CHAPTER II

REVIEW OF RELATED LITERATURE

This chapter discusses some lite related to the construct validity of TOEFLlike test. The related literature covered validity, test, TOEFL-like test at EIP, intensive English course program, language development center (P2B) of UIN Sunan Ampel Surabaya and some previous studies.

A. Validity

1. Definition of validity

Brown states that validity is the degree to which a test measures what it claims, or purports, to be measuring¹. Validation is an important enterprise especially when the test is a high stakes one. Admission tests for universities or other professional programs, certification exams, or citizenship tests are all high-stakes assessment situations.² According to Messick , if the validity of a test is not known it might have undesirable consequences for the society at large.³ One validates not a test, but 'a principle for making inferences'.⁴

¹ James Dean Brown, *Testing in Language Program* (Upper Saddle River, Nj: Prentice Hall Regent, 1999),193.

² C. Roever, "Web-based language testing". *Language learning and technology*, Vol: 5 No:2, 2001, 87.

³ S. Messick - H. Wainer, & H. Braun (Eds.), *The once and future issues of validity: Assessing the meaning and consequences of measurement* (Hillsdale, NJ: Erbaum, 1998), 35.

⁴ L. J. Cronbach & P. E. Meehl, "Construct validity in psychological tests". *Psychological Bulletin*. 52, 1955, 297.

2. 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.⁵Basically, content validity depends on the extent to which an empirical measurement reflects a specific domain of content.⁶ 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.⁷ 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.

b. Construct Validity

The word 'construct' can be defined as psychological construct such as proficiency and ability.⁸ 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.

⁵ Arthur Hughes, *Testing for Language Teachers Second Edition* (Cambridge: Cambridge University Press, 2003), 26.

⁶ Edward Carmines & Richard Zeller, *Reliability and Validity Assessment* (London: Sage University Press, 1987), 17.

⁷H. Douglas Brown, *Language Assesment: Principles and Classroom Practice* (Longman: California, 2003), 22.

⁸ James Dean Brown. "What is construct validity?" *Shiken: JALT Testing & Evaluation SIG Newsletter*. Vol: 4 No: 2, 2000, 9.

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.⁹ The result on the test agrees with some independent and highly dependable assessment of the candidate's ability.¹⁰ A criterionrelated validity can be proven if the notion of "criterion" of the test has actually been reached.¹¹ Criterion -related validity can be divided into two categories: concurrent and predictive validity.

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.¹² Thus, a test is said to have face validity if it looks as if it measures what it is supposed to measure.

⁹ J.C. Nunally, *Psychometric Theory*. (New York: Mc Graw Hill, 1978), 87.

¹⁰ Arthur Hughes, *Testing for Language Teachers Second Edition* (Cambridge: Cambridge University Press, 2003), 27.

¹¹ H. Douglas Brown, *Language Assessment: Principles and Classroom Practice* (Longman: California, 2003), 24.

¹² Sayyed Abbas Mousavi, An Encyclopedic Dictionary of Language Testing Third Edition. (Taiwan: Tuang Hua Book Company, 2002), 125.

3. Construct Validity

a. Definition of Construct

Arthur Hughes states that the word 'construct' refers to any underlying ability (or trait) that is hypothesized in a theory of language ability.¹³ In TOEFL-like test for example, the ability that being measured is listening, reading and structure. On the other hand, Glenn states that constructs are the abilities of the learner that we believe underlie their test performance, but which we cannot directly observe.¹⁴ We are not able to directly observe the students' language ability, therefore the test are made in case of measuring it.

b. Definition of Construct Validity

James Dean Brown defines construct validity as the experimental demonstration that a test is measuring the construct it claims to be measuring.¹⁵ In simply, it means that the construct validity of a test can be proven if the test is able to measure what it tends to assessed. In TOEFL-like test, the construct validity can be proved if the listening section, for example, is able to measure the students' listening ability. The TOEFL-like test can be said that it has good construct validity if each items of the test can surely measure the skill of language that need to be measured. For

¹³ Arthur Hughes, *Testing for Language Teachers Second Edition* (Cambridge: Cambridge University Press, 2003), 31.

¹⁴ Glenn Flucher, *Practical Language Testing* (UK: Hodder Education, 2010), 96.

¹⁵ James Dean Brown. "What is construct validity?" Shiken: JALT Testing & Evaluation SIG Newsletter. Vol: 4 No: 2, 2000, 9.

example, the questions' items in reading section should be able to measuring the students' reading skill. Moreover, Cronbach and Meehl observed that construct validity must be investigated whenever no criterion or universe of content is accepted.¹⁶

c. The strategies used to gather construct-related evidence.

