The Authenticity of EFL Summative Test-Task Items at a Senior High School in West Seram, Maluku Province, Indonesia

Hardianto Hitimala, Universitas Pendidikan Indonesia, Indonesia
Susi Septaviana Rahkmawati, Universitas Pendidikan Indonesia, Indonesia

The Asian Conference on Education 2024
Official Conference Proceedings

Abstract
The objective of this study was to find out the authenticity of the EFL summative test at one of the Senior High Schools in Seram, Maluku, Indonesia. The research design was a descriptive quantitative study. The data was collected from the documents of summative test items that consist of two major parts; forty multiple-test items and three open-ended questions. The instruments rubric was constructed to assess test authenticity. The analysis focused on test task authenticity, covering the setting, structure of the communicative event, input, and expected response. The rubric was constructed based on the characteristic of authenticity proposed by Bachman and Palmer (1996), and Brown and Abeywickrama, (2018). The finding showed that the summative test consists of 13 tasks using *9 short reading texts, 3 open-ended questions, and 2 short dialogue texts* that served different social functions such as s*hort message, self-introduction, recount text, announcement, narrative, argumentative text, invitation, recount text of personal experience, business letter, greeting,* and *turn taking*. Concerning task authenticity, there were 32 out of 43 items on all task components were identified as high while 11 items were low. The cumulative analysis of all tasks indicated that 77% (10 tasks covering 32 items) were highly authentic, while 23% (3 tasks covering 11 items) were low authentic.

Keywords: Test Task, Authenticity, EFL, Summative Test

1. Introduction

Evaluation is an integral part of teaching English in schools, with testing being the primary instrument teachers use. Testing encompasses several interrelated aspects, including goal, and activity. The goal refers to the overall objective of the lesson at the end of the semester, while the activity pertains to classroom activities. Test policymakers have implemented testing to effect educational change (Shohamy, 2001 in Broad, 2003). Testing is an ongoing process employed by teachers to monitor and guide progress through quizzes, homework, and informal tests (Airasian, 2001). Like patients requiring a doctor's diagnosis, students need tests to assess their academic progress. Tests serve as a policy tool to measure student achievement and the success of teaching-learning programs and diagnose students' strengths and weaknesses. In designing and constructing tests, teachers must ensure that the level of difficulty is appropriate for their students (Hughes, 2003).

Generally, in Indonesian schools, a summative test is frequently conducted to show the standard, which the students have reached with other students at the same stages. The test is used at the end of the semester term to measure what has already been achieved both by groups and by individuals including tests, projects, and formal tests (Rajhy, 2014, cited in Sugianto, 2017). Besides, formulating authentic content is very important when designing tests for students, especially for EFL students. Bachman and Palmer (1996) argue that authenticity is the degree of correspondence between the characteristics of a given language test task and the features of the language task. It means that a language should connect to the real world. Furthermore, Bachman (1991) and Liu (2005) categorize test authenticity into two aspects: test text authenticity and test task authenticity. These aspects cover elements such as setting, test rubric, input, expected response, and the relationship between input and response (Bachman & Palmer as cited in Liu, 2005). These elements mean that test items should reflect the naturalness of language, and the relevance of the topic, be engaging, and represent real-world scenarios (Liu, 2005). A test task is realistic when we find it is authentic. In fact, many kinds of test items do not mimic what people do in real life. They might be fake or unnatural because they focus on a grammar rule or a word. The order of questions that have nothing to do with each other is not authentic. It is easy to find passages in skill tests that are not like real-world texts. (Brown & Abeywickrama, 2018). Concerning the authenticity of a test, Fauziah (2019), asserted that in designing effective and efficient tests, Indonesian teachers tend to consider aspects of validity, reliability, and practicality, rather than concerned about authenticity aspects, which produce tests that are not sufficient to fulfill the pedagogical elements.

Studies on summative test items analysis in Indonesia have been explored by (Bernasela, 2014; Ardhian et al, 2016; Sugianto, 2017; Semiun, & Luruk, 2020, Wisrance, & Napitupulu, 2022). There have been only a handful of studies (Fauziah, 2019), scrutinizing the authenticity of summative assessment in avocational schools in Bandung City. However, it is very rare for research related to summative test items analysis from an authenticity perspective. Those previous researchers focused on validity and reliability rather than authenticity. Moreover, the researcher was drawn to study the authenticity of English test items on summative tests at the Senior High School of West Seram, Maluku, Indonesia, to obtain more information about the quality of the test items. Some students at this school reported problems with the test content and stated that the questions were not suitable for their abilities. Many students complained about the suitability and relevance of the language test. The problem could be seen directly from the students' explanations through the researcher’s preliminary study (short interview). Most of them complained about the

suitability, relevance, and easy language of the test. The students said that the test still lacked authenticity and made it difficult for them to succeed. Therefore, in further research, the researcher is concerned with analyzing the authenticity of the EFL summative test designed according to one principle of language assessment, mainly focusing on the test task authenticity items on the English teacher test made (Bachman & Palmer, 1996; Brow & Abeyickrama, 2018). The researcher used descriptive quantitative. The researcher was curious about how far the effectiveness of tests was created for students. Therefore, the research questions are formulated as follows:

1. How authentic is the construction of the test task found in the EFL summative test?

2. Literature Review

2.1 Summative Test

Assessment has two purposes: to support learning and to summarize learning. The formative assessment is used to guide teaching and learning. On the other hand, the summative assessment is used to record and report (Allen, 2004). The summative test focuses more on students' achievement. The outcome of a summative test will be used to give grades to students. The summative test involves gathering evidence about students' achievement in a systematic way to be reported at a specific time based on teachers' professional judgment (Harlen, 2004). Summative assessment is a way of measuring or summarizing what students have learned. It evaluates how well students have achieved their goals, but it does not necessarily help them improve in the future (Brown, 2003). Spolsky and Halt (2008) also explain that summative assessment, or assessment of learning, is less detailed and aims to assess the outcomes of educational programs or students. Therefore, summative assessment is used to test different language skills and learners' achievements.

2.2 Authenticity

According to Bachman and Palmer (1996), the term authenticity as used in the context of testing can be understood to mean the degree to which a given test and set of materials corresponds to 'real life' context and interactions (Shomoossi & Tavakoli, 2010). Authenticity is an important quality for test development (Lynch, 1982). Morrow (1991) points to the overriding importance of authenticity, and Wood (1993) considers it as one of the most important issues in language testing. Also, Bachman and Palmer (1996) see authenticity as a critical quality of language tests (Shomoossi & Tavakoli, 2010). It means that when making and choosing tests, authenticity should be a priority in the practical phase of test creation. The main focus should be on controlling how authentic it is, which means how well a language test matches a real-life task.

2.3 Test Task Authenticity

Widdowson (1979; 1978) and Skehan (2003) point out that task authenticity entails "the learner's reaction or response". Morley (2000) elaborates further by stating that task authenticity is contingent on whether learners are engaged by the task. Therefore, a task may be authentic in relation to real-world situations, but it may seem inauthentic to some groups of learners. Task authenticity is a concept that relates to how well a language test simulates the real-world situations and tasks that test-takers will encounter in their target language use (TLU) domain. Task authenticity is important for measuring the test takers' ability to use language for communicative purposes and to engage with meaningful and relevant content.

Task authenticity can be divided into two types: situational authenticity and interactional authenticity.

2.3.1 Situational Authenticity

Situational authenticity is the perceived relevance of the test method characteristics to the features of a specific target language use situation (Bachman & Palmer, 1996, cited in Purpura & Kunnan, 2024). Thus, for a test task to be perceived as situationally authentic, the characteristics of the test task need to be perceived as corresponding to the features of a target language use situation. For example, one set of test method characteristics relates to certain characteristics of vocabulary (e.g., infrequent, specialized) and topics (e.g., academic, technical) included in the test input. If test takers were specialists in engineering, the inclusion of technical terms and topics from engineering would likely tend to increase the situational authenticity of the test. In contrast, we define the situational authenticity of a given test task in terms of the distinctive features that characterize a set of target language use tasks. Thus, in designing a situationally authentic test, we do not attempt to sample actual tasks from a domain of non-test language use but rather try to design tasks that have the same critical features as tasks in that domain. Language testers and teachers alike are concerned with this kind of authenticity, for we all want to do our best to make our teaching and testing relevant to our students' language use needs. For a reading test, for example, we are likely to choose a passage whose topic and genre (characteristics of the test input) match the topic and genre of material the test user is likely to read outside of the testing situation. Or, if the target language use situation requires reciprocal language use, then we will design a test task in which reciprocity is a characteristic of the relationship between test input and expected response.

