International Journal of Instruction *January 2021 ● Vol.14, No.1*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp. 555-566

Article submission code: 20191018164240
Received: 18/10/2019 Revision: 11/07/2020
Accepted: 02/08/2020 OnlineFirst: 31/10/2020

Development of Two-Tier Multiple Choice Test to Assess Indonesian Elementary Students' Higher-Order Thinking Skills

Peduk Rintayati
Dr., Universitas Sebelas Maret, Indonesia, *pedukrintayati@staff.uns.ac.id*

Hafizhah Lukitasari
Sekolah Dasar Negeri Combongan 01, Indonesia, *hafizhahlukitasari@gmail.com*

Ahmad Syawaludin
Universitas Sebelas Maret, Indonesia, *syawaluns@gmail.com*

Assessment of higher-order thinking skills (HOTS) provides few opportunities for students to develop more in-depth knowledge, serving students' ability to identify and solve their problems. One type of instrument for measuring HOTS objectively is the two-tier multiple-choice test (TTMCT). This research is part of the research and development stage, which is the implementation phase of the TTMCT instrument as a developed product. The objective of the study was to determine the feasibility of the TTMCT instrument developed to assess Indonesian elementary students' higher-order thinking skills (HOTS) on learning science in the concept of force, motion, and energy This study used to research and development method with collection technique using a questionnaire, test, and interview. The subjects in this research were 227 students, 65 teachers, 5 principals, 8 lectures. The results of the study show that TTMCT worthy development results were 25 items, had good content validity, "very high" reliability, "quite difficult" level of difficulty, and "very good" difference power. TTMCT applies to Indonesian Elementary School with high and medium criteria. This study concludes that the TTMCT instrument developed is considered feasible by senior teachers and users to measure higher-order thinking skills in the concept of force, motion, and energy in elementary schools.

Keywords: higher-order thinking skills, two-tier multiple-choice test, thinking skills, elementary school, students

INTRODUCTION

The Competency Standards of Elementary School Graduates suggests that each student is expected to build and apply information or knowledge in logical, critical, creative, and innovative ways; demonstrate the ability to think logically, critically, creatively, and innovatively in decision making; as well as demonstrate the ability to analyze and solve complex problems (Peraturan Menteri Pendidikan Nasional No. 54 of 2013). These

Citation: Rintayati, P., Lukitasari, H., & Syawaludin, A. (2021). Development of Two-Tier Multiple Choice Test to Assess Indonesian Elementary Students' Higher-Order Thinking Skills. *International Journal of Instruction*, *14*(1), 555-566. https://doi.org/10.29333/iji.2021.14133a

competencies refer to higher-order thinking skills (HOTS) which is seen as students' reading, writing, speaking and listening skills; in addition to increase the likelihood of providing better reasons in all subjects; support correct decision-making and problem-solving; establish critical analysis and conclude and assess students' emotions; as well as help students to making smart choices in relationships with other fellow humans (Schraw & Robinson, 2011; Brookhart, 2010; O'Dowd's, 2007). Therefore, high-order thinking skills in schools is needed to give student competence in seeking for an alternative resolution to the problems faced.

Higher-order thinking (HOTS) is a cluster of elaborative mental activities requiring nuanced judgment and analysis of complex situations according to multiple criteria or find a possible answer in a perplexing situation (Widana, 2017; Lewis & Smith, 1993; Resnick, 1987). HOTS include critical, logical, reflective, metacognitive, and creative thinking. Brookhart (2010) classified HOTS into three contexts of understanding, includes (1) higher-order thinking as a transfer (students can apply their knowledge and skills which they can further develop into a new context); (2) higher-order thinking as critical thinking (express self-reasoning, responding, and decision making without teacher's intervention); and (3) higher-order thinking as a problem solving (serving students' ability to identify and solve their problems in the work and daily lives).