The result of construct validity is not in form of validity coefficient. ¹⁷ Therefore, in measuring the construct validity we should gather the evidence in which the construct validity existed. Donald Ary et al porpose some strategies that can be used in gathering the evidence of construct validity as follows: ¹⁸

1) Related measures studies

The aim of this strategy is to show that the test in question measures the construct it was designed to measure and not some other theoretically unrelated construct. There are two types of evidence based on relations to other variables: convergent and discriminant. Convergent evidence related to the relationships between test scores and other measures, whereas relationships between test scores and measures purportedly of different constructs provide discriminant

¹⁶ L. J. Cronbach & P. E. Meehl, "Construct validity in psychological tests". *Psychological Bulletin*. 52, 1955, 285.

¹⁷ Saifuddin Azwar, *Reliabilitas dan Validitas Edisi 4* (Yogyakarta : Pustaka Belajar), 4.

¹⁸ Donald Ary, Lucy Cheser Jacobs and Asghar Razavieh, *Introduction to Research in Education* (USA: Cangage Learning, 2010), 231.

evidence.¹⁹ Campbell and Fiske used a multitrait– multimethod matrix (MTMM) of correlation coefficients in evaluating convergent and discriminant validity of a construct. Campbell and Fiske assumed that the same construct should correlate with each other even if they use different methods (convergent validity), and the measures of different constructs should not correlate with each other even if they use the same method (discriminant validity).²⁰

2) Known-groups technique

In this technique, the researcher compares the performance of two groups already known to differ on the construct being measured. The researcher then make a hypothesis that the group known to have a high level of the construct will score higher on the measure than the group known to have a low level of the construct. If the hypothesis is proven, it is showed that the test is measuring construct.

3) Intervention studies

Another strategy for measuring construct validity is to apply an experimental manipulation and determine if the scores change in the hypothesized way. For example, the researcher may expect that the

¹⁹ American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for educational and psychological tests*. (Washington, DC: Author, 1999), 19.

²⁰ Campbell, D. T., & Fiske, D. W., "Convergent and discriminant validation by the multitrait - multimethod matrix". *Psychological Bulletin*, 56, 1959, 82.

scores on a scale designed to measure anxiety will be increase if the subjects of the research are put into an anxiety-provoking situation. The scores of a control group which is not involves in the experimental manipulation should not be affected. If anxiety was manipulated in a controlled experiment and the resulting scores change in the predicted way, this is show that the scale is measuring anxiety and the research has good construct validity.

4) Internal structure studies

This procedure involves showing that all the items making up the test or scale are measuring the same thing. A procedure called factor analysis provides a way to study the constructs that underlie performance of a test. The extent to which the observed item intercorrelations agree with the theoretical framework provides evidence concerning the construct being measured. More detailed explanation about factor analysis will be given on the next sub chapter because this strategy is used in doing this research.

5) Studies of response processes

Another way to obtain evidence about how well a test is measuring the construct of interest is to look at the evidence based on response processes of individuals actually taking the test. Questioning test takers about the mental processes and skills that they use when responding to the items of a test can provide information about what construct is being measured.

e. Internal structure studies

In this research, the construct validity is being examined through internal structure studies. As the researcher has said above, this strategy invloves a procedure called factor analysis. Johann and Steyn state that factor analysis is often used in constuct validation researchs.²¹ They also stated that the aim of factor analysis is to establish whether the test measure the proposed factors. According to Musavi, the goal of factor analysis is to construct the underlying factors and decompose the score variance in terms of correlation of factor analysis is present in the next subchapter.

f. Factor Analysis

1) Definition of Factor Analysis

Donald Ary states that factor analysis calculates the correlations among all the items and then identifies factors by finding groups of items that are correlated highly with one another but have low correlations with other groups.²² Factor analysis is a set of mathematical complex procedure

²¹ Johann L. van der Walt and H.S Steyn, "The validation of language tests" Stellenbosch papers in linguistics, Vol. 38, 2008, 200.

² Donald Ary, Lucy Cheser Jacobs and Asghar Razavieh, *Introduction to Research in Education* (USA: Cangage Learning, 2010), 234.

which use in analyzing the correlation between variables and clarifying the correlation in form of limited variable which is called factor.²³ Factor analysis, or exploratory factor analysis, is a family of techniques used to detect patterns in a set of interval-level variables. ²⁴ Factor analysis begins with a table of pairwise correlations (Pearson r's) among all the variables of interest; this table is called a correlation matrix.²⁵ The purpose of the analysis is to try to reduce the set of measured variables to a smaller set of underlying factors that account for the pattern of relationships. The search follows the law of parsimony, which means that the data should be accounted for with the smallest number of factors. This reduction of the number of variables serves to make the data more manageable and interpretable.

