JPII 7 (4) (2018) 483-490

Jurnal Pendidikan IPA Indonesia

http://journal.unnes.ac.id/index.php/jpii

THE DEVELOPMENT AND VALIDATION OF CONCEPTUAL KNOWLEDGE TEST TO EVALUATE CONCEPTUAL KNOWLEDGE OF PHYSICS PROSPECTIVE TEACHERS ON ELECTRICITY AND MAGNETISM TOPIC

Rahmawati*[1,2], N. Y. Rustaman[3], I. Hamidah[4], D. Rusdiana[5]

[1]Department of Physics Education, Universitas Muhammadiyah Makassar,Indonesia
[2,3,4,5]Department of Science Education, Post-graduated Studies Universitas Pendidikan Indonesia , Indonesia

DOI: 10.15294/jpii.v7i4.13490

Accepted: July 4th 2018. Approved: December 10th, 2018. Published: December 29th, 2018

ABSTRACT

The conceptual knowledge test is an efficient way to measure the conceptual knowledge of physics prospective teachers on electrical and magnetism topic. The employed instrument was physical questions in the form of multiple-choice options. The process of developing and validating the conceptual knowledge test consisted of 5 steps: (1) content analysis; (2) construction of multiple-choice items; (3) readability test and expert validation; (4) limited tryout; and (5) large-scale application. The instrument validation test through trials was conducted in order to obtain the data related to difficulty index, discriminating power, distractor functionality, and reliability coefficient value that was then analyzed using ITEMAN version 3.0 program. The participants were 215 physics prospective teachers of a university in Makassar city. The instrument validation resulted in 40 items that consisted of 26 items for electricity and 14 items for magnetism. The instrument is called Conceptual Knowledge Test-Electricity and Magnetism (CKT-EM). The value of the reliability coefficient (α) (Alpha Cronbach) of 0.87 indicated that the instrument of conception test on electrical and magnetism topics was valid and sufficient to measure students' conception on electrical and magnetism topic.

Keywords: conceptual knowledge-test, electricity and magnetism, readability test, validity, reliability

INTRODUCTION

Educational research related to the development of students' conceptual knowledge in the world of science education (physics) has been done for the past decades since Conceptual knowledge has been a primary review in learning science (Boorman & Rushworth, 2009; Duschl, 2008; Rittle-Johnson & Star, 2009; Fang et al., 2016). Conceptual knowledge is an important element to solve problems (Streveler et al., 2008). There are numerous previous studies about conceptual knowledge related to subject-matter or content in physics such as electricity and magnetism (Dega et al., 2013; Gok, 2012; Sadaghiani, 2011; Stelzer et al., 2010). Furthermore, conceptual knowledge is often approached from the viewpoint of semantic networks, because all retrieval and inference are based on traversing such networks (Koponen & Nousiainen, 2018). Electricity and magnetism are the important concepts that have to be learned by students as it plays a central role

*Correspondence Address
E-mail: rahmawatisyam@unismuh.ac.id

in determining the structure of the natural world and is the foundation of most current and emergent technology (Dega et al., 2013; Tiruneh et al., 2017). Therefore, it is important to explore students' conceptual knowledge of electricity and magnetism topic.

A variety of tests have been developed to measure students' conceptual knowledge of electricity and magnetism. Unfortunately, an extensive literature review on electricity and magnetism conceptual knowledge indicated that there are very few valid and reliable instruments applicable to measure students' conceptual knowledge of electricity and magnetism. Some examples of instruments that have been widely used in a number of previous studies focused on static electric for electricity topic. They are conceptual Survey of Electricity and Magnetism (CSEM), Brief Electricity and Magnetism Assessment (BEMA), Electricity and Magnetism Conceptual Assessment (EMCA), Colorado Upper-division Electrostatics (CUE), and Critical Thinking Skills in Electricity and Magnetism (CTEM) (Tiruneh et al., 2017). These instruments have been used to measure students' difficulties in some conceptual areas from electricity and magnetism (Gok, 2012; Pollock, 2009; Sadaghiani, 2011; Tiruneh et al., 2017). The Conceptual Survey of Electricity and Magnetism (CSEM) is a multiple-choice test that consists of 32-question. This instrument comprises eleven categories of question about conductor and insulators, Coulomb's law, force and field superposition, force, work, electric potential, magnetic force, Faraday's law, and Newton's third law. Based on the characteristics of the CSEM instrument, it tests electrical content, particularly on static electricity. In the CSEM instrument, no single item is found which tests dynamic electrical content needed in this study. Therefore, this study focused on developing instruments on dynamic electricity and magnetism topic.