The concept of higher-order thinking skills derived from Bloom's Taxonomy (1956). There are six orders of Bloom's Taxonomy, consist of memorizing (C1), understanding (C2), applying (C3), analyzing (C4), evaluating (C5), and creating (C6) (Afandi & Sajidan, 2017). Bloom's Taxonomy classified thinking skills into higher and lower-order thinking skills. Memorizing, understanding, and applying as lower-order thinking skills and analyzing, evaluating, and creating as higher-order thinking skills (Wilson, 2016; Anderson et al., 2010; Airasian & Miranda, 2002). Feedback and assessment of the learning process and the existing formative assessments provide few opportunities for students to develop more in-depth knowledge (Cullinane & Liston, 2011; Limbach & Waugh, 2010). The development of formative assessment alternatives is needed to help students empowering higher-order thinking skills.

Evaluation is a systematic process determining the extent to which instructional objectives are achieved by students which reflect procedures for obtaining information on student learning (Mallett, 2015; Mardapi, 2012; Purwanto, 2010; Miller et al, 2009). Teachers should be able to choose appropriate assessment procedures to make learning decisions and use assessment results to make educational decisions (Kusaeri & Suprananto, 2012). Thus, the assessment should be well implemented, as an assessment is a major component of student personal development for personal students and classroom. Assessment of higher-order thinking skills can improve students' achievement and motivation (Brookhart, 2010).

The selected response and short answer are one of a method in learning assessment. In this test type, students choose the most correct answer among the already provided alternatives (Butler, 2018; Stiggins et al, 2004, Thorndike & Hagen, 1977). The selected-response assessment uses a scoring technique that calculates the proportion of right and wrong answers to learners. In this study, the type of assessment developed is

the selected response assessment, where this assessment has an objective nature. The multiple-choice item is one of the most widely applicable test items for measuring achievement (Linn & Gronlund, 2000). The multiple-choice test is comprehensive, objective scoring and easy checking, in addition to high item reliability, it can measure different levels of ability including higher-order thinking skills, the type of item can be arranged in such way that requires the ability of the test participants to distinguish different index of truth at once, it grains difficulty level that can be set by simply changing the homogeneity of alternative answers, and the information related to students' thinking skills can be more translated for teachers.

Assessment procedures must provide opportunities to students and teachers to engage in discussions on the assessed works (Cullinane & Liston, 2011). An alternative assessment that can be developed is a modified multiple-choice question form of a two-tier multiple-choice test. Two-tier multiple-choice test (TTMCT) is modifications of multiple-choice form which belongs to a kind of objective test. TTMCT developed consists of two levels of questions, the first tier is the content of the main question or item that has two answer choices and the second tier is the reason for the answer given based on the first tier. The existence of reasons at the second tier aims to improve thinking skills and see students' ability to reason (Bayrak, 2013; Cullinane & Liston, 2011; Treagust, 2006). TTMCT was applicable as an alternative formative assessment, to assess students' understanding, asking the student to use higher-order thinking skills in giving reasons in the second tier, and identify misconception that students may have, and (Adodo, 2013; Sesli & Kara, 2012; Cullinane & Liston, 2011; Treagust, 2006; Sampson, 2006). TTMCT can be used as an insight into making a form of assessment that challenges students' knowledge, providing a technique to assess students' concepts, especially in classroom learning.

The observation results in Indonesian elementary schools indicated that most teachers have shared positive perceptions and being aware of the importance of higher-order thinking skills in Elementary Schools, however, it remains difficult for teachers in establishing assessment instruments that apply to measure students' HOTS. Among the difficulties faced by teachers are: 1) the difficulty in developing assessment which not only measures lower-order thinking skills but also the higher-order thinking skills; 2) higher-order thinking skills would be better measured using objective tests, such as multiple-choice tests; 3) it is most often found that teachers use multiple-choice test, however, they also realize on the difficulty in establishing distractors or effective deceiving tests; 4) the items of the multiple-choice test also has a limit. It is not able to distinguish which students answer earnestly involves higher-order thinking skills and which students answer based on guesswork. The data found in the field showed the higher-order thinking skills test is quite rarely found in the teachers' test items bank, both formative and summative assessments.