However, basically it involves searching for the clusters of variables that are all correlated with each other. The factor is represented as a score, which is generated for each subject in the sample. Next, a correlation coefficient is computed between subjects' factor score and their score on the particular variable entered into the factor analysis. This correlation between a variable and a factor is called the factor loading. The higher its loading, the more a variable contributes to and defines a

²³ Saifuddin Azwar, *Reliabilitas dan Validitas Edisi 4* (Yogyakarta : Pustaka Belajar), 121

²⁴ Spicer, J., Making sense of multivariate data analysis (Thousand Oaks, CA: Sage, 2005), 195

²⁵ Donald Ary, Lucy Cheser Jacobs and Asghar Razavieh, *Introduction to Research in Education* (USA: Cangage Learning, 2010), 364.

particular factor. A factor loading is interpreted like a correlation coefficient: The larger it is (either positive or negative), the stronger the relationship of the variable to the factor. The result of the factor analysis is a factor matrix, which shows the number of important underlying factors and the weight (loading) of each original variable on the resulting factors. The square of the factor loading is the proportion of common variance between the test and the factor.

To get the right number of factor the first criterion is that all the factors should be interpretable; an un-interpretable factor serves no practical or theoretical function.²⁶Second, the factors should account for a satisfactory amount of shared variance in the data. What is "satisfactory" is defined by the researcher. Some writers suggest that the analysis keep extracting factors as long as a factor accounts for at least another 10 percent of the variance. "There is general agreement that over factoring is preferable to under factoring". ²⁷

2) Types of factor analysis

a) Definition of Confirmatory Factor Analysis

Bachman said that in the confirmatory mode, we begin with hypotheses about traits and how they are related to each other and

 ²⁶ Spicer, J., *Making sense of multivariate data analysis* (Thousand Oaks, CA: Sage, 2005),195
²⁷ Ibid, 95.

attempt to either confirm or reject these hypotheses by examining the observed correlations".²⁸ Thus, in this type of factor analysis, the researcher wants to confirm whether the test items are assessed the proposed factors or not.

b) Definition of Exploratory Factor Analysis

Bachman says, "In the exploratory mode, we attempt to identify the abilities, or traits that influence performance on tests by examining the correlations among a set of measures".²⁹ As the main issue of this study, the researcher gives more detailed explanation on the next sub chapter.

3) Exploratory Factor Analysis

Donald Ary et al define factor analysis as a family of techniques used to detect patterns in a set of interval-level variables. ³⁰ Exploratory factor analysis was used to investigating how many factors needed in describing the correlation between a set of indicators by observing the factor loadings. The correlation between a variable and a factor is called the factor loading. The higher it is loading, the more a variable contributes to and defines a particular factor. A factor loading is

²⁸ Ibid, 260.

²⁹ Lyle F Bachman, Fundamental Considerations in Language Testing (Oxford: OUP, 1990), 259.

³⁰ Donald Ary, Lucy Cheser Jacobs and Asghar Razavieh, *Introduction to Research in Education* (USA: Cangage Learning, 2010), 361.

interpreted like a correlation coefficient: The larger it is (either positive or negative), the stronger the relationship of the variable to the factor.

According to Nuroris, there are four important steps in conducting exploratory factor analysis, they are:

a) Initial solution

In this first step, the data analysis adequacy is being examined as the requirement for doing factor analysis. The criteria in which the data is appropriate for factor analysis is based on Kaisr-Meyer-Olkin (KMO) and Bartlett's Sphericity test. Based on Subhash Sharma, the table of KMO is presented as follow:³¹

	The KMO score	Recommendation
	≥ 0.9 0	Very good
	≥ 0.80	Good
	≥ 0.70	Fair
	≥ 0.60	Less
	\geq 0.50	Enough
	Under 0.50	Rejected

b) Extracting the factors

The extraction process is aimed to get fewer factors (eigenvalues factor) from the number of variable and the contribution of each factor to all variance (total variance explained). Kaiser recommended retaining all factors with eigenvalues greater than 1

³¹ Subhash Sharma, Applied Multivariate Technique, (New York: John Willey & sons Inc., 1996), 107.

which means that the variable can be said as factor if its eigenvalues greater than $1.^{32}$

c) Rotating the factors

An Gie Yong and Sean Pearce state factors are rotated for better interpretation since unrotated factors are ambiguous.³³ Moreover, Rummel said that the goal of rotation is to attain an optimal simple structure which attempts to have each variable load on as few factors as possible, but maximizes the number of high loadings on each variable. ³⁴ Catell states that the simple structure attempts to have each factor define a distinct cluster of interrelated variables so that interpretation is easier.³⁵ For example, variables that relate to vocabulary should load highly on vocabulary items but should have close to zero loadings on other items.

d) Naming the factors

The last step of factor analysis is naming the factors which formed in the extraction and rotation process. The name of factor is given based on the similarities of the items in a factor.

