SCRIPT JOURNAL
Journal of Linguistic and English Teaching
http://jurnal.fkip-uwgm.ac.id/index.php/Script
P-ISSN: 2477-1880; E-ISSN: 2502-6623
October 2023, Vol. 8 No. 2

Problems Faced by English Teacher Candidates in Developing Test Kits for Assessing Students' Learning

Ari Purnawan[1], Asfar Arif Nurharjanto[2], and Annisa Nurul Ilmi[3]
Faculty of Languages, Arts, and Culture, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
Email Correspondence: ari_purnawan@uny.ac.id

Abstract

Background:
Learning assessment has been left behind compared to other components of teaching such as instructional methods or media development. This paper aims to describe problems faced by English teacher candidates studying at the English Education Department of a university in Java, Indonesia in developing test kits for assessing students' learning

Methodology:
This study is qualitative in nature. The investigation involved 28 teacher candidates taking two sequential courses on testing namely English Learning Assessment and English Language Learning Test Development who were then asked to write a set of English learning achievement tests as their end-of-class mini project. A total of 1052 multiple choice questions were analyzed by three investigators to reveal the teacher candidates' problems in developing test kits. Problems are reported and described quantitatively.

Findings:
The study reveals that the students face difficulties in formulating test kit formats, stems, options, and texts.

Conclusion:
Recommendations include education for writing test questions and improving or revising the learning syllabus and materials of the above-mentioned courses on language learning assessment

Originality:
The number of questions or participants can be improved as well as the scope of analysis to make the study more reliable.

Key words: learning assessment; teacher candidate; test kits

DOI	:	https://doi.org/10.24903/sj.v8i2.1441
Received	:	September 2023
Accepted	:	October 2023
Published	:	October 2023
How to cite this article (APA)	:	Purnawan, A., Nurharjanto, A, A., & Ilmi, N, A,. (2023). Problems Faced by English Teacher Candidates in Developing Test Kits for Assessing Students' Learning. *Script Journal: Journal of Linguistic and English Teaching*, 8(2), 214-225. https://doi.org/10.24903/sj.v8i2.1441

1. INTRODUCTION

One of the characteristics of a good teacher is having the ability to develop tools for measuring student learning outcomes. This is in line with the opinion that testing is an integral part of teaching activities (Brown & Abeywickrama, 2018). No matter how good the ability to make students learn, develop lesson plans, develop learning media, and prepare teaching materials, a teacher still cannot be called a good teacher if he does not have the ability to design, compile and carry out measurements, and use the results for various academic affairs.

Some items are known in teaching such as assessment, test, and evaluation. Assessment is intended to measure certain competencies or abilities for activities that have been carried out in learning activities (Hosnan, 2014). The test is part of the assessment that is a method to measure a person's ability, knowledge or performance in a particular domain. The method must be explicit and structured, for example, multiple choice with choices and one correct answer, a writing test with a scoring rubric, an interview test with an interview guideline or a list of answers that must be selected by the interviewer. A good test must meet the principles of practicality, validity, reliability, authenticity, and washback or test impact (Brown & Abeywickrama, 2018). Evaluation is giving a decision on the results of a test which is given at the end of the course, broadly known as summative.

A good measurement must be aligned with the material being taught so that it can show the progress of student growth and development. In the process of compiling assessment instruments, English teachers and English teacher candidates should recognize and master the competencies of the course that will be assessed. The competencies are usually reflected in the curriculum used by the schools. During the long history of education in Indonesia, there have been a number of different curricula used in the schools. At present, there is a transition between two curricula: the 2013 Curriculum and the Merdeka Curriculum. Given this, teacher candidates in the future will possibly teach students within those two curricula which both have their distinct assessment characteristics and challenges. A number of schools still apply the 2013 Curriculum, and some others are still in the process of transitioning towards the Merdeka curriculum. The 2013 curriculum was a response to changes from several previous curricula, for example, the KTSP curriculum where the National Examination was the focus resulted in teachers only preparing students for the exam rather than teaching communicative skills (Putra, 2014). This curriculum has competency references that are structured on the basis of the Communicative Language Teaching approach in which developing students' English communication skills is the main goal (Celce-Murcia, 2007). Therefore, the teacher will then facilitate students in learning various types of text that are closely related to their daily lives and also the social functions, text structure, and grammar of the text. In addition, students must also learn English as a language of interpersonal, transactional, and functional communication. Student learning experiences are also based on students' socio-cultural and cognitive backgrounds (Widodo, 2016). In addition, starting in 2019, The Ministry of Education, Culture, Research, and Technology decided to end the 2013 curriculum and began mandating the Merdeka Curriculum. This curriculum relies heavily on student-centered where students can

