International Social Sciences and Humanities
UMJember Proceeding Series (2024) Vol. 3 No 2: 390-403

KOPI X LATTE 2024

Developing Summative Grammar Test for Intermediate Level Students of English Education Program

Indri Astutik[1], Nurkamilah[2], Anam Fadlillah[3], and Yayah Ikhda[4]
[1]Universitas Muhammadiyah Jember; indri@unmuhjember.ac.id
[2]Universitas Muhammadiyah Jember; nurkamilah@unmuhjember.ac.id
[3]Universitas Muhammadiyah Jember; anam.fadlillah@unmuhjember.ac.id
[4]Universitas Muhammadiyah Jember; yayah@unmuhjember.ac.id

*Correspondence: Anam Fadlillah
Email: anamfadlillah@unmuhjember.ac.id

Published: May, 2024

Abstract: The research is conducted based on the fact that teachers rarely develop tests especially for English Grammar tests. The reason underpinning the condition is because developing a test is not an easy task for teachers regarding the process that takes much time and has to follow certain standards of constructing good test. Teachers tend to adopt and adapt available tests from sources available in the market although these kinds of tests usually constructed using common criteria and neglecting specific condition of their students. The objective of the research is how to develop a summative test to measure students' achievement on grammar subject. The specific objective of the research is how to develop a valid, reliable, and practical summative English grammar test for intermediate level students of English education program. The findings, hopefully, can bring theoretical and practical significance for teachers/researchers in general and teachers/researchers of grammar subject in particular. Theoretically, all the references in the article can be used as significance sources when they construct their own test. Practically, the result of the research can be administered to their classes or their research, or it can inspire them to develop similar test for their classes.

Keywords: Grammar test, development, Jember.

INTRODUCTION

The recent development of language teaching indicates that language programs be tailored to a specific group of teachers and students with specified objectives in a specific setting (Kumaravadivelu, 2001). As a result, an essential thing in language instruction is program assessment which provides context-bound responses to typical concerns in the field (Yıldırım & Topkaya, 2020). Grammar as one of the subject courses in language teaching program, also has to include assessment in the teaching learning process. Regarding to the course assessment in the teaching learning process, a test as one of the subsets of assessment instruments, comprises three components i.e., method, measure, and individual's ability. A test uses a method defined as a set of techniques, procedures, or items performed by test takers. A test has to be able to measure abilities—general or specific ones. A test measures individual's ability, knowledge, or performance resulting in scores (Brown 2004).

In a course, there must be distinction of summative and formative tests. A summative test is one that contributes to a student's final grade in a course. Summative tests must select content from a wide range of cognitive domains, and they are often quite extensive. On the other hand, throughout the course, students are given formative quizzes to help them track their progress. Formative quizzes are used to help students learn, but they can also be used to determine a course grade (Haladyna, 2018).

Referring to summative and formative tests in a course, teachers have the responsibility to build test instruments to measure the development of the students' progress following the course. This is also due to the fact that there are not any standardized national examinations for higher education. Thus, students' achievements are measured through tests built by teachers. However, developing good test instruments/items is not an easy task for teachers since teachers have to pay attention to the processes of test development. As a result, teachers' best option is to beg, borrow, adopt, modify, and create items, which they can then save in an item bank on their computer (Haladyna, 2018).

Although the aspects of test development are essential in the academic discipline of educational and psychological assessment, concerns and methods related to test development have received little scholarly scientific attention in the past; most study and writing has concentrated on the more statistical aspects of testing. Nonetheless, test development, as well as all of the difficulties and practices related with creating accomplishment and ability tests, play a vital and crucial part in any testing program (Downing & Haladyna, 2011).

For higher education classes there are two sorts of test item formats to choose from: selected response (SR) and constructed response (CR) (Haladyna, 2018). Other terms for such tests are open-ended test (free response) and closed (objective or forced choice) test (Urbina, 2004). Both are quite beneficial. Urbina (2004) described open-ended test including writing samples, free oral responses, or performance of certain tasks while Haladyna (2018) mentioned student performance, research paper, essay, experiment, oral report, observation, simulation, creative work, project, portfolio, and demonstration are examples of constructed-response formats. They are useful for investigating concepts, circumstances, and phenomena because they provide a broader range of options, deeper samples of the individual's behavior, and allow for the emergence of their own qualities. However, scoring of such tests is more difficult and time-consuming, and such exams have lower reliability and validity than closed answer (forced-choice) assessments (Urbina, 2004). In contrast, closed tests (Fenn et al. (2020) gives the respondent a limited number of options from which to pick in the forced-choice format. Multiple choice, true/false statement, rating, ranking, and matching, as well as rearrangement of the options supplied, can all be used to make a decision. Further, Haladyna (2018) stated that the popular multiple choice (MC) and true-false (TF) formats are among the SR formats.

The first step in constructing a test for students is to figure out what the test will be used for. Determining the purpose will assist teachers in selecting the appropriate test and focusing on the test's definite objective which refers to test type. Relating to test type, there are five test types i.e., language aptitude tests, language proficiency tests, placement tests, diagnostics tests, and achievement tests. Among the five types of the tests, language aptitude and proficiency tests are not commonly designed by teachers. In contrast, the three other types—placement tests, diagnostics tests and achievement tests are the test types that are familiar for teachers (Brown, 2004).