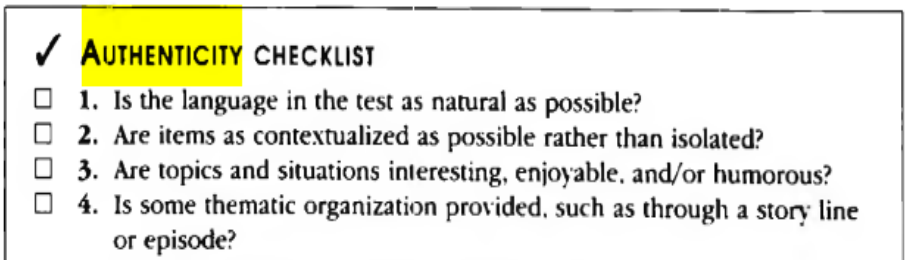
2.3.2 Interactional Authenticity

Interactional authenticity is essentially a function of the extent and type of involvement of task takers' language ability in accomplishing a test task (Widdowson, 1978). Assessing interactional authenticity and designing tasks that are interactionally authentic, however, is more complex, since this requires us to consider both the characteristics of the test task and the components of the test taker's language ability.

2.4 Theoretical Framework

Bachman (1990) and Bachman and Palmer (1996) developed a framework to describe language tasks, incorporating five key features. First is *the setting*, which refers to the environment where the task occurs, including details like location, participants, and timing. Second is *the rubric*, which includes the task instructions, detailing the situation, what students are expected to do, and how they will be evaluated. Third is *the input*, which encompasses the material that students need to process, whether it's auditory, visual, verbal, or nonverbal. Fourth is *the expected response*, which outlines what students are supposed to do with the given input. Lastly, *the relationship between input and response* is examined, considering factors such as the level of interaction, the amount of information to be processed, and the reliance on prior knowledge. This framework ensures that language tasks are realistic and effective in assessing language skills. In relation to this, Bachman and Palmer simplified the five points above into a checklist that can be used to analyze test task items.

- **Setting (Are items as contextualized as possible rather than isolated)?**
  Good test items are set within a certain context, not just stand-alone sentences or words. This helps students understand meaning within a broader situation. Example: A question asks students to complete a conversation in a restaurant, rather than translating random words.
- **Structure of communicative event (Is some thematic organization provided, such as through a storyline or episode?)**
  Questions that follow a storyline or theme make the test more structured and easier to follow. Example: A test includes a series of questions that tell a story about someone's journey, from planning a trip to returning home.
- **Input (Is the language in the test as natural as possible?)**
  The language used in the test should sound like everyday speech or writing, not overly formal or artificial language. Example: A question asks students to write an email to a friend, instead of a rigid formal letter.
- **Expected Response (Are topics and situations interesting, enjoyable, or humorous?)**
  Engaging or enjoyable topics make students more motivated to complete the test. Example: A question is about a vacation or popular movie instead of overly technical or boring content.



✓ AUTHENTICITY CHECKLIST

☐ 1. Is the language in the test as natural as possible?
☐ 2. Are items as contextualized as possible rather than isolated?
☐ 3. Are topics and situations interesting, enjoyable, and/or humorous?
☐ 4. Is some thematic organization provided, such as through a story line or episode?

Figure 1: Task Authenticity Guideline

3. Method

3.1 Research Design

The current study employed a quantitative research design to investigate the authenticity of the EFL summative test construction at the Senior High School in West Seram, Maluku Province, Indonesia. The investigation was done by studying the entire summative test focusing on the test tasks.

3.2 Data Source

The data used in this research was the summative test items document prepared by a teacher for first-grade students at the Senior High School of West Seram, Maluku, Indonesia, for the 2023-2024 academic year. The summative test has 43 question items, consisting of 40 multiple-choice test items and 3 open-ended questions. They were constructed by an English teacher at the school to measure students' English competence. The multiple-choice test is presented through three big tasks, consisting of short reading passages, short dialogues, and open-ended questions. Eight short reading texts indicate different social functions such as 1) a short message from a friend, 2) self-introduction (introduce name, address, age, favorite subject, and hobby), 3 and 4) recount texts (recount text 1&2), 5) announcement, 6) Folktale (Narrative) 7) argumentative text on 'smoking', 8) Invitation, 9) business letter (order good).

Meanwhile, the test has six *short dialogue* texts related to speech acts such as greeting, self-introduction, and turn-taking. Three *open-ended questions* are also constructed to test students' writing skills. The test items are classified in the following table.

Table 1: Reading Text (Monologue) Tasks

No	Theme	Item number of test					Total item
1	Short message	1	2	3	4		4
2	Self Introduction (introduce name, address, age, hobby)	8	9	10	11	12	5
3	Recount Text 1 (personal experience)	16	17	18	19	20	5
4	Recount Text 2	36	37	38			3
5	Announcement	25	26	27			3
6	Narrative (tale)	28	29	30			3
7	Argumentative text on " smoking"	31	32	33			3
8	Invitation	34	35				2
9	Business letter (order good)	39	40				2
Total Items							30

Table 2: Short Dialogue Text Tasks

No	Theme	Item number of test				Total item
1	**Greetings** Focus on opening and closing and the meaning of the phrasal expression (language function)	5	6	7		3
2	**Greeting (introducing others)**: language function, meaning, comprehension.	13	14	15		3
3	**Turn-taking** (matching)	21	22	23	24	4

Table 3: Open-Ended Question (Essay) Tasks

No	Theme	Item	Total
1	Self-description, Story genre, Describe school, and friends	1-3	3

3.3 Data Collection

In collecting the data, we used checklists as a rubric that was compiled according to the theory of authentic tests proposed by Bachman and Palmer (1996), and adjusted to the test specifications of the EFL summative test of the Senior High School of West Seram, Indonesia. The checklist or rubric was constructed to assess the authenticity of the test task,

consisting of 1) the contextualization (setting), 2) thematic organization (structure of communicative event), 3) the natural language use (input), and 4) interesting and enjoyable topics and situations (Expected response). These components of authenticity are employed to match the test items and ensure whether or not the items are suitable for authenticity. Table 4 is an example of the checklist used. It has four columns, the first for the test items, and the second for the element authenticity and description. The third and fourth columns were for the level of authenticity (Yes or No) utilized to match the component of authenticity.

Table 4: Sample of Checklist of the Test Task Framework

Test Items	**Theme task authenticity**	**Simplification**	**Authenticity level (Yes/No)**
Test Item 1	**Are items as contextualized as possible rather than isolated?** *(Setting)*	Contextual Items	**-**
	Is some thematic organization provided, such as through a storyline or episode? *(Structure of communicative events)*	Thematic Organization	**-**
	Is the language in the test as natural as possible? *(Input)*	Natural Language	**-**
	Are topics and situations interesting, enjoyable, and/or humorous? *(Expected response)*	Interesting topic	**-**

(Source: Adjusted from Bachman and Palmer 1996; Brown and Abeywickrama, 2018)

3.4 Data Analysis

As shown in checklist Table 4, it is utilized to check the content authenticity of the test items following task characteristics. To simplify the computation, the criteria for each element found in the test items were given "Yes/No" which was interpreted as (Yes: 2 which means high authentic & No: 1 which means low). This is enough to assess whether the items were authentic enough (Bachman & Palmer, 1990, 1991; Purpura & Kunan, 2024). This is also a common and frequent scoring and rating used in research, especially for data calculation and interpretation (Krippendorff, 2018) to quantify and capture the overall authenticity level of each item in a quantifiable way, making it easier to determine whether the item generally reflects authentic use. All the scoring data were then calculated using descriptive statistics manually to summarize and organize the data's characteristics by looking at measures like frequency and percentage (Bland, 2015). The analysis results were shown in tables and percentages, followed by explanations. The following table is the analysis example.

Table 5: Sample of Analysis

Components	Description	Low Authentic (1)	High Authentic (2)
Setting	Are items as contextualized as possible rather than isolated?	1	2
Structure of communicative event	Is some thematic organization provided, such as through a storyline or episode?	1	2
Input/Feature of context	Is the language in the test as natural as possible?	1	2
Expected Response	Are topics and situations interesting, enjoyable, and/or humorous?	1	2
Total Score		4 (Low)	8 (High)

As shown in Table 5, there are four criteria employed to measure authentic test tasks. Each criterion was given a score, 1 for low and 2 for high (Bachman & Palmer, 1990, 1991; Purpura & Kunan, 2024). All scores were analyzed using descriptive statistical analysis to summarize characteristics of a data set by using measures of frequency and percent (Bland M., 2015). To reach the level of authenticity in the form of a percentage, the researcher put the range category as follows: *4-6 is Low authentic and 7-8 is High authentic.* The data analysis process starts by examining each item in the test tasks using the given scoring criteria (See Table 5). After analyzing all the items, the researcher calculates the authentic results based on three task categories: *reading text, short dialogue text,* and *open-ended questions.* The calculation process involves counting the number of tasks that fall within a specified score range and then classifying how many tasks are high and how many are low. Then, the results of the analysis were presented in the form of tables, descriptions, and percentages.

