rusdiana, 2018).

4.1. Preparation stage

In the preparation stage, previous studies were reviewed (Morales, 2012; Önder, 2016; Rahmawati et al., 2016; Sadaghiani & Pollock, 2015), and results from the initial survey conducted through questionnaires (Rahmawati et al., 2018), open-ended test (Brown & Singh, 2021; Sadaghiani & Pollock, 2015) and interview (Önder, 2016; Sadaghiani & Pollock, 2015) were also analyzed to determine the students' prevalent difficulties or misconceptions in physics concepts. Physics books (Brown & Singh, 2021; Rahmawati et al., 2018) and course syllabi (Morales, 2012) were also examined.

4.2. Development stage

After all inputs are collected, this will serve as base data for the development stage. A test blueprint is developed to provide a framework for the desired test attributes, which entailed scope and complexity, format, testing duration, and determining the weights of different subtopics (Brown & Singh, 2021; Li & Singh, 2017). Formulation of specific test objectives and scope takes place (Brown & Singh, 2021; Morales, 2012) with the consultation of previous instructors (Brown & Singh, 2021; Li & Singh, 2017). Using the specific objectives of each topic, a table of contents was prepared together with the level of cognitive complexity desired, which uses Bloom's taxonomy is examined by the instructors and professors handling the subjects (Brown & Singh, 2021; Li & Singh, 2017; Morales, 2012). During this process, even the kind of questions in homework and quizzes are scanned (Brown & Singh, 2021). Some modifications, like weights assigned to various concepts, were made appropriately based on the feedback.

Interestingly, in Li and Singh (2017) study, they constructed a table specification along with the criteria for good performance. The performance criteria were used to convert the description of conditions/contexts within which the concepts would be tested to make free-response questions. These free-response questions were given to the students, and individual interviews were conducted during this stage to provide students with reasoning for their responses.

In the item test construction, the wording of the questions took advantage of the existing findings from literature regarding student difficulties or misconceptions, input from students' written responses and interviews, and input from physics instructors who teach the subject (Brown & Singh, 2021; Önder, 2016; Sadaghiani & Pollock, 2015). More items in the cognitive domain that promote higher-order-thinking skills were prepared (Morales, 2012), meaning most questions require reasoning. Very few questions can be answered simply by rote memory (Brown & Singh, 2021).

In terms of the number of choices, a 4-choices format is utilized as it is the most common in standardized tests in the Philippines as well as teacher-made tests (Morales, 2012), but other foreign studies utilized a 5-choices format where the four are distractor or alternative or incorrect choices (Brown & Singh, 2021; Li & Singh, 2017). This is done to increase the discriminating properties of the test items (Li & Singh, 2017). The distractors were developed from the students' most frequent incorrect responses in free-response questions/open-ended tests (Brown & Singh, 2021; Li & Singh, 2017; Sadaghiani & Pollock, 2015). Even at this point, Sadaghiani and Pollock (2015) conducted a follow-up interview to understand better students' reasoning for the answers to the open-ended questions. This ensures that the language is a minor obstacle for the students. Nevertheless, results from interviews in any period of the process serve as a guide for making the distractor choices (Brown & Singh, 2021; Li & Singh, 2017; Önder, 2016).

4.3. Validation stage

To examine the appropriateness and relevance of the test instrument, it is subjected to validation in the face and content validity contexts. Face validity is the extent to which a test is subjectively viewed as covering the concept it purports to measure (Sadaghiani & Pollock, 2015). Descriptive validation is commonly done for face validity, and this is done through investigation of the comments, remarks, or suggestions from the experts (Morales, 2012; Sadaghiani & Pollock, 2015). Also, in the study of Sadaghiani and Pollock (2015), additional face validity was established during student interviews after alpha and beta testing. Refining and modifying the wording of some of the questions based on the interview helped improve the validity of the instrument.