Referring to the background above, the kind of research proposed is development research. The research intended to design or construct a test for grammar class. The main objective of the research is how to develop a summative test to measure students' achievement on grammar subject. The specific objective of the research is how to develop a valid, reliable, and practical summative English grammar test for intermediate level students of English education program. The project was proposed based on the fact that teachers rarely (instead of saying never) construct their own test for their students or their research. Hopefully, the finding can give contributions for teachers and researchers of similar field of study. Theoretically, all the references in the article can be used as significance sources when they construct their own test. Practically, the result of the research can be administered to their classes or their research, or it can inspire them to develop similar test for their classes or research.

The description of students' level is taken from the course description of BA or undergraduate catalogue of English Language Education (ELE), Department of English, Faculty of Letters, a prominent state university in Malang, East Java Province. This catalogue is used as a reference to construct the test because of some reasons: 1) the researcher has got permission to use the catalogue; 2) the course description of the subject is clear; 3) the objectives of the course are also well stated and very specific. Thus, these aspects are beneficial resources for the researcher to develop the test for the course.

Based on the catalogue, there are three grammar courses for BA students—Basic English Grammar, which is integrated in the Intensive Course subject, Intermediate English Grammar and Advanced English Grammar. Intermediate English Grammar is the second grammar subject taking 4 credits and 4 hours per week. The objective of the course is "to provide students with sound knowledge English Grammar and the ability to apply this knowledge in comprehension and production, and to provide the students with practice in taking the Structure and Written Expression part of the TOEFL" (English Department, 2020 p. 63). The materials of the Intermediate English Grammar comprise Gerunds and Infinitives, Noun Clauses, Adjective Clauses, and Adverbial Clauses (English Department, 2020).

One of the objectives of the course clearly mentioned that it provides students with practice in taking Structure and Written Expression part of the TOEFL. Consequently, this summative test is designed as TOEFL-like format to give students opportunity to practice TOEFL. It also points out that multiple choice (MC) test with four options is the model of the test. Even though according to theory, research, and practicality, three options are adequate (Haladyna & Downing, 1993; Rodriguez, 2005), the test was designed with four options because TOEFL is multiple choice test which has four options.

A test is supposed to provide reliable information as an instrument for measuring students' achievement meaning that it should be able to provide information that is as close to the reality as possible. This is significant since the information from the test is utilized to make decisions on student regulations as well as teaching and learning activities in general (Nurgiantoro, 2001). In other words, the constructed test should have good criteria. A good test as an instrument to measure students' achievement should be valid, reliable and practical.

A valid test should meet the extent to which inferences drawn from the test results are relevant, meaningful, and valuable in terms of the assessment's objective (Gronlund, 1998). Weir (1990) stated when a test measures what it claims to measure, it is said to be valid. Brown (2004) illustrated that there is no final, absolute measure of validity of a test. In some circumstances, it may be necessary to look at how closely a test requires achievement to reflect that of the subject or unit of course being assessed. In other circumstances, it could be about how successfully a test evaluates whether or not students have achieved a set of objectives or reached a certain level of competency. Another commonly recognized kind of proof is statistical correlation with other relevant but independent measurements. To sum up, there are five types of validity—content-related evidence which is popular called as content validity (Hughes, 2003; Mousavi, 2002), criterion related evidence (concurrent and predictive validity), construct-related validity (construct validity), consequential validity, and face validity (Brown 2004). Bachman (1990) affirmed that on the logical argument basis, validity is divided into two types: construct and content validity, and on the empirical research there are concurrent and predictive validity. In other words, logical validity may be determined by creating a table of specifications, and empirical validity can be determined by comparing the results of the two tests and calculating the correlation coefficient.

A good test should also expose reliability. A reliable test is one that is constant and dependable. If the same test is administered to the same student or students who are matched on two separate times, the results should be identical. There are some kinds of reliability i.e., student-related reliability, rater reliability, test administration reliability, test reliability (Brown, 2004). The most common learner-related reliability issue is inconsistency. Temporary illness, exhaustion, 'a bad day', anxiety, and other physical or psychological circumstances might cause an observed-score to differ from one's true score, causing reliability issues. Factors like a test taker's 'test-wiseness' or test-taking methods are also covered in this category (Mousavi, 2002, p. 804). The scoring procedure may be influenced by human error, subjectivity, and bias. The inconsistent scores caused by these factors influence the reliability which is called as rater reliability. Rater reliability takes two forms i.e., inter-rater reliability and intra-rater reliability. Inter-rater reliability arises when two or more scores for the same exam produce inconsistent results, maybe due to a lack of focus on scoring criteria, incompetence, recklessness, or even preexisting prejudices. In other words, two or more scorers do not use the same criteria. On the other hand, because of vague criteria, exhaustion, prejudice toward certain good and bad students, or simple carelessness, intra-rater reliability is a regular occurrence for classroom teachers. Because writing competency entails multiple features that are difficult to describe, rater reliability is particularly challenging to attain in tests of writing skills. However, carefully specifying an analytical scoring tool can improve rater reliability (Brown, 2004; Brown, 1991). The conditions under which the test is administered may also cause unreliability. Photocopying variances, the amount of light in different parts of the room, temperature fluctuations, and even the state of the desks and chairs are all sources of inaccuracy. These are the sources of test administration reliability (Brown, 2004). The last is test reliability. The nature of the test might sometimes lead to measurement inaccuracies. If an exam is excessively long, test-takers may grow tired by the time they get to the last few items and react erroneously in haste. Students who do not score well on a timed test may be discriminated against in timed exams. Another source of test unreliability could be poorly written test items (that are confusing or have more than one valid answer) (Brown 2004). Above all, consistency is a key measure of trustworthiness. It means that if a test result is constant from one test to the next, the test result has a high level of reliability. The coefficient correlation index, which ranges from 0.00 to 1.00, is a statistical value that indicates reliability (Weir, 1990).