Another similar instrument that can be used to measure students' conceptual knowledge of electricity and magnetism topics is Brief Electricity and Magnetism Assessment (BEMA). BEMA was developed by Ding (2014). This instrument is a 30-item multiple-choice test which covers the main topics discussed in both the traditional calculus-based E&M physics curriculum and the matter and interactions curriculum (Matter & Interactions II: Electric and Magnetic Interactions). It was re-evaluated through Rash Analysis (Ding, 2014). This instrument is specifically designed for students' knowledge of electrical and magnetic concepts that have completed electrical and magnetic courses. Electrical and magnetic contents tested in BEMA have a more complex level of difficulty than the CSEM instrument. All items of the BEMA instrument are taken from the curriculum and the electrical course materials given in the fourth or fifth semester in college. Unlike the CSEM instrument, all items of this instrument are basic electrical and magnetic contents which are part of basic physics subject. Nevertheless, the BEMA instrument was not in accordance with the needs of this study because the content provided focused on more complex electrical and magnetic materials and there are no items intended to test dynamic electrical materials.

McColgan et al. (2017) has developed an instrument to assess students' conceptual knowledge of electricity and magnetism. This instrument was named electricity and Magnetism Conceptual Assessment (EMCA). It includes 30 multiple-choice questions with a completion time of 30–40 minutes. Some topics covered in this instrument include electrostatics, electric fields, circuits, magnetism, and induction. This instrument is inspired by the CSEM and BEMA, yet it focuses on assessing the content of electricity and magnetism in the introductory physics course. This instrument is in accordance with the content that will be tested in this study. However, the weakness in this instrument lies in the content which mainly talks about static electricity instead of dynamic electricity. On the other hand, the basic physics course curriculum in the second semester discusses more dynamic electrical. Based on the weaknesses of previous studies about the development of conceptual knowledge tests, we designed a form of test that could assess the conceptual knowledge of pre-service physics teachers on electricity and magnetism topic with the distribution of sub-concepts based on the results of the initial survey conducted through questionnaires administered to a number of physics prospective teachers.

Electricity and magnetism review in Basic Physics course was generally given in the first year. The scope of electricity and magnetism is broad enough. Therefore, there were some physics prospective teachers had difficulties to master the materials though they have passed it. The results of an initial survey conducted on pre-service physics teachers who have contracted Basic Physics course showed that they experienced difficulties in some electrical concepts especially in dynamic electrical and magnetism subtopic (Rustaman et al., 2017). This finding was then followed up by conducting a survey of prospective teachers' conceptual knowledge related to electricity (dynamic electricity) and magnetism materials to support the prior findings. In other words, an instrument that measures physics prospective teachers' conceptual knowledge of electricity and magnetism topic was required in particular content areas.

An instrument is said to be good if it had three characteristics; valid, reliable, and usable (Gronlund, 1985; Secolsky & Denison, 2012). There were several approaches that can be used to perform instrument validity namely content validity, construction validity, and criterion validity (Gronlund, 1985; Mehrens & Lehmann, 1991).

The purpose of this study was to describe the results of the development and validation of the conceptual knowledge test to measure physics prospective teachers' conception of electricity and magnetism topic, certainly in sub-topics that considered being difficult by them according to the questionnaires. The instrument design of conceptual knowledge on electricity and magnetism topic should not only assess physics prospective teachers' conceptual knowledge but also detect their difficulty level about electricity and magnetism.

METHODS

This was developmental research intended to validate test. Clifton & Schriner (2010) explained that the step of developing and validating test items consist of five steps including content, format, writing the stem, and writing the choice. Furthermore, Erdogan et al. (2009) revealed that the development and validation of instruments or tests require eight steps, they are literature review, development of item pool, validation of item pool, constructing initial draft, taking expert opinion, pilot testing, administration of the instrument, and calculation of validity and reliability. Therefore, the research method used in this research was developmental as it in accordance with the research purpose that was to describe the development and validation of conceptual knowledge test on electricity and magnetism topic. The test took place in one of the universities in Makassar. A number of 105 physics prospective teacher consisting of 62 woman and 43 men participated in this research. They were second year physics prospective teachers who have contracted Basic Physics courses.