learn at their own pace, characteristics, and potential through several learning phases that the government has designed in the curriculum. Schools and teachers then facilitate students' diverse needs by providing teaching learning that suits their characters. The objective of this curriculum is to build students' Pancasila Profile. In terms of assessment, the National Examination is substituted by a Minimum Competency Assessment focuses on students' level of literacy and numeracy.

However, the implementation of this curriculum faces obstacles. Learning assessment processes in the 2013 curriculum are considered quite complicated by teachers so that they hinder them in implementing this curriculum (Maba, 2017). Institutions or universities that create teacher candidates should review the learning evaluation course to strengthen the understanding of teacher candidates in the future so that the ability to develop assessment tools by teachers can be even better (Arrafii & Sumarni, 2018).

Nowadays, the regulations of the government require teachers to be able, not only to teach material but also to implement the concept of Higher-Order Thinking Skills (HOTS) in the learning process. This HOTS capability needs to be developed to prepare students to develop 21st-century skills such as critical thinking skills, problem solving, decision making, and innovation. In its application, teachers face various challenges such as limited ability to understand HOTS concepts and construct HOTS-based questions, limited sources of learning materials owned by teachers, and diversity of students' cognitive abilities which makes it difficult to arrange proper instruction in class (Tyas et al., 2019).

In the learning process at the English Education Study Program, there are two courses that students must take sequentially before carrying out internship activities in the form of Educational Practice courses (6 credit units). These courses are English Language Assessment (2 credit units) and English Language Learning Test Development (2 credit units). In these two courses, students are expected to be able to design tests, put these designs into test prototypes, test the quality of these tests, test them in learning situations in the field, interpret the results, and use these results to take actions or recommend follow-ups for other parties. These two courses are related to various theories of measurement and the practice of developing measuring instruments which are often being a problem for some students. This research is aimed at uncovering problems that are often faced by students who are English teacher candidates in Educational Practice, especially in the process of developing an assessment tool in the form of a test.

2. METHODS

This study aimed to describe teacher candidates' problems in constructing test-kit for assessing students' learning. The data were in the form of quantitative and qualitative. The study involved 28 English teacher candidates taking the English Learning Assessment and English Language Learning Test Development course offered at the English Language Study Program at a university in Java, Indonesia . The students developed a multiple-choice test kit as their end-semester-mini project. The kit consisted of 30 - 40 multiple choice questions along

with its key answers and blueprint. The English teacher candidates created their test blueprint designed for students who were in junior and senior high school. A total of 1052 multiple choice questions were gathered from 28 students, and then they were analyzed by 3 investigators to look for problems frequently encountered in the process of developing the test kit. The result of the analysis was presented in descriptive statistics through percentages and enriched with qualitative descriptions.

3. RESULTS AND DISCUSSION

The results of the study reveal some information on teacher candidates' problems in developing test kits for assessing students' learning in schools. The problems are related to test kits formats, stems, options, and texts. The third year pre-service English teachers, aged 19-21 years old, consisting of 18 women and 10 men created objective items requiring the test takers to select the correct response from several alternatives. The objective items or the questions are in the form of reading comprehension, grammar mastery, completing a missing word or expression in a short dialog and text, and completing a missing word in a sentence. Twenty-eight English teacher candidates of English Education study program designed 30-40 questions as the final project in their English Language Learning Test Development class.