Practicality is another issue of a good test. A test is said to be practical if it meets some aspects: 1) It is not too costly; 2) It adheres to reasonable time constraints; 3) It is relatively simple to use; and 4) It has a particular and time-saving scoring/evaluation mechanism (Brown, 2004). Practicality refers to test administration, scoring, and interpretation of test findings, as well as the financial aspects of test administration ((Djiwandono, 1996). To summarize, a practical test is a test which is easy to administer with clear instruction, easy to score with clear scoring criteria, is not too expensive and reusable, and does not take long time allotment to do and to score.

METHOD

The research is development research in which the researcher develops a test to measure students' achievement on grammar course. The objective of the research is how to develop English grammar test. The specific objective is how to develop a valid, reliable, and practical English grammar test for intermediate level students of English education program. The research was planned to involve a number of undergraduate students from English educational programs of faculty of teacher training and education from a few universities in East Java Province. The data would be got from the tryout test used to determine the validity, reliability, and practicality of the test. However, since the time was very limited, the tryout test had not been able to be conducted. Thus, the quality of the test and test items used the judgement of validators. The validators were three lecturers of a private university in Jember who have taught English for more than five years. The validation and judgment from them were used as source to define the quality of the test. The quality judgment of the test was also based on theories underlying the constructing a good and qualified test.

To develop a valid, reliable, and practical test, the researcher followed a standard process of developing a test. In this research, the researcher used a model of Standard Step of Developing a Test designed by Azwar (1996). The model consists of nine steps: 1) identifying the objective; 2) 2a. identifying content description and 2b. identifying competence limitation; 3) developing blue print (specification); 4) 4a. developing items, and 4b. reviewing items; 5) conducting preliminary tryout; 6) conducting field test/item analysis; 7) arranging the test and instructional development; 8) estimating the reliability of the test; and 9) presenting final form (ready to use) test items (Azwar, 1996). The steps of developing the test shown in Figure 1.

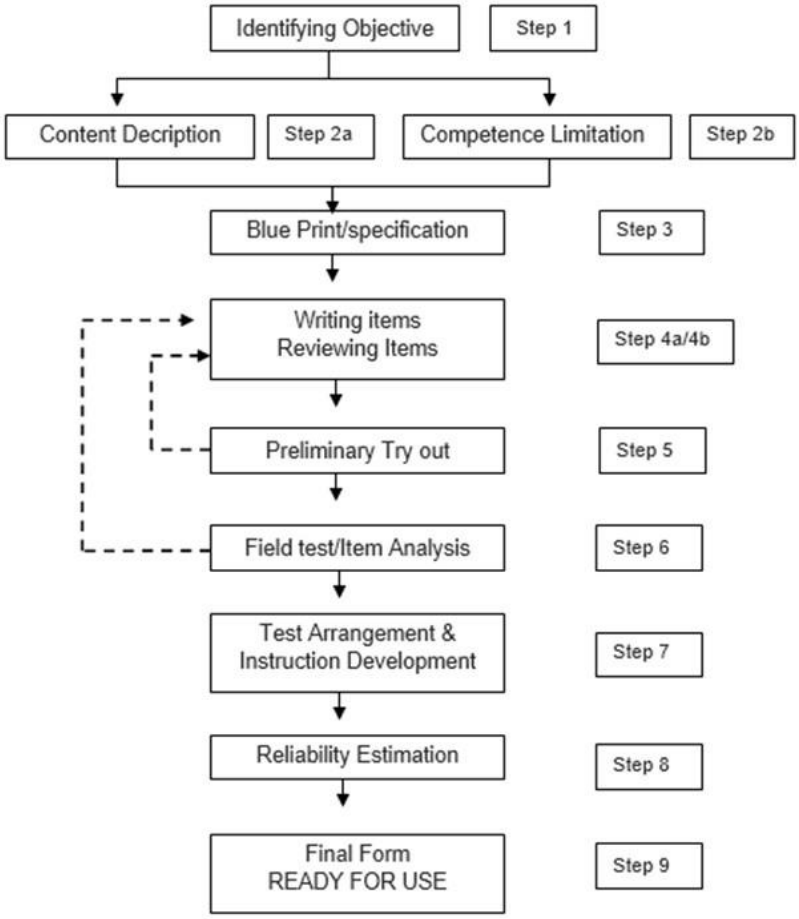

Figure 1. Model of Test Development

(Azwar, 1996, p. 54)

Unfortunately, because of limited time, this research was only able to finish four stages—stage 1 to 4. The stages are: 1) identifying the objective; 2) 2a. identifying content description and 2b. identifying competence limitation; 3) developing blue print (specification); 4) developing items and reviewing items. Other stages would be continued to get empirical data of validity and reliability statistically. For clearer order of the stages, Figure 2 describes the stages in more comprehensive way.