³² Kaiser, H. F, "The application of electronic computers to factor analysis". *Educational and Psychological Measurement*, 20, 145.

³³ An Gie Yong and Sean Pearce, "A Beginner's Guide to Factor Analysis: Focusing on Exploratory Factor Analysis". Tutorials in Quantitative Methods for Psychology 2013, Vol. 9(2), 80.

³⁴ Rummel, R.J. Applied factor analysis. (Evanston, IL:Northwestern University Press, 1970), 135.

³⁵ Cattell, R.B. *Factor analysis*. (Westport, CT: Greenwood Press, 1973), 220.

EFA is particularly useful for investigating patterns of convergence, or commonality, among many different measures.³⁶ In the factor analysis model, the variance in each observed variable is explained in terms of the factor leadings on a number of different factors, as in the following equation:

Equation : $z_i = b_1F_1 + b_2F_2 + \dots + b_nF_n + Ui$, where

 z_i is the normal deviate for a variable;

b is the factor loading on a given factor;

F is a factor; and

U is the amount of variation that is unique to a particular variable.

Therefore, the researcher will use this method as the only way to measuring the construct validity of TOEFL-like test in this study.

B. Test

1. Definition of Test

Brown states that a test is a method of measuring a person's ability knowledge, or performance in a given domain.³⁷ At the first, a test is a method which means that a test is an instrument which requires performance on the part of the test-taker. Second, a test should be able to measure a person's

 ³⁶ Lyle F Bachman, *Statistical Analyses for Language Assessment* (Cambridge: Cambridge University Press, 2004), 279.
³⁷ H Douglas Brown Language Assessment: Principles and Classroom Practice (Longman):

³⁷ H. Douglas Brown, *Language Assessment: Principles and Classroom Practice* (Longman: California, 2003), 3.

ability. It is mean that the tester-makers need to understand who the test-takers are. A test measures performance, but the results imply the test-taker's ability. Finally, a test measures a given domain. In the case of proficiency test, even though the actual performance on the test involves only a sampling of skills, that domain is overall proficiency in a language-general competence in all skills of a language.

2. Test Type

a. Proficiency Test

Proficiency test are designed to test people's ability in a language, regardless any training they may have had in that language.³⁸Moreover, brown states that proficiency test is not limited to any one course, curriculum, or single skill in the language; rather it tests overall ability.³⁹ One good example of proficiency test is TOEFL test. In TOEFL test, almost all of English skill is tested. This is the main topic of this study that will be more elaborate on the next sub chapter.

b. Placement Test

The purpose placement tests are to place a student into a particular level or section of a language curriculum or school. A

³⁸ Arthur Hughes, *Testing for Language Teachers Second Edition* (Cambridge: Cambridge University Press, 2003), 11.

³⁹ H. Douglas Brown, Language Assessment: Principles and Classroom Practice (Longman: California, 2003), 44.

placement test 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.⁴⁰A pre-test of Intensive English Program that should be taken by all new students of UIN Sunan Ampel Surabaya is can be categorized as placement test because the result of the test use as the consideration on placing the students on the class.

c. 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 ad should therefore become part of a curriculum.⁴¹ In simple way, this kind of test is used to identify the strengths and weaknesses of learners.⁴²

d. 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

⁴⁰ Ibid., 45.

⁴¹ Ibid, 46.

⁴² Arthur Hughes, *Testing for Language Teachers Second* Edition (Cambridge: Cambridge University Press, 2003), 15.

objective.⁴³ 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.

3. Proficiency Test

Proficiency test are designed to test people's ability in a language, regardless any training they may have had in that language⁴⁴. A proficiency test is not limited to any one course, curriculum, or single skill in the language; rather it tests overall ability.⁴⁵ The content of proficiency test, therefore, is not based on the content or objectives of language courses that test-takers may have followed. Rather, it is based on a specification of what candidates have to be able to do in the language. Proficiency test consist of standardized multiple choice on grammar, vocabulary, reading comprehension and aural comprehension.⁴⁶ Proficiency test are always summative and norm-referenced.⁴⁷ Proficiency test is categorized as summative assessment because the aims of this test is to measure, or summarize, what a student has grasped, and typically occurs at the end of a course or unit of instruction.⁴⁸ This test also categorized as norm-referenced testing, the purpose in such tests is to

⁴³ Arthur Hughes, *Testing for Language Teachers Second Edition* (Cambridge: Cambridge University Press, 2003), 13.

⁴⁴ Arthur Hughes, *Testing for Language Teachers Second Edition* (Cambridge: Cambridge University Press, 2003), 11.