In content validation, three or more experts will examine the test instrument. Its appropriateness and relevance can be examined using quantitative content validation using a 5-

point Likert evaluation scale checklist (Morales, 2012; Rahmawati et al., 2018) or using a form that experts filled out to mark the items as ‘Necessary,’ ‘Useful but not necessary,’ or ‘Unnecessary’ (a technique developed by Lawshe in 1975) (Önder, 2016) besides from the comments and suggests from the experts (Brown & Singh, 2021; Li & Singh, 2017). In addition to the mean values of experts using the Likert scale checklist, the content validity coefficient per checklist item using Aiken’s content validity coefficient formulas is used to ensure that the test is rated as a content-valid test (Morales, 2012; Rahmawati et al., 2018). Whereas, to determine which questions would remain or be removed in the test, the Content Validity Ratio (CVR) must determine then the calculation of the Content Validity Index (CVI) of the test as a whole if the technique is to be followed in the study of Önder (2016).

Usually, content validation by experts is done after revisions of the test instrument (Morales, 2012) and prior to testing to detect ambiguity, but Sadaghiani and Pollock (2015) conducted their content validation in three different stages such as during the construction of the test format and after both alpha and beta testing. Moreover, Rahmawati et al. (2018) study included a readability test conducted on selected pre-service teachers. Based on their result, some questions need to be revised in terms of the use of more familiar words, communicative sentences, and complete and detailed illustrations. On the other hand, in the study by Li and Singh (2017), several students were asked to answer the survey questions individually in an interview in which they were asked to think aloud before the pilot testing. This is to ensure that students understand the questions and the wording is clear.

4.4. Pilot-testing

After validating the content of the test instrument, the instrument validation process the test was followed by an instrument test for a group of undergraduates in a specific academic year (Morales, 2012; Rahmawati et al., 2018) or a combination of different groups in successive years (Önder, 2016). A wide range of participants can also be utilized, like in the study of Sadaghiani and Pollock (2015), where they selected 12 classes at ten institutions nationwide. On the other hand, Brown and Singh (2021) and Li and Singh (2017) involved PhD students aside from the undergraduates majoring in physics for comparison.

4.5. Statistical analysis

Most item analyses for the developed multiple-choice conceptual test for physics include item categorization based on difficulty and discrimination indices. Based on the statistical test result for the difficulty index, the items were categorized as easy, moderately difficult, or very difficult. In contrast, the discrimination index items were classified as accepted, revised, or rejected. Few studies utilized the test items’ point biserial coefficients (Brown & Singh, 2021; Li & Singh, 2017; Önder, 2016). This is to measure a test item’s consistency with the whole test. Ferguson’s delta (δ), or the coefficient of test discrimination, is also utilized by Sadaghiani and Pollock (2015) to measure the discrimination power of a test instrument by investigating how broadly the total scores of a sample are distributed over the possible.

On the other hand, Rahmawati et al. (2018) included the distractor functionality of test items to identify whether the provided distractor is a reasonable distractor that can be selected by at least 5% of all test takers. Morales (2012) utilized distractor analysis to identify the alternative conception of the students concerning a particular concept being assessed by an item. This is done by grouping the students with high or low score categories and determining the frequency of students answering each choice. The most frequent distractor chosen by students may include data or phrases with the students’ alternative conception. Furthermore, to determine the test

instrument's internal consistency, the Kuder-Richardson reliability index or Cronbach alpha was used. This is a measure of the overall correlation between items. Also, Önder (2016) included the average inter-item correlations to analyze internal consistency reliability. The value obtained in various statistical analyses will indicate whether the conceptual test instrument is valid and sufficient to measure students' understanding of physics concepts.

Overall, developing and validating the physics conceptual test instrument involved a series of processes and a protocol for developing a good quality test (Morales, 2012). The previous studies take into serious consideration in formulating effective distractors (Brown & Singh, 2021; Sadaghiani & Pollock, 2015) because the analysis of the distractors is important in providing diagnostics of student naive ideas and difficulties, evaluating teaching methods, and comparing curricula (Morales, 2012; Sadaghiani & Pollock, 2015). According to Haladyna and Rodriguez (2013), one of the ways to develop plausible distractors can be identified by evaluating the students' answers to response questions. This is done by the studies of Brown and Singh (2021), Li and Singh (2017), and Sadaghiani and Pollock (2015), where they utilized the students' most frequent incorrect responses in free-response questions/open-ended tests or interviews. This approach is very useful as it extracts the actual responses of the students, but students' responses may also produce a broad spectrum of plausible distractors and keys (Sadaghiani & Pollock, 2015). So, in constructing a multiple-choice conceptual test utilizing answers from open-ended questions, it is best to organize common core themes among similar students' ideas and use them as a plausible distractor.