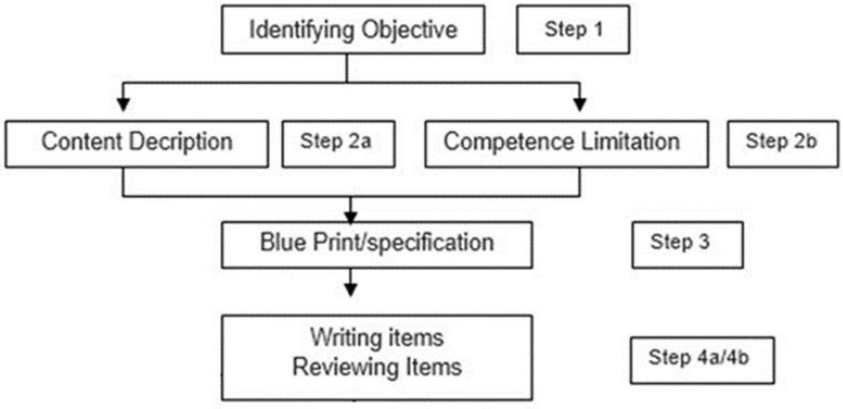

Figure 2 The Model of Current Test Development

Stage 1 Identifying the Goal and Instructional Objectives of Intermediate English Grammar

As stated in Figure 2, the first step of developing the test and test items is identifying the goal and the instructional objectives of intermediate English grammar course. The goal of the course can be seen from the description of the course. The catalogue of English Department, from Malang university states the description of the course (English Department, 2020, p. 42) as: "This is the second part of a three-part English Grammar course which provides the students with a sound knowledge of English Grammar and the ability to apply this knowledge in comprehension and production. Intermediate English Grammar focuses on gerunds and infinitives, noun clauses, adjective clauses, and adverbial clauses. In addition, it provides the students with practice in taking the Structure and Written Expression part of the TOEFL."

The syllabus (course profile) was proposed referring to the instructional design theories (Brown, 1995; Finney, 2002; Marzano & Kendall, 2007; Nation & Macalister, 2010) based on the course description. Following that, the instructional objectives were aligned with the course description, ensuring the content validity. The main goal of this course is to develop students' knowledge on English grammar dealing with gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses and their ability to apply the knowledge in comprehension and production, and provide them practice in taking the Structure and Written Expression part of the TOEFL as well.

Goal

The students are able to apply their knowledge of English Grammar on the topics of gerunds and infinitives, noun clauses, adjective clauses, and adverbial clauses in comprehension and production as well as in facing TOEFL for Structure and Written Expression at intermediate level.

Instructional Objectives

a) Identify the forms and uses of gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses.
b) Use the correct forms of gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses in comprehending structure and written expression as well as in production such as speaking and writing.
c) Use correct connectors to analyze and produce structure and written expression using noun clauses, adjective clauses, and adverbial clauses
d) Use correct word order to analyze and produce structure and written expression using noun clauses, adjective clauses, and adverbial clauses.
e) Use correct connectors to analyze and produce reduced structure and written expression using noun clauses, adjective clauses, and adverbial clauses
f) Apply the rules to analyze correct structure and written expression using noun clauses, adjective clauses, and adverbial clauses.
g) Choose the correct/incorrect forms/uses of gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses in structure and written expression.

These instructional objectives indicate the indicators of students' ability in identifying, producing, analyzing, and choosing the correct/incorrect structure and written expression using gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses at the intermediate level.

Stage 2 Content Description and Competence Limitation

2a Content Description

The content of the test referring to the course description of Intermediate English Grammar course covers gerunds & infinitives, noun clauses, adjective clauses and adverbial clauses. In the test, the students are expected to apply their knowledge and understanding about gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses in analyzing and choosing correct/incorrect structure and written expression.

The test was designed in the TOEFL Structure and Written Expression format to give students experience to face TOEFL test. Thus, the forms of the test covered structure consisting of 15 items where the students are asked to analyze, apply their knowledge, and choose the correct answer for each item, and written expression consisting of 25 items in which the students are challenged to analyze and choose the wrong part of each item.

2b Competence Limitation

Based on the description of the course, the goal, and the instructional objectives, the test is intended to measure students' achievement on English grammar at the intermediate level which comprises gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses. Further, the instructional objectives clearly mention specific competences that were expected from the students where they are to: a) identify the forms and uses of gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses; b) use the correct forms of gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses in comprehending structures and written expressions as well as in production such as speaking and writing; c) use correct connectors to analyze and produce structures and written expressions using noun clauses, adjective clauses, and adverbial clauses; d) use correct word order to analyze and produce structures and written expressions using noun clauses, adjective clauses, and adverbial clause; e) use correct connectors to analyze and produce reduced structures and written expressions using noun clauses, adjective clauses, and adverbial clauses; f) apply the rules to analyze correct structures and written expressions using noun clauses, adjective clauses, and adverbial clauses; g) choose the correct/incorrect forms/uses of gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses in structures and written expressions.