3.1 Problems with Test Kits Formats

Test kits formats relate to the availability of blueprint, test instruction, and text readability. A test blueprint requires test developers to identify and select particular skills or objectives to be tested and how those will be specifically distributed and weighted to the entire test items. The students formulated the distribution of items or questions in terms of learning indicators, HOTS and LOTs questions, level of difficulty questions, and the order of key answers. Out of 28 English teacher candidates, 6 of them did not provide a test blueprint. Though it is important to make sure the questions or items meet content requirements in the test blueprint (Gierl et al., 2017). Moreover, given that it is hard to decide whether every test item will be suitable for the intended learning purposes. AlFallay (2018) experienced a similar situation when conducting a study on English teachers' ability to construct test items. The results of AlFallay's study showed that a large number of respondents in the study started their item writing for formative or summative in-class tests with no clear ideas of what to write in the complete test. There was no sign of planning the language elements and language components to be included. In this situation, it is compulsory that the instructor or lecturer always monitors the process from the very beginning of the process.

Test kits formats also relate to test instruction and text readability. Five students failed to provide test instruction before the appearance of text that became a source of stems and answers. The participants simply put a text without explaining the text for which numbers. In addition, 8 students did not provide text readability. It seems that they did not realise the importance of text readability in creating test kits so that they only copy pasted from other sources without considering adaptation process to suit the test takers' level. When a test item's

readability level exceeds the test-takers' reading proficiency, it is likely that the item is not measuring the construct of interest (the subject matter), but rather the test-takers' reading proficiency (Wray & Janan, 2013). The following diagram illustrates the students' problems with test kit formats.

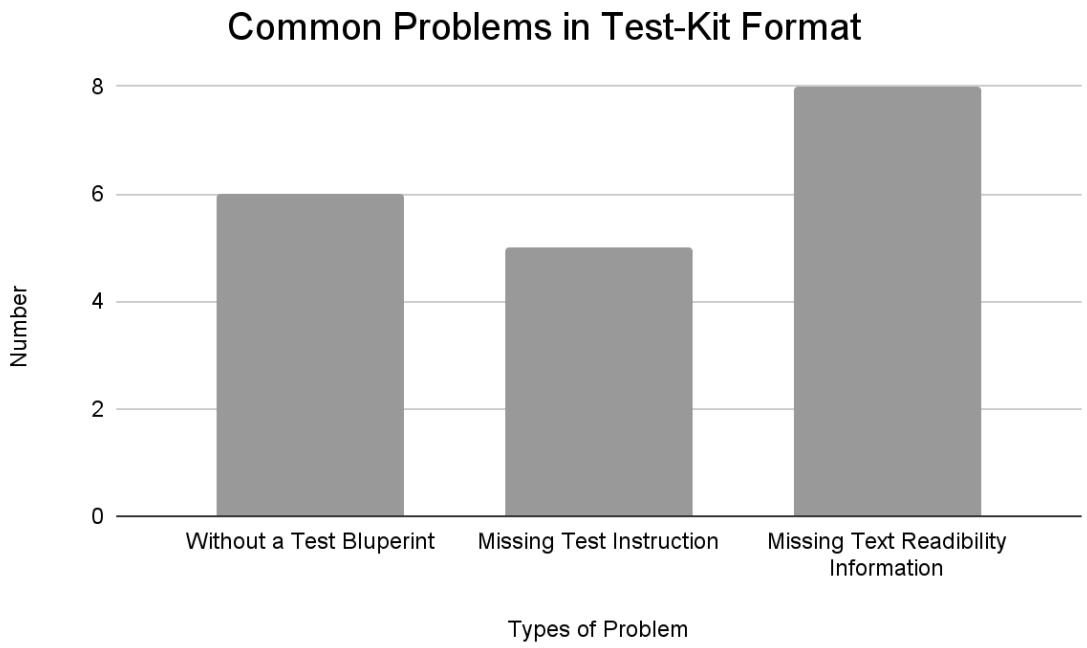

Figure 1. Common Problems in Test-Kit Format

3.2 Problems with the Test-kits' Stems and Alternatives

The results of the study also show 33% of items missed the learning indicators although they can be easily taken from the government competence standard. This situation garbled the three investigators to decide whether every item was appropriate for each indicator. This further demonstrates that the candidates actually were missing an essential part of the test item which is determining learning objectives. When it comes to items involving more complex thinking skills, it seems that there were very few students challenged to construct such items. This is in line with the findings of a study conducted by Scully (2019) revealing that MC items measuring complex cognitive processes are simply rarely constructed. This kind of item, commonly referred to as a HOTS item, has been very crucial in preparing students to face the 21st century challenges. As Wilson & Narasuman (2020) state, in the classroom setting, information collected from consistent monitoring by the authorities will support teachers; efforts in implementing the complex cognitive processing in the assessment procedure.