The CKT-EM instrument covered 45 items before being validated. After validation, the number of items reduced by 5 becoming 40 consisting of 26 items about electricity and 14 items about magnetism. The form of CKT-EM instrument was multiple choice with five choices of answers provided along with the reasons. The developed electrical material distributions included electric current, electric potential difference, electromotive force, resistance, energy and electrical power, dc circuit, and Kirchhoff I & II law. Meanwhile, the magnetic material distributions included forces in electric charge moving in magnetic fields caused by straight wires, and coils. Readability validation was measured based on the consideration of 5 physics prospective teachers. Furthermore, the content validation was performed by three expert judgments using assessment rubric. The empirical validation was done through tryout using physics prospective teachers' responses which then were analyzed on ITEMAN Version 3.0 program.

The instruments used in this study were questionnaires, readability assessment sheet, and validation sheet for expert judgment. First, the questionnaires intended to identify the materials of Basic Physics course that were perceived as difficult by physics prospective teachers. Second, readability assessment sheets were to get the responses of 5 students about the degree of instrument readability. Generally, a test's readability validation is examined from 3 main components: (1) the clarity of words and sentences; (2) grammar; and (3) the clarity of instrument guidance (Brookhart, 2010). Scaling is a method used to assess each aspect, and this research employed the Likert Scale. In Likert, subject read every statement in the questionnaire and evaluate the question based on the categorized answer (Clifton & Schriner, 2010; Svetina & Levy, 2014; Tarrant et al., 2009). The score ranged from 1 to 5; 1 indicates 'very bad', 2 is bad, 3 is fair, 4 is good, and 5 is excellent. The data were analyzed both qualitatively and quantitatively. Furthermore, the results of quantitative data analysis were converted into qualitative form. The categorization of scoring is explained in Table 1.

Table 1. Conversion of Average Score Interval

The Average Score Interval	Categorize
$X \geq \bar{X} + 1.SBx$	Very good
$\bar{X} + 1.SBx > X \geq \bar{X}$	Good
$\bar{X} > X \geq \bar{X} - 1.SBx$	Bad
$X < \bar{X} - 1.SBx$	Very Bad

Based on Table 1, the interpretation of the readability test average score after being converted into the rating scale is shown in Table 2.

Table 2. Average Score Interpretation of Instrument Readability Test

Average Score Interval	Category of test readability	Judgment
≥ 4	Very clear	Without revision
$4 > X \geq 3$	Clear	Few revision
$3 > X \geq 2$	Enough	Full revision
$X < 2$	Not Clear	Rejected

The validation sheet was employed to validate the test contents through expert judgment by 3 experts. There were 3 elements which should be assessed by the experts: (1) the accuracy of content; (2) the accuracy of problem construction; and (3) language aspect. The used assessment technique was a rating scale system in which validators provide a score on the validation sheet. The validation results from all validators were analyzed to prove instrument content validation. To assess each aspect of the statement, the Likert scale was employed. For the favorable statement, 1 is very bad, 2 is not good, 3 is fair, 4 is good, and 5 is very good. The content validation resulted were examined quantitatively in the form of validators' agreement level using an inter-rater method. To find out this agreement, the researchers utilized the validity index proposed by Aiken (Aiken, 1988). The content validity index of each item was calculated using Aiken's Formula Index as follows.

$$V = \frac{\Sigma s}{n(c-1)} \ldots \quad (1)$$

Description:
V = rater agreement index about content validity; s = total score; n = number of validator; c = number of categories which can be chosen by validators

The value of s was calculated using the formula s = r - 10, where r = category selection scores rater and 10 was the lowest score in the scoring category. Index V values ranged from 0-1. Criteria for determining validity were categorized by the coefficient of the V content validity index (Aiken, 1988). The categorization of the content validity index (V) is shown in Table 3.

Table 3. Expert Validity Category

Value of Validity Index (V)	Category
$V_{account} \leq 0.40$	Low
$0.40 < V_{account} \leq 0.80$	Moderate
$V_{account} > 0.80$	High

RESULTS AND DISCUSSION

The results and discussion of this study are divided into three parts: (1) The development of Conceptual Knowledge Test; (2) The Validation of Conceptual Knowledge Test; and (3) The Reliability of Conceptual Knowledge Test.

