CHAPTER II

REVIEW OF RELATED LITERATURE

In this chapter, the researcher will explicate several theories through reviewing some literatures related to this study. This theoretical construct deals with three main areas, language testing, tests, and reliability.

A. Theoretical Foundation

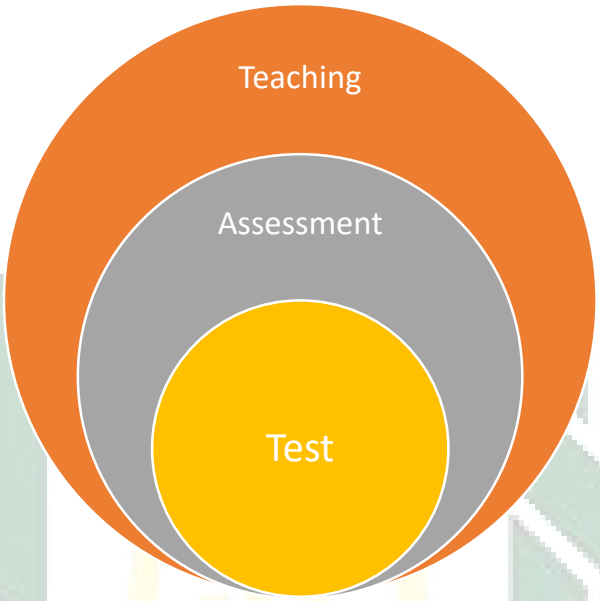
1. Language Testing

Language testing, like all educational assessment, is a complex social studies[1]. However, many experts define language testing in difference. Alan Davis, a Professor of Applied Linguistic, describes it as the activity of developing and using language test as well as a psychometric activity, that language testing traditionally concerned with the production, development and analysis of tests[2]. Meanwhile, Carol Chapelle and Geoff Brindley describes language testing as the act of collecting information and making judgments about a language learner's knowledge of a language and ability to use it[3]. Based on those experts' definitions, the researcher himself defines language testing as a theoretical formal study concerns on measuring four basic language domains output.

[1] Fulcher, G. (2010). *Practical Language Testing.* London: Hodder Education. Page: 1
[2] Davis, A. (1990). *Principles of Language Testing*. Cambridge: Blackwell Pub. Page: 6
[3] Chapelle, C. and G. Brindley. (2002). Assessment. In N. Schmidt (ed.) *An Introduction to Applied Linguistics.* London: Longman. Page 267.

Figure 2.1

Language Testing Field of Study



Source: Language Assessment Principles and Classroom Practices, Brown D.

Common people does not really recognize the scope of language testing. They tend to think that it only deals with the evaluation. However, the misconception needs to be comprehensively corrected. Language testing field of study, described as figure 2.1, includes teaching, assessment, and test as well.

Teaching sets up the practice of language learning: the opportunities for learners to listen, think, take risks, and set goals.[4] Assessment is an ongoing process that encompass students' respond of question, offers comment, or tries out a new

[4] Brown, D. (2004). *Language Assessment Principles and Classroom Practices*. New York: Longman Press. Page: 5

word or structure[5]. A test is a method of measuring a person's ability, knowledge, or performance in a given domain. The deeper explanation about test is discussed on the section below.

The purpose of language testing, based on J. B. Carol, is to render information to aid in making intelligent decisions about possible courses of action[6]. Nevertheless, Glenn Fulcher denies the statement by arguing: the purpose of such testing is primarily related to the needs of the teachers and learners working within a particular context[7]. Even though both statements tend to be similar in intention, researcher inclines to be more on Glenn's side, for the goal of the study is to fully comprehend the measurement, and also scale language learning development of each basic skills and competence domains.`

Language testing experts mostly agree that test is unquestionably the best way to assess learning process. Alan Davis emphasizes on recent critical and ethical approaches to language testing that have placed more stressing on the uses of language tests[8]. He also emphasizes on recent critical and ethical approaches to language testing that have placed more stressing on the uses of language tests.

[5] Brown… Ibid. Page 4
[6] Carroll, J. B. (1958). *Notes on the Measurement of Achievement in Foreign Languages*. Mimeograph: Library of the Iowa State University of Science and Technology. Page: 314.
[7] Fulcher, G. (2010). *Practical Language Testing.* London: Hodder Education. Page: 5
[8] Davis, A. (1990). *Principles of Language Testing*. Cambridge: Blackwell Pub. Page: 8.

2. Test

A test is a method of measuring a person's ability, knowledge, or performance in a given domain. It is an instrument: a set of techniques, procedures, or items that requires performance of the test takers.

Test is a method that must be described explicit and structured: multiple-choice questions with prescribed correct answers, a writing prompt with a scoring rubric, or an oral interview based on a question script[9]. In short, the whole part of a test such as questions, instructions, and scoring rubric needs to be described clearly so that the test takers comprehend what they are going to answer.

In order to judge the effectiveness of any test it is sensible to lay down criteria against which the test can be measured as valid and reliable[10]. In short, valid means a test that is supposed to test what it is supposed to test. It is not valid, for example, to test writing ability with an essay that requires some specific knowledge such as history or mathematics. Reliable means consistent. A good test gives consistent result. The deeper discussion about validity and reliability is given in the section after the explanation of test itself first.

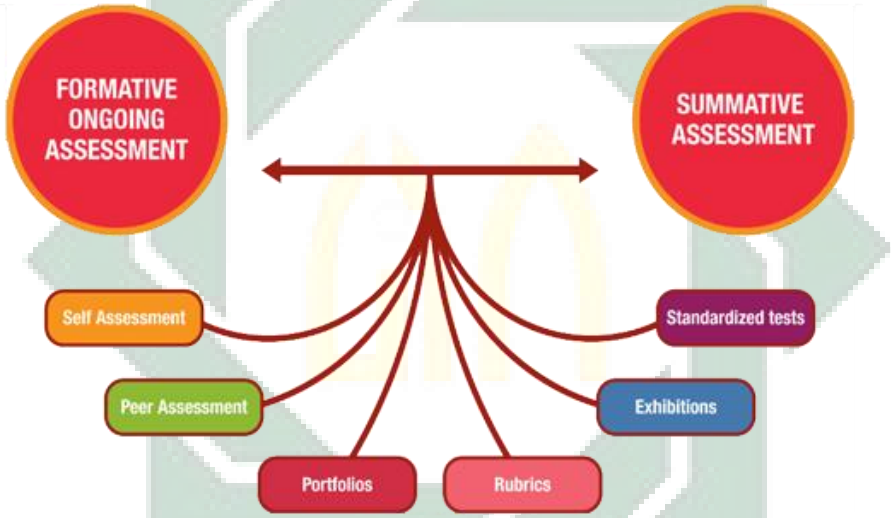
Tests has to measure specific individual ability[11]. There are 4 language skills that needs to be assessed: listening, reading, writing, and speaking. Discussing