The analysis of stems covers mechanics, grammar, and the appropriateness of stems with texts. Stems identify the question or problem. Mechanical mistakes on stems consist of the use of punctuation, capitalization, spelling, and space. The students lacked composing stems accurately both in grammar, counting up to 8% and also in mechanical, counting up to 13%. Meanwhile, 1% of the total item mismatched the text used. One of the mistakes related to mechanics is described as follows.

Please, complete the following dialogue to compliment!

The above excerpt depicts inaccurate punctuation. Instead of using an exclamation mark, it is advised to use a full stop at the end of the stem. Also, a stem does not need the word 'please'. With some modification, the above stem can be best written with '*Complete the following dialogue with a compliment expression'.* In terms of grammar, some mistakes could be found in the teacher candidate students' works. One of the examples is described as follows.

When the meeting will be held?

The correct stem should be 'When will the meeting be held?'

Mistakes in formulating stems which are concerned with punctuation and grammar are in line with a study by (Khan et al., 2013) who investigated Identification of technical item flaws leads to improvement of the quality of single best multiple-choice questions. The study reveals that final exam questions that were collected from faculty members placed grammar, spelling, and punctuation as the three highest technical items flaws in multiple choice questions. In addition, developing a multiple-choice test kit requires candidates to recheck and to put the kit into a review to ensure that the items are not only applicable to the indicators but also have appropriate format and grammar. As item writing in multiple choice tests often faces imperfections, test developers should be aware of the points where writers often commit mistakes. The study conducted by Tarrant and Ware (2008) in Przymuszała et al. (2020) showed that multiple choice questions that are poorly developed tend to damage talented test candidates of students. This may prevent those students from demonstrating their full ability and potential, and therefore there might be some biases of the test, that is the score does not reflect the actual condition of the test takers.

The analysis of options or alternatives covers grammatical aspects, logical or illogical options, parallel with all alternatives, bad distractors, and alternatives' length. Five percent of test items lacked appropriate alternatives especially in terms of grammar. Meanwhile, 4% of the items did not have parallel alternatives, and 3 % of them had bad distractors. One percent of items had illogical alternatives and inappropriate alternative's length. These should actually be problems that the students can avoid. They can work collaboratively with others to review and check their final work. They can also utilize writing correction tools to help them ensure the alternatives are in check. This finding is consistent with the results of a study conducted by Nedeau-Cayo et al. (2013) who found that inexperienced test writers are prone to write bad or flawed items as they might not be too familiar with the guidelines for writing test questions or opt to write the quicker or easier one.

Aside from that, some items are also poorly constructed as the alternatives are either unparalleled or are bad distractors. Without having good distractors, it is hard for a test to truly measure learners' abilities. However, some students seemed to have difficulties creating a good one. This is similar to the work of Quaigrain and Arhin (2017) in that to have four working

distractors in a test is such a difficult task to do for a teacher thus they might sometimes seek help from their peers when needed. The distractors should be able to derive test-takers from choosing the real answer and each available distractor should work equally. Furthermore, Gajjar et al. (2014) also emphasise removing or substituting bad distractors for they are implausible or serve little use as a decoy, and to ensure the quality of the test. One of the mistakes related to bad distractors from the students' work is described as follows.

From the text it is known that Surabaya comes from 2 animal names including?

a. Lion and cat
b. Monkey and chicken
c. Shark and crocodile
d. Fox and tiger

The excerpt above shows that the answer was obvious and the test-maker did not provide other alternatives involving repeating words. Hence, those who had background knowledge about Surabaya would easily choose Shark and crocodile as the correct answer. In addition, one of the examples of poor parallel alternatives can be seen as follows.