The Development of Conceptual Knowledge Test

The instrument of knowledge of electrical concept and magnetism was developed from several sources (Nugroho & Setiawan, 2009; Thohir et al., 2013). The initial number of items was 45 numbers in the form of multiple choice with the reasons. The comparison of electrical matter and magnetism item number was 1: 2. Material distribution and item number are shown in Table 4. Furthermore, all of the items on the CKT-EM test were validated. Instrument validation results eliminated 5 items; the item number 8, 16, 21, 30, and 33. The final design of this CKT-EM test produced 40 items.

Table 4. The Blue Print of CKT-EM

Materials	Indicators (Learning goals)	Number of Items
Electric current	Applying electric current concept in circuit	1, 2 3, 4, 5
Potential difference		6, 7, 8, 9
Electromotive Force	Applying electric charge conservation (electric current conservation) in circuit	10, 11 12, 13
Resistance	Applying potential difference in circuit	14, 15, 16
Electric Energy and power		17, 18 19, 20, 21
Direct current circuit	Applying electromotive force in a direct current circuit	22, 23, 24 25, 26 27, 28

Kirchhoff's Rules		29, 30
The force of charged particles moving in a magnetic field	Defining battery's potential difference when it is connected to lamp and other loads	31, 32, 33, 34, 35 36, 37
Magnetic force acting on a current-carrying conductor	Defining wire's size and temperature affects its resistance	38, 39, 40, 41 42
Coil carried a steady current	Applying the electric power concept in circuit	43, 44, 45

The Validation of CKT-EM

The process of test validation consisted of 3 stages: (1) the readability of instrument, (2) the expert validation, and (3) the validity of instrument

Readability test was performed by five second-grade physics prospective teacher who have passed Basic Physics course. There were 3 elements developed into 9 assessment aspects (Brookhart, 2010; Gronlund, 1985). The nine criteria were: (1) the use of brief, clear, and assertive action verbs or command; (2) a series of sentence questions and answer choices is a necessary statement only; (3) choice of logical answers in terms of material and in accordance with the command questions; (4) the length of choice of answers is relatively the same; (5) the use of language in accordance with Indonesian rules; (6) words are easily understood (not ambiguous); (7) the use of sentences/ communicative statements; (8) suitability of illustrations (in the form of tables, graphics, drawings, diagrams etc.); and (9) clear and legible illustrations of tables, graphs, drawings, diagrams or others. The results of the readability test showed that the average score for all items was in the range of 3.2 to 4.8 with a score ranging from 1 to 5. According to Table 2 related to the average interpretation of the test instrument assessment scores, the results of the readability test categorized as "very clear" and "clear". The summary outline of item classification based on the readability categories appears in Table 5.

Table 5. The Category of Instrument Readability Test

Rating scale	Category	Number of Items (%)	Judgment
≥ 4	Very clear	24 (53.33)	Without revision
$4 > X \geq 3$	Clear	21 (46.67)	Few revision
$3 > X \geq 2$	Fair	0	Full revision
$X < 2$	Not clear	0	Rejected

The result of students' responses to the readability test instrument could be classified into three elements as seen in Table 6.

Table 6. Students' Responses to the Readability Test

Response Aspects	Number of Test Items	Total of Test Items (%)
The use of unfamiliar words	33,34,35,41,42	5 (11.11)
Sentence ambiguity	12,14,16,23,26, 31,32,33,36,40, 41,43	12 (26.67)
The unclear illustration of picture, symbol, table, graphic,	19,21,23,27,33, 38,40,42,44,45	10 (22.22)

Table 6 shows that there are a number of questions that need to be revised both in terms of words, sentences, and illustrations. In addition, from this input, it can also be said that there are some issues that need to be revised in the mild revision category.

Instrument' validation data from three experts was analyzed with descriptively quantitative through the determination of content validity index using Aiken's Formula. The results of determining the content validity index of all items were further categorized based on their validity level referring to Table 3. The summary of the items' content validity level is presented in Table 7.