⁴⁵ H. Douglas Brown, Language Assessment: Principles and Classroom Practice (Longman: California, 2003), 44.

 ⁴⁶ H. Douglas Brown, *Language Assessment: Principles and Classroom Practice* (Longman:California, 2003), 44.

⁴⁷ Ibid, 44.

⁴⁸ Ibid, 45.

place test-takers along a mathematical continuum in rank order.⁴⁹ Mathematical continuum in rank order means that the score of this test is in form of numerical score (for example, 400 out of 450) and a percentile rank (such 90 percent, which means that the test-taker's score was higher than 90 percent of the total number of test-takers, but lower than 10 percent in that administration). The TOEFL-like test is categorized as a proficiency test because this test measured the overall skill of English, has its own standard of measurements and can be considered as summative assessment and norm-referenced testing.

C. TOEFL-like test of UIN Sunan Ampel Surabaya

TOEFL-like test is an English proficiency test which produces by P2B of SIU Sunan Ampel Surabaya which use TOEFL as the standard in giving the scores and making the questions. P2B do not make the test items by themselves, they take the questions from valous references such as TOEFL test. This test is divided into three sections: listening, grammar and reading. Every new students should take this test as their requirement to pass from intensive English program. The minimum score of this test is 400. If they fail to pass on the first test, they can take the second test 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.

⁴⁹ Ibid, 46.

1. Section of TOEFL-like 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 office is already closed
- B. The report is due tomorrow
- C. He can't finish all these tasks today
- D. He will be able to mail the report today
- b) Long Dialogues

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. But one long dialogue may also have seven to eight questions. Each long dialogue usually consists of 140 to 290 words and 40 to 80 seconds time for listening it. The test-takers are not allowed to take 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 Lectures

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. The 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 the 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 chosing 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. planned
- D. had planned

b) Finding the grammatical errrors

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

Thousands of settlers gone west after the Civil War ended in



c. 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.

2. The categorizaton of each section of TOEFL-like Test

As the researcher has said above, TOEFL-like test consists of listening section, structure and written expression section and reading section. The categorization at the analyzed TOEFL-like test is checked again Cliff's TOEFL preparation by Michaele A Pyle.⁵⁰

⁵⁰ Michael A Pyle, *Cliff's Test Preparation* (New York: IDG Books Worldwide, 2001), 39.

- a. Listening section
 - 1) Detail information

In this sub question types, the test-takers' ability in finding the detail information will be tested.

2) Main ideas

In this sub question types, the test-takers' ability in finding the main idea/main topic will be tested.

3) Implication

In this sub question types, the test-takers' ability in finding the detail information will be tested.

b. Structure and Written Expression Section⁵¹

1) Sentence structure

The sentence structure questions test more than a word or two; they test your ability to make a sentence complete. A sentence must have a subject, verb, and perhaps a complement. Sentence structure questions also test your understanding of subordinate clauses, which must not be independent clauses.

2) Word order

Word order questions are generally more detail-oriented than sentence structure questions. They test, for example, your

⁵¹ Michael A Pyle, *Cliff's Test Preparation* (New York: IDG Books Worldwide, 2001), 63.

understanding that an adjective should appear before the noun it modifies, not after it.

3) Word choice

The word choice type of question tests your understanding of idiomatic expressions, of which prepositions to use with certain words, of problem words that are sometimes confused, and so on.

- c. Reading section⁵²
 - 1) Main idea

In this type of question, the test-takers will be asked to identify the main idea of a passage or to indicate what an appropriate title for the passage.

2) Detail information

In this type of question, the test-takers will be asked to identify the detail information of a passage.

3) Vocabulary

In this type of question, the test-takers will be tests the understanding of particular words within passage.

4) References

In this type of question, the test will test the test-takers' ability in identifying antecedents of pronouns used in the passage.

⁵² Ibid, 153.

5) Inferences

In this type of question, the test will test the test-takers' ability in making logical conclusion based on the passage.

D. Intensive English Program (IEP) of UIN Sunan Ampel Surabaya

1. Profile

Intensive English Program (IEP) is one of program in Foreign Language Competence Development Program (P2KBA) held by Language Developmant Center at UIN Sunan Ampel Surabaya. It is an intesive program of English learning for first year students of UIN Sunan Ampel Surabaya in either indoor or oudoor class which concern on psychomotor, cognitive and affective aspects.⁵³ This program is non-credits, but it will be a form of certificate for students who pass in this program. The certificate of IEP becomes a requirement for students who will take a thesis. This program is directed for all first year or freshmen students who are enrolled in Bachelor Degree at UIN Sunan Ampel Surabaya.