For kind of elaborated matter, this study attempted to implement higher-order thinking skills in two-tier multiple-choice test forms. The objective of the study was to determine the feasibility of the TTMCT instrument developed to assess Indonesian elementary

students' higher-order thinking skills (HOTS) on learning science in the concept of force, motion, and energy.

METHOD

Research design

This research is a research and development stage, which is the implementation phase of the TTMCT instrument as a developed product. The objective was to determine the feasibility of the TTMCT instrument developed to assess Indonesian elementary students' higher-order thinking skills (HOTS) on learning science in the concept of force, motion, and energy. This study used to research and development method with a collection technique using questionnaires, tests, interviews, and observation. The subjects in this research were 227 students, 65 teachers, 5 principals, 8 lecturers. The data of the research is analyzed quantitatively and qualitatively (Arikunto, 2016; Cresswell, 2012). Quantitative data were collected by the survey results of the assessment needs of 49 teachers and the results of the validity, reliability, and item analysis of the two-tier multiple-choice test instrument involving 227 students. Qualitative data about the implementation of a two-tier multiple-choice test was obtained from an interview with 16 teachers and 5 principals and also observations in 5 different Elementary Schools in Purbalingga Regency, Central Java, Indonesia. The schools involved are schools that are categorized as high, medium, and low criteria.

Data collection techniques used questionnaires, tests, and interviews. The questionnaire was conducted to find out the needs and problems in the assessment of HOTS in elementary schools. Questionnaire results are used to investigate the testing and feasibility of TTMCT. A test used to find out the TTMCT validity, reliability, and item analysis. The validity test used was the content validity test based on Aikens' Formula involving lecturers as experts in Primary Education. The test aimed to analyze the ability of each item question in the two-tier multiple-choice test to measure indicators of higher-order thinking skills. Reliability testing and item analysis were analyzed by inputting the data of students' work into Iteman 3.0. Interviews were conducted involving teachers and principals, involved consisted of 16 Elementary Schools teachers who had at least 10 years of teaching experience in a higher class, and a principal who has a master education degree in education. The purpose of in-depth interviews with teachers and principals was to find out about the teachers' responses to the practical implications of implementing two-tier multiple-choice tests. Qualitative data analysis uses interactive analysis with data collection step, data reduction, display data, and conclusion (Miles & Huberman, 1994).

TTMCT scoring is not too much different from scoring on the multiple-choice test which refers to the correct answer in the first tier and correct answer in the second tier (Adesoji & Omilani, 2012). TTMCT scoring in this study refer to Yamtinah (2015) can be seen in Table 1.

Table 1
Scoring two-tier multiple-choice test

1st Tier (Answer)	2nd Tier (Reason)	Score
Correct	Correct	3
Correct	Incorrect	2
Incorrect	Correct	1
Incorrect	Incorrect	0

The scoring based on table 1 provides a different scoring process of the students who answered incorrectly on the first tier but answered correctly on the second tier and answered wrongly on both tiers. The item analysis is performed using Microsoft Excel 2010 and Iteman 3.0 software.

FINDINGS

The development of the two-tier multiple-choice test used in this study aims to measure the success of achieving cognitive indicators in higher-order thinking skills developed by Anderson et al. (2001) and Airasian & Miranda (2002) that covering the skills of analyzing, evaluating, and creating. Stem writing questions adapted to operational verbs that represent cognitive higher-order thinking skills. The reasoning as a second tier of the question is given directly under the question.

The two-tier multiple-choice test instruments that had been developed consisted of 25 items covering 3 competency indicators for 5th-grade students at elementary school such as identifying the gravity force, frictional force, and magnetic force and its utilization in everyday life; analyzing the relationship between force, motion, and energy; formulating problem-solving related to gravity force, friction force, and magnetic force or simple machine. Then, the three indicators of competence were described into 25 indicators of test forms which combine operational verbs of higher-order thinking skills as follows: identifying, analyzing, describing, and defining features (C4); criticizing, clarifying, and interpreting (C5); making a generalization, connecting hypothesis, predicting, and proposing hypothesis (C6).