Table 7. The Category of Items' Content Validity

Validity index (V)	Category	Number of item (%)
$V_{count} \leq 0.40$	High	0
$0.40 < V_{count} \leq 0.80$	Moderate	42 (93.33)
$V_{count} > 0.80$	Low	3 (6.67)

Table 7 showed that item 16, 22 and 36 were in a low category. The low index of content validity (V) of the three items was then cross-checked with notes in the form of comments, suggestions and input validator. The results of re-checking on item 16 and 22 to the input of the three validators indicated that there were some errors in the process of composing the problem, namely: (1) less communicative sentences; (2) less precise use of words including at some points of the answer choice; and 3) discrepancies between items and indicators to be measured. Meanwhile, the error found in item 36 was a less obvious picture.

After validating the test instrument contents, instrument trials to some Physics prospective teacher as the empirical validation stage was carried out. The empirical validation results were then analyzed using the ITEMAN version 3.0 to determine the discriminating power, difficulty level, and distractor functionality of test. In addition, the need to test the instrument was to determine reliability coefficient value (α) of the test instrument showing the level of instrument' validity and reliability whether it was feasible to use or not.

A test's discriminating power is a measurement of items' effectivity in distinguishing between high and low scores of a test (Aiken, 1988). The quality of multiple-choice items could be determined by discriminating power coefficients. Nitco (1983) and Damayanti et al. (2018) established a decision rule of an item that is divided into three types based on the value of the distinguishing power coefficient; accepted, revised, and rejected. The three types of decisions are: (1) if the discriminating power coefficient is greater than 0.3, then the decision on the item is accepted; (2) The item is revised when the discriminating power coefficient is between 0.10 and 0.29; and (3) the item is rejected if the discriminating power coefficient is below 0.10. Interpretation of the discriminating power analysis data of the items was presented in Table 8.

Table 8. The Distribution of Items' Criterion based on Discriminating Power Index

Discriminating Power (p value)	Criterion	Number of Items (%)
> 0.3	Accepted	14 (30.11)
0.10 to 0.29	Revised	26 (58.78)
< 0.10	Rejected	5 (11.11)

Table 8 showed that the total number of items worthy of use as a test instrument was 90% and 10% of the items were not feasible in terms of items' discriminating power coefficient.

The difficulty level of items is generally defined by the term percentage or proportion that responds correctly (Anastasi & Urbina, 1997). In this case, the proportion of correct answers was the comparison between a numbers of test participants who answered correctly on the items analyzed compared to the total number of test participants. The main reason for using problem-level analysis of difficulty level was to select a number of items that have the appropriate degree of difficulty level (Aiken, 1988). The difficulty level category is determined by a coefficient (p) and divided into three categories: (1) difficult, if the value of coefficient p is greater than 0.3; (2) medium, if coefficient p is between 0.3 and 0.7; and (3) easy, if the p-value is greater than 0.7. The results of ITEMAN version 3.0 item analyses for the difficulty level of items showed the varying p values spread into the three categories of difficulty levels. Furthermore, the difficulty level category was presented in Table 9.

Table 9. The Distribution of Items based on Difficulty Level Index

Difficulty level (p)	Criterion	Number of Items (%)
p <0.3	Difficult	24 (54)
$0.3 \leq p \leq 0.7$	Medium	19 (42)
p > 0.7	Easy	2 (4)

One of the important elements in the analysis of multiple-choice item is the functionality distractor. The purpose of functional distractor analysis was to know whether the provided distractor made sense or not. The distractors should be reasonable so that it would probably be selected by at least 5% of all test takers (Chavda et al., 2015). Based on this statement, Mukherjee & Lahiri (2015) distinguished the effectivity of distractor level over two types of non-functional distractors; Non-Functional Distractors (NFDs) and Effective Distractor (ED) distractors. The distractors of a question item categorized as NFD if the number of selected distractors was <5% of the total number of test takers. Conversely, the distractors were categorized as ED when selected by 5% or more of test participants. Next, the researchers determined the efficiency distractor (distractor efficiency / DE) of each item based on the number of NFD items in %. If a given item contains three or more NFDs, then its DE value is 0%. Furthermore, if the item consists of two, one, or zero NDF, then DE is 33.3%, 66.6%, and 100% respectively. The distribution of item clustering based on the characteristics of the distractors was shown in Table 10.