Additionally, pilot testing is essential as it reduces the measurement of errors during full-scale data collection. All cited studies conducted pilot testing, and necessary changes were made before the actual data collection. If changes are substantive, a second round of pretesting is possible (Grimm, 2010). As Peat, Mellis, Williams, and Xuan (2001) mentioned, pilot testing ensures that materials are complete and function well in practice. Procedures such as discarding all unnecessary, difficult, or unclear questions, recording the time to complete the questionnaire, and deciding whether it is reasonable will help improve the internal validity of a questionnaire. Noting that follow-up individual interviews with students after limited tryouts of the instruments were also done to improve question validity (Sadaghiani & Pollock, 2015).

Development and validation of a multiple-choice conceptual test is one of the ways to determine conceptual understanding with large numbers of students. Multiple-choice tests are chosen because of their validity, reliability, and practicality (Soeharto, Csapó, Sarimanah, Dewi, & Sabri, 2019). Conceptual tests were explicitly designed due to thorough research in the field, and each distracter usually gives evidence about specific student misconceptions and difficulties (Gurel, Eryilmaz, & McDermott, 2015). A test time duration is one of the drawbacks of a multiple-choice type of instrument, especially the four-tier multiple-choice tests (Caleon & Subramaniam, 2010; Soeharto et al., 2019). Noting that using this type of multiple-choice conceptual test measures misconceptions free from errors and lack of knowledge in a valid way (Gurel et al., 2015).

5. Conclusions & recommendations

Therefore, constructing a high-quality multiple-choice test to assess the conceptual understanding of physics involves a series of processes from preparation, development, validation, and pilot testing and the conduct of various and appropriate statistical analyses. The process of validating test instruments through trials was widely conducted to obtain various data indicating whether the conceptual test instrument was valid and reliable. A stable, usable, and promising form test instrument can help instructors evaluate the effectiveness of innovative pedagogies and,

importantly, the effectiveness of the new BSEd – Science curriculum to develop a solid grasp of Physics concepts if a plausible distractor in a four-tier multiple-choice test is well formulated. Conceptual multiple-choice test instrument has limitations; hence, a combination of written and oral instruments may be utilized to toughen the analysis of the obtained data and eliminate weaknesses found in an instrument.

References

- Alonso-Tapia, J. (2006). Knowledge assessment and conceptual understanding. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 389-413). Dordrecht, Netherlands: Springer.
- Brown, B., & Singh, C. (2021). Development and validation of a conceptual survey instrument to evaluate students' understanding of thermodynamics. *Physical Review Physics Education Research*, 17(1), Article 10104. doi:10.1103/PhysRevPhysEducRes.17.010104
- Caleon, I., & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, 32(7), 939-961. doi:10.1080/09500690902890130
- CHED. (2017). *Policies, standards and guidelines for Bachelor of Secondary Education (BSEd)*. Retrieved October 10, 2022, from <https://www.ptonline.com/articles/how-to-get-better-mfi-results>
- Department of Education. (2019). *Policy guidelines on the K to 12 basic education program*. Retrieved October 10, 2022, from https://www.deped.gov.ph/wp-content/uploads/2019/08/DO_s2019_021.pdf
- Ely, L. L. (2019). Mastery learning of chemistry competencies through the spiral progression approach in curriculum. *International Journal of Educational Science and Research (IJESR)*, 9(5), 9-28.
- Engelhardt, P. V. (2009). An introduction to classical test theory as applied to conceptual multiple-choice tests. *Getting Started in PER*, 2(1), 1-40.
- Grimm, P. (2010). Pretesting a questionnaire. In J. N. Sheth & N. K. Malhotra (Eds.), *Wiley international encyclopedia of marketing* (p. 2010). doi:10.1002/9781444316568.wiem02051
- Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science and Technology Education*, 11(5), 989-1008. doi:10.12973/eurasia.2015.1369a
- Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items* (1st ed.). New York, NY: Routledge.
- Hartini, T. I., Liliyasi, S., Setiawan, A., & Ramalis, T. R. (2020). Enhancing conceptual understanding and critical thinking skills pre-service of physics using Mechanics Multi Representation (MMR). *International Journal of Scientific & Technology Research*, 9(4), 2994-2998.
- Kaltakci-Gurel, D. (2023). Exploring pre-service teachers' conceptual understanding and confidence in geometrical optics: A focus on gender and prior course achievement. *Education Sciences*, 13(452).

