IJEE

Indonesian Journal of English Education

VALIDITY, RELIABILITY, AND PRACTICALITY OF TEST ITEMS OF INTERNET-BASED ENGLISH PROFICIENCY TEST FOR ENGLISH LABORATORY

Nuri Emmiyati[1]* Sardian Maharani Asnur[2] Ahmad Ahmad[3]
[1,2,3] State Islamic University of Alauddin, Makassar
(nuri.emmiyati@uin-alauddin.ac.id)

Received: December 2023; Revised: September 2024; Accepted: December 2024

ABSTRACT

One way to evaluate English language skills is through the English Proficiency Test (EPT). This research aimed to develop the Internet-Based English Proficiency Test (IB-EPT) by evaluating its test items' validity, reliability, and practicality. The study followed a research and development approach incorporating qualitative and quantitative methods. The sample consisted of 24 students from UIN Alauddin Makassar. The developed English language test media received positive feedback from the expert validation and field trials. The results revealed that the validity of test items was confirmed, with a coefficient of 0.462 (r table = 0.44). Similarly, using Spearman-Brown product-moment correlation, the test items were reliable, with an index of 0.632, exceeding the critical value of the product-moment table (r table = 0.44). In addition, the students' evaluation of the product's practicality yielded a score of 83.26%, categorized as "very good," indicating the product was feasible. It is recommended that higher education policymakers should prioritize equipping campuses with comprehensive facilities, particularly information and communication technology, and provide specialized training for lecturers to develop effective computer-based and online assessments.

Key Words: English proficiency test; internet-based test; practicality; reliability; validity

ABSTRAK

Salah satu cara untuk mengevaluasi kemampuan bahasa Inggris adalah melalui English Proficiency Test (EPT). Penelitian ini bertujuan untuk mengembangkan Internet-Based English Proficiency Test (IB-EPT) dengan mengevaluasi validitas, reliabilitas, dan kepraktisan butir-butir tesnya. Penelitian ini mengikuti pendekatan penelitian dan pengembangan, yang menggabungkan metode kualitatif dan kuantitatif. Sampel terdiri dari 24 mahasiswa dari UIN Alauddin Makassar. Secara keseluruhan, media tes bahasa Inggris yang dikembangkan menerima umpan balik positif dari validasi ahli dan uji coba lapangan. Hasil penelitian mengungkapkan bahwa validitas butir tes terkonfirmasi, dengan koefisien 0,462 (r tabel = 0,44). Demikian pula, dengan menggunakan korelasi product moment Spearman Brown, butir-butir tes ditemukan reliabel, dengan indeks 0,632, melebihi nilai kritis tabel product moment (r tabel = 0,44). Selain itu, evaluasi mahasiswa terhadap kepraktisan produk menghasilkan skor 83,26%, dikategorikan "sangat baik," yang menunjukkan produk tersebut layak. Disarankan agar pembuat kebijakan pendidikan tinggi memprioritaskan melengkapi kampus dengan fasilitas yang komprehensif, khususnya teknologi informasi dan komunikasi, dan memberikan pelatihan khusus bagi dosen untuk mengembangkan penilaian berbasis komputer dan daring yang efektif.

Kata Kunci: *tes kemampuan Bahasa Inggris; tes berbasis internet; kepraktisan; keandalan; validitas*

How to Cite: Emmiyati, N., Asnur, S.M., Ahmad. (2024). Validity, Reliability, and Practicality of Test Items of Internet-Based English Proficiency Test for English Laboratory. IJEE (Indonesian Journal of English Education), Vol(number), pages 197-206. doi: 10.15408/ijee.v11i2.37276

*Corresponding author

P-ISSN: 2356-1777, E-ISSN: 2443-0390 | DOI: http://10.15408/ijee.v11i2.37276

INTRODUCTION

Effective communication skills, particularly in English, are essential for students to compete globally. English proficiency is critical for participating in student exchange programs, pursuing higher education, or studying abroad. This proficiency is typically assessed through standardized English Proficiency Tests (EPT) (Arikan et al., 2019; Gultom & Oktaviani, 2022). The English Proficiency Test is a comprehensive assessment of language proficiency that is only sometimes designed or conducted in accordance with a specific educational program that has been taught in the past (Henning, 2001). It is a standardized test.

In Indonesia, universities are increasingly adopting policies that set English proficiency benchmarks for students to meet before graduation. At Alauddin State Islamic University in Makassar, the English Language and Literature Department Laboratory regularly administered face-to-face, paper-based proficiency exams. Nonetheless, the Covid-19 pandemic has forced the department and laboratory to conduct online tests.

Online English proficiency tests, however, have many challenges, including difficulty using test platforms or media, question management, and controlling test takers to avoid cheating. The online system automatically configured the 120-minute test length for the web-based media platform, including Google Forms, Google Drive, and timify.com.

Another alternative to see students' English skills is to direct students to take a standardized English Proficiency Test from an official institution, of which, according to Brown (2010) there are four, namely: official TOEFL from ETS; MELAB from the University of Michigan; IELTS from the University of Cambridge, UCLES, British Council or IDP Education Australia; and TOEIC from The Chauncey Group International with much more expensive costs required.

On this basis, researchers see the importance of developing an online English language proficiency test system at Universitas Islam Negeri (UIN) Alauddin Makassar. In addition, online English language proficiency tests are necessary for accessible, standardized, and efficient assessment of students' language skills. It allows institutions to assess English competency more efficiently, meeting varied student demands and avoiding logistical issues like paper-based tests. An internet-based system enables large groups to be tested concurrently and collects data for continuing test analysis to verify validity, reliability, and practicality (Peng et al., 2020). A system that follows worldwide digital education trends prepares students for academic and professional communication demands.

Several studies on the Internet-Based English Proficiency Test were previously conducted by Putria & Lestari (2021) and Ananda (2019). The research that developed the English language proficiency test presented by Parwanto (2011) has contributed to online tests. Then, several other test development studies only developed tests for one skill, as designed by Azmi (2020) and Hesti (2021), who developed the Reading test. Meanwhile, Endarto & Subekti (2020) developed an online vocabulary test. However, this research develops an internet-based test and assesses the practicality, validity and reliability of the English Proficiency Test questions.

By carrying out development-based research, research outputs can be developed and used by the English Language and Literature Department UIN Alaudin Makassar and the wider community by providing a prediction test for those who want to measure their English language skills.

The urgency of this research lies in addressing the growing global demand for accessible and accurate English proficiency assessments in an increasingly digitalized world. Given that English is essential for academic and professional achievement, providing an Internet-based test is a practical approach to addressing time effectiveness and limiting cost. This research develops a valid, trustworthy, and useful tool that aligns with the fast digital change in education and supports UIN Alaudin Makassar's objective to foster innovation.

To know whether the developed test is good or not, Brown (2007) suggests the three criteria for testing a test: validity, reliability, and practicality. Therefore, this study aims to develop an Internet-based English Proficiency Test by measuring its validity, reliability, and practicality.

METHODS

*Corresponding author

P-ISSN: 2356-1777, E-ISSN: 2443-0390 | DOI: http://10.15408/ijee.v11i2.37276

Research design

The main aim of this research is to develop an Internet-based English Proficiency Test (IB-EPT). This study adopts a Research and Development (R&D) design, following Richards' (2022) stages for material development, including (1) Writing a draft of the Internet-Based English Proficiency Test, (2) a Trial of the Internet-Based English Language Proficiency Test, and (3) Make final revisions to the test material. These stages are iterative, where the initial draft is revised based on expert validation for content, construct, and language validity. Subsequently, the revised version undergoes reliability testing to ensure consistency. Practicality tests are carried out in the final stage.

Research site and participants

This study was conducted at Adab and Humanities Faculty, UIN Alauddin Makassar, with a population of 721 English Language and Literature students. A purposive sampling technique was employed, selecting 24 participants who met the following criteria: the first-year BSI FAH UIN Alauddin students with a minimum TOEFL PBT score of 400. 12 students were selected as the samples representing elementary and low intermediate levels and for the final-year BSI FAH UIN Alauddin students with a TOEFL PBT score of at least 500. 12 students were selected as the samples representing high intermediate and advanced levels.

Data Collection and Instrument

The IB-EPT was developed using the following steps: designing question blueprints, expert validation, revisions, reliability testing, online deployment, and practicality testing. Three methods were used to conduct those procedures for example, surveys, interviews, and tests. The primary instrument was the IB-EPT test itself, complemented by 1) a validation rubric (to evaluate the content, construct, and linguistic quality of the test, filled out by two expert validators), 2) an Interview guide (to gather qualitative feedback on practicality from students and instructors), 3) EPT questions (to test reliability using internal consistency measures and test-retest methods).

Validity Testing

To establish content validity, the two English language experts reviewed the test items to ensure alignment with the target skills, including listening, structure and written expression, and reading comprehension. Expert feedback was collected using a rubric that focused on the clarity of the questions, their relevance, and their alignment with the test objectives. To determine the validity of the items, validity scores were calculated using a percentage agreement approach, with an acceptability threshold set at ≥80%. The revised version of the English Proficiency Test (EPT) questions was then used to assess item validity based on two key criteria: the difficulty index and item discrimination levels.

Reliability Testing

Reliability was assessed through two primary methods: internal consistency and test-retest reliability. For internal consistency, Cronbach's alpha was used to measure item coherence and reliability coefficients were calculated for each test section, with the following thresholds: 0.80–0.89 for Listening Comprehension and 0.90–0.99 for Structure and Reading Comprehension. Additionally, test-retest reliability was evaluated by administering the test to the same sample twice within two weeks to assess score consistency. A Pearson correlation coefficient of ≥0.85 was considered acceptable for demonstrating reliability.

Data Analysis

Quantitative data, such as reliability scores, were analyzed using SPSS 25, with content validity percentages and reliability coefficients computed to assess the test's quality. In addition, qualitative feedback gathered from interviews was thematically analyzed using Miles and Huberman's framework. This analysis involved three stages, data reduction, data display, and conclusion drawing, to identify key themes and insights from the feedback.

FINDINGS AND DISCUSSION

Findings

In this study, researchers have obtained data that will answer and support the answers to research questions. The researcher made a grid of easy to difficult questions in the first step. After that, the researcher made the items by referring to the types of questions used in the TOEFL test, mainly from Barron and Longman, which were then adjusted to the needs of the English language proficiency test. A single draft of questions was created and included three components: Listening Comprehension (50 items), Structure and Written Expression (40 items), and Reading Comprehension (50 items).

The validation test was carried out by professionals knowledgeable in English language materials after these things were created. Validation of the products that had been created was carried out by two specialists who were responsible for the process. After the experts had validated the instructions and question items, the researchers revised them. Following the assessment of validity, the question items underwent a reliability test. A reliability test was conducted on 24 samples on two separate occasions. After being reliable, all the items were entered into an internet-based application, namely Google Forms combined with quilqo.com.

Furthermore, the reliability test was repeated using the internet-based test on the same 24 students. After obtaining reliable results, a practicality test was carried out, and the final product was produced after revisions were made based on the results of the practicality test. The following are the results of validation, reliability, and practicality.

English Language Test Design

Blue Prints

The formulation of examination items should be grounded on fundamental proficiencies, indicators, and explanations of the subject matter that has been instructed (Burhan, 2010). The researcher's test material comprises three components: Listening, Structure and Written Expression, and Reading Comprehension. The program's unit structure is well-organized and logical, reflecting a specific understanding of language and the learning process (Grabe & Kaplan, 2014).

Additionally, the researcher assessed the textbooks previously utilized by the students as a determining factor in the development of the blueprint. A 140-item multiple-choice examination was designed in adherence to the specifications outlined in the blueprint.

Validation Results on the Expert Assessment Rubric

The material expert validation activity is a stage of testing the product's feasibility before it is tested. According to experts, the feasibility of the product is a benchmark for whether the product can be used or needs to be further revised (Serevina et al., 2018). The material expert validators in this study are individuals who possess the necessary aptitude and capability in the specific subject of study.

The product validation process involved soliciting input from specialists through direct consultation, during which the product was presented, elucidated, and demonstrated. Subsequently, the validators evaluated the product by responding to a questionnaire. Moreover, the questionnaire includes a section for recommendations that can be used to assist in the revision process before the product is tested.

Material validation is accomplished by presenting question items generated and collated by researchers during the study process. This is the conclusion that was reached by the expert evaluation:

Table 1 Validation Result from 1st Validator

No.	Validated Item	Score
1.	Content	
	a. Questions are in accordance with the language skills tested (Listening Comprehension, Structure and Written Expression, and Reading)	5
	b. Statements are relevant to the aspects to be measured (Listening Comprehension, Structure and Written Expression, and Reading)	5
	c. Questions are clearly formulated	4
2.	Construction	
	a. Instructions are clearly formulated	3
	b. Questions do not contain other meanings	5
	c. Answer options are clearly formulated	5

3	Language	
	a. Clarity of sentences	5
	b. Appropriateness of diction	5
	c. Use of sentences that are in accordance with English grammar	5
	Total Score	42

Table 1 above displays the findings of the initial evaluation conducted by material experts. It reveals that one criterion received a score of 3, another criterion received a score of 4, and seven criteria achieved the highest score, indicating an excellent performance. The initial validator's cumulative score is 42. The questionnaire consists of nine elements, resulting in a maximum total score of 45.

The initial material expert assessment of the product yielded a percentage value of 93.33%. This value falls inside the "Very Good" category on the product feasibility interpretation table, which ranges from 80% to 100%. The box below illustrates the image.

Value QUOTE x 100 = 93,33%

Additionally, table 2 below displays the validation results obtained from the second material expert.

Table 2 Validation Result from 2nd Validator

No.	Validated Item	Score
1.	Content	
	a. Questions are in accordance with the language skills tested (Listening Comprehension, Structure and Written Expression, and Reading)	5
	b. Statements are relevant to the aspects to be measured (Listening Comprehension, Structure and Written Expression, and Reading)	5
	c. Questions are clearly formulated	4
2.	Construction	
	a. Instructions are clearly formulated	4
	b. Questions do not contain other meanings	5
	c. Answer options are clearly formulated	5
3	Language	
	a. Clarity of sentences	5
	b. Appropriateness of diction	5
	c. Use of sentences that are in accordance with English grammar	5
Total Score		43

Table 2` above displays the outcomes of the second validator examination, indicating that two criteria received a score of 4, while eight additional criteria obtained a score of 5. The second validator has a total of 43 scores. The percentage of test feasibility from the second validator can be calculated using the percentage formula as follows:

Value QUOTE x 100 = 95,56%

From the validation results of the second validator, the percentage of test feasibility was 95.56%, also in the "Very Good" category. The average score from the first and second validators is 42.50 in the percentage of test 94.44% and categorized as "Very Good".

Based on the outcomes of the two validators mentioned earlier, the viability of the English language proficiency test items can be concluded. The product's practicality is highly promising and compelling for implementation in the exam. This questionnaire serves as a means to gather input and identify areas for development to enhance the product. Overall, the two material experts recommended improving the clarity of the question instructions to minimize any potential confusion or uncertainty among the examinees.

Following the revision of the product based on the questionnaire results and input from material specialists, the researchers proceeded to conduct trials to assess the item validity according to the criteria of difficulty index, item discrimination levels, and reliability of the planned tests. Then, the practicality evaluation was conducted by analyzing the students' feedback during practical trials.

Question Validity

The validity was measured by assessing the item validity according to the criteria of difficulty index and item discrimination levels.

Question Item Difficulty Level

Assigning an index to the probability of answering a question regarding a specific ability corresponds to the question's difficulty level. The range of the total difficulty index is from 0.00 to 1.00. The test becomes easier as the question's difficulty increases. An item of difficulty (TK) of 1 indicates that every participant has responded correctly. The condition "TK = 0" indicates that no candidate has provided a valid response on the test.

The ratio of participants who answer correctly to the total number of participants is sometimes used to determine the test items' difficulty level. Criteria for question difficulty index level:

0.30 = Difficult
0.31 - 0.70 = Moderate
0.71 - 1.00 = Easy

The test results indicate that in the Listening section, there are 3 questions categorized as simple, 33 as medium, and 14 as challenging. The section on Structure and Written Expression consists of 6 simple questions, 22 moderate questions, and 12 difficult questions. The Reading Comprehension section consists of 6 questions categorized as easy, 34 as medium, and 10 as challenging. The easy level consists of 15 questions, the intermediate level consists of 89 questions, and the difficult level consists of 36 questions, totalling 140 questions.

Level of Discrimination of Question Items

The potential of a question to differentiate between test candidates who have mastered the subject matter and those who have not is called question discrimination. As an index, item discrimination is expressed. As the item discriminator's index increases, it becomes more capable of differentiating test candidates who have comprehended the material from those who have not. As the item discrimination increases, the query becomes more robust and superior. A negative index (<0) indicates that a greater proportion of the bottom group (constituting students who lack comprehension of the material) provides an accurate response to the query in comparison to the upper group (proficient candidates).

Any item with an index value above 0.50 can be confidently classified as a strong discriminating item during item selection. Conversely, items with an index value below 0.20 can be promptly disregarded. Items falling between these two thresholds warrant further examination for potential adjustment. Classification of item discrimination:

0.40 to 1.00 = Acceptable/good question
0.30 to 0.39 = Question is acceptable but needs improvement
0.20 to 0.29 = Question corrected
0.19 = Question not used / Discarded

The calculations reveal that 11 questions in the Listening section fall under the "Good" or "Acceptable" category, while 39 questions are classified as "Excellent." Within the Structure and Written content domain, 23 questions were classified as "Good," and 17 questions were classified as "Excellent." The Reading Comprehension section consisted of 14 questions categorized as "Excellent," while the remaining 36 questions were classified as "Good" or Satisfactory without any need for improvement.

The research findings indicated that the multiple-choice English test had a validity coefficient of 0.462, indicating its general validity. The researcher presents a concise summary of the overall test validity to provide further information.

Table 3 Analysis of Validity

Correlation	Table	Status
0,462	0,44	Valid

As indicated in the description, Table 3 comprises three columns. The initial column presents data on the validity analysis outcomes. The second column presents details on the product moment critical value table, which has a significance level of 95%. Additionally, the validity status is detailed in the third column.

The overall validity test is deemed valid based on the analysis results, which indicate that the R-values exceed the product moment table. Furthermore, the researchers incorporated findings data that demonstrated the items' validity according to the validity classification, revealing that certain items are deemed to be flawed. Conversely, the researcher was solely concerned with the multiple-choice English test's overall validity.

Question Reliability

Based on the research data, the English test demonstrates a reliability index of 0.82. At a significance level of 95%, an item is deemed reliable if the correlation coefficient of each item is equal to or greater than the product moment critical value table. The researcher furnishes the following table of reliability analysis for further information:

Table 4 Reliability Analysis

Correlation	Table	Status
0,82	0,44	Reliable

The correlation is detailed in the first column of the table presented above. The product moment critical value table, with a significance level of 95%, is detailed in the second column. Furthermore, details regarding the reliability status are presented in the third column. The researchers determined the reliability of this evaluation by utilizing product moment.

Question Practicality

The developed, validated and revised English examination was tried in the Faculty of Adab and Humanities at UIN Alauddin's English laboratory. Twenty-four students were involved in this procedure. Students were instructed to operate and attempt the online test after the researcher had provided an overview of the study. Students responded to the queries with apparent enthusiasm during this procedure. Students are more motivated to respond to subsequent questions when they can ascertain whether their answers are correct or incorrect upon clicking the submit icon. The audio and visuals also captured their attention .

Researchers identified several aspects of the product that required improvement during the trial with online students. Resolving a few of the audios in the listening section and more subtle aspects require correction, as some exhibit distortion and short duration.

As shown above, the average value of the students' product evaluations is 83.26 per cent. This value falls within the "Very Good" category, which spans from 80% to 100%, as indicated by the previously established interpretation table of product feasibility.

Discussion

The overall validity of the test has been confirmed based on the results of expert validation and item analysis. The high average validation scores (>93%) provided by two expert validators indicate that the test items align well with the measured English language proficiency constructs. These results reinforce the ability of the IB-EPT to provide accurate score interpretations and to assess the intended skills, consistent with the framework of Sürücü & Maslakci (2020) and Chapelle & Lee (2021). Validity ensures that the test accurately reflects the linguistic competencies of the examinees and provides meaningful insights into their English proficiency.

The research data shows that out of 140 items, there are fifteen easy items, eighty-nine medium items, and thirty-six difficult items. These findings highlight the balance of question levels, which ensures the test caters to a range of proficiency levels, from elementary to advanced. According to Brown and Abeywickrama (2004), a well-constructed test maintains this balance, enabling participants to experience a sense of progression and engagement without feeling discouraged by overly

*Corresponding author

P-ISSN: 2356-1777, E-ISSN: 2443-0390 | DOI: http://10.15408/ijee.v11i2.37276

challenging items. This makes participants aware of and takes note of the characteristics of the tests given to them, whether they are easy, medium, or difficult. Thus, the test questions are in accordance with English standards and meet the characteristics of a good test.

In terms of reliability, the IB-EPT demonstrated strong consistency across its sections. The reliability coefficient of 0.82 for Listening Comprehension and 0.90–0.95 for Structure and Reading Comprehension sections exceeded the benchmarks Hughes (2008) outlined for high-quality tests. This indicated that the test items were reliable, as determined by the Spearman Brown Product Moment analysis.

Essentially, it refers to the degree of consistency with which a test delivers the intended measurement. When the same test paper is graded by two or more distinct examiners or the same examiner on multiple occasions, reliability is the degree to which the same mark or grade is assigned. These findings confirm the stability of the test, ensuring that repeated administrations would yield consistent results, as emphasized by Eldridge et al. (2021). Reliability not only enhances the credibility of the test but also establishes its usability in diverse academic contexts.

Most participants agreed on the role of online English exams and the importance of improving their performance. They agreed that the online test can help them with knowledge, experience and academic issues.

Based on the analysis findings and participant feedback, the online test will enhance students' academic performance. Additionally, the online test will facilitate students' transition to the TOEFL examination in the future by equipping them with the ability to complete English tests rapidly, thereby improving their ICT proficiency. This examination may serve as an overall gauge of the English proficiency that students have acquired during their time on campus.

According to a number of earlier studies, online examinations have the potential to yield positive outcomes. These outcomes include the identification of students' strengths and weaknesses, the provision of data that aid in decision-making, and the utilization of this information to assist individuals in improving their performance (Korkmaz & Öz, 2021; Zboun & Farrah, 2021).

The practicality of the IB-EPT was also well-received, with over 83% of participants rating it as "very practical." Feedback highlighted the test's clarity, ease of use, and efficiency of the Quilgo-integrated Google Forms platform. Participants noted that the test format was engaging and aligned with contemporary digital testing standards. However, minor revisions, particularly in the audio quality of the Listening Comprehension section, were necessary to further enhance user experience.

The findings have significant implications for developing online English proficiency tests in higher education. The validated, reliable, and practical nature of the IB-EPT demonstrates its potential as a scalable and cost-effective alternative to standardized tests like TOEFL or IELTS.

This comprehensive assessment instrument effectively evaluates English language competency while offering an interesting and pragmatic testing experience. The implications of this research exceed the immediate results, providing a framework for future advancements in language assessment techniques and facilitating informed teaching practices.

CONCLUSIONS AND SUGGESTION

According to the results of both expert validation and field trials, the English test media that has been developed has, on the whole, received positive response. This product of the IB-EPT underlines its importance as a reliable, valid, and practical assessment tool that effectively measures English proficiency levels. By applying these insights in educational settings, institutions can enhance their language programs, support student success, and equip learners for future opportunities in academic and professional contexts. Engaging with the data derived from the IB-EPT will enable continuous improvement, ensuring that educational assessments align with the evolving needs of students in a digital world.

The findings of this study may also be applied in other universities for student placement, progress assessment, evaluation of English language courses, and assessment of students' language proficiency for graduation or advanced studies. Universities can employ the IB-EPT to assist students in preparing for high-stakes assessments and improving their English language proficiency scores. IB-EPT may be integrated into standardized test preparation courses, necessitating adjustments to curriculum and instruction. The overall efficacy of the language program will improve. Universities must promote digital assessment tools to prepare students for online assessments.

REFERENCES

Ananda, K. R. (2019). *Developing English Proficiency Test Media in Junior High School* [Institut Agama Islam Negeri Parepare]. http://repository.iainpare.ac.id/958/1/14.1300.010.pdf

Arikan, S., Kilmen, S., Abi, M., & Üstünel, E. (2019). An example of empirical and model based methods for performance descriptors: English proficiency test. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(3), 219–234. https://doi.org/10.21031/epod.477857

Azmi, U. (2020). Developing web-based reading tests for the students of English language education. *Journal of Applied Linguistics, Translation, and Literature*, *1*(2), 92–104.

Brown, H. D. (2010). *Language Assessment Principles and Classroom Practices*. Pearson Longman.

Brown, H.D. (2001) *Teaching by Principles: An Interactive Approach to Language Pedagogy.* 2nd Edition. Pearson Longman

Brown, H. D., & Abeywickrama, P. (2004). Language assessment. *Principles and Classroom Practices. White Plains, NY: Pearson Education.*

Burhan, N. (2010). *Penilaian Pembelajaran Bahasa Berbasis Kompetensi.* Yogyakarta: BPFE-Yogyakarta.

Chapelle, C. A., & Lee, H. (2021). Conceptions of validity. In *The Routledge handbook of language testing* (pp. 17–31). Routledge.

Eldridge, H., De Donno, M., & Champod, C. (2021). Testing the accuracy and reliability of palmar friction ridge comparisons–a black box study. *Forensic Science International*, *318*, 110457.

Endarto, I. T., & Subekti, A. S. (2020). Developing a Web-Based Vocabulary Size Test for Indonesian EFL Learners. *Teknosastik*, *18*(2), 72. https://doi.org/10.33365/ts.v18i2.492

Grabe, W., & Kaplan, R. B. (2014). *Theory and practice of writing: An applied linguistic perspective*. Routledge.

Gultom, S., & Oktaviani, L. (2022). the Correlation Between Students' Self-Esteem and Their English Proficiency Test Result. *Journal of English Language Teaching and Learning*, *3*(2), 52–57. https://doi.org/10.33365/jeltl.v3i2.2211

Henning, G. (2001). *A Guide to Language Testings: Development, Evaluation, and Research.* Heinle & Heinle Publishers

Hesti, S. W. (2021). *The Influence Of Using Problem Based Learning Towards Students Writing Ability On Analytical Exposition Text* [UIN Raden Intan Lampung]. http://repository.radenintan.ac.id/id/eprint/16617

Korkmaz, S., & Öz, H. (2021). Using Kahoot to improve reading comprehension of English as a foreign language learners. *International Online Journal of Education and Teaching*, *8*(2), 1138–1150.

Meadows, M., & Billington, L. (2005). A review of the literature on marking reliability. *London: National Assessment Agency*.

Parwanto, T. (2011). Perancangan Aplikasi Simulasi Toefl (Test of English As Foreign Language). *Program Studi Teknik Informatika Fakultas Sains Dan Teknologi Universitas Islam Negeri Syarif Hidayatullah*, 100. https://repository.uinjkt.ac.id/dspace/handle/123456789/3767

Putria, A., & Lestari, W. (2021). Need Assessment For Developing An Online English Proficiency Test Instrument In Higher Education During Covid-19 Pandemic. *JPP (Jurnal Pendidikan Dan Pembelajaran)*, *27*(2), 78–82. https://doi.org/10.17977/um047v27i22020p078

Richards, J. C., & Pun, J. (2022). Teacher strategies in implementing English medium instruction. *ELT Journal*, *76*(2), 227–237. https://doi.org/10.1093/elt/ccab081

Serevina, V., Astra, I., & Sari, I. J. (2018). Development of E-Module Based on Problem Based Learning (PBL) on Heat and Temperature to Improve Student's Science Process Skill. *Turkish Online Journal of Educational Technology-TOJET*, *17*(3), 26–36.

Sürücü, L., & Maslakci, A. (2020). Validity and reliability in quantitative research. *Business & Management Studies: An International Journal*, *8*(3), 2694–2726.

Tavakol, M., & Wetzel, A. (2020). Factor Analysis: a means for theory and instrument development in support of construct validity. International Journal of Medical Education, 11, 245.

Zboun, J., & Farrah, M. (2021). Students' perspectives of online language learning during corona pandemic: Benefits and challenges.

*Corresponding author

P-ISSN: 2356-1777, E-ISSN: 2443-0390 | DOI: http://10.15408/ijee.v11i2.37276