Table 10. Item Classification based on Functional Distractor Level

Parameter	Number of Items' NDF			
	0	1	2	≥ 3
Total (%)	27 (60)	10 (22.22)	5 (11.11)	3 (6.67)
DE (%)	100	66.60	33.30	0

Table 10 shows that the number of questions that all the answer choices worked well there was 60% of the 45 items.

The Reliability of CKT-EM

One of the important aspect considered in a measuring instrument is the test reliability, which refers to the consistency of scores obtained by the same test participants when re--tested by the same test in different situations or from one measurement to another (Anastasi & Urbina, 1997; Thorndike, 1971). Therefore, the reliability level of a test or measuring instrument is a very important consideration.

The determination of the reliability coefficients of this test employed the ITEMAN version 3.0 program. The results of test instrument reliability analysis showed that the coefficient value of Alpha (α) was 0.56 from 60 prospective physics teachers. The value of the test reliability coefficient was 0.56, which according to Remmrers et al. (1965), it was applicable for research purposes. The value of the reliability coefficients interprets that the instrument was acceptable. Therefore, the CKT-EM test instrument was reliable to measure students' conceptual knowledge.

The revised CKT-EM instruments were further tested to 215 prospective physics teachers. The result of a larger application revealed that the coefficient of Alpha was 0.87. This value is sufficient based on a report by Thorndike (1971) that alpha coefficients in the range 0.58-0.81 indicated that the instrument had satisfactory reliability for scales containing five items each.

CONCLUSION

Research on the development and validation of the concept knowledge test instrument on electrical and magnetism topic has produced 40 items. The test instrument validation stages ranged from validation of instrument legibility, content validity, and test instrument testing have resulted in valid and reliable instruments with a reliability coefficient of 0.87. Therefore, it concluded that the concept of a knowledge test instrument on electricity and magnetism topics was feasible to use.

REFERENCES

Aiken, L. R. (1988). *Psychological Testing and Assessment* (Sixth Edit). United State of America: Allyn and Bacon, Inc.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (Seventh Ed). United State of America: Prentice-Hall, Inc.

Boorman, E. D., & Rushworth, M. F. S. (2009). Conceptual Representation and the Making of New Decisions. *Neuron*, *63*(6), 721–723.

Brookhart, S. M. (2010). *How to Assess Higher-Order Thinking Skills in Your Calssroom*. Virginia USA: ASCD Alexandria.

Chavda, P., Misra, S., & Duttaroy, B. (2015). Item Analysis of Multiple Choice Questions based Undergraduate Assessment in Community Medicine. *South East Asian Journal of Medical Education*,*9*(1), 66–69.

Clifton, S. L., & Schriner, C. L. (2010). Assessing the quality of multiple-choice test items. *Nurse Educator*, *35*(1), 12-16.

Damayanti, A., Wennyta, W., & Munawwaroh, K. (2018). An Analysis on The Items Difficulty Level of English Semester Test at The Tenth Grade Students of SMAN 3 Jambi City Academic Year 2016/2017. *JELT: Journal of English Language Teaching*, *2*(1), 141-152.

Dega, B. G., Kriek, J., & Mogese, T. F. (2013). Students' Conceptual Change in Electricity and Magnetism Using Simulations: A Comparison of Cognitive Perturbation and Cognitive Conflict. *Journal of Research in Science Teaching*, *50*(6), 677–698.

Ding, L. (2014). Seeking Missing Pieces in Science Concept Assessments: Reevaluating the Brief Electricity and Magnetism Assessment through Rasch Analysis. *Physical Review Special Topics - Physics Education Research*, *10*(1), 1–15.

Duschl, R. (2008). Science Education in Three-Part Harmony: Balancing Conceptual, Epistemic, and Social Learning Goals. *Review of Research in Education*, *32*(1), 268–291.

Erdogan, M., Özel, M., Uşak, M., & Prokop, P. (2009). Development and Validation of An Instrument to Measure University Students' Biotechnology Attitude. *Journal of Science Education and Technology*, *18*(3), 255–264.

Fang, S. C., Hsu, Y. S., Chang, H. Y., Chang, W. H., Wu, H. K., & Chen, C. M. (2016). Investigating The Effects of Structured and Guided Inquiry on Students' Development of Conceptual Knowledge and Inquiry Abilities: A Case Study In Taiwan. *International Journal of Science Education*, *38*(12), 1945-1971.