The developed two-tier multiple-choice test instrument was tested using a content validity test by eight validators consisting of five Lecturers in the Elementary School Teacher Education Study Program and three teachers who have had more than ten years of teaching experience. The content validity procedure used in this analysis was the Aiken's V formula for calculating content coefficient validity based on the results of validator assessment on each item to know how exactly the item describes the indicator being measured (Azwar, 2012). On the table of coefficient validity with 8 validators and 4 rating scale, the item was valid if its coefficient validity is $\geq 0{,}75$. Test results in Table 2 showed 17 valid items without having required revision while 8 items were valid after the revision.

Table 2
Expert validation uses the content validity formula

Items	Validators' assessments								V	Description
	1	2	3	4	5	6	7	8		
1	3	4	4	3	3	3	3	3	0,75	valid after revision
2	3	3	4	2	4	4	3	4	0,79	valid after revision
3	3	4	4	3	4	3	3	3	0,79	valid
4	3	4	4	3	4	4	4	4	0,92	valid
5	3	4	4	2	4	4	4	4	0,88	valid after revision
6	3	4	4	3	3	3	3	3	0,75	valid
7	3	3	2	3	4	4	3	4	0,75	valid
8	4	4	4	3	4	4	4	3	0,92	valid
9	4	4	3	3	4	3	3	3	0,79	valid
10	4	4	3	3	4	4	4	3	0,88	valid
11	3	4	3	3	4	4	4	3	0,83	valid
12	3	4	3	2	4	4	4	4	0,88	valid after revision
13	4	3	3	3	4	4	3	3	0,79	valid
14	4	1	4	3	4	4	4	3	0,75	valid
15	3	4	4	3	4	4	4	3	0,88	valid after revision
16	3	4	4	2	4	4	4	4	0,88	valid
17	3	4	4	3	3	3	3	3	0,75	valid
18	4	3	4	4	3	3	3	3	0,79	valid
19	3	3	3	3	4	3	4	3	0,75	valid after revision
20	4	3	3	3	4	3	3	4	0,79	valid after revision
21	3	4	3	3	4	4	4	3	0,83	valid
22	3	4	4	3	4	4	4	3	0,88	valid
23	4	4	4	2	4	3	3	3	0,79	valid
24	4	4	4	3	4	4	4	3	0,92	valid
25	3	3	4	3	4	3	3	3	0,75	valid after revision

The two-tier multiple-choice test tool analyzed the suitability of its implementation separately based on students' ability level using Item 3.0. out of 100 students who took on the two-tier multiple-choice test in operational field testing, 30 students came from predefined "high ability" Elementary School, 40 students from predefined "medium-ability" elementary school, and 30 students from predefined "low ability" elementary school. School predicate was determined based on accreditation and National Examination rank in Purbalingga Regency in the academic year of 2016/2017. It aimed to identify the two-tier multiple-choice test whether or not it applied to schools with appropriate categories.

Table 3
Analysis result items based on the elementary school's criteria

Analysis		School Criteria					
		High-Ability		Medium-Ability		Low-Ability	
		Result	Interpretation	Result	Interpretation	Result	Interpretation
Reliability	Answer (1st tier)	0,892	Reliable	0,788	Reliable	0,527	Not Reliable
	Reason (2nd tier)	0,903	Reliable	0,801	Reliable	0,709	Not Reliable
Mean P	Answer (1st tier)	0,428	Fairly Difficult	0,369	Fairly Difficult	0,198	Difficult
	Reason (2nd tier)	0,492	Fairly Difficult	0,449	Fairly Difficult	0,264	Fairly Difficult
Mean Item-Tot	Answer (1st tier)	0,525	Excellent	0,380	Good	0,282	Fairly enough
	Reason (2nd tier)	0,542	Excellent	0,564	Excellent	0,344	Good

The reliability coefficient, difficulty index (Mean P), and determination power (Mean Item-Tot) were analyzed separately based on school criteria and it gets different and significant results. The test instrument is reliable when applied to students in schools with "high" and "medium" criteria, while for schools with "low" criteria it is "unreliable" to apply. Judging from the mean of the difficulty index (Mean P) value, for schools with "high" and "medium" criteria, the test instrument has a "Fairly Difficult" interpretation, while for "low" schools criteria, the test instrument has a "difficult" interpretation for the answer and "quite difficult" for the reason. Based on Mean Item-Tot (Determination Index) two-tier multiple-choice test has an "Excellent" interpretation, however, on different schools with "low" criteria it has a "Fairly Good" interpretation. Table 4 provides information that the implementation of the two-tier multiple-choice test instrument can be implemented in schools with "medium" and "high" criteria.

