

An Analysis of Validity and Reliability of a Teacher-made Test on Descriptive Text: Case Study to Tenth Grade Office Administration Students at SMKN 3 Pontianak

Nurwinda Laili Puspitasari ^{1*}, Nur Ulyasari ²,
Yohanes Gatot Sutapa Yuliana ³, Clarry Sada ⁴

¹⁻⁴ Universitas Tanjungpura, Pontianak, Indonesia

* winda.laili@gmail.com

Abstract

This study aimed to analyse the connection between the test items' difficulty and discrimination indices in order to investigate the validity and reliability of a teacher-made test on descriptive text in SMKN 3 Pontianak. The purpose of this analysis was to ascertain if the test items accurately assessed the students' critical thinking abilities. This study used case research design. The test used in this study was created by an English teacher for office administration students in the tenth grade. The test consisted of 20 multiple-choice questions with a descriptive text. Documentation served as the research tool in this study. The researchers aimed to evaluate the degree to which the test items effectively assessed the students' abilities and comprehension of the descriptive text by looking at the test items' levels of difficulty and discrimination. The analysis's findings can shed light on the reliability of the test items as well as how well they assess students' language and critical thinking abilities. The results of the study also support the continuous efforts to increase the reliability of assessments created by teachers and to better evaluate students' learning outcomes in the context of EFL training.

Keywords: *validity, reliability, critical thinking, teacher made-test*

Introduction

With the introduction of modern English as a Foreign linguistic (EFL) instruction, there is an increasing focus on evaluating students' critical thinking and linguistic skills. The creation and use of exams created by teachers is one of the key components of this evaluation process. Richards (1988) agreed that today, English is the world's most widely studied foreign language. It is accepted that the English lesson must be conducted in the classroom. The paractical English itself should be assessed as the indicator of successful learning objectives. These assessments, created by teachers, are intended to gauge how well pupils are progressing and comprehending the subject matter. Exams are crucial for determining an EFL learner's level of language competency. Nunan et al., (2003) mentioned that evaluation has to do with how well students have mastered the objectives o f the course and how effectively the course has met their needs. They assist in assessing students' proficiency in a range of language-related domains, such as speaking, listening, reading, and writing. In fact, assessing students should be done as an ongoing process. How the students learn every day in the classroom should be assessed to see the progress of their learning.

Brown (2003) mentioned that assessment, on the other hand, is an ongoing process that encompasses a much wider domain. However, in order to guarantee accurate and impartial assessments of students' performance, the validity and reliability of these exams are essential. Validity and reliability are prerequisites for test efficiency. Furthermore, it is widely acknowledged that if these two requirements are met, we can obtain an accurate assessment of the students because they are crucial to the testing process. It is not unexpected, then, that the variables used to assess a test's validity and reliability can occasionally yield a variety of outcomes. Assessment is perhaps one of most difficult and important parts of our jobs as teachers. Ideally, it should be seen as a means to help us guide students on their road to learning. No single procedure can meet the needs of all learners and situations, so we need to remember to incorporate a variety of tools to help our students know how they are progressing and to gauge the effectiveness of our own methodology and materials. Brown (2003) stated that the well-constructed test is an installment that provides an accurate measure of the test-taker's ability within a particular domain. The definition sounds fairly simple, but in fact, constructing a good test is a complex task involving both science and art.

A good test items will help the students not only gain their score but also gain their understanding of the lesson. Students may get a high score but still not master the lesson. Dhia & Alsalihi (2019) agreed that good test item should have high validity and reliability. Validity, in the context of testing and assessment, refers to the degree to which a test measures what it claims to measure. It is a fundamental concept in psychometrics, which is the field of study concerned with the theory and technique of psychological measurement. In other words, a test is considered valid if it accurately and appropriately assesses the construct or trait it is designed to measure. Brown (2003) mentioned that validity is the extent to which inferences made from assessments results are appropriate, meaningful, and useful in terms of the purpose of the assessment. Whereas, reliability is to what the test is consistent and dependable.

There are 4 types of validity: content validity, construct validity, criterion-related validity, and face validity. One kind of validity that focuses on how well a test assesses the complete topic or domain it is meant to measure is called content validity. Put another way, it determines if the test's items adequately reflect the larger subject matter or idea that the test is intended to measure. Brown (2003) mentioned that content validity can clearly define the achievement that you are measuring. The good assessment should meet the needs of learning objectives. A sort of validity called construct validity evaluates whether a test captures the underlying theoretical notion or characteristic that it claims to capture. The degree to which the test faithfully reflects the target psychological notion or trait is what matters. To show that a test is assessing what it purports to measure at a theoretical level, construct validity is essential. A sort of validity known as "criterion-related validity" looks at the connection between test results and an outside standard or result. It evaluates how well a test correlates or predicts a particular criterion, demonstrating the test's efficacy in predicting or making decisions about a person's performance, behavior, or traits. Concurrent validity and predictive validity are the two primary categories of criterion-related validity. Face validity is the simplest and most straightforward form of validity, and it refers to the extent to which a test appears, on the surface, to measure what it is intended to measure. It is a subjective judgment made by individuals who are not experts in the field, often based on a cursory examination of the test items. It's a subjective assessment made by people who

aren't subject matter experts, frequently based on a quick scan of the test questions. Brown (2003) agreed that face validity refers to the degree to which a test looks right, and appears to measure the knowledge or abilities it claims to measure. McCowan et al., (1999) mentioned that Many tests are flawed because they focus on insignificant or unrelated information, disproportionately test one segment of curriculum and ignore other sections, or include poorly written, confusing test items. Every kind is essential to guaranteeing the validity and precision of an assessment in gauging an individual's skills, knowledge, or characteristics.

The test items for the students should be well-prepared in order to make an effective test of knowing the process of learning from the students. This study was conducted in order to explore whether the test made teacher at SMK Negeri 3 Pontianak is valid, reliable and was able to evaluate their critical thinking. Here the researchers explored the multiple choices test items of descriptive text which consisted of twenty questions answered by twenty students. This research hopefully can help the teachers to measure their validity and reliability of their test, in the same direction with taxonomy bloom that indicates the level of students critical thinking implicitly in indicators. Assessing CT is essential for the English language teaching (ELT) in higher education classrooms since it can help teachers to know their students' English proficiency and CT skill level as both are interconnected (Andania, 2021). Besides, the material given in the classroom should be authentic and based on real world situation so the students can convey the meaning relate to their prior knowledge. Harmer (2001) mentioned that the language learning will be succeed if the teacher manages the activities directly involving the learners solving communication problems.

While reliability relates to the consistency of results, validity refers to how well a test measures what it is meant to measure. Both are essential to guaranteeing any assessment tool's reliability. It would be helpful to restate the main ideas on validity and reliability in the conclusion, highlighting their significance in the planning and execution of assessments.

Method

This research used case study as the support of the methodology in terms of analysing the test score from teacher made test at SMK 3 Pontianak. A case study is a type of research methodology that is used to investigate a difficult problem in its real-world environment in great detail and with many facets. It entails the in-depth examination of a particular subject, such as an individual, group, occasion, establishment, or phenomena, and is frequently employed in a variety of domains, including business, social, educational, and clinical research. The case study technique, which frequently integrates both qualitative and quantitative data collection methods, enables the examination of the case's salient features, consequences, and meanings. It can be used to explain, compare, assess, and comprehend various parts of a study problem. It is especially helpful when attempting to gain a thorough understanding of a particular real-world subject. Cohen, Manion, & Morrison (2007) mentioned that case study: an investigation into a specific instance or phenomenon in its real-life context. Case studies are important research instruments that provide comprehensive, contextual insights into difficult problems, directing further investigation and aiding in the creation and testing of ideas. Creswell (2012) stated that cases are bounded by time and activity, and researchers collect detailed information using a variety of data collection procedures over a sustained period of time. This research used the score of

students in X MPLB 2 at SMK 3 Pontianak on one of their assessment in descriptive text which in form of multiple choice. Therefore, the researcher explored the level of difficulty, item of discrimination, validity, and reliability involved in the test which also calculated using Excel.

Results

In this section, the researchers presented the result of the study conducted. The test items distributed has several indicators that develop critical thinking based on bloom's taxonomy including students are able to evaluate the topic and main idea, analyze the certain information, and create the ideas of descriptive text contextually and correctly. The questions were in form of multiple choice which can be accessed in gform as follow: <https://forms.gle/T34bCy6q8HbsVQwX8>

After the teacher distributed tests for the students, it came to the researcher exploration to the students scores that indicates difficulty level of the items. The index itself was taken from the distribution of the students who answered correctly with the total number of the students. It is provided in table.2 followed by the scale below.

Tabel 1. Index

Level	Interval
Difficult	0,1 – 0,3
Standard	0,4 – 0,7
Easy	0,8 - 1

Tabel 2. Level of difficulty

Name	Items																				Total	Score
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
D	E	E	D	B	E	C	A	A	D	B	D	C	C	A	B	A	C	A	A	C	14	70
O	E	E	D	C	B	B	A	D	D	A	A	C	B	A	A	A	C	B	A	C	17	85
N	E	E	A	B	A	A	A	A	A	E	B	E	C	A	A	D	C	B	C	C	9	45
L	A	E	D	B	A	A	A	D	D	E	A	E	B	A	A	A	A	B	E	C	11	55
M	B	A	A	E	A	A	D	D	D	B	A	C	A	A	C	A	A	E	B	C	7	35
NY	D	B	D	C	B	B	D	A	D	A	A	C	B	A	A	B	C	B	A	C	12	60
A	D	E	D	B	E	E	A	D	D	E	A	C	C	A	A	A	C	B	D	C	14	70
NYL	E	E	D	C	B	B	A	D	D	A	A	C	B	A	A	A	C	B	A	C	17	85
LV	E	E	D	B	B	B	A	D	D	E	A	C	C	A	B	A	C	B	A	C	19	95
J	E	E	D	B	B	B	A	D	D	E	A	C	C	A	B	A	C	B	A	C	19	95
RY	E	E	D	B	E	E	A	A	D	E	A	C	C	A	A	A	A	B	A	C	14	70
Z	E	E	D	B	D	D	A	C	D	B	D	C	C	A	B	A	A	B	B	C	13	65
S	E	E	D	C	B	B	A	D	D	E	A	E	B	A	A	A	A	B	C	C	13	65
APR	E	E	D	B	B	B	A	D	D	A	A	C	C	A	B	A	C	E	B	C	18	90
AGL	E	E	D	B	B	B	A	D	D	A	A	C	C	A	A	A	C	B	A	C	19	95
K	B	B	D	B	C	C	A	D	D	E	B	E	B	A	A	A	A	B	D	C	9	45
TLT	E	E	C	B	E	E	A	A	D	D	A	C	C	A	B	A	C	B	C	C	14	70
LN	D	B	D	C	A	A	A	B	D	E	D	C	A	A	A	B	A	B	C	C	7	35
NZ	A	B	A	B	B	A	A	D	D	E	A	E	C	A	A	A	A	A	A	C	11	55
MD	E	E	D	B	B	B	A	D	D	A	B	C	C	A	B	A	C	C	A	C	18	90
BENAR	13	15	16	14	10	9	18	13	19	6	14	15	12	20	7	17	12	15	10	20		
LOD	0,65	0,75	0,8	0,7	0,5	0,45	0,9	0,65	0,95	0,3	0,7	0,75	0,6	1	0,35	0,85	0,6	0,75	0,5	1		

From Table 1, it can be seen that most items had standard difficulties about 55%, but students still were required to achieve the indicators at the end of the study. The easy items covered 35% and on difficult level were for the rest.

As difficulties on different items existed, it reflected to the discrimination on students' abilities. The relationship between the item discrimination index and difficulty index is

statistically significant, as the difficulty level of an item increases, the discrimination index also increases (Kunjappagounder et al., 2021).

The level discrimination power was taken from the number of students in the upper group who answered the item correctly minus the number of students in the lower group who answered the item correctly then divided into the half of the total number of students themselves. The researcher provided the discrimination in table 3 together with its index as below.

Tabel 3 Index of discrimination

Categories	Interval
Excellent	0.7 – 1.00
Good	0.40-0.70
Satisfactory	0.20 - 0.40
Poor	< 0.20
Very poor	Negative value on D

Tabel 4. Level of Discrimination

Name	The items																				Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
LV	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	19
J	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	19
AGL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	19
APR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	18
MD	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	18
O																					17
NYL																					17
D																					14
A																					14
RY																					14
TLT																					14
Z																					13
S																					13
NY																					12
L																					11
NZ	0	0	0	1	1	0	1	1	1	0	1	0	1	1	0	1	0	0	1	1	11
N	1	1	0	1	0	0	1	0	0	0	0	0	1	1	0	0	1	1	0	1	9
K	0	0	1	1	0	0	1	1	1	0	0	0	0	1	0	1	0	1	0	1	9
M	0	0	0	0	0	0	0	1	1	0	1	1	0	1	0	1	0	0	0	1	7
L	0	0	1	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	0	1	7
DISCRIMINATION	0,8	0,8	0,6	0,4	0,8	1,2	0,4	0,4	0,2	1,6	0,6	0,6	0,6	0,6	1,4	0,4	0,8	0,4	0,8	0,8	0

From table above, researchers found that there were 40% item categorized excellent in term of its discrimination, followed by 40% good items, satisfactory items were 10% in account and the rest items categorized poor. The relationship between the difficulty level and discriminations on the test would impact to how teacher will construct the newest assessment.

The parallel between discrimination and score distribution on a test was that the items with high discrimination index tend to be more effective in distinguishing between high and low scoring students. Here was the score distribution on the test which showed the comparison among students performance level.

Tabel 5. Score distribution

Name	Items																				Total	S2
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
D	1	1	1	1	0	0	1	0	1	0	0	1	1	1	1	1	1	0	1	1	14	196
O	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	17	289
N	1	1	0	1	0	0	1	0	0	0	0	0	1	1	0	0	1	1	0	1	9	81
L	0	1	1	1	0	0	1	1	1	0	1	0	0	1	0	1	0	1	0	1	11	121
M	0	0	0	0	0	0	0	1	1	0	1	1	0	1	0	1	0	0	0	1	7	41
NY	0	0	1	0	1	1	0	0	1	1	1	1	0	1	0	0	1	1	1	1	12	144
A	0	1	1	1	0	0	1	1	1	0	1	1	1	1	0	1	1	1	0	1	14	196
NYL	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	17	289
LV	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	19	361
J	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	19	361
RY	1	1	1	1	0	0	1	0	1	0	1	1	1	1	0	1	0	1	1	1	14	196
Z	1	1	1	1	0	0	1	0	1	0	0	1	1	1	1	1	0	1	0	1	13	169
S	1	1	1	1	0	1	1	1	1	0	1	0	0	1	0	1	0	1	0	1	13	169
APR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	18	324
AGL	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	19	361
K	0	0	1	1	0	0	1	1	1	0	0	0	0	1	0	1	0	1	0	1	9	81
TLT	1	1	0	1	0	0	1	0	1	0	1	1	1	1	1	1	1	1	0	1	14	196
LN	0	0	1	0	0	0	1	0	1	0	0	1	0	1	0	0	0	1	0	1	7	49
NZ	0	0	0	1	1	0	1	1	1	0	1	0	1	1	0	1	0	0	1	1	11	121
MD	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	18	324
BENAR	13	15	16	14	10	9	18	13	19	6	14	15	12	20	7	17	12	15	10	20	275	4069
SALAH	7	5	4	6	10	11	2	7	1	14	6	5	8	0	13	3	8	5	10	0		
p	0,65	0,75	0,8	0,7	0,5	0,45	0,9	0,65	0,95	0,3	0,7	0,75	0,6	1	0,35	0,85	0,6	0,75	0,5	1		
q	0,35	0,25	0,2	0,3	0,5	0,55	0,1	0,35	0,05	0,7	0,3	0,25	0,4	0	0,65	0,15	0,4	0,25	0,5	0		
p.q (p,q)	0,22	0,19	0,16	0,21	0,25	0,24	0,09	0,22	0,05	0,21	0,21	0,19	0,24	0	0,22	0,13	0,24	0,19	0,25	0	3,51	

Discussion

Based on the result on the validity and reliability, the researchers ensured that the assessments provide accurate and meaningful information about students' learning and achievement. The quality of assessment is essential as the concerned point in learning progress. Validity and reliability in every question/tess play a crucial role to the quality itself which essential to prove the consistency, accuracy and the appropriateness of the test which can lead teacher to make a good decision making about whether the students achieve the indicators. Also, the level of critical thinking on blooms' taxonomy implicitly applied on test items which required students to apply knowledge and skills in more critical level.

The provided result above showed that the test for Q.1 has standard difficulty because the LOD is 0,65 and its discrimination categorized excellent because the D is 0,8, Q.2 has easy difficulty because the LOD is 0,75 and its discrimination categorized excellent because the D is 0,8, Q.3 has easy difficulty because the LOD is 0,8 and its discrimination categorized good because the D is 0,6, Q.4 has standard difficulty because the LOD is 0,7 and its discrimination categorized good because the D is 0,4, Q.5 has standard difficulty because the LOD is 0,5 and its discrimination categorized excellent because the D is 0,8, Q.6 has standard difficulty because the LOD is 0,45 and its discrimination categorized excellent because the D is 1, Q.7 has easy difficulty because the LOD is 0,9 and its discrimination categorized satisfactory because the D is 0,2, Q.8 has standard difficulty because the LOD is 0,65 and its discrimination categorized good because the D is 0,4, Q.9 has easy difficulty because the LOD is 0,95 and its discrimination categorized satisfactory because the D is 0,2, Q.10 has difficult level because the LOD is 0,3 and its discrimination categorized excellent because the D is 1, Q.11 has standard difficulty because the LOD is 0,7 and its discrimination categorized good because the D is 0,6, Q.12 has standard difficulty because the LOD is 0,75 and its discrimination categorized good because the D is 0,6, Q.13 has standard difficulty because the LOD is 0,6 and its discrimination categorized good because the D is 0,6, Q.14 has easy

difficulty because the LOD is 1 and its discrimination categorized poor because the D is 0, Q.15 has difficult level because the LOD is 0,35 and its discrimination categorized excellent because the D is 1, Q.16 has easy difficulty because the LOD is 0,85 and its discrimination categorized good because the D is 0,4, Q.17 has standard difficulty because the LOD is 0,6 and its discrimination categorized excellent because the D is 0,8, Q.18 has standard difficulty because the LOD is 0,75 and its discrimination categorized good because the D is 0,4, Q.19 has standard difficulty because the LOD is 0,5 and its discrimination categorized excellent because the D is 0,8, Q.20 has easy difficulty because the LOD is 1 and its discrimination categorized poor because the D is 0.

Conclusion

Based on the findings and discussion, it can be concluded that the study analyzed the difficulty and discrimination of twenty items of multiple choice in the descriptive text contextually and was found they varied in their levels of difficulty and discrimination. The results of this study can be used to inform the development of future assessments and to improve the quality of existing ones. Teacher can use the findings to ensure that test items are appropriately challenging and that they effectively discriminate between high and low performers which can lead to the validity and reliability. Overall, this study highlights the understanding of the difficulty and discrimination levels of test items that teachers can use to create questions that accurately measure students' knowledge and skills which ultimately contributing to the improvement of educational assessment.

References

- Andania, R. A. (2021). Assessing the English As a Foreign Language (Efl) University Students' Critical Thinking. *Education and Human Development Journal*, 6(2), 77–87. <https://doi.org/10.33086/ehdj.v6i2.2316>
- Brown, H. D. (2003). Language Assesment Principles and Classroom Practice. *Pearson Longman*, 324.
- Cohen, L., Manion, L., & Morrison, K. (2007). Quantitative data analyses. *Research Methods in Physical Activity and Health*. <https://doi.org/10.4324/9781315158501-17>
- Creswell, J. W. (2012). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research* (Fourth edi).
- Dhia, H., & Alsalihi, A. (2019). Characteristics of a Good Test. *University of Baghdad, October*, 2–4. <https://doi.org/10.13140/RG.2.2.21716.76169>
- Harmer, J. (2001). Book Review : The Practice of English Language Teaching. *RELC Journal*, 32(1), 135–136. <https://doi.org/10.1177/003368820103200109>
- Kunjappagounder, P., Doddaiiah, S., Basavanna, P., & Bhat, D. (2021). Relationship between difficulty and discrimination indices of essay questions in formative assessment. *Journal of the Anatomical Society of India*, 70(4), 239–243. https://doi.org/10.4103/jasi.jasi_170_20
- McCowan, R. J. R. (1999). Item Analysis for Criterion-Referenced Tests.
- Nunan, D., Terrell, T. D., & Brown, H. D. (2003). *Practical English Language Teaching*.
- Richards, J. C. and T. S. R. (1988). *Approaches and Methods in Language Teaching: a description and analysis* . Cambridge: Cambridge University Press