4. Result and Discussion

4.1 Result

This section provides the information to answer the following research questions: How authentic is the construction of the test task found in the EFL summative test? The authenticity of a test item can be measured using the task authenticity framework by Bachman and Palmer (1996); and Brown and Abeywickrama, (2018). The detailed analysis of task authenticity within the EFL summative test reveals varying levels of authenticity across different themes. Each task was evaluated based on four elements of authenticity: *Setting (Contextualization), Structure of Communicative Events (Thematic Organization), Input (Natural Language),* and *Expected Response (Interesting Topic).* The analysis of the data is presented in the following descriptive statistic tables.

4.1.1 Reading Text (Monologue) Task

The analysis of task authenticity within Reading text items shows that the nine themes or tasks in this text reveal varying levels of authenticity.

Table 6: Result of Reading Text Findings

No	Theme	Authenticity Level (Overall Average)
1	Short message	**6.5** (Low)
2	Self Introduction (introduce name, address, age, hobby)	**5.8** (Low)
3	Recount Text 1 (personal experience)	**8** (High)
4	Recount Text 2	**6.66** (High)
5	Announcement	**6.66** (High)
6	Narrative (tale)	**8** (High)
7	Argumentative text on " smoking"	**6.66** (High)
8	Invitation	**8** (High)
9	Business letter (order good)	**6** (High)
N=9 Tasks		78 % High Authentic (7 Tasks)
		22 % Low Authentic) (2 Tasks)

As shown in Table 6, out of nine reading text tasks, seven (78%) were rated as highly authentic. These included tasks based on *recount texts, narrative texts, argumentative texts, announcements, invitations,* and *business letters*. Such tasks scored high in authenticity because they incorporated real-world language use and were contextually relevant. However, two tasks, *short message,* and *self-introduction,* were rated as low in authenticity (23%), indicating areas where the test content may not fully align with real-life language scenarios.

In terms of *recount text,* this theme consists of recount text 1 on personal experience (five items, each achieving a perfect score of 8, indicating high authenticity), and recount text 2 (three items achieving a score of 6.66, indicating moderate to high authenticity). The tasks in Recount 1 are well-c*ontextualized*, with a clear *thematic organization* and use of *natural language*, reflecting real-life personal experiences. In Recount 2, two (items 36 & 37) out of three, scored high (8/High), showing effective *contextualization* and *natural language* use. However, item 38 scored low (4/Low), suggesting a lack of *thematic structure* and *natural language* use. The average score for this theme is 6.66, indicating a moderate to high level of authenticity. In the *Narrative Text,* all three tasks in this theme received high scores (8/High) across all elements, showing strong contextual relevance, interesting topics, and natural

language. In Argumentative Text, two items (31 and 32) scored high (8/High), demonstrating thematic organization and natural language suitable for argumentative discourse. However, item 33 scored low (4/Low), lacking thematic relevance and natural language. The average score is 6.66, indicating moderate authenticity. Concerning, Announcement text Items, this theme includes three tasks, with two items (25 and 26) rated as highly authentic (8/High). They are contextualized and align with the thematic structure of real-world announcements. Item 27, however, scored low (4/Low), due to less natural language and less interesting content. The average score of 6.66 indicates moderate authenticity. In relation to Invitation Text, both items in this theme received high authenticity scores (8/High), suggesting effective contextualization, thematic structure, and engaging content that reflects real-world invitation contexts. The average score is 8, showing a high level of authenticity. In Business Text, the two tasks in this theme show contrasting results. Item 39 received a high authenticity score (8/High), reflecting appropriate contextualization and natural language typical of business communication. However, item 40 scored low (4/Low), due to limited thematic organization and lack of natural language. The average score is 6, indicating a moderate level of authenticity.

Meanwhile, in *short massage text*, two items (1 and 2) received the maximum score of 8, indicating high authenticity, as they provide contextualized scenarios, thematic structure, natural language, and interesting content. However, items 3 and 4 scored lower (5/Low) due to lacking *natural language* and *thematic organization* consistency. The average score across these items is 6.5, suggesting a moderate level of authenticity for this theme. In Self-Introduction, the five items on self-introduction vary in authenticity, with three items (9, 10, & 12) achieving a high authenticity score of 7. These items use *natural language* and *interesting and relatable topics* for students. However, items 8 and 11 scored lower (4/Low), indicating limited *contextualization and thematic connection*. The overall average for this theme is 5.8, highlighting a generally low to moderate authenticity level.

4.1.2. Short Dialogue Text

Table 7: Result of Short Dialogue Text Findings

No	Theme	Authenticity Level
1	**Greetings** Focus on opening and closing and the meaning of the phrasal expression (language function)	7.33 (High)
2	**Greeting (introducing others)**: language function, meaning, comprehension.	8 (High)
3	**Turn-taking** (matching)	5 (Low)
N= 3 Tasks		66 % High Authentic (2 Tasks)
		33 % Low Authentic (1 Task)

As shown in Table 7, among the three short dialogue tasks, two (66%) were rated as highly authentic, specifically, those *focused on greetings* and *introducing others*, which align well with everyday conversational situations. The *"Turn-taking"* task, however, received a lower

authenticity score (33%), suggesting it may lack situational relevance or sufficient interactional context to reflect real-world communication.

For the *Greeting text items*, three items were assessed. Items 5 and 6 achieved high scores across all four elements, each receiving a score of 8, indicating strong contextualization, clear thematic structure, use of natural language, and engaging topics. These items were well-designed to simulate real-life greeting scenarios, allowing students to experience relevant language use. Item 7 also achieved a high authenticity score of 6, though it scored slightly lower in *input (natural language)* and *expected response*, suggesting that while it was contextually appropriate, it lacked the full *natural language* flow seen in real conversations. Overall, the greeting tasks have an average authenticity score of 7.33, indicating that this theme successfully incorporates realistic language use and context, making it one of the stronger areas of authenticity in the dialogue section. In the *Short Dialogue of Self-Introduction tasks*, all three items (13, 14, & 15) received the maximum score of 8, reflecting high authenticity across all elements. These items effectively use *natural language*, t*hematic organization,* and *contextual settings* that mimic real-life self-introduction situations.

The *Turn-Taking* tasks display a wider range of scores, with mixed results across the four items. Item 21 achieved a high score of 8, showing strong contextualization and natural language, as it closely reflects real-life conversation exchanges where turn-taking is essential. However, the remaining items (22, 23, & 24) scored low (each receiving a score of 4), indicating weaknesses in thematic organization and natural language. These items lack the dynamic and interactive language elements typical of natural turn-taking conversations, which may make them feel less authentic and disconnected from real-world dialogue. This theme has an average score of 5, suggesting limited authenticity in capturing the essence of turn-taking, as the tasks may feel more mechanical and less engaging for students.

4.1.3 Open Ended-Questions (Essay)

Table 8: Result of Open-Ended Question Findings

No	Theme	Authenticity Level
1	Self-description, Story genre, Describe school, and friends	8 (High)
N=1		100 % High Authentic

As shown in Table 8, this section contained a single task focused on self-description, story genre, and descriptions of school and friends. It received a high authenticity rating, as it aligns well with students' personal experiences and allows for expressive language use relevant to real-world contexts. This section includes three items (1, 2, and 3), each receiving a very high score of 8, indicating that these test items are highly authentic in reflecting real-world contexts. Items 1, 2, and 3 show strengths across all four elements. The high score in *Setting (Contextual Items)* indicates that these questions are highly relevant to real-life situations, creating a realistic context for the test takers. The score of 2 for *Structure of Communicative Events* for each item reflects that the thematic organization of the questions is clear and logical, creating an easy-to-follow communication pathway. In terms of *Input (Natural Language),* all three items use natural and easily understandable language, consistent with how people communicate in everyday life. Finally, the high score in *Expected Response (Interesting Topic)* suggests that the topics presented in these items are engaging, encouraging test takers to provide thoughtful and relevant responses. An average score of 8/High (100%), indicates that the authenticity level of these test tasks is very high.

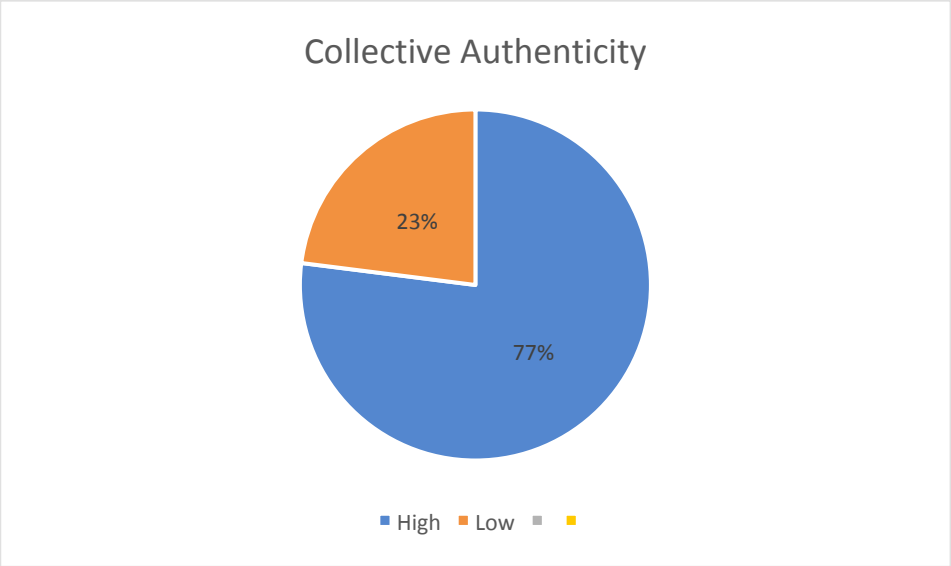


Figure 2: Collective Authenticity

Collectively, 10 tasks out of 13 tasks (77%) were at the level of high authenticity, while only 3 tasks (23%) were at the level of low authenticity.

4.2 Discussion

Based on the findings of this study on the authenticity of English summative test items at a Senior High School in West Seram, Maluku, Indonesia, the majority of test items exhibit a high level of authenticity. Approximately 77% of the test items received high authenticity scores, suggesting they align well with real-world language use contexts and meet the criteria of contextual relevance, thematic organization, natural language use, and engaging content. However, 23% of the test tasks still exhibit low authenticity, indicating areas for improvement in test design.

Most of the highly authentic tasks align positively with authenticity principles outlined by Bachman and Palmer (1996) and Brown and Abeywickrama (2018), who emphasize that language tests should closely replicate real communication situations (Liu, 2005). For example, tasks like "Self-introduction" and "Narrative" resonate with real-life scenarios where students introduce themselves or recount personal experiences, fulfilling both situational and interactional authenticity. Open-ended tasks, such as self-description or descriptions of school and friends, received particularly high authenticity scores, reflecting their suitability for evaluating students' language abilities in practical contexts. These types of tasks not only assess language proficiency but also increase student motivation by allowing them to relate test content to their own lives. Hood (1984) in Joy (2011) have placed the importance of authentic text on language tests as they believe that the authentic language of the text is natural and hence can easily connect students to the real world.

However, certain tasks, particularly those involving isolated dialogues or contextually less relevant themes, did not meet high authenticity criteria. For instance, tasks that emphasize rote grammar or vocabulary without embedding these elements into a meaningful communicative framework fall short in terms of authenticity, making them less relevant and potentially less engaging for students. Additionally, tasks like "Turn-taking" in dialogue settings received lower authenticity scores, which may stem from a lack of situational context or thematic connection. To improve these, more natural conversational scenarios or

interactive elements reflecting real-life language dynamics could be incorporated, as Skehan (2003) suggests regarding task authenticity.

Given these findings, test designers are encouraged to review and refine tasks that fall short of authenticity standards, ensuring that each task replicates real-world language use as closely as possible. Emphasizing thematic continuity, contextually relevant input, and engaging topics can further enhance the test's usefulness. While authenticity is essential, balancing it with other factors such as reliability and practicality is also necessary, especially in the high-stakes context of summative assessments. Nevertheless, ensuring tasks are both meaningful and communicatively relevant can significantly enhance their pedagogical value, making the test more relatable and manageable for students (Brown & Abeywickrama, 2018).

These findings somewhat contradict students' statements from the preliminary study, where they expressed that the test topics and language were difficult to understand. This discrepancy suggests that students' literacy and vocabulary skills in English are still limited, which likely affects their ability to comprehend test materials effectively. Consequently, there is a need to enhance students' English vocabulary and reading comprehension skills to better prepare them for assessments and help bridge the gap between test content and students' understanding. Strengthening vocabulary acquisition and literacy skills will enable students to approach test items with greater confidence and improve their overall performance in language assessments.

5. Conclusion

The analysis revealed that the test included 13 tasks based on 9 short reading texts, 3 open-ended questions, and 2 short dialogue texts that had different social purposes such as short message, self-introduction, recount text, announcement, narrative, argumentative text, invitation, recount text of personal experience, business letter, greeting, and turn taking. Regarding task authenticity, 32 items out of 43 in all task components were classified as high authentic, and 11 items were low authentic. The overall analysis of all tasks showed that 77% were highly authentic, while 23% were low authentic. These findings imply that the authenticity of test questions needs to be enhanced by revising their design to be more contextual, using natural language, and reflecting real-life situations that students encounter. Teachers should also receive specialized training to design questions based on authenticity principles as Bachman and Palmer proposed (1996). This process should involve ongoing training for teachers and be part of professional development programs to ensure consistency in applying authenticity in schools. Collaboration with language assessment experts is also a strategic step to evaluate and refine test questions. This can be achieved by adopting a more systematic and evidence-based evaluation approach. Integrating authenticity into national standards can enhance the relevance and impact of learning outcomes.

This study identifies several limitations and offers practical suggestions to improve the test quality and relevance. This research focuses exclusively on a single high school in West Seram, Maluku, Indonesia, which means its findings may not be fully applicable to schools with varying social and cultural contexts. Additionally, data collection occurred over just one exam period, limiting understanding of how test authenticity may shift over time or adapt to changing educational objectives. To address these, future research should involve multiple exam periods to observe changes in authenticity, use standardized evaluation criteria to reduce scoring subjectivity, and apply advanced statistical methods for a more detailed analysis. Expanding test tasks to include realistic scenarios from students' daily lives would

also enhance authenticity, as would teacher training on assessment design aligned with authenticity frameworks like those by Bachman and Palmer. Regular evaluations and collaboration with language assessment experts can ensure tests reflect real-life language use, and literacy programs in vocabulary and reading comprehension could better prepare students to engage with authentic test content.

Acknowledgment

This work was supported by Lembaga Pengelola Dana Pendidikan (Indonesian Education Endowment Fund for Education) under the Ministry of Finance of the Republic of Indonesia (2023).

References

Allen, M. (2004). *Assessing academic programs in higher education bolton*. Anker Publishing Company, Inc.

Broad, (2003). The power of tests, a critical perspective on the uses of language texts. *Journal of Writing Assessment.*

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

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly, 25*(4): 671-704.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*(4): 453-476.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.

Brown, H. D. (2003). *Language assessment principles and classroom practices*. Oxford University Press.

Brown & Abeywickrama (2018). *Language assessment principles and classroom practices*. Pearson.

Fauziah (2019). The analysis of authenticity in summative assessment in a vocational school in Bandung. *Jurnal Sains Riset. 7*. https://www.researchgate.net/publication/345334436

Husna H.H. & Fachrurrazy (2017). An analysis of an English summative test for 6th grade students in three public elementary schools in udanawu district, Blitar Regency. *State University of Malang*

Kim, K., &, H. (2017). *Authentic Assessment*. Oxford Research Encyclopedias, Education. https://doi.org/ 10.1093/acrefore/9780190264093.013.22

Liu, J. (2005). Authenticity in language testing: A case study of a communicative language test for Chinese English majors (Doctoral dissertation, University of Hong Kong)

Lopez & Whitehead (2013). *Sampling data and data collection in qualitative research*. Nursing *and Midwifery research 4E.*

Purpura, J.E., & Kunnan, A.J. (2024). *The writings of Lyle F. Bachman: Assuring that "What We Count Counts" in Language Assessment (1st ed.).* Routledge. https://doi.org/10.4324/9781315765211

Purwandani, A.D, Raja. P, &Suparma. (n. d.) The analysis of the authenticity of authentic reading materials in students' text book. *University of Lampung.*

Salaria (2012). Meaning of the term-descriptive survey research method. *International Journal of Transformations in Business Management. (IJTBM), 1* (6). http://www.ijtbm.com/

Semiun, T. T., Wisrance, M. W., & Napitupulu, M. H. (2022). English summative test: The quality of its items. *English Education: Journal of English Teaching and Research*, *7*(2), 119-127. https://doi.org/10.29407/jetar.v7i2.18347

Semiun, T. T., & Luruk, F. D. (2020). The quality of an English summative test of a public junior high school, Kupang-NTT. *English Language Teaching Educational Journal*, *3*(2), 133–141. https://doi.org/10.12928/eltej.v3i2.2311

Setiyana (2016). Analysis of summative tests for English. *English Education Journal (EEJ), 7*(4), 433-447.

Shomoossi &. Tavakoli (2010). Designing assessment tools: The principles of language assessment. *Mukogawa women's univ. humanities and social sci., 60.* 41－49.

Spolsky, B., & Halt, F. M. (2008). *The Handbook of Educational Linguistics*. Blackwell.

Sugianto. (2017). Validity and reliability of English summative test for senior high school. *Indonesian EFL Journal: Journal of ELT, Linguistic, and Literature.*

Tilfarlioğlu. (2017). Testing and Evaluation in ELT Methodology. *ResearchGate.* Gaziantep University.https://www.researchgate.net/publication/321155250

Widdowson, H. G. (1978). *Teaching language as communication.* Oxford University Press

Yon (2016). Validity and reliability of English summative test items designed for SMP students in depok. *Dialectical Literature and Education Journal (DLEJ)*, *(1)* 1.

Contact emails: hardyantohitimala@upi.edu
susisr@upi.edu