Table 4
The recommendation of two-tier multiple-choice test implementation

No.	School Criteria	Feasibility (Implementation)			
		Reliability	Difficulty Index	Different Strength	Conclusion
1.	High	√	√	√	applicable
2.	Medium	√	√	√	applicable
3.	Low	-	-	√	inapplicable

Teacher's responses which were gained through the interview was related to the two-tier multiple-choice test instrument were identifiable as follows: 1) Teachers were well-received development of two-tier multiple-choice test instruments although some adjustments were still needed; 2) developing questions which measure the high-level thinking skills, it might lead to providing assessment for students' concepts of understanding and identifying students' learning meaningfulness as well as their ability

to relate the received materials according to the contextual surrounding environment; 3) being a more functional measuring alternative of multiple choice in general since it reduced students 'chance in guessing the tests answers and simultaneously measured their ability in understanding the concept and its relation to the environment; 4) applicable through taking into account the learning methods given to the students, the scope of material, and the students' sufficient average ability if only requiring optimal results.

DISCUSSION

Two-tier multiple-choice questions are modifications of the form of multiple-choice tests usually grouped into objective test types. The multilevel multiple-choice test form used in research development evaluation instruments was adapted from Treagust (2006). The form of the developed test consists of two levels of questions, the first level is the contents of the questions or main questions which have two choices of answers and the second level is the reason for the answers given based on the first choice. The inclusion of reason at the second level aims to improve thinking skills and see students' abilities in giving reasons (Cullinane & Liston, 2011).

The results of the experts' assessment showed that the TTMCT instrument developed was good in terms of language, material, and construction with several revisions. Revisions made include correcting the concepts that were still improper, improving the suitability of Bloom's taxonomy in the problem, improving the relationship between the problem stems with the answer reasons, fixing the answer keys to the questions that were still improper, paying attention to the allocation of work time, simplifying the writing of the questions, and correcting writing errors.

The implementation of TTMCT in this study yielded several findings. The TTMCT instrument developed was able to measure higher-order thinking skills because the fulfillment of characteristics including the development of TTMCT rests on indicators of high-level thinking skills from Anderson et al. (2001) including the skills of analyzing, evaluating, and creating; valid questions in content as well as language, material, and construction; have reliability with a minimal interpretation of "good"; has a difficulty level with an average interpretation of "quite difficult"; and has a minimum power difference of "very good". The use of this instrument provides a stimulus or stimulation to students to think at a high level, the response to the stimulus is the response given by students by selecting the available options. The advantages of two-tier multiple-choice test among others to measure the level of high-order thinking skills (analysis, evaluation, and creation) which are commonly difficult carried out by common double choice (Cullinane & Liston, 2011; Tuysuz, 2009; Haladyna & Downing, 1989; Treagust, 2006); scoring becomes easy, fast, and objective, besides, to apply to determine the teachers' learning effectiveness (Cullinane &Liston, 2011); applicable to measuring both problem-solving skills and critical thinking (O'Dowd, 2007); applicable to diagnose material understanding and detect possible misconceptions that students can make (Cullinane & Liston, 2011; Sampson, 2006).

Science learning teaches students to understand the natural environment scientifically. One of the basic competencies of science learning that must be achieved by 5th-grade

students is to describe the relationship of force, motion, and energy (gravitational force, friction force, magnetic force). Thus, an effective test instrument is needed to measure students' thinking abilities in the study of force, motion, and energy. TTMCT is an alternative test instrument because it presents multiple-choice tests within a short period that can effectively cover broad material and numerous test items (Susetyo, 2015; Adodo, 2013). It can effectively measure student understanding and various types of complex learning outcomes (Bayrak, 2013; Linn & Gronlund, 2000). In this study, TTMCT facilitates students to identify gravitational, frictional, and magnetic forces as well as their use in daily life, analyze the relationships between forces, motion, and energy, and formulate problem-solving related to gravity, friction, and magnetic force or simple machine.

Once it is implemented in Elementary Schools, the two-tier multiple-choice test provides several practical impacts and implications, including: 1) the assessment results using two-tier multiple-choice test applicable as material for evaluation and follow-up in an individual profile form of mapping high-ability thinking skills from several activities such as analyzing, evaluating, and creating; 2) applying two-tier multiple-choice test can be a simple representation of the meaningful teaching and learning processes within a classroom as well as showing undergoing learning effectiveness. The representation will be more illustrated if the test instrument is used as a training medium to streamline the learning objectives; 3) a two-tier multiple-choice test applicable as a basis for developing schools' teaching materials which encompass what appropriate teaching materials should be provided to student's needs, what materials should be deepened, not only at the knowledge level but also at technical mastery level; 4) the result of a two-tier multiple-choice test is one of the alternatives in collecting data based on follow-up planning in solving efforts form according to the already identified problems or difficulties after applying test instruments, thus at least it not only measures students' high-order thinking skills but also a diagnostic test of learning difficulties and misconceptions towards a concept.

CONCLUSIONS

The conclusions are the results of using TTMCT can be presented in the individual profiles from regarding students' thinking skills mapping at the level of analyzing, evaluating, and creating. The results of the study show that TTMCT worthy development results amounted to 25 items, has good content validity, "very high" reliability, "quite difficult" level of difficulty, and "very good" difference power. TTMCT applies to Indonesian Elementary School with high and medium criteria. This study concludes that the TTMCT instrument developed is considered feasible by senior teachers and users to measure higher-order thinking skills in the concept of force, motion, and energy in elementary schools. Technically this approach requires further guidance and development.

Learning practice at the Indonesian elementary schools' level, both students and teachers have not yet accustomed to use a two-tier multiple-choice test assessment. The two-tier multiple-choice test implementation needs being well-prepared since at the stage of preparing instructional planning, learning indicators and assessment indicators, model

selection and learning methods, and formative authentic assessment which apply to stimulate students' thinking skills. Teachers were not accustomed yet to compiling and using two-tier multiple-choice tests, therefore, guidance, development, and direction were required if a similar assessment needede to be applied.

REFERENCES

Adesoji, & Omilani. (2012). A Comparison of Secondary Schools Students Levels of Conception of Qualitative and Qualitative Inorganic Analysis. *American Journal of Scientific and Industrial Research*, *3*(2), 56-61.

Adodo, S. O. (2013). Effects of two-tier multiple choice diagnostic assessment items on students' learning outcome in basic science technology (BST). *Academic Journal of Interdisciplinary Studies*, *2*(2), 201-201.

Afandi, & Sajidan. (2017). *Stimulasi Keterampilan Berpikir Kritis*. Surakarta: UNS Press.

Airasian, P., & Miranda, H. (2002). The Role Assessment in The Revised Taxonomy. *Theory Into Practice, 41*(4), 249-254.

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives (Abridged ed.).* New York: Longman.

Arikunto, S. (2016). *Prosedur Penelitian Suatu Pendekatan Praktik*. Jakarta: Rineka Cipta.

Bayrak, B. K. (2013). Using Two-Tier Test to Identify Primary Students' Conceptual Understanding and Alternative Conceptions in Acid Base. *Online Submission*, *3*(2), 19-26.

Bloom, B. S. (1956). *Taxonomy of Educational Objective: The Classification of Educational Goals Handbook Cognitive Domain.* London: Longmans Inc.

Butler, A. C. (2018). Multiple-Choice Testing in Education: Are the Best Practices for Assessment Also Good for Learning? *Journal of Applied Research in Memory and Cognition*, *7*(3), 323-331.