Stage 3 Blue Print/Specification

The blue print of intermediate English grammar test describes the purpose of the test, type of the test, grammar components that are tested, grammar skills and the number of items for each component. It is presented in Table 1.

Table 1 The Blue Print of Intermediate English Grammar Test

The purpose of the test:
The test is intended to measure students' achievement on Intermediate English Grammar, which comprises gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses at intermediate level.

Type of test: Multiple Choice

Grammar Components	Grammar Skills	Number of Item
Gerund & Infinitives	●Use gerunds/infinitives correctly ●Be careful with object of preposition	10 items

Noun Clauses	●Use noun clause connectors correctly ●Use noun clause connector/subjects correctly ●Invert the subject and verb with question words in noun clauses	10 items
Adjective Clauses	●Use adjective clause connector/subjects correctly ●Use reduced adjective clauses correctly ●Invert the subject and verb with place expressions ●Recognize active and passive meaning in reduced adjective clauses	10 items
Adverbial Clauses	●Use adverb time and cause connectors correctly ●Use other adverb connectors correctly ●Use reduced adverb clauses correctly ●Recognize active and passive meaning in reduced adverbial clauses	10 items

Stage 4 Writing Items and Reviewing Items

4a. Writing Items

As explained in the background of the study, the form of the test is multiple choice (MC) with four options referring to TOEFL formats. This form of test is used because of some reasons: 1) the course description clearly mentions that one of the course's objectives is to provide students with practice in taking the Structure and Written Expression part of TOEFL, thus, MC is the only format that is used in the Structure and Written Expression part of TOEFL; 2) MC is the test format that is easy to administer and score; 3) most students are familiar with MC for grammar tests.

The test consisted of 40 items covering two MC forms. First, structural test, consisting of 15 items in which the students are challenged to choose the correct answers. Second, written expression, consisting of 25 items where in the test the students are asked to choose the incorrect part of the written expression. The test is set for 60 minutes by considering the level of the students. Real TOEFL test takes 25 minutes to do the Structure and Written Expression part. However, it is considered very hard for the students to complete the test in 25 minutes because of the students' level that is intermediate level. Thus, 60 minutes is thought appropriate time allotment for the test by considering that each item takes one and a half minutes.

In developing test items, some considerations have to put in order that the test has qualifications as a good test. Some experts have proposed how to write a good MC test. Among them are Butler and Haladyna. Butler (2018) mentioned that there are five principles and an extra principle of developing an MC test i.e., a) avoid employing complicated item types, techniques, or answers, b) make items that necessitate the use of specific cognitive processes, c) avoid to create answers with none-of-the-above and all-of-the-above, d) offer three option answers, e) create a demanding but not overly difficult multiple-choice test, and the extra principle is giving feedback. In line with Butler's principles, Haladyna (2018) proposed guidelines to develop test items of MC format that was extracted from latest version of Haladyna & Rodriguez (2013). The guidelines include content, format, writing stem, writing option, and the stem shell (2018).

The content consists of six guidelines i.e., 1) Each item should focus on a single sort of information and a single cognitive demand. 2) Make the test item out of something new. 3) Each item's content should be distinct from those of other items. 4) Avoid text that is too particular or too generic; instead, stress significant content over minor stuff. 5) avoid giving test takers to express their opinion unless it is so qualified. 6) Avoid trick items that mislead the test takers into selecting the incorrect answer (Haladyna, 2018).

The format offers three guidelines i.e., 1) It is preferable to format items vertically rather than horizontally. Although horizontal formatting allows for more things to fit on the page, reading selections horizontally might be difficult or confusing. 2) Editing and proofreading should be done

on all items. 3) The level of linguistic difficulty of the items should be appropriate for the students' level (Haladyna, 2018).

In the writing stem, there are three aspects that should be considered, i.e., 1) Minimize the number of words in the item's stem and avoid using window dressing. Clarity is a value as long as the essence of the topic need to be tested is communicated. 2) State the primary notion in the stem, not in the options. 3) Use positive language instead of negative ones like "not," "except," or "false." The usage of negative keywords in the stem, according to the authors of testing textbooks and certain studies, has a detrimental influence on student replies (Haladyna, 2018).

Dealing with writing options, there are eight guidelines that should be considered, i.e., 1) distractors should be realistic incorrect answers, rather than absurd or otherwise improbable options; 2) there should only be one correct answer in any of the SR forms; 3) the correct answer should appear in a different spot each time; 4) options should be listed in numerical order if numbers are utilized; 5) options should be distinct from one another in terms of quality and substance; avoid options that are similar in quality or content; 6) term like "all of the above" and "none of the above" should be avoided; 7) positive, not negative, language should be used in the options; and 8) avoid hints that direct to the correct answer (Haladyna, 2018).

The last guideline is dealing with the item shell. Avoid humor is a principle in item shell. A hollow item stem is referred to as an item shell. It gives you a starting point for composing your item. Many item designers struggle with writer's block, and item shells can help them overcome it (Haladyna, 2018).