[9] Brown, D. (2004). *Language Assessment Principles and Classroom Practices*. New York: Longman Press. Page: 3
[10] Harmer, J. (2006). *The Pratice of English Language Teaching*. Essex: Pearson Education Limited. Page 322.
[11] Brown…. Ibid. Page 3

deeper, Brown also adds that test must measure a common concept in the field of linguistic competence. Linguistic competence means knowledge about defining a vocabulary item, reciting a grammatical rule, or identifying a rhetorical feature in written discourse[12]. Furthermore, Bethan Marshall divides the test into two categories[13]:

Figure 2.2

Formative and Summative



Source: www.cmu.edu

a. Formative test

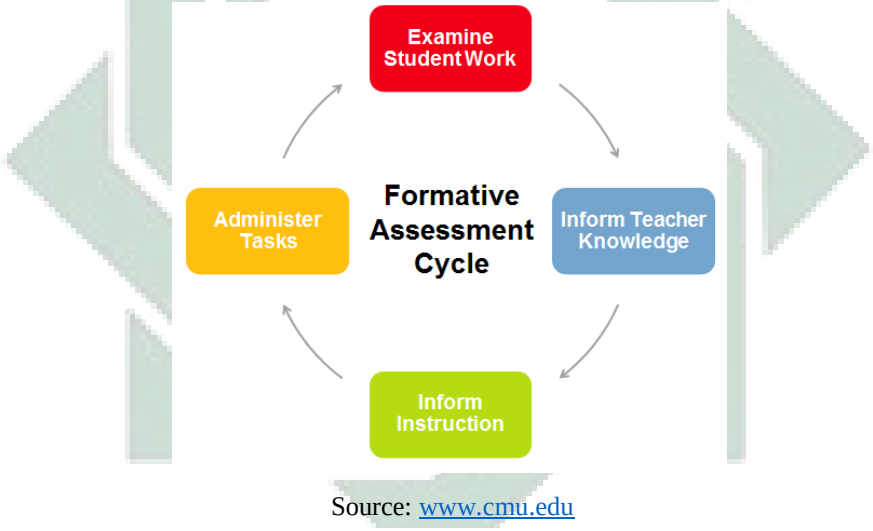
Most of classroom test, is evaluating students in the process of forming their competencies and skills with the goal of helping them to continue that growth

[12] Brown…. Ibid. Page 4
[13] Marshal, B. (2011). *Testing English Formative and Summative Approaches to English Assessment.* London: Continuum International Publishing Group. Page: 11

process. The key to such formation is the teacher's delivery and student's internalization of appropriate feedback on performance, with an eye toward the future continuation or formation of learning.[14] The example of formative test are quick quizzes, portofolio, self-assessments, role play, mapping, and practice quiz. Brown also gives diagram to map the deeper key attributes of formative assessment concept:

Figure 2.3

Formative Assessment Flow



Source: www.cmu.edu

1. A planned process

Formative assessment involves a series of carefully considered, distinguishable acts on the part of instructors or students or both.[15] The

[14] Brown, D. (2004). *Language Assessment Principles and Classroom Practices*. New York: Longman Press. Page: 6
[15] Brown, D. (2004). *Language Assessment Principles and Classroom Practices*. New York: Longman Press. Page: 5

researcher considered this step as concept of educational assessments such as play and poetry reading preparation.

2. Instructional adjustments

Formative assessment is to improve students' learning[16]. One of the most obvious ways to do this is for instructors to improve how they're teaching. Accordingly, one component of the formative assessment process is to adjust their ongoing instructional activities. Relying on assessment-based evidence of students' current status, such as test results showing that students are weak in their mastery of a particular cognitive skill, an instructor might decide to provide additional or different instruction related to this skill.[17] Researcher believes that the formative assessment process deals with ongoing instruction, modifications in educational activities that focus on students' mastery of the learning objectives reached.

3. Students' Learning Tactic Adjustments

Within the formative assessment process, students also take a look at assessment evidence and, if need be, make changes in how they're trying to learn[18]. The process dealing inside the students. The objectives of this process is to encourage, and also guide students to find their passion and develop it themselves. Teacher acts only as a guide.

[16] Bruner, J. (1996). *The Culture of Education*. Harvard University Press: Cambridge, MA. Page 156
[17] Bruner… *Ibid*. Page 157.
[18] Bruner… *Ibid*. Page 159.

The example of formative test is placement and diagnostic test:

– The Placement Test

The Placement test is to place a student into a some level or section of a language curriculum or school. The placement test also usually includes a sampling of the material to be covered in the various courses in a curriculum; a student's performance on the test should indicate the point at which the student will find material neither too easy nor too difficult but appropriately challenging[19]. A pre-test of Intensive English Program that should be taken by all new students of UIN Sunan Ampel Surabaya can be categorized as placement test because the result of the test use as the consideration on placing the students on the class.

– Diagnostic Test

A diagnostic test is designed to diagnose specific aspects of a language. A testing pronunciation, for example, might diagnose the phonological features of English that are difficult for learners and should therefore become part of a curriculum[20]. In shirt, this kind of test is used to identify the strengths and weaknesses of learners.

[19] Brown… Ibid Page 45
[20] Hughes. A. (2003). *Testing for Language Teachers Second Edition.* Cambridge: Cambridge University Press. Page 15

b. Summative test

Summative test aims to measure, or summarize, what a student has grasped, and typically occurs at the end of a course or unit of instruction[21]. A summation of what a student has learned implies looking back materials and learning process. Summative test answers questions raised from teachers to their students:

Figure 2.4

Summative Assessment Questions Requirements

Source: www.emaze.com

[21] Brown… *Ibid*. Page: 7.

Are there any common gaps in the learning? The answering of this questions is the summative score itself. If the highest score is far different from the middle or even lowest, this must be deal with 3 estimation: students' knowledge gaps, validity of the test, or the reliability.

Is there a need to develop further reteaching or enrichment activities? If the average score is below standard, teacher needs to improve the teaching by adding future reteaching or enrichment activities to enlarge students' knowledge.