Gok, T. (2012). The Impact of Peer Instruction on College Students'beliefs about Physics and Conceptual Understanding of Electricity and Magnetism. *International Journal of Science and Mathematics Education*, *10*(2), 417-436.

Gronlund, N. E. (1985). *Measurement and Evaluation in Teaching*. United Kingdom: Macmillan College.

Koponen, I. T., & Nousiainen, M. (2018). Concept Networks of Students' Knowledge of Relationships between Physics Concepts: Finding Key Concepts and Their Epistemic Support. *Applied Network Science*, *3*(1), 14-28.

McColgan, M. W., Finn, R. A., Broder, D. L., & Hassel, G. E. (2017). Assessing Students' Conceptual Knowledge of Electricity and Magnetism. *Physical Review Physics Education Research*, *13*(2), 020121-020132.

Mehrens, W. & Lehmann, I. (1991). *Measurement and Evaluation in Education and Psychology* (Fourth Edi). United State of America: Wadsworth Publishing Company.

Mukherjee, P., & Lahiri, S. K. (2015). Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from An Assessment In A Medical College of Kolkata, West Bengal. *IOSR J Dent Med Sci*, *1*, 47-52.

Nitco, J. (1983). *Educational Test and Measurement An Introduction*. New York: Harcourt Brace Jovanovich, Inc.

Nugroho, S. E., & Setiawan, A. (2009). Analisis Kognitif Konsepsi Medan Listrik dan Magnetik melalui Respon Jawaban Spontan Mahasiswa Calon Guru. *Jurnal Pendidikan Matematika dan Sains*, *14*(1), 31-36.

Pollock, S. J. (2009). Longitudinal Study of Student Conceptual Understanding in Electricity and Magnetism. *Physical Review Special Topics-Physics Education Research*, *5*(2), 020110-020122

Rustaman, N. Y., Hamidah, I., & Rusdiana, D. (2017, February). The Use of Classroom Assessment to Explore Problem Solving Skills Based on Pre-Service Teachers' Cognitive Style Dimension in Basic Physics Course. In *Journal of Physics: Conference Series* (Vol. 812, No. 1, p. 012047). IOP Publishing.

Remmers, H. H., Gage, N. L., & Rummel, J. F. (1965). *A practical introduction to measurement and evaluation*. New York: Harper & Row.

Rittle-Johnson, B., & Star, J. R. (2009). Compared With What? The Effects of Different Comparisons on Conceptual Knowledge and Procedural Flexibility for Equation Solving. *Journal of Educational Psychology*, *101*(3), 529-544.

Sadaghiani, H. R. (2011). Using multimedia learning modules in a hybrid-online course in electricity and magnetism. *Physical Review Special Topics-Physics Education Research*, *7*(1), 010102-010114

Secolsky, C., & Denison, D. B. (Eds.). (2012). *Handbook on Measurement, Assessment, and Evaluation in Higher Education.* United State of America: Routledge.

Stelzer, T., Brookes, D. T., Gladding, G., & Mestre, J. P. (2010). Impact of Multimedia Learning Modules on An Introductory Course on Electricity and Magnetism. *American Journal of Physics*, *78*(7), 755-759.

Streveler, R. A., Litzinger, T. A., Miller, R. L., & Steif, P. S. (2008). Learning Conceptual Knowledge In The Engineering Sciences: Overview and Future Research Directions. *Journal of Engineering Education*, *97*(3), 279-294.

Svetina, D., & Levy, R. (2014). A framework for dimensionality assessment for multidimensional item response models. *Educational Assessment*, *19*(1), 35-57.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An Assessment of Functioning and Non-Functioning Distractors in Multiple-Choice Questions: A Descriptive Analysis. *BMC Medical Education*, *9*(1), 40-48.

Thohir, M. A., Wasis, W., & Sugimin, W. W. (2017). Peningkatan Keterampilan Berpikir Kritis melalui Pembelajaran Metode Penemuan Terbimbing dalam Upaya Remediasi Miskonsepsi Materi Listrik Dinamis. *JPPS (Jurnal Penelitian Pendidikan Sains)*, *1*(2), 62-67.

Thorndike, R. L., & Angoff, W. H. (1971). *Educational measurement*. United State of America: American Council on Education.

Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2017). 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*(4), 663-682.