Scoring

Scoring is also an important part of designing a test. This research used two systems of scoring. The first part, structure, which consists of 15 items, scores 2 for each item, so the total score is 30. The second part, written expression, which consists of 25 items, scores 3 for each item, so the total score is 70. All in all, the total score for the test is 100.

4b Reviewing Item

Reviewing item was done in two stages, self-reviewing and peer reviewing. Self-reviewing was done many times until the test was considered well designed and appropriate to be applied. After the test items considered as well prepared and appropriate to use, then peer review was conducted by sending the test to three lecturers of English Education Program, Faculty of Teacher Training and Education, a private university in Jember, East Java province.

The peer lecturers employed in the peer review have taught English in the department for more than 10 years. Thus, they are considered as experts in the field. The review and validation of the test dealt with conformity among the description of the course, the goal, instructional objectives, and the blue print and the test items. It also dealt with practicality of the test and the quality of the test items based on the guidelines referring to the content, format, writing stem, writing option, and item shell.

RESULTS AND DISCUSSION

The test designed was Intermediate English Grammar achievement test with MC format with four options of the TOEFL-like with 40 items. The test is in two models taking Structure and Written Expression of TOEFL's part. Structure takes 15 items, and written expression takes 25 items. The materials for the test are gerunds and infinitives, noun clauses, adjective clauses, and adverbial clauses. Each material of the grammar component has 10 items. The level of competence as the name and the course description is intermediate level. Table 2 shows the specification of the test.

Table 2 Multiple Choice Test Specification of Intermediate Grammar Test

Grammar component	Grammar Skills	Number of Item	Test Item Number

Gerund & Infinitives	● Use gerunds/infinitives correctly ● Be careful with object of preposition	10 items	4, 8, 15, 17, 20, 25, 28, 33, 35, 39
Noun Clauses	● Use noun clause connectors correctly ● Use noun clause connector/subjects correctly ● Invert the subject and verb with question words in noun clauses	10 items	2, 5, 6, 9, 16, 23, 27, 30, 38, 40
Adjective Clauses	● Use adjective clause connector/subjects correctly ● Use reduced adjective clauses correctly ● Invert the subject and verb with place expressions ● Recognize active and passive meaning in reduced adjective clauses	10 items	1, 3, 12, 14, 18, 22, 26, 31, 34, 37
Adverbial Clauses	● Use adverb time and cause connectors correctly ● Use other adverb connectors correctly ● Use reduced adverb clauses correctly ● Recognize active and passive meaning in reduced adverbial clauses	10 items	7, 10, 11, 13, 19, 21, 24, 29, 32, 36,
Total		40 items	

Findings of the Preliminary Try-out

Preliminary try-out conducted by involving three English lecturers of English Education Program, Faculty of Teacher Training and Education of the university in Jember. The three lecturers are experienced ones because they have taught English in the department for more than ten years. They were asked to review the conformity among the course description, goal, instructional objectives, blue print and the test items. They were also asked to evaluate the practicality dealing with time spent, funding, and administering and scoring the test, and the quality of the test items regarding the content, format, writing stem, writing options, and the item shell.

Validity of the Test

The three reviewers stated that the test items have been matched with the course description, goal, instructional objectives and the blue print. It can be said that the test has had content validity since the test items have reflected the content of the curriculum or syllabus of the course namely Intermediate English Grammar course. As stated by Brown (2004) that there is no final, absolute measure of validity of a test. In some circumstances, it may be necessary to look at how closely a test requires achievement to reflect that of the subject or unit of course being assessed. However, the test could not be said to have empirical validity since there was not any statistical evidence presented. It was due to time limitation in developing the test that the researcher was not able to conduct try-out. In short, the test met the content validity, but not the empirical validity. Further processes should be conducted to get the empirical validity.

Reliability of the Test

Reliability refers to the consistency of a test. It can be achieved by administering test-retest. Reliability needs a long process because the test should be administered more than one time. Since the test had not been administered, the test could not be said as reliable test because there was not any evidence presented to support the reliability. Thus, the test needs further processes/stages to prove the reliability. It opens chances to be proved in the future.

Practicality

Dealing with practicality, a good test should be practical. The test could be said as a practical test in term of time spent, funding and administering and scoring. The time spent for the test is 60 minutes for 40 items. Previously, the test was set for 90 minutes. The reviewers suggested to change the time allotment because 90 minutes was too long for a 40-item-test. They suggested the time spent became 60 minutes assuming that students have already had enough time to spend to analyze each item for one and a half minutes. It is enough for the

intermediate level students. Another issue of practicality is funding. The reviewers considered that the test is practical because it consists of 40 items and it is written in 5 pages, so it is not costly for photocopying the test. Moreover, the test can be administered many times since the answer sheet is given separately. The test can also be given online which makes it more economical. Thus, it can be said that the test is practical in term of funding. The last issue of practicality is ease to administer. Since the form of the test is multiple choice test, the reviewers believed that it is very easy to administer and score. Consequently, the test can be said practical in every aspect of practicality—time spent, funding and administering. The findings matched with the theories of practicality of a test to be said as a good test proposed by Brown (2004) and Djiwandono (1996). A test is said to be practical if it meets some aspects: 1) It is not too costly; 2) It adheres to reasonable time constraints; 3) It is relatively simple to use; and 4) It has a particular and time-saving scoring/evaluation mechanism (Brown, 2004). Practicality refers to test administration, scoring, and interpretation of test findings, as well as the financial aspects of test administration ((Djiwandono, 1996).