2. Aims

The IEP has aims to develop the English competency of students both receptive and productive skills for enhancig students' academic qualification. Specifically, its aim can be stated in detail as follows:

 ⁵³ Pusat Pengembangan Bahasa (P2B), *Pedoman Program Kompetensi Berbahasa Asing* (Surabaya: Pusat Pengembangan Bahasa (P2B), 2012), 4.

- a. Students can interpret the English spoken language in certain academic activities.
- b. Students are able to communcate by English in academic forum, such as speech and master of ceremony.
- c. Students are capable in both academic and non-academic writing.
- 3. Targets

In directing Englsih cours for first year students, P2B has targets toward IEP:

- a. Have students who are competence in English.
- b. Apply all in one learning system (listening, speaking, reading, and writing in one course.
- c. Attain TOEFL-like test score 400 at minimum.
- d. Produce English writing (fiction ir essay).
- 4. Agendas

IEP has some agendas toward the course of English learning:

- a. IEP is held every Tuesday and Thursday morning at 6 7.45 a.m.
- b. IEP uses course outline.
- c. IEP give pre and post-test.
- d. IEP holds TOEFL-like test.

e. IEP has workshop of language learning and learning evaluation for instructors.

5. System

IEP is held for two semester in first year study. In the first semester, IEP is focused on General English while in the second semester, students are taught specifically in TOEFL. Before the learning is conducted, IEP gives pretest as placement test fr students to recognize the English competence level of each students. It becomes the consideration of the institution to give appropiate material and methods based on each level. After knowing the level of studenst from the result of pre-test, the students are placed in class based on their level. This placement is conducted in each faculties at UIN Sunan Ampel Surabaya, thus one class will be many students from different majors.

6. Method

Method use by IEP for English learning is communicative approach. It is used based on the consideration of implement the English curriculum in UIN Sunan Ampel Surabaya.

7. Evaluation

For the first semester, students are assessed by test of General English. Second semester, students are directed to do TOEFL-like test by the requirement score is 400 at minimum. If students do not reach the minimum score of TOEFL-like tes standardized by P2B, students will stated they do not pass the IEP and should re-study at IEP in next year or re-test TOEFL-like test.

E. Language Development Center (P2B) of UIN Sunan Ampel Surabaya

Language Development Center (P2B) of UIN Sunan Ampel Surabaya is the organizer of English intensive program. P2B is coordinated with each faculty and had responsibility to the Rector of UIN Sunan Ampel Surabaya cq. Vice Rector I of academic field.

1. The duty of P2B of UIN Sunan Ampel Surabaya

P2B has some duties and responsibilities, they are:

- a. Providing textbook, media and sources of learning
- b. Preparing the detail curriculum, SAP and lesson plans
- c. Preparing the hardware which is needed for the learning and teaching process of English intensive program such as laboratory, classroom and ect.
- d. Recruting the lecturers of English intensive program
- e. Preparing the schedule for English intensive program
- f. Evaluating the process of English intensive program activities
- g. Giving periodic report about English intensive program to the Rector of UIN Sunan Ampel Surabaya cq. Vice Rector I sixmonthly.
- 2. The authority of P2B of UIN Sunan Ampel Surabaya

P2B has some authorities in conducting English intensive course program, they are presented as follows:

- a. Recruting the lecturers of English intensive program based on the inquiry of each faculties.
- b. Reviewing and revising the materials and another sources for learning and teaching process.
- c. Evaluating the IEP lecturers' performance.
- d. Making the final examination test
- e. Managing the financial of IEP

F. Review of Previous Study

In this part, the researchers will explore the previous study conducted by other researchers that have similar focus with this study.

There are some research that has been conducted in investigating the construct validity of a test. The first is, the journal entitled "Construct Validity of The Pearson Test Of English Academic: A Multitrait-Multimethod Approach by Hye K. Pae, Ph.D. University of Cincinnati, U.S.A." In this research, the researcher examine the construct validity value of an English academic test held in the pearson university. He use a multitrait-multimethod approach to measuring the construct validity. The MTMM approach is largely used to examine how different traits (multitraits) of language abilities and methods (multimethods) of testing materials influence the student's performance.⁵⁴ Then, the result of MTMM approach matrix is analyses using confirmatory factor analysis. A confirmatory factor analysis (CFA) can be used

⁵⁴ Hye K. Pae, Ph.D. "Construct Validity of The Pearson Test Of English Academic: A Multitrait Multimethod Approach" September 2012,1.

to overcome the limitation of a simple comparison of MTMM correlations, as CFA models are a powerful and direct means to test the relative contributions of traits and methods to test-takers' performance and to explain underlying relationships.⁵⁵The approach used in measuring the construct validity of the test is very different from this research. In this research, the researcher uses an exploratory factor analysis in term of measuring the construct validity.