- Laureano, R. A., Espinosa, A. A., & Avilla, R. A. (2015). Effects of “Grade 9 science learner’s material” on students’ self-regulation and achievement in chemistry. *Electronic Journal of Science Education*, 19(8), 28-58.
- Li, J., & Singh, C. (2017). Developing and validating a conceptual survey to assess introductory physics students’ understanding of magnetism. *European Journal of Physics*, 38(2), 1-25. doi:10.1088/1361-6404/38/2/025702
- Morales, M. P. E. (2012). Development and validation of a concept test in introductory physics for biology students. *The Manila Journal of Science*, 7(2), 26-44.
- Önder, F. (2016). Development and validation of the photoelectric effect concept inventory. *European Journal of Physics*, 37(5), 1-18. doi:10.1088/0143-0807/37/5/055709
- Peat, J. K., Mellis, C., Williams, K., & Xuan, W. (2001). *Health science research: A handbook of quantitative methods* (1st ed.). London, UK: Routledge.
- Rahmawati, Rustaman, N. Y., Hamidah, I., & Rusdiana, D. (2018). The development and validation of conceptual knowledge test to evaluate conceptual knowledge of physics prospective teachers on electricity and magnetism topic. *Jurnal Pendidikan IPA Indonesia*, 7(4), 483-490. doi:10.15294/jpii.v7i4.13490
- Rahmawati, Rustaman, N. Y., Hamidah, I., Rusdiana, D., Li, J., Singh, C., Önder, F., ... Singh, C. (2016). Developing and validating a conceptual survey to assess introductory physics students’ understanding of magnetism. *European Journal of Physics*, 11(1), 1-14. doi:10.1088/0143-0807/37/5/055709
- Sadaghiani, H. R., & Pollock, S. J. (2015). Quantum mechanics concept assessment: Development and validation study. *Physical Review Special Topics - Physics Education Research*, 11(1), 1-14. doi:10.1103/PhysRevSTPER.11.010110
- Soeharto, Csapó, B., Sarimanah, E., Dewi, F. I., & Sabri, T. (2019). A review of students’ common misconceptions in science and their diagnostic assessment tools. *Jurnal Pendidikan IPA Indonesia*, 8(2), 247-266. doi:10.15294/jpii.v8i2.18649
- Takaoğlu, Z. B. (2017). Challenges faced by pre-service science teachers during the teaching and learning process in Turkey. *Journal of Education and Training Studies*, 5(2), 100-110. doi:10.11114/jets.v5i2.2132
- The Official Gazette. (2013). *Implementing rules and regulations of the enhanced basic education act of 2013*. Retrieved October 10, 2022, from <https://www.officialgazette.gov.ph/2013/09/04/irr-republic-act-no-10533/>
- Viennot, L. (2009). Learning and conceptual understanding: Beyond simplistic ideas, what have we learned? In M. Vicentini & E. Sassi (Eds.), *Connecting research in physics education with teacher education* (pp. 1-17). Retrieved October 10, 2022, from <http://web.phys.ksu.edu/ICPE/Publications/teach2/Viennot.pdf>
- Winarno, N., Widodo, A., Rusdiana, D., Rochintaniawati, D., & Afifah, R. M. A. (2019). Pre-service science teachers’ conceptual understanding of integrated science subject: A case study. *Journal of Physics: Conference Series*, 1204(1), 1-8. doi:10.1088/1742-6596/1204/1/012104