They … to school yesterday.
What is the appropriate word to fill in the blank space?

a. tent
b. went
c. gone
d. going

The above alternatives are not parallel as one of the alternatives is noun. All alternatives should be in verb form. The following diagram describes the type of problems concerning formulating stems and alternatives among the students.

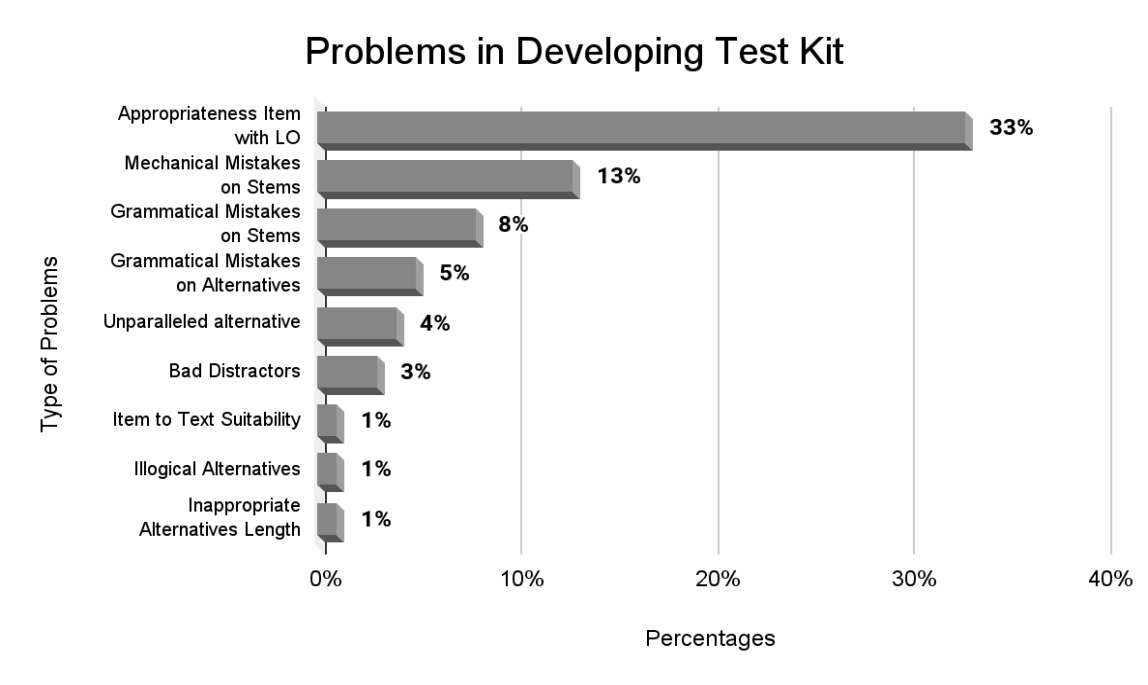

Figure 2. Type of Problems concerning to Stems and Alternatives

3.3 Problems with the Texts

The next analysis is related to the texts used as the sources of stems and alternatives. This study collected in total of 208 texts in the form of narrative, description, procedure, report, and dialogue. One hundred forty nine out of 208 were in the form of adopted text. This means that candidates preferred to take text that is readily available on the internet rather than to adapt it to suit student levels or objectives. The main problem is that they did not mention the sources where they took the texts. The texts were usually taken from websites that provide exercises for language learners or any other courses. In addition, 59 texts were adapted from English materials websites with some modifications that can suit students' levels. The candidates were also requested to count the readability of each text used in the item. Of the total of 208 texts, 58 of them did not cite the readability index. This could be that the candidates forgot to put one or they did not have the resources or time to adapt the text to suit the learners' level or the objectives.