The second is a journal entitled "The Construct Validity of a Test: A Triangulation of Approaches" by Mohammad Salehi at Sharif University of Technology. In his research he uses three different kind of approach in determining the construct validity of the test. The first approach is factor analysis, the researcher use a factor analysis in investigating whether the test items in the 'Reading Comprehension' sections of the UTEPT distinctly measure various sub-skills or not. The result shows that this test is good in measuring various sub-skills of reading. The second is multitrait-multimethod approach which use in finding the correlation between traits being tested and the methods of testing. The result of this approach shows that there are correlation between traits being tested and the methods of testing. The last approach is protocol analysis. The use of this protocol analysis is to distinguish the degree of correlation among sub-parts of the test. The approach use in his journal is different from the approach will be used in this research. As the researcher has said above, this research will be investigated using factor analysis.

⁵⁵ Lyle F Bachman, *Statistical Analyses for Language Assesment* (Cambridge: Cambridge University Press, 2004), 279.

The third journal is "The Construct Validity of Student Engagement: A Confirmatory Factor Analysis Approach" by Steven M. LaNasa, Alberto F. Cabrera and Heather Trangsrud. This research is focus on examine the construct validity of the National Survey of Student Engagement (NSSE) which increasingly use in some universities in observing the student engagement on campus.⁵⁶ As institutions seek to promote student engagement on campus, the National Survey of Student Engagement (NSSE) is increasingly being used to chart progress and compare results using the Five Benchmark Scores. While recent research has begun to decompose the five benchmarks in a variety of ways; few research studies have sought to explore the underlying structure of these five benchmarks, their interdependence, and the extent to which the items do reflect those five dimensions. This study begins to address the instrument's construct validity by submitting a single, first-time freshman cohort's NSSE responses to a confirmatory factor analysis, and proposes as an alternative, eight "dimensions" of student engagement that fit this set of data slightly better and in a more useful way. Results have practical implications for institutions utilizing NSSE, but also contain conceptual implications pertaining to the application of these benchmarks. Confirmatory factor analysis was used in conducting the research. However, the result of this study is showed that the institution or universities need to re-check the National Survey of Student Engagement (NSSE) in proving the student

⁵⁶ Steven M. LaNasa, Alberto F. Cabrera and Heather Trangsrud, "The Construct Validity of Student Engagement: A Confirmatory Factor Analysis Approach". 2.

engagement because it less proven. The approach in this journal is different from the research will be examine, this research use an exploratory factor analysis.

The fourth is a journal with title "Construct Validation of TOEFL-iBT (as a Conventional Test) and IELTS (as a Task-based Test) among Iranian EFL Testtakers' Performance on Speaking Modules" created by Elham Zahedkazemi. The aim of his research is to examine the construct validity of the speaking section of two well-known international English proficiency tests; TOEFL-iBT and IELTS. The researcher seeks how IELTS and TOEFL iBT tap the same construct validity on the speaking proficiency of their candidates and moreover about the similarity of results for both IELTS and the TOEFL speaking tests.⁵⁷ The result of his study showed that both TOEFL-iBT and IELTS have similar result in measuring the speaking ability of the test-takers which means that these two tests have strong correlation. Thus, his research will be focus on the construct validity of each item on a TOEFL-like test while his research focus on investigating the construct validity by seeking the correlation between TOEFL-iBT and IELTS.

The fifth is a research by Ebrahim Khodadady, with the title "Construct Validity of C-tests: A Factorial Approach". The C-Tests designed by Klein-Braley (1997) were changed into two other types of tests called Spelling Test and Decontextualised C-Test and administered along with the disclosed Test of English as

⁵⁷ Elham Zahedkazemi, "Construct Validation of TOEFL-iBT (as a Conventional Test) and IELTS (as a Task-based Test) among Iranian EFL Test-takers' Performance on Speaking Modules" *Theory and Practice in Language Studies*, Vol. 5, No. 7, July 2015, pp. 1513-1519.

a Foreign Language (TOEFL), a semantic schema-based cloze multiple choice item test (S-Test) and a lexical knowledge test (LKT). The research focus on finding what factor structure is for the C-Tests, the Decontextualised C-Test, the Spelling Test and what factor structure is if the TOEFL, the lexical knowledge test and S-Test are included. The application of principal component analysis (PCA) to the responses of the participants on the three tests, i.e., C-Tests, Spelling Test and Decontextualised C-Test, revealed two components called *language proficiency* and *direction specificity* in this study. While the inclusion of the S-Test, TOEFL and the LKT in the PCA yielded the same two components, their rotation brought about the highest loadings of the included tests as well as the moderate loadings of the C-Tests on the first component, validating them as proficiency measures of language. However, they loaded the highest on the second component along with the Spelling Test and Decontextualised C-Test and thus confirmed their spelling and direction specificity. The difference between the study of C-test and this research is the focus of research. This research focuses on examining the construct validity of TOEFL-like test using factor analysis procedure.