Quality of the Test

The quality of the test means how the development of the test follows guidelines of constructing a good test based on theories and best practices. The guidelines of developing/constructing follows two experts, Butler's (2018) and Haladyna's (2018) principles. However, the guidelines were taken from Haladyna's (2018) mostly because the guidelines are presented more detail. The principles consist of content, format, writing stem, writing option, and item shell.

Content

The reviewers found that the test fulfills some parts of content's principle, i.e., each item focuses on a single type of information and single cognitive demand because each item is constructed to a particular component of grammar tested—gerunds & infinitives, noun clauses, adjective clauses, or adverbial clauses. It can be seen from Table 2 that every number presents a particular grammar component and cognitive processes (Butler, 2018; Haladyna, 2018). Each item displays different content from other items. It goes with the third principle which says each item's content should be distinct from those of other items (Haladyna, 2018). The items constructed do not give any test takers to express their opinion since they just have to select the option without adding anything on the answer sheet. It goes to the fifth principle which says avoid giving test takers to express their opinion unless it is so qualified (Haladyna, (2018). The items do not use trick item since all of them require test takers to apply their knowledge about the subject matters. It matches with the sixth principle—avoid trick items that mislead the test takers into selecting the incorrect answer (Haladyna, 2018).

Format

Regarding the format, the reviewers expressed that the items has already followed the principle of a good format. The format follows the first principle of formatting that says it is preferable to format items vertically rather than horizontally (Haladyna, 2018). The test also edited many times and proofread by the reviewers, so it supports second principle of formatting i.e., editing and proofreading should be done on all items (Haladyna, 2018). The test has also been suited for intermediate level as stated in the course description, thus it follows the third principle—the difficulty level of the test should be appropriate for the students' level (Haladyna, 2018).

Writing Stem

For writing stem, the reviewers considered that the test has also adhered to first principle of writing stem that says minimize the number of words in the item's stem and avoid using window dressing. Each stem was designed as concise as possible to meet the principle, hence it does not lose the extent. Clarity is a value as long as the essence of the topic need to be tested is communicated (Haladyna, 2018). For structure part, the stem design had followed the second principle—state the primary notion in the stem, not in the options (Haladyna, 2018). However, some items in the written expression might not follow the principle since the stem also contains the options, thus some of them are very long. All the items do not use negative language which goes with the third principle, use positive language instead of negative ones like "not," "except," or "false." The usage of negative keywords in the stem, according to the authors of testing textbooks and certain studies, has a detrimental influence on student replies (Haladyna, 2018).

Writing Option

In writing option, the reviewers affirmed that some parts of the principle have confirmed the guidelines. The distractors were considered realistic which suites the first principle—distractors should be realistic incorrect answers, rather than absurd or otherwise improbable options (Haladyna, 2018). The distractors have also followed the second principle that says there should only be one correct answer in any of the SR forms (Haladyna, 2018) since every item has only one correct answer. However, some correct answers were not distributed in different spot each time that discorded the third principle, i.e., correct answer should appear in a different spot each time (Haladyna, 2018). Consequently, the researcher revised the distribution of the correct answers in different item each time and made the number of correct answers of the option equal for each component tested. The reviewers criticized that some options were similar that they needed revision because of the similarity with other option. It can be said that it did not follow the fifth principle—options should be distinct from one another in terms of quality and substance; avoid options that are similar in quality or content (Haladyna, 2018). The reviewers said that the options have followed the sixth principle, option should avoid term like "all of the above" and "none of the above" (Butler, 2018; Haladyna, 2018) since there is no option stating this way. All the option have also used positive language not the negative one which follows the seventh principle—positive, not negative, language should be used in the options (Haladyna, 2018). The options have also followed the eighth principle which says avoid hints that direct to the correct answer (Haladyna, 2018) because there is no hint for every option. Test takers have to analyze carefully before choosing the correct or incorrect parts.

Item shell

The only principle of item shell is to avoid humor. The reviewers thought that all the items and options have been constructed carefully to avoid carelessness in constructing the items and the options. Thus, the test had met the principle of item shell by avoiding humor (Haladyna, 2018) in the items and options.

Based on the review given by three reviewers, some stems and options have to be revised to meet the principles/guidelines of constructing a good test. the revision is presented in Table 2.

Table 2 The Revision after Reviewed

Item (Number)	Reason	Option (Number)	Reason
12, 13, 14, 21, 24, 25, 28,	To make fair distribution of the number of items for each component (gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses. Each should have 10 items based on the blue print.	16c, 16d, 21a, 21d, 22d, 22b, 24a, 24c, 25c, 25b, 28b, 28d, 38b, 38c	To make fair distribution of the options to meet different option in different spot.
16	It does not relate with any of the grammar component tested (contains passive voice)		

CONCLUSION

The findings revealed that the test had showed content validity. The test items designed have fulfilled the main objective of the course, i.e., to provide students with a sound knowledge and apply the knowledge of Intermediate English grammar in comprehension and production. The test had also reflected expected materials of Intermediate English Grammar course comprising gerunds & infinitives, noun clauses, adjective clauses, and adverbial clauses. The test has been designed to meet the students' level—intermediate level that have met the standard competence of the students. The test has also been designed as TOEFL-like to meet the additional objective of the course to give students experience to face Structure and Written Expression part of TOEFL.

The test was considered practical based on the review given by the three lecturers selected as the peer reviewers. The time allotment is 60 minutes that is said enough for 40 items. The test consists of 5 pages that is considered as economical (not costly) for photocopying the booklet. It is also reusable since the answer sheet

is given separately, and it can also be administered online that make it more economical. It is also easy to administer and score the test. it can be administered by any teacher/lecturer who teaches the same level with the same material. It also does not spend to much time to score because it just needs a few minutes to score each answer sheet. Thus, the test is practical in term of time spent, funding and administering.

The test was considered as well-designed test following guidelines of constructing a good test (Butler's (2018) and Haladyna's (2018) principles). It has followed 4 out of 6 principles of the content guidelines. It has supported all the three principles of format guidelines. It has reinforced all the three principles of writing stem guideline with some revisions on some parts of the stem (item number 12, 13, 14, 16, 21, 24, 25, and 28). These items have already been revised to get fairness in the distribution of the materials of the test. The test has maintained 5 out of 8 principles of writing option guidelines. The revision on the distribution of the option has already been made (7 options i.e., 16c, 16d, 21a, 21d, 22d, 22b, 24a, 24c, 25c, 25b, 28b, 28d, 38b, 38c) to keep different correct answer in different spot. It also has followed the last principle of item shell by avoiding humor. To sum up, it can be said that the test is well designed and appropriate as a summative test instrument to measure students' achievement on Intermediate English Grammar course.

However, the test needs further processes/stages to meet empirical validity and reliability. Thus, future research is suggested to conduct try-out to measure the empirical validity and reliability. It opens opportunities for future research to give evidence or process to next stages of the test to provide its empirical validity and reliability.

REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press, Ltd.

Brown, H. D. (2004). *Language assessment principles and classroom practices*. Pearson Education Inc.

Brown, J. D. (1991). Do English faculties rate writing samples differently? *TESOL Quaterly, 25*, 587-603.

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. Heinle & Heinle Publishers.

Butler, A. C. (2018). Multiple-choice testing in education: Are the best assessment practices for also good for learning? *Journal of Applied Research in Memory and Cognition, 7*(3), 323-331. https://doi.org/10.1016/j.jarmac.2018.07.002.

Djiwandono, S. (1996). *Tes Bahasa dalam Pengajaran*. Penerbit ITB Bandung.

Downing, S. M. & Haladyna, T. M. (Eds.). (2011). *Handbook of test development*. The Taylor & Francis e-Library.

English Department. (2020). *Catalogue of department of English 2020 edition*. Faculty of Letters, Universitas Negeri Malang.

Fenn, J., Tan C.-S., & George, S. (2020). Development, validation and translation of psychological tests. *BJPsych Advances 33*, 1-10. doi: 10.1192/bja.2020.33.

Finney, D. (2002). The ELT curriculum: A flexible model for a changing world. In J. C. Richards & W. A. Renandya (Eds.). *Methodology in language teaching: An anthology of current practice* (pp. 69-79). Cambridge University Press.

Gronlund, N. E. (1998). *Assessment of student achievement*. Sixth Edition. Allyn and Bacon.

Haladyna, T. M. (2018). *Developing test items for course examinations*. IDEAedu.org. Arizona State University.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item. *Educational and Psychological Measurement, 53*(4), 999–1010. https://doi.org/10.1177/0013164493053004013.

Haladyna, T. M., & Rodriguez, M. R. (2013). *Developing and validating test items*. Routledge.

Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.

Kumaravadivelu, B. (2001). Toward a postmethod pedagogy. *TESOL Quarterly, 35*(4), 537–560.

Marzano, R. J. & Kendall, J. S. (2007). *The new taxonomy of educational objectives*. Second Edition. Sage Publications Inc.

Mousavi, S. A. (1999). *Dictionary of language testing*. Second Edition. Rahnama Publications.

Mousavi, S. A. (2002). *An encyclopedic dictionary of language testing*. Third Edition. Tung Hua Book Company.

Nation, I. S. P., & Macalister, J. (2010). *Language curriculum design*. Routledge.

Nurgiantoro, Burhan. (2001). *Penilaian dalam Pengajaran Bahasa dan Sastra*. BPFE- Yogjakarta.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice test items: A meta- analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2) 3–13). https://doi.org/10.1111/j.1745-3992.2005.00006.x.

Urbina, S. (2004). *Essentials of psychological testing*. John Wiley & Sons.

Weir, C.J. (1990). *Communicative language testing.* Prentice Hall.

Yıldırım, B. & Topkaya, E. Z. (2020). Clarificative evaluation of elementary level grammar program of a tertiary level preparatory class in Turkey. *Studies in Educational Evaluation 65*, 100862. https://doi.org/10.1016/j.stueduc.2020.100862