Importance of measurement in research

Measure is important in research. Measure aims to ascertain the dimension, quantity, or capacity of the behaviors or events that researchers want to explore. According to Maxim (1999), measurement is a process of mapping empirical phenomena with using system of numbers.

Basically, the events or phenomena that researchers interested can be existed as domain. Measurement links the events in domain to events in another space which called range (Figure 1). In another words, researchers can measure certain events in certain range. The range is consisting of scale. Thus, researchers can interpret the data with quantitative conclusion which leads to more accurate and standardized outcomes. Without measure, researchers can’t interpret the data accurately and systematically.

Quantitative Measurements

Quantitative Measurement is a quantitative description of the events or characteristics which involves numerical measurement. For example, the description made as “ There are three birds in the nest”. This description includes the numerical measurement on the birds. Quantitative measurement enables researchers to make comparison between the events or characteristics. For example, researchers tend to know who the tallest person in a family is. So, they use centimeter to measure their height and make comparison between all the family members.

Levels of Measurement

Level of measurement refers to the amount of information that the variable provides about the phenomenon being measured (McClendon, 2004). For all

variables, they should include exhaustive attributes and mutually exclusive attributes. These two attributes are related to the accuracy and precious measurement in a study.

Exhaustive and mutually exclusive attributes. For the exhaustive attributes, it assign a full range of attributes which possessed by all of the subjects (people) in the study. It enables all of the subjects in study can answer their preferred answer in each question. For instance, a question ask the marital status to all subjects in a study with four options, which are (a)married; (b)divorced; (c)widowed; and (d)never been married. However, subject who is in the legally separated status unable to choose the options that provided by researchers. Thus, this question is not applying the exhaustive attributes. The categories are not exhaustive may lead to missing data in the study and lastly affect the outcomes of the study.

For mutually exclusive attributes, it stated that researchers should assign only one attribute for each person in a study. For example, a question asked how much of the monthly income of each subjects in the study with four options, which are (a) RM0-RM500; (b) RM500-RM1000; (c) RM1000-RM1500; and (d) RM1500-RM2000. The four options in above is not mutually exclusive because each category contains the same income with another category. Subjects who have monthly income with RM1000 may feel confuse to choose option (a) or (b).

Nominal measurement. There are four levels of measurement, which are nominal, ordinal, interval, and ratio measurement. Nominal measurement is a process of assigning numerals to categories. In another words, nominal is

in name or form only (McClendon, 2004). Researchers can’t describe and differ the cases or events with using the adjective such as higher than, lower than, more than, less than, and others. So, it is the lowest level in measurement. Percentages, central tendency, chi square are appropriate used in this level of measurement.

One of an experiment in Social Psychology which conducted by Stanley Schachter in year of 1957 aimed to measure the affiliation when people feel worry toward a situation. Experimenter told all subjects that they will assigned into two situations which are intense electrical shock and mild electrical. Then, experimenter asked whether they prefer to wait together with others or enter the room alone, and which situation that they prefer to engage in. The result showed that, subjects in intense electrical shock group prefer to wait together with others and there are no preferences in the subjects who in mild electrical shock group. Experimenter measure the reactions of subjects with using the nominal measurement, which were (1) prefer mild shock; (2) prefer intense shock; (3) prefer wait together with other and, (4) no preferences to wait together with others. The conclusion made by the experimenter was depending on the nominal measurement.

Ordinal measurement. Ordinal measurement permits researchers to make comparisons like “ greater than”; “ less than”, “ higher than”, and “ lower than” but not “ how much” (McClendon, 2004). For example, researchers can make a comparison according to one of the example in table 1 as people who strongly agree with legalize abortion is higher than people who disagree with it.

Besides that, attributes also be rank-ordered. The academic rank if students such as freshman, junior, and senior is an example in this measurement. We can’t use arithmetic operations in this level of measurement because the distances or intervals between the attributes are unknown.

Some process involved in this level of measurement such as precedence or preference (Maxim, 1999). aPb, bPc, and aPc indicates that a precedes b, b precedes c, and a precedes c. These processes involve the measurement such as “ greater than”, “ higher than”, and other. For example, four students’ result ranking make by using their marks, they are 97, 81, 79, and 70. If 97 as “ a”, 81 as “ b”, 79 as “ c”, and 70 as “ d”, we can make a conclusion that aPb, aPc, aPd, bPc, bPd, and cPd.

Interval measurement. Characteristics of this level of measurement are the attributes are ordered and the distances between attributes are equal. However, it doesn’t have true zero point. The Fahrenheit and Celsius temperature scales always used as examples in this level of measurement. Fahrenheit and Celsius temperature scales don’t have true zero point because the zero temperature does not mean “ no temperature”.