The sixth is a study by Neil J. Anderson, Lyle Bachman, Kyle Perkins and Andrew Cohen entitled "An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources". Reading, while recent thinking in language testing has recognized the importance of gathering information on test taking processes as part of construct validation. And while there is a growing body of research on test-taking strategies in language testing, as well as research into the

relationship between item content and item performance, no research to date has attempted to examine the relationships among all three - test taking strategies, item content and item performance. This study thus serves as a methodological exploration in the use of information from both think-aloud protocols and more commonly used types of information on test content and test performance in the investigation of construct validity. The study is examining the construct validity of the reading comprehension test by investigating the relationship among test taking strategies, item content and item performance.58 In investigating the construct validity of TOEFL-like test, the researcher will only use one data resource, test item. Types of data analysis were performed, providing both qualitative and quantitative data. Each data analysis task reflects part of the triangulation of data sources to examine the construct validity of the reading comprehension test. The first task involved reviewing each of the think-aloud protocols and coding them for the use of reading and test-taking strategies. The second data analysis task involved the content analysis of each test item on both forms of the test from two perspectives: that of the test designer and of Pearson and Johnson's (1978) taxonomy of relationships between texts and test items. Third, the data from the think-aloud protocols were then submitted to chi-squared analyses. The result of chi-square analyses indicate that these subjects are using these strategies differently, depending on the type of question

⁵⁸ Neil J. Anderson, Lyle Bachman, Kyle Perkins and Andrew Cohen, "An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources". *Language Testing*, Vol.8 No.2, 1991, 41.

that is being asked. The results of the chi-square analysis also show that the subjects tended to use some strategies fairly consistently for the different types of items.

The seventh is a journal entitled "Testing the Convergent and Discriminant Validity of the Decisional Balance Scale of the Transtheoretical Model Using the Multi-Trait Multi-Method Approach" by Boliang Guo, Paul Aveyard, and Antony Fielding of University of Birmingham and Stephen Sutton of University of Cambridge. The authors extended research on the construct validity of the Decisional Balance Scale for smoking in adolescence by testing its convergent and discriminant validity. Hierarchical confirmatory factor analysis multi-trait multi-method approach (HCFA MTMM) was used with data from 2,334 UK adolescents, both smokers and non-smokers. They completed computerized and paper versions of the questionnaire on 3 occasions over 2 years. The results indicated a 3-factor solution; Social Pros, Coping Pros, and Cons fit the data best. The HCFA MTMM model fit the data well, with correlated methods and correlated trait factors. Subsequent testing confirmed discriminant validity between the factors and convergent validity of both methods of administering the questionnaire. There was, however, clear evidence of a method effect, which may have arisen due to different response formats or may be a function of the method of presentation. Taken with other data, there is strong evidence for construct validity of Decisional Balance for smoking in adolescence, but evidence of predictive validity is required. The difference between the research above and this research is that this research use factor analysis procedure in case of measuring the

construct validity of this TEOFL-like test while the research use MTMM in case of measuring the discriminant validity which is the part of construct validity.

The eight is a journal entitled "Does the text matter in a multiple choice test of comprehension? The case for the construct validity of TOEFL's minitalks" by Roy Freedle and Irene Kostin. The current study addresses a specific construct validity issue regarding multiple choice language-comprehension tests by focusing on TOEFL's minitalk passages: Is there evidence that examinees attend to the text passages in answering the test items? To address this problem, we analysed a large sample (n = 337) of minitalk items. The content and structure of the items and their associated text passages were represented by a set of predictor variables that included a wide variety of text and item characteristics identified from the experimental language-comprehension literature. Stepwise and hierarchical regression techniques showed that at least 33% of the item difficulty variance could be accounted for primarily by variables that reflected the content and structure of the whole passage and/or selected portions of the passage; item characteristics, however, accounted for very little of the variance. The pattern of these results was interpreted, with qualifications, as favoring the construct validity of TOEFL's minitalks. Our methodology also allowed a detailed comparison between TOEFL reading and listening (minitalk) items. Several criticisms concerning multiple-choice languagecomprehension tests were addressed. The difference between this research and the research which the researcher worked for is the method in examining the construct validity. As the researcher has said above, this research uses regression in case of measuring the construct validity of TOEFL's minitalks. The research which the researcher worked for is using factor analysis, exploratory analysis for specifically for measuring the construct validity of TOEFL-like test.