What information did the students routinely masters? Students 'mastery can be seen from students' test handling.

What are the strengths and weaknesses in the instructional plans? The score of the test answers the strengths and weaknesses of the lesson plan. Teacher needs to evaluate the lesson plan continuously.

The example of summative tests are Proficiency Test and Achievement Test:

– Proficiency Test

Proficiency test are designed to test people's ability in a language, regardless any training they may have had in that language[22]. Moreover, brown states that proficiency test is not limited to any one course, curriculum, or single skill in the language; rather it

[22] Hughes. A. (2003). *Testing for Language Teachers Second Edition.* Cambridge: Cambridge University Press. Page 11.

tests overall ability[23]. The most well-known English proficiency tests are TOEFL, TOEIC, and IELTS

- TOEFL

TOEFL stands for Test of English as Foreign Language. The TOEFL test is the most widely respected English-language test in the world, recognized by more than 9,000 colleges, universities and agencies in more than 130 countries, including Australia, Canada, the U.K. and the United States[24].

Figure 2.5

TOEFL by ETS Original Logo

Source: www.ets.org

ETS, the formal authority holding license of TOEFL, is introduced as nonprofit organization that passionate about

[23] H. Douglas Brown, *Language Assesment: Principles and Classroom Practice* (Longman: California, 2003), 44.
[24] https://www.ets.org/toefl accessed on 27th of January '17.

advance quality and equity in education for all people worldwide. ETS provide innovative and meaningful measurement solutions that improve teaching and learning, expand educational opportunities, and inform policy[25]. The real TOEFL is a little bit expensive. The rate is about $300 each test. Some people are looking for TOEFL like-test to prepare themselves before taking the ‘’real’’ TOEFL. Those TOEFL like-tests have many famous name such as TOEFL Preparation, TOEFL Equivalent Test, and Mirror TOEFL.

- TOEIC

Figure 2.6

TOEIC by ETS Original Logo

Source: www.ets.org

[25] https://www.ets.org/about accessed on 27th of January ’17.

TOEIC stands for Test of English as International Communication. Different with TOELF, TOEIC highlights direct communicative skills such as listening. For more than 30 years, the TOEIC has set the standard for assessing English-language skills used in the workplace[26]. TOEIC test scores are used by nearly 14,000 companies, government agencies and English Language Learning programs in 150 countries, and more than seven million TOEIC tests were administered in 2013[27]. Based on ETS official web, there are advantages taking TOEIC tests: 1) Help businesses build a more effective workforce, 2) Give job seekers and employees a competitive edge, and 3) Enable universities to prepare students for the international workplace.

[26] https://www.ets.org/toeic/succeed accessed on 29 Januari '17
[27] https://www.ets.org/... *Ibid*, accessed on 29 January '17

- IELTS

Figure 2.7

IELTS Logo

Source: www.ielts.org

IELTS stands for International English Language Testing System. IELTS measures the language proficiency of people who want to study or work where English is used as a language of communication. It uses a nine-band scale to clearly identify levels of proficiency, from non-user (band score 1) through to expert (band score 9)[28]. IELTS is a variety of test that

[28] https://www.ielts.org/what-is-ielts/ielts-introduction accessed on 27th of January '17.

accepted world-wide as TOEFL. The variety goes as well as the IELTS types. There are two types that can be taken:

The first is IELTS Academic. The IELTS Academic test is for people applying for higher education or professional registration in an English speaking environment. It reflects some of the features of academic language and assesses whether you are ready to begin studying or training.[29]

The second is IELTS General Training. The IELTS General Training test is for those who are going to English speaking countries for secondary education, work experience or training programs. It is also a requirement for migration to Australia, Canada, New Zealand and the UK. The test focuses on basic survival skills in broad social and workplace contexts.[30]

In addition, British Council is introduced as the most known formal authority that hold IELTS. The British Council is the United Kingdom's international organization for educational opportunities and cultural relations. The British Council creates international opportunities for the people of the UK and other countries and builds trust between them worldwide[31]. The

[29] https://www.ielts.org/about-the-test/test-format accessed on 27th of January '17.
[30] https://www.ielts.org/... Ibid. Aaccessed on 27th of January '17.
[31] https://www.britishcouncil.in/about accessed on 28th of January '17.

British Council has more than 75 years' experience teaching and testing English. British Council has 500 test locations around the world and more than 90 countries.[32]

– Achievement Test

Achievement tests are directly related to language courses, their purpose being to establish how successful individual students, groups of students, or the courses themselves have been in achieving objective.[33] These tests are limited to particular material addressed in a curriculum within a particular time frame and are offered after a course has focused on the objectives in question. The example of this tests in Indonesia are daily examination (UH), mid semester examination (UTS), and final examination (UAS).

3. Validity

Brown states that validity is the degree to which a test measures what it claims, or purports, to be measuring[34]. Validation is an important enterprise especially when the test is a high stakes one. Admission tests for universities or

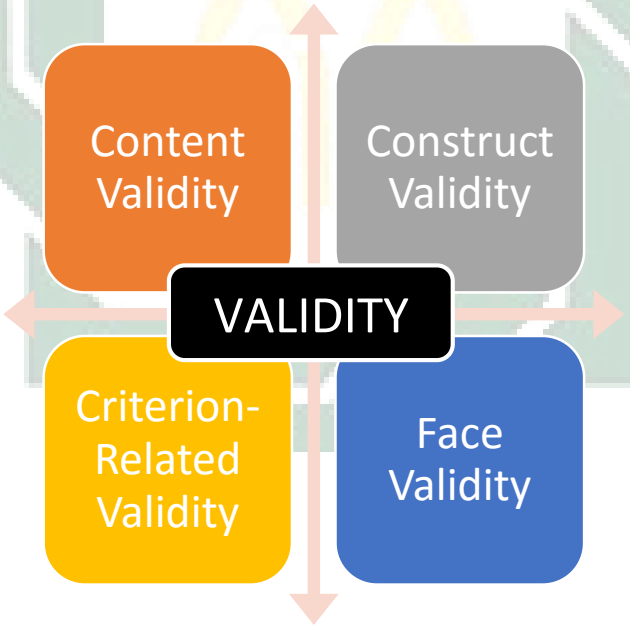
[32] https://www.britishcouncil.in/exam/ielts accessed on 28th of January '17.
[33] Hughes. A. (2003). *Testing for Language Teachers Second Edition.* Cambridge: Cambridge University Press. Page 13.
[34] Brown. J. D. (1999). *Testing in Language Program.* Upper Saddle River, Nj: Prentice Hall Regent. Page 193.

other professional programs, certification exams, or citizenship tests are all high-stakes assessment situations[35].