The candidates mostly preferred to adopt rather than to adapt the text used in the test kits. The findings imply that the students might find adjusting authentic texts difficult especially when they need to match the text within appropriate readability to suit the learner levels as well as the objectives. Adjusting text requires students to change the length of words, vocabulary and also the sentence structure which sometimes are meticulous and arduous. However, it is necessary to do as authentic texts can be too hard sometimes for particular students while unauthentic texts might not provide as much natural target language use. This is similar to Beresova's (2015) in that teachers should be able to prepare or adjust authentic material to be

beneficial for learners despite how demanding it is. The findings also imply that students need to do more practice in searching, and adjusting authentic text to match the difficulties to the learners' level.

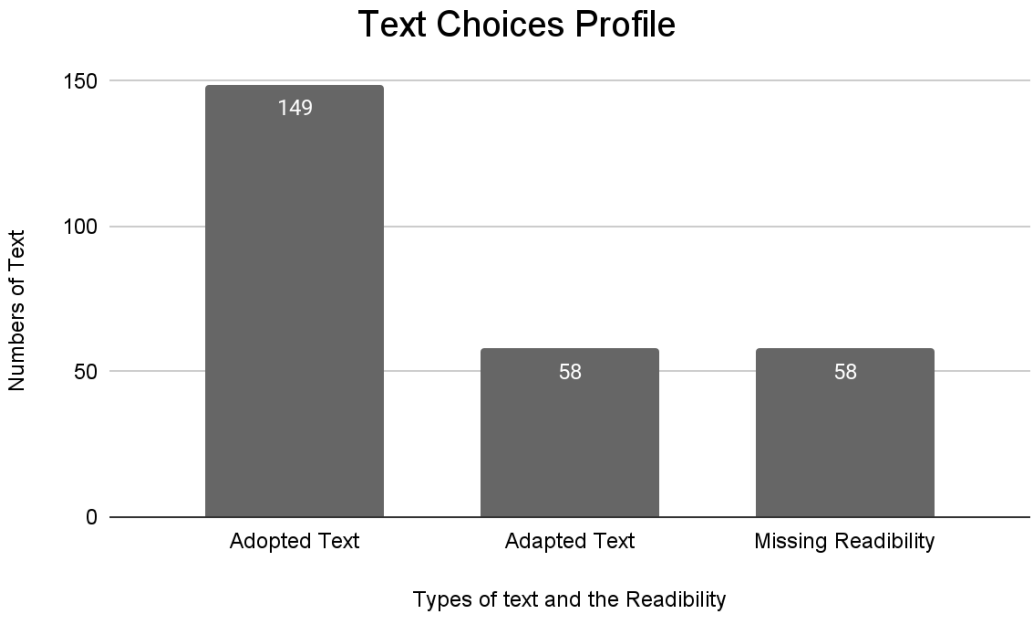

Figure 3. Text Choices Profile

4. CONCLUSION

This present study aimed at finding students' main difficulties in constructing multiple-choice questions test kits. The result revealed that the students face difficulties in writing test-items which truly match and measure the learning objectives, good distractors, minimal grammatical errors both in stems and alternatives, and choosing and adjusting text for the test-items. This study suggests that teacher candidates, teacher or beginner test writers may conduct a test-item review prior to using their own developed test to their students. In addition, beginner-test writers should also prepare time and needed resources when it comes to creating well-written multiple-choice questions concerning the complexity of the process and the long-list guidelines to adhere to. It is hoped that courses related to language assessment consider addressing test review as part of their curricula as it is an essential skill to have for their students.

The study also comes with some limitations. First, there was potential inconsistency between the test investigators during the analysis due to different interpretations of each item, and fatigue. Second, HOTS and LOTS aspects were excluded to limit the discussion in the analysis. Future research is suggested to take account of HOTs aspects as it may provide thorough understanding of quality test-items.

5. REFERENCES

AlFallay, I. S. (2018). Test specifications and blueprints: Reality and expectations. *International Journal of Instruction*, *11*(1), 195–210. https://doi.org/10.12973/iji.2018.11114a

Arrafii, M. A., & Sumarni, B. (2018). Teachers' understanding of formative assessment. *Lingua Cultura*, 12(1), 45-52. https://doi.org/10.21512/lc.v12i1.2113

Beresova, J. (2015). Authentic Materials – Enhancing Language Acquisition and Cultural Awareness. *Procedia - Social and Behavioral Sciences*, *192*, 195–204. https://doi.org/10.1016/J.SBSPRO.2015.06.028

Brown, H. D., & Abeywickrama, P. (2018). *Language Assessment Principles and Classroom Practices* (3rd ed.). Pearson Education.

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. *Intercultural Language Use and Language Learning*, 41–57. https://doi.org/10.1007/978-1-4020-5639-0_3/cover

Gajjar, S., Sharma, R., Kumar, P., & Rana, M. (2014). Item and test analysis to identify quality multiple choice questions (MCQS) from an assessment of medical students of Ahmedabad, Gujarat. *Indian Journal of Community Medicine*, *39*(1), 17–20. https://doi.org/10.4103/0970-0218.126347

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082-1116. https://doi.org/10.3102/0034654317726529

Hosnan, M. (2014). *Pendekatan saintifik dan kontekstual dalam pembelajaran abad 21: Kunci sukses implementasi kurikulum 2013* (1st ed.). Ghalia Indonesia.

Khan, H. F., Danish, K. F., Awan, A. S., & Anwar, M. (2013). Identification of technical item flaws leads to improvement of the quality of single best Multiple Choice Questions. *Pakistan Journal of Medical Sciences*, *29*(3), 715. https://doi.org/10.12669/PJMS.293.2993

Maba, W. (2017). Teacher's Perception on the Implementation of the Assessment Process in 2013 Curriculum. *International journal of social sciences and humanities*, 1(2), 1-9. https://doi.org/10.29332/ijssh.v1n2.26

Nedeau-Cayo, R., Laughlin, D., Rus, L., & Hall, J. (2013). Assessment of item-writing flaws in multiple-choice questions. *Journal for Nurses in Professional Development*, *29*(2), 52–57. https://doi.org/10.1097/NND.0B013E318286C2F1

Przymuszała, P., Piotrowska, K., Lipski, D., Marciniak, R., & Cerbin-Koczorowska, M. (2020). Guidelines on Writing Multiple Choice Questions: A Well-Received and Effective Faculty Development Intervention. *SAGE Open*, *10*(3). https://doi.org/10.1177/2158244020947432/asset/images/large/10.1177_2158244020947432-fig1.jpeg

Putra, K. A. (2014). The implication of curriculum renewal on ELT in Indonesia. Parole: *Journal of Linguistics and Education*, 4(1 April), 63-75. https://doi.org/10.14710/parole.v4i1%20April.63-75

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1). https://doi.org/10.1080/2331186X.2017.1301013

Scully, D. (2019). Constructing Multiple-Choice Items to Measure Higher-Order Thinking. *Practical Assessment, Research, and Evaluation*, *22*(1), 4. https://doi.org/https://doi.org/10.7275/swgt-rj52

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*(2), 198–206. https://doi.org/10.1111/J.1365-2923.2007.02957.X

Tyas, M. A., Nurkamto, J., Marmanto, S., & Laksani, H. (2019). Developing Higher Order Thinking Skills (HOTS) – Based Questions: Indonesian EFL Teachers' Challenges. *Proceedings of the International Conference on Future of Education*, *2*(1), 52–63. https://doi.org/10.17501/26307413.2019.2106

Widodo, H. P. (2016). Language Policy in Practice: Reframing the English Language Curriculum in the Indonesian Secondary Education Sector. *Language Policy(Netherlands)*, *11*, 127–151. https://doi.org/10.1007/978-3-319-22464-0_6/COVER

Wilson, D. M., & Narasuman, S. (2020). Investigating Teachers' Implementation and Strategies on Higher Order Thinking Skills in School Based Assessment Instruments. *Asian Journal of University Education*, *16*(1), 70–84. https://doi.org/10.24191/AJUE.V16I1.8991

Wray, D., & Janan, D. (2013). Exploring the Readability of Assessment Tasks: The Influence of Text and Reader Factors. *Multidisciplinary Journal of Educational Research*, *3*(3), 69–95. https://doi.org/10.4471/remie.2013.04