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

Cresswell, J. W. (2012). *Research Design: Qualitative, Quantitative, and Mixed Methods.* Yogyakarta: Pustaka Pelajar.

Cullinane, A., & Liston, M. (2011). *Two-tier Multiple Choice Question: An Alternative Method of Formatif Assessment for First Year Undergraduate Biology Students*. Linmark: National Center for Excellence In Mathematics and Education Science Teaching and Learning (NCE-MSTL).

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37-50.

Kusaeri & Suprananto. (2012). *Education Measurement and Evaluation.* Yogyakarta: Graha Ilmu.

Lewis, A., & Smith, D. (1993). Defining Higher-Order Thinking. *Theory Into Practice, 32*(3), 131-137.

Limbach, B. & Waugh, W. (2010). Developing Higher-Order Thinking. *Journal of Instructional Paedagogies, 3*(1), 1-9.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching (8th ed.)*. Upper Saddle River, N J: Prentice-Hall.

Mallett, C. A. (2015). *The school-to-prison pipeline: A comprehensive assessment*. Springer Publishing Company.

Mardapi, D. (2012). *Education Measurement, Assessment, and Evaluation.* Yogyakarta: Nuha Medika.

Miles, M.B., & Huberman, A.M. (1994). *Qualitative data analysis.* Newbury Park, CA: Sage.

Miller, M. D., Linn, R., & Gronlund, N. (2009). *Measurement and Evaluation in Teaching (12th ed.).* New York: Merrill Press Edu Inc.

O'Dowd, G. V. (2007). Developing Higher-Order Thinking Skills in Medical Students. *Hamamatsu University School of Medicine*, 21, 21-33.

Peraturan Menteri Pendidikan Nasional No 54 Tahun 2013 tentang Standar Kompetensi Lulusan Pendidikan Dasar dan Pendidikan Menengah.

Purwanto, N. (2010). *Prinsip-prinsip dan Teknik Evaluasi Pembelajaran*. Bandung: Remaja Rosdakarya.

Resnick, L. B. (1987). *Education and Learning to Think*. Washington DC: National Academy Press.

Sampson, V. (2006). *Two-tier Assessment*: *Teacher Toolkit*. Arizona: College of Education at Arizona State University in Tempe.

Schraw, G., & Robinson, D. H. (Eds.). (2011). *Assessment of Higher Order Thinking Skills*. IAP.

Sesli, E., & Kara, Y. (2012). Development and application of a two-tier multiple-choice diagnostic test for high school students' understanding of cell division and reproduction. *Journal of Biological Education*, *46*(4), 214-225.

Stiggins, R.J, Arter, A., Chappuis, J., Chappuis S. (2004). *Classroom Assessment for Student Learning: Doing it Right -- Using it Well.* Assessment Training Institute.

Susetyo, B. (2015). *Prosedur Penyusunan dan Analisis Tes untuk Penilaian Hasil Belajar Bidang Kognitif.* Bandung: Refika Aditama.

Thorndike, L. R., & Hagen, E. P. (1977). *Measurement and Evaluation in Psychology and Education.* John Wiley Publication.

Treagust, D. F. (2006). Diagnostic Assessment in Science as A Means for Improving Teaching, Learning, and Retention. *Invited Speaker Paper in Universe Science Assessment Symposium Proceedings,* 1-12. The University of Sydney, 28 September 2006.

Tuysuz, C. (2009). Development of Two-Tier Diagnostic Instrument and Asess Students Understanding in Chemistry. *Scientific Research and Essay, 4*(6), 626-631.

Widana, I. W. (2017). Higher order thinking skills assessment (HOTS). *JISAE*, *3*(1), 32-44.

Wilson, L. O. (2016). Anderson and Krathwohl–Bloom's taxonomy revised. *Understanding the New Version of Bloom's Taxonomy*.

Yamtinah, S. (2015). Development of Diagnostic Instruments Learning Difficulties in Chemistry Learning in Higher-Education. *Jurnal Penelitian dan Evaluasi Pendidikan, 19*(1), 69-91.