Validity is by far known as the most complex criterion and arguably the most important principle of a test quality[36]. He also adds that there is no final and absolute measure of validity, but several different kinds of evidence may be invoked in support.[37] In order to discover its bias, an expert helps define validity by his own way.

Figure 2.8

Validity Types



Source: Language Assessment Principles and Classroom Practices, Brown D.

[35] C. Roever, "Web-based Language Testing". *Language Learning and Technology,* Vol: 5 No:2 , 2001, 87.
[36] Brown, D. (2004). *Language Assessment Principles and Classroom Practices*. New York: Longman Press. Page: 22.
[37] Brown… Ibid. Page 22.

According to Messick , if the validity of a test is not known it might have undesirable consequences for the society at large[38]. One validates not a test, but 'a principle for making inferences[39]. There are 4 types of validity:

a. Content Validity

Hughes said that a test can be said to have content validity if its content constitutes representative sample of the language skills, structures, etc[40]. Basically, content validity depends on the extent to which an empirical measurement reflects a specific domain of content[41]. The test can be said to have a good content validity if the test actually samples the subject matter about which conclusions are to be drawn, and if it requires the test-takers to perform the behavior that is being measured[42]. In simply, content validity is related to the meant/content of the test. Such as in structure section, the test items should be made up by the correlating knowledge of structure.

[38] S. Messick - H. Wainer, & H. Braun (Eds.). (1998). *The Once and Future Issues of Validity: Assessing The Meaning and Consequences of Measurement.* Hillsdale, NJ: Erbaum, , 35.
[39] L. J. Cronbach & P. E. Meehl, "Construct Validity in Psychological Tests". *Psychological Bulletin*. 52, 1955, 297.
[40] Arthur Hughes, *Testing for Language Teachers Second Edition* (Cambridge: Cambridge University Press, 2003), 26
[41] Edward Carmines & Richard Zeller, *Reliability and Validity Assessment* (London: Sage University Press, 1987), 17.
[42] H. Douglas Brown, *Language Assesment: Principles and Classroom Practice* (Longman: California, 2003), 22.

b. Construct Validity

The word 'construct' can be defined as psychological construct such as proficiency and ability[43]. For example, the "overall English proficiency" is a construct. Then, a test can be said to have good construct validity if the test can surely measures what it claims to measured. This is the main topic of this study so the researcher will give more detailed information on the next sub chapter.

c. Criterion-Related Validity

Nunnally defines the criterion-related validity as when the purpose is to use an instrument to estimate some important form of behavior that is external to the measuring instrument itself, the latter being referred to as the criterion[44]. The result on the test agrees with some independent and highly dependable assessment of the candidate's ability[45]. A criterion- related validity can be proven if the notion of "criterion" of the test has actually been reached. Criterion -related validity can be divided into two categories: concurrent and predictive validity.

[43] James Dean Brown. "What Is Construct Validity?" *Shiken: JALT Testing & Evaluation SIG Newsletter*. Vol: 4 No: 2, 2000, 9.
[44] J.C. Nunally, *Psychometric Theory*. (New York: Mc Graw Hill, 1978), 87.
[45] Arthur Hughes, *Testing for Language Teachers Second Edition* (Cambridge: Cambridge University Press, 2003), 27.

d. Face Validity

Mousavi stated 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, based on the subjective judgment of the examinees who take it, the administrative personnel who decide on its use, and other psychometrically unsophisticated observers[46]. Thus, a test is said to have face validity if it looks as if it measures what it is supposed to measure.

4. Reliability

Reliability is one of the most important elements of test quality[47]. In language testing issue, different experts have different redaction in defining reliability. Fulcher explains reliability as the center of a test enterprise[48]. Another expert, Stainback, states that reliability is often defined consistency and stability of data[49]. Roever also adds that reliability is a must thing have which every single test should insist on, especially if the test is the high-stakes one. High-stakes assessment situations are admission tests for universities or other professional programs, certification exams, or citizenship tests[50]. More specifically, reliability concerns to the extent to which a test, or any measuring procedure yields the same results on repeated trials. The measurement of any phenomenon always contains a certain

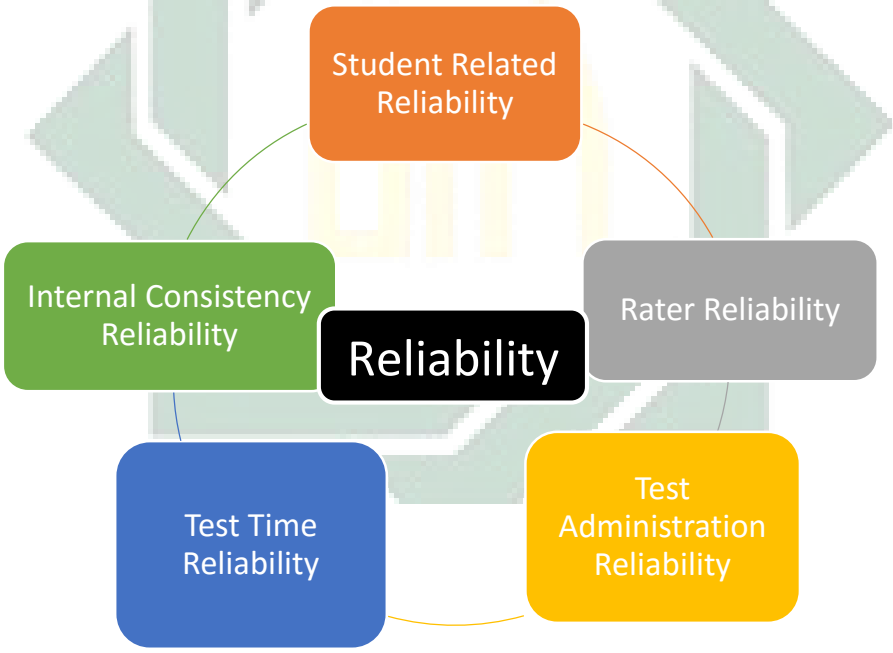
[46] Sayyed Abbas Mousave, An Encyclopedic Dictionary of Language Testing Third Edition. (Taiwan: Tuang Hua Book Company, 2002), 125.
[47] Professional Testing Incorporated. (2006). *Test Reliability*. Page 1
[48] Glenn F. (2010). *Practical Language Testing*. United Kingdom: Hodder Education. Page 19.
[49] Stainback, S. (2007). *Research and Statistics*: Cambridge: Cambridge University Press. Page 67
[50] C. Roever. (2010). "Web-based Language Testing". Language Learning and Technology. Vol. 5 No. 2, Page 86.

amount of chance error[51]. In short, a reliable test is consistent and dependable. If the same test to the same students or matched students on two different occasions, the test should yield the similar result.

In practice, reliability is enhanced by making the test instructions absolutely clear, restricting the scope for variety in the answers, and making sure that test conditions remain constant[52].

Figure 2.9

Reliability Types



Source: Language Assessment Principles and Classroom Practices, Brown D.

[51] Carmines, Edward & Zeller, Richard… *Ibid*. Page 11-12
[52] Harmer, J. (2006). *The Pratice of English Language Teaching*. Essex: Pearson Education Limited. Page 322.

The issues of reliability of a test may best be addressed by considering a number of factors that may contribute to the unreliability of a test[53]. The following consideration possibilities may fluctuate the result: student-related reliability, rater reliability, test administration reliability, test reliability[54] and internal consistency reliability[55].

a. Student-Related Reliability

The most common learner-related issue in reliability is caused by temporary illness, fatigue, a bad day, anxiety, and other physical or psychological factors, which may make score deviate from the true one.[56]

There are some cases that students feel illness during the test. But the consequences are individuals'. Therefore, students need to have proper physical and psychological preparation for encouraging fit condition before taking the exam.

b. Rater Reliability

Rater reliability deals with human error, subjectivity, and bias that may enter into the scoring process. The correction application tends to have more attention for this step may cause unreliability bias. This principal specifically

[53] Brown… *Ibid*. Page: 21
[54] Mousave, S. A. (2002). An Encyclopedic Dictionary of Language Testing Third Edition. Taiwan: Tuang Hua Book Company. Page: 801
[55] McMillan, J. (2014). *Research in Education: Evidence-Based Inquiry. Facts101*: Textbook Outline. Page 181.
[56] Mousave… *Ibid*. Page 804

divided into two categories by Brown: Inter-rater reliability and Intra-rater reliability[57].

Inter-rater reliability occurs when two or more scores yield inconsistent scores of the same test, possibly for lack of attention to scoring criteria, inexperience, inattention, or even preconceived bias[58].

Intra-rater reliability is a common occurrence for classroom teacher because of unclear scoring criteria, fatigue, bias toward terms good and bad students, or simple carelessness[59]. The careful specification of an analytical scoring instrument, however, can increase rater reliability[60].

As rater reliability takes place in the end of the test correction, some aspects such as subjectivity and human error must be totally avoided to keep the quality of the test result.

c. Test Administration Reliability

Unreliability may also result from the conditions in which the test is administered[61]. It deals with the practicality stuff such as class condition, the quality of tape recorder, the clearness of question sheet, paper thickness, light

[57] Brown… *Ibid*. Page: 21
[58] Brown… *Ibid*. Page: 21
[59] Brown… *Ibid*. Page: 21
[60] Brown, J. D. (1991). *New Ways of Classroom Assessment*. Alexandria, VA: Teachers of English to Speakers of Other Languages. Page: 289
[61] Brown, D. (2004). *Language Assessment Principles and Classroom Practices*. New York: Longman Press. Page: 21

adequateness, classroom temperature, and the arrangement off desks and chairs[62].

Unclear tape recorder, dull lighting, or even dirty class may cause students feel irritable and discomfort during answering the test. In order to increase the test administration reliability, the authority of the test holder needs to pay attention on this case. The test holder has to try the audio quality before it is played, the adequate lighting before the students come to the class, and make sure that the class used is clean and tidy.

d. Test Time Reliability

Nature of the test itself can cause measurement errors. If a test is too long, test-takers may become fatigued by the time they reach the later items and hastily respond incorrectly[63]. In addition, Gleen Fulcher also emphasizes on the length of the test as the number of items is correlated with time given[64]

e. Internal Consistency Reliability

Internal consistency reliability is an assessment of how reliably survey or test items are designed to measure the same construct. In specific, a construct is an underlying theme, characteristic, or skill such as reading comprehension or customer satisfaction[65]. There are a wide variety of internal consistency

[62] Brown…. Ibid. Page: 21
[63] Brown…. Ibid. Page: 22
[64] Hughes, A. (2003). *Testing for Language Teachers Second Edition*. Cambridge: Cambridge University Press. Page: 57.
[65] Brown, D. (2004). *Language Assessment Principles and Classroom Practices*. New York: Longman Press. Page: 124

measures that can be used[66] such as Kuder Richardson 20, KR 21, Anova Hoyt Variants Analysis, and Spearman-Brown formula. However, every formula has requirements that must be fulfilled. This research uses Spearman-Brown Formula as the data is in the shape of total score.

Internal consistency reliability analysis results a value that can be generalized into test quality standard. As internal consistency reliability means the test items' consistency and dependency, it does affect the output result. The higher internal consistency reliability value of a test, the more a test generates the same score as the previous ones.

There are two ways to obtain internal consistency reliability value: Average Inter-item Correlation and Split-half Method[67].

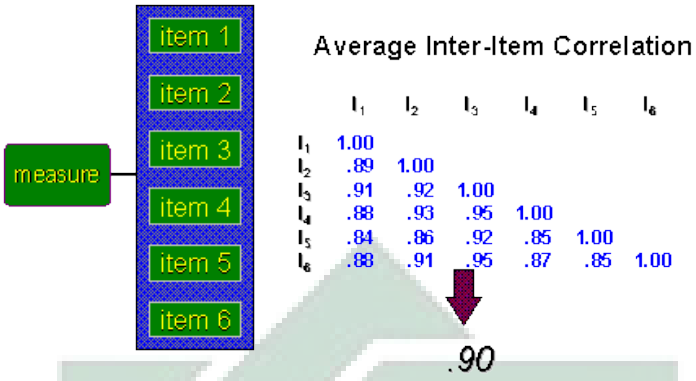
•. Average Inter-item Correlation

Average Inter-item correlation is obtained by taking all of the items on a test that probe the same construct, determining the correlation coefficient for each pair of items, and finally taking the average of all of these correlation coefficients[68]. Inter-rater reliability is also known as *inter-observer reliability* or *inter-coder reliability*.

[66] Brown… Ibid Page 125.
[67] McMillan, J. (2014). *Research in Education: Evidence-Based Inquiry. Facts101*: Textbook Outline. Page 181.
[68] Cozby, C. (2001). *Measurement Concepts. Methods in Behavioral Research*. California: Mayfield Publlishing Company. Page 231

Figure 2.10

Average Inter-item Correlation



Source: www.socialresearchmethods.net

This is the best way of assessing reliability when using observation, as observer bias very easily creeps in. However, this method needs re-test which the researcher believes is inefficient because of the practical consideration. Inefficient practical consideration means the impossibility for both P2B to give the second test just in order to measure the internal reliability. Therefore the split-half reliability is the most efficient way to test the internal reliability itself.

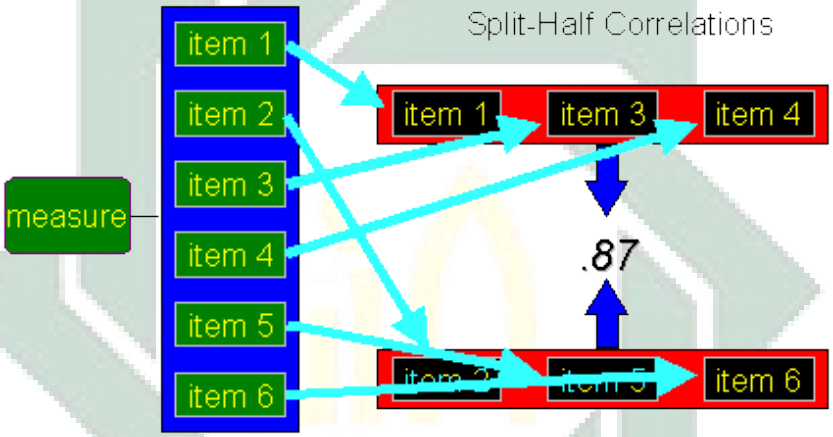
•. Split Half

Split-half is another subtype of internal consistency reliability. The process of obtaining split-half reliability is begun by ‘’splitting in half’’ all items of a test that are intended to probe the same area of knowledge in order to form two ‘’sets’’ of items. The entire test is administered to a group of

individuals, the total score for each set is computed, and finally the split-half reliability is obtained by determining the correlation between the two total ''set'' scores[69]. In short, this is done by comparing the results of one *half* of a test with the results from the other *half*.

Figure 2.11

Split-Half Correlations



Source: www.socialresearchmethods.net

Researcher believes that this way is the most effective and efficient way to get the data for internal consistency reliability analysis. International standard used for analyzing the score data is Spearman-Brown Formula which measure internal consistency precisely[70]. In addition, Pearson Product Moment is absolutely needed because the Spearman-Brown Formula

[69] James. D. B (2009). *What Is Internal Consistency Reliability*?" Shiken: JALT Testing & Evaluation SIG Newsletter. Vol 4: No: 2 Page 9.
[70] James, H., Millan M.C., Schumacher S. *Research in Evidence Based Inquiry*. Pearson: Commonwealth University: Pearson. Page. 181

requires coefficient correlation value. The value is obtained by examining the correlation between two total set scores.

• Pearson Product Moment Formula

Pearson Product Moment or Pearson Correlation Coefficient is a statistical tool that takes function to examine the relationship between two variables. Correlation itself is based on two words, “co—“and “relation”. The word ‘‘co’’ means gather as one, pair, and the same level. The word ‘’relation’’ can be synonymized as effect or connection. Thus, the definition of correlation based on the statistics is a method as to know the connection between factors, or variables being examined. In mathematics, the correlation is symbolized as ‘’rxy’’.

Pearson Product Moment is a basic method that is often used to examine connection between variables and factors. However, there are two requirements before using Pearson Product Moment:

1) Sample is gathered by Random Sampling Technique

Another data sampling such as snowball sampling, or multi clustered sampling is not allowed. This is because the data that will be examined the connection need to be fair gathered.

2) Data must be homogeny.

Data that will be processed has to be homogeny. It means that the data can be generalized.

Furthermore, the basic linear correlation is used to assess the direction of two variables. This is the formula of Pearson Product Moment.

$$r_{XY} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2}\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$$

rxy: Pearson Correlation Coefficient

n : Total data

∑X: Variable X total score

∑Y: Variable Y total score

‘’rxy’’ as Pearson Correlation Coefficient is needed before applying the Spearman-Brown Formula. Symbol ‘’n’’ is the total data. ∑X is the total score of variable x, while ∑Y is the total score of variable Y.

To ease the formula’s counting, helper table is needed. Helper table is a common table filled with total data, X, Y, X^2, Y^2, XY.

Table 2.12

Helper Table

No. Data	**X**	**Y**	**X^2**	**Y^2**	**XY**

After the helper table is filled with data, and the formula is processed, the rxy is obtained. Therefore, the Spearman-Brown formula as the final step for examine the internal consistency reliability value can be applied.

• The Spearman-Brown Formula

The Spearman-Brown Formula is used for examining internal consistency reliability. The formula is used to increase split-half reliabilities to estimate what the correlation would be for whole test[71].

. $$\rho = \frac{2r}{1+r} =$$

SPLIT_HALF (R1, R2) = split half coefficient (after Spearman-Brown correction) for data in ranges R1 and R2

SPLITHALF (R1, *type*) = split-half measure for or the scores in the first half of the items in R1 vs. the second half of the items if *type* = 0 and the odd items in R1 vs. the even items if *type* = 1.

After the ‘’ρ’’ (internal consistency reliability) is found, it needs to be assessed with this rubric criteria.

[71] James… Ibid Page: 181

Table 2.13

Standard Criteria

r	Interpretation
0	No correlation
0,01 – 0,20	Very low correlation
0,21 – 0,40	Low correlation
0,41 – 0,60	Quite low correlation
0,61 – 0,80	Quite correlation
0,81 – 0,99	High correlation
1	Very high correlation

Source: Pengantar Statistika, Usman H. and Setiady A.

4. TOEFL Equivalent Test in UIN Sunan Ampel Surabaya

TOEFL Equivalent Test is an English proficiency test which is produced by P2B of UIN Sunan Ampel Surabaya which use TOEFL as the standard in giving the scores and making the questions. P2B does not make the test items by themselves, they take the items from various references such as Cliff's TOEFL Preparation Standardization by Michaele A. Pyle and Longman[72]. This test is divided into three sections: listening, grammar and reading. The minimum score of this test is 400. If students fail to pass on the first test, they can take the second test

[72] Aina… Ibid. Page 1.

and so on until they are able to reach the score. The certificate of TOEFL-like test is also used as one of the requirement for participating in thesis examination.

a. Section of TOEFL Equivalent Test

Based on the book entitled Road to English Proficiency test of UIN Sunan Ampel Surabaya, one of the material resource in intensive English program, the TOEFL-like test consists of three sections, they are:

A) Listening Comprehension

1) Definition of Listening Comprehension

This section tests the test-takers' ability in listening to dialogue or short lecture on English through tape recorder or others media which prepared by P2B. This section consists of 50 questions and forty minutes for doing it.

2) Sections of Listening Comprehension

a) Short Dialogues

In this short dialogue, the test-takers will hear the part A. The test-takers do not need to understanding the whole dialogue in answering the questions. The most important thing is focusing on some key words which can be in form of noun and verb. The key words is often said by the second speakers. Here is the example of short dialogue question.

Woman : Can you have this report written, typed, copied, and mailed before the post office closes today?

Man : Today?

What does the man mean?

A. The post is already closed

B. The report is due tomorrow

C. He can't finish all these task today

D. He will be able to mail the report today

b) Long Dialogue

Long dialogues are categorized as part B of listening section. Commonly, the test-takers will hear two dialogues with three to four questions for each dialogue. However, one long dialogue may also have seven to eight questions. Based on Road to English Proficiency Book, each long dialogue usually consists of 140 to 290 words and 40 to 80 seconds time for listening it.[73] The test-takers are not allowed to take

[73] I.W. Harits & M. Kurjum (2009). *Road to English Proficiency Test*. Surabaya: IAIN Sunan Ampel Language Center. Page 7.

note while listening to the long dialogues and the questions. The example of the long dialogue question.

Woman : I've registered for all my classes, and fortunately I'm happy with my professors. Now, all I need to do is buy my books.

Man : Let's go over the list you've been given, and I'll direct you to the shelves where you can find them.

What will probably be the main topic of this conversation?

a. How to register for classes

b. The best professors on campus

c. Where to locate required classroom books

d. how to use the library

c) Long Lecture

In Part C, the test-takers wil hear some short lectures which usually called "talks". In this part, the theme of talks is usually about first year college student orientation, lectures, and also about the college students' life. The duration of the talks is not more than 2 minutes. The vocabularies used in the talks is more specific so it is more

difficult to understand the talks. The example of long lectures question is presented as follow:

''Today we'll continue our study of space exploration. If you remember, last week we discussed the first lunar module and what plans for future lunar landings. Today, we'll look at the most recently develop spacecraft, the shuttle craft, which replaced the wasteful-single use rockets and spacecraft of the past.''

What will probably the main topic of this lecture?

a. Wasteful policies past space programs
b. The importance of lunar landing
c. Current and future space exploration programs
d. The characteristics of the space shuttle

B) Structure and Written Expression

1) Definition

This section test the test-takers' ability in understanding structure and written expression of English as well as able to use and know the misused of it. This section consists of forty questions and twenty five minutes for doing it.

2) Section of Structure and Written Expression

a). Sentence Completion

This kind of question is an incomplete sentence, for example a sentence which the place of verb or to be is empty. So the test-takers need to fill the blank space by choosing the right answers. Here is the example of sentence completion question:

The company had dumped waste into the river for years and it _____ to continue doing so.

a. Plans
b. Planning
c. Planed
d. Had planned

b). Finding Grammatical Errors

In this kind of question, there will be find four words or phrases which being underlined. The test-takers need to choose one the underlined word / phrase which might having the grammatical errors. Here is the example of this question:

Thousands of settlers gone west after the Civil War ended

A B C D

B) Reading Comprehension

This section test the test-takers' ability in comprehending various academic reading related to the topic, main idea, reading content, word meaning, or word classification and detailed information of it. This

section consists of fifty questions and fifty five minutes for doing it. The example of the questions as presented below:

''The next artist in this survey of American artists is James Whistler. He is included in this survey of American artists because he was born in the United States, although the majority of his artwork was completed in Europe. Whistler was born in Massachusetts in 1834, but nine years later his father moved the family to St. Petersburg, Russia, to work on the **construction** of a railroad. The family returned to the United States in 1849. Two years later Whistler entered the U.S military academy at West Point, but he was unable to graduate. At the age of twenty one, Whistler went to Europe to study art despite familial **objections**, and he remained in Europe until his death.

Whistler worked in various art forms, including **etchings** and lithographs. However, he is most famous for his paintings, particularly *Arrangement in Gray and Black No. 1: Portrait of the Artist's Mother or Whistler's Mother*, as it is more commonly known. This painting shows a side view of the portrait with his mother seated off – centre, is highly characteristic of Whistler's work.''

1. The paragraph preceding this passage most likely discusses ....
a. a survey of eighteenth century article.
b. Whistler's other famous paintings.
c. the work of European artists.
d. a different American artists.

2. Which of the following best describes the organization of the information in the passage?
a. One artist's life and works are described.
b. Various paintings are contrasted.
c. Whistler's family life is outlined.
d. Several artists are presented.

3. The word "objections" in line 8 is closest in meaning to ....
a. agreements.
b. protests.
c. battles.

d. goals.

4. In line 9, the word “etchings” refers to ....
a. an art form introduced by Whistler.
b. an art form involving engraving.
c. the same as lithograph.
d. a type of painting.

5.Whistler is considered an American artist because ....
a.he created most of his famous art in America.
b.he spend most of his life in America.
c.he served in the U.S military.
d.he was born in America.

6.It is implied in the passage that Whistler’s family was ....
a.highly supportive of his desire to pursue art.
b.very influential in U.S military academy.
c.considered as a working class family.
d.unable to find any work in Russia.

7.Which of the following is NOT true according to the passage?
a.*Whistler’s Mother* is not the official name of his painting.
b.Whistler’s Mother is painted in sombre tones.
c.Whistler worked with a variety of art forms.
d.Whistler is best known for his etchings.

B. Review of Previous Study

Here, the researcher reviews some researchers which were related to this research, as follows:

Related to this research, there were some similar researches which have relationship with this research; the first and newest study was done by Qory Aina, UIN Sunan Ampel Surabaya in 2016. The title was *"AN ANALYSIS OF CONSTRUCT VALIDITY OF TOEFL-LIKE TEST IN ENGLISH INTENSIVE COURSE PROGRAM OF UIN SUNAN AMPEL SURABAYA*"[74]. Qorry Aina measured the construct validity of TOEFL-like test. The setting of the study was in UIN Sunan Ampel Surabaya and the subjects are 183 student and. This study used descriptive method. The data in this study were the question's sheet and the students' answers of TOEFL-like test. The instrument of this research is in form of documents. The result stated only minor items are not valid[75]. The items are 20, 102, 111, 106, 112, 62, 66, 67, 85, and 83 which counted total as 10 out 140. The rotation of test items shows that the test items are not able to measure the indicators.

The other similar study was done in 2014, entitled "*AN ANALYSIS OF TEST-TAKING STRATEGIES USED IN TOEFL EQUIVALENT TEST BY SIXTH SEMESTER STUDENTS OF ENGLISH TEACHER EDUCATION DEPARTMENT*

[74] Aina, Q. (2016) *AN ANALYSIS OF CONSTRUCT VALIDITY OF TOEFL-LIKE TEST IN ENGLISH INTENSIVE COURSE PROGRAM OF UIN SUNAN AMPEL SURABAYA.* Undergraduate thesis, UIN Sunan Ampel Surabaya

[75]Ibid… Aina. Page 57

UIN SUNAN AMPEL SURABAYA" conducted by Elis Rahmawati[76]. Here, the researcher discussed about TOEFL. The TOEFL is divided into TOEFL by ETS and TOEFL by Language Development Center. This study gives much definition and clear statement to my research.

. Another research was done by Althafurrahman Wafi in 2016 with the research entitled "*A PREDICTIVE VALIDITY ANALYSIS ON "SELECTION TEST" OF FOREIGN LANGUAGE DEVELOPMENT INSTITUTE OF NURUL JADID, PAITON, PROBOLINGGO*"[77]. The researcher attempted to find out the predictive validity of Foreign Language Development Institute (FLDI) of Nurul Jadid. His study focused on descriptive quantitative that analyze document analysis as the instrument. Even though validity and reliability is different things, but they are considered as one entity that cannot be separated. They are in the field study of language assessment. The result is the high value of selection test's predictive validity.

The fourth study was done by Ullia Dwi Agustina entitled "AN ANALYSIS OF THE TEST ITEMS IN ENGLISH TRY-OUT TEST FOR UN 2010/2011

[76] Rahmawati, E. (2014) *AN ANALYSIS OF TEST-TAKING STRATEGIES USED IN TOEFLEQUIVALENT TEST BY SIXTH SEMESTER STUDENTS OF ENGLISH TEACHER EDUCATION DEPARTMENT UIN SUNAN AMPEL SURABAYA*. Undergraduate thesis, UIN Sunan Ampel Surabaya

[77] Wafi, Althafurrahman (2016) *A PREDICTIVE VALIDITY ANALYSIS ON "SELECTION TEST" OF FOREIGN LANGUAGE DEVELOPMENT INSTITUTE OF NURUL JADID, PAITON, PROBOLINGGO.* Undergraduate thesis, UIN Sunan Ampel SurabayaUIN Sunan Ampel Surabaya, 2016)

PUBLISHED BY DIKNAS SURABAYA"[78]. The analysis went deep into test items that was analyzed in the way of face and content validity were constructed. The research was in field of language assessment. This study resulted the analysis of each item presented in the table analysis.

The fifth study is "INTERNAL CONSISTENCY, RETEST RELIABILITY, AND THEIR IMPLICATIONS FOR PERSONALITY SCALE VALIDITY"[79] by Robert R, McCrae, John E. Kurtz, Shinji Yamagata, and Antonio Terraciano. This research examined psychometric properties such as ages, cultures, and methods of measurement and the relationship with validity criteria associated with different scales of reliability.

"AN ASSESSMENT OF THE INTERNAL CONSISTENCY OF MEASURES OF CONSTRUCTS USED TO REVISE THE INNOVATION DECISION FRAMEWORK[80]" by Raja Peter and Vasanthi Peter becomes the sixth studies. The study analzed internal consistency of diffused and multi literature. The internal consistency reliability analysis approach is adopted in the study for allowing identification of variables which had more than one measurement constructs. Using

[78] Agustina, U. D., (2011). *AN ANALYSIS OF THE TEST ITEMS IN ENGLISH TRY –OUT TEST FOR UN 20120/2011 PUBLISHED BY DIKNAS SURABAYA*. Undergraduate thesis, UIN Sunan Ampel Surabaya.
[79] Mc Crae, R.R., Kurtz, J.E., Yamagata S., and Terracciano, A. (2011). *INTERNAL CONSISTENCY, RETEST RELIABILITY, AND THEIR IMPLICATION FOR PERSONALITY SCALE VALIDITY*. Pers Soc Psychol Rev.
[80] Peter, R., and Peter, V. (2008). *AN ASSESSMENT OF THE INTERNAL CONSISTENCY OF MEASURES OF CONSTRUCTS USED TO REVISE THE INNOVATION DECISION FRAMEWORK*. Academy of World Business, Marketing, and Management Devleopment. Volume 3 No. 1

Annova, the result of the study is the variance of three internal consistency reliability value.

The last previous study was in the form of journal "INTERNAL CONSISTENCY: DO WE REALLY KNOW WHAT IT IS AND HOW TO ASSESS IT?" by Wei Tang, Ying Cui, and Oksana Babenko[81]. This research focuses on meanings of theoretical and practical concept of internal consistency. The analysis goes deep in difficulties, interpretation, and redefinition of the complex context of internal consistency. The researchers also adds new and better indices for measurement. In addition, built on the review of various meanings and measurement, the study attempted to provide an explicit definition of internal consistency, added with recommendation of appropriate measures for assessment.

Seeing from the studies that have been conducted before, the researcher concludes that all previous studies have the similarity and different areas of study. Those previous studies could be the foundation of conducting this research. The previous studies mostly focusing on the language assessment, TOEFL in Language Development Center of UINSA and the validity and reliability study, while in this research, the researcher focuses on the internal consistency of TOEFL in Language Development Center.

[81] Tang W., Cui Y., Babenko O. (2014). *INTERNAL CONSISTENCY: DO WE REALLY KNOW WHAT IT IS AND HOW TO ASSESS IT?* American Research Institute for Policy Development: Journal of Psychology and Behavioral Science. Vol. 2, No. 2.