In this level of measurement, researchers can make a description that 40-50 degree is same as 80-90 degree because it has equal distances between the categories. However, they can’t make a description like 80 degree is twice as 40 degree because it doesn’t have true zero point. For example, researchers can’t say that temperature in city A is twice hot than city B. As the Table 1 showed, 80 degree in Fahrenheit actually is same as 27 degree in

Centigrade, and 40 degree in Fahrenheit Degree is same as 4 degree in Centigrade, and 27degree is not twice of 4 degree.

Conceptualization and Operationalization in Measurement

Conceptualization is a process of specifying a term or concept that researchers want to measure. In Deductive research, it helps researchers to specify the theory and come out with a specific variable that can place in a hypothesis. For the Inductive research, it helps researchers to have an idea about what related behaviors or events that need to be observed.

For example, researchers tend to measure the influences of social status towards academic performances among adolescents. At first, researchers should define what " social status" is. In conceptualization aspect, social status can be defined as power, prestige, and privilege. Another example is deviant behavior. Researchers tend to study the relationship of deviant behavior and academic performance. At the first, researchers have to understand the meaning of deviant behavior is. Thus, they defined it as the behavior of smoking, fighting, underage drinking, and threatening. At here, a concept will defined without using any quantitative methods.

Operationalization defined as a process of defining a concept by measure it (Maxim, 1999). Operationalization is specifying that how a concept in a research be measured. Scoring, coding, and scaling may used in operationalization measure of concept. For example, researchers tend to study the relationship between students happiness and school performances. School performances can include some performances such as students' examination results, on-time submission of assignment, and attendance. In

the study, researchers aim to focus on these three performances. So, they start to make operationalization towards it.

They plan to measure students' examination result by using the Cumulative of Grade Point Average (CGPA) and they assume that high happiness may have high CGPA in school. Then, they measure " on-time submission of assignment" by using frequency as how many times that they have submitted assignment on-time in a month and assume that high level of happiness may have high frequency of on-time submission assignment. Lastly, they measure students' attendance depends on the percentage they have attended to class in a month and assume that high level of happiness may have high percentage to attend class in a month.

Difference between Conceptualization and Operationalization in Measurement

Conceptualization is a process of defining the concept without operates any quantitative methods or others methods that can indicates the values of a variable. Conceptualization only can make researchers and population understand what does a term or concept means. For operationalization, it define a term or a concept and also operates some methods especially the methods involve quantitative to indicate the values of a variable.

Indexes and Scales

Indexes and scales are measuring instruments or devices. Both of them used to measure variables or concept that researchers interested. Scale is a cluster of items that arranged into a unitary dimension or single domain of

behavior, attitudes, and feelings. Scales are more specific than indexes do. Scales can predict outcomes such as behavior, attitudes, and feelings because it measures the underlying traits. For example, a scale tends to measure more specific variable such as Introversion. Thus, Introversion scale should consist of the items that related to Introversion only. The items in Scale used to measuring Introversion such as 1) I blush easily; 2) At parties, I tend to be a wallflower; 3) Staying home every night is all right with me; 4. I prefer small gatherings to large gatherings, and 5) When the phone rings, I usually let it ring at least a couple of times. Likert scale such as " Strongly Agree……Strongly Disagree" will used. For another example, Hare Self-Esteem Scale includes three specific sub-scales which are Peer Self-Esteem Scale, Home Self-Esteem Scale, and School Self-Esteem Scale. These three sub-scales are used to measure the concept of self-esteem.

For indexes, it is a set of items that consist multiple aspects of dimensions which are interrelated. These entire dimensions will be made into single indicator or score. Index is more general than scales. It is also designed for exploring the relevant causes or underlying symptoms of traits. Indexes tend to measure a concept depends on what happens in the real world.

An index tends to measure life satisfaction of college students. Due to the reason that life satisfaction may consist a lot of dimensions or categories, the index should includes items related to all categories. For example, life satisfaction should include satisfaction of career, satisfaction of family relationship, satisfaction of peer relationship, and satisfaction of marital relationship. Researchers total up the scores of all items and the scores will reflex the level of life-satisfaction.

RELIABILITY AND VALIDITY

Reliabilityis important because it enables researchers to have some confidence that the measure they taken are close to the true measure. Validity is important because it tell researchers that the measure they taken is actually measures what they hope it does. So, if researchers want to know how good the measurement is, they should depend on the reliability and validity of a measurement.

Reliability is synonym of repeatability and consistency. Reliability defined as the degree to which test scores are free from errors of measurement (AERA et al., 1999, p. 180 in Neukrug & Fawcett, 2006). The degree of reliability can decide whether the scores or data that researchers obtained can be relied to measure a variable or construct.

Measurement error. An unreliable measurement is caused by error source of variability. There are two types of error which are Systematic Measurement Error and Unsystematic Measurement Error. Systematic measurement error is the factors that affect measurement systematically across the time. It is predictable and can be eliminated if it gets identified. It is also related to validity of a measurement. Systematic measurement error arises when researchers unknown to the test developer and a test measure something others than the trait that researchers tend to measure. These may seriously influence the validity of a test.

Unsystematic measurement error is the effects or errors that unpredictable and inconsistent. It is related to reliability of a measurement. Item selection,

test administration, and test scoring are examples of unsystematic measurement error.

Item selection means that error happened in the instrument itself. The example of this error such as instrument which includes not valid questions or items, contents can’t fair to all respondents even though it is already considered as good, and there are too many items inside the test. Test administration error includes uncomfortable room, dim lighting, noise in room, fatigue, nervous, and others which may influence respondents’ performances. For the test scoring error, it happened when the format of test not using machine-score multiple-choice items. Subjective judgment in scoring occurred especially for the projective test and essay questions. Rorschah Inkblot Test, Sentence Completion Test, and Thematic Apperception Test are related to subjective judgment.

Types of reliability. There are two major types of reliability which are Reliability as Temporal Stability and Reliability as Internal Consistency. Reliability as Temporal Stability is related to the times to collect data. Reliability as Temporal Stability includes Test-retest and Alternate-forms Reliability. Internal Consistency includes Split-half, Coefficient Alpha, and Interscorer Reliability.

Test-retest reliability defined as the relationship between scores from one test given at two different administrations (Neukrug & Fawcett, 2006). Alternate-forms Reliability is the relationship between the scores from two version of same test. In this type of reliability, everything in the different version test such as the difficulty level, number of items, and content should

be same. Split-half reliability defined as correlating one-half of the test to the other half. Researchers can divide the test into two parts which are first half and second half. They also can divide the items by odd numbers and even numbers of the items. Spearman-Brown used when the numbers of items in test is short. Spearman-Brown is more accurate when the numbers of items is few.

Coefficient Alpha and Kuder Richardson determined by correlating the scores of each item with total scores on the test. Kuder Richardson used when the items need to be answered by “ yes” and “ no”. Interscorer Reliability defined as correlating the scores from two or more observers’ rating to the same phenomenon. Observers should be trained to rating on the events or behaviors of respondents.

Test-retest is appropriate be used when researchers aim to measure the behaviors of respondents across times. Coefficient Alpha is appropriate to be used in both unidimensionality tests. Split the test by odd and even numbers is appropriate to be used when the difficulties of items have carefully ordered. If the difficulties level of items is not carefully orders, the method of split the test to first half and second half is appropriate. Interscorer reliability used when the test involves subjectivity of scoring.

Validityrefers to an accuracy of a measure. A measurement is valid when it measures what the researchers suppose to measure (Gregory, 2007). For example, IQ tests are supposed to measure intelligence and depression tests are supposed to measure depression level or symptoms of respondents.

Normally, the inferences drawn from a valid test are appropriate, meaningful, and useful.

Types of validity. There are three types of validity which are Content Validity, Criterion Validity, and Construct Validity. For the Criterion Validity, it includes Predictive Validity and Concurrent Validity. For the Construct Validity, it includes Convergent and Discriminant Validity.

Content validity determined by the degree to which the questions, tasks, or items on a test are representative of the universe of behavior the test was designed to sample (Gregory, 2007). The appropriateness of content of a measurement is determined by experts. Researchers make a judgment on whether the items in a measurement have covered all domains that they want to measure. For example, teacher would like to develop a test which tends to measure the understanding of students toward a subject from chapter 1 to 5. The type and number of questions are designed. Sixty multiple-choices questions and 60 minutes are given to the students to do the test. Ten questions will cover each chapter and the rest questions will cover chapter five which considered as the most important chapter in the test.

Validity of content also can be made by the experts' rating towards each item to decide whether the items can indicate the content or not. Two experts evaluate each item on the four-point scale. The rating of each expert on each item can be dichotomized into weak relevance of content (rating of 1 and 2) and strong relevance of content (rating of 3 and 4). If both experts agree that the item is strongly relevance, then the item will be put in cell D;

if both experts agree that the item has weak relevance, the item will be put in cell A. Cell B and C involved the items that agreed by one expert and disagreed by another expert (Figure 2).

For the Criterion validity, both Predictive and Concurrent validity will made by comparing them with others criterion. Concurrent validity correlates test scores with criterion scores and these two types of scores are obtained in the same time. For example, researchers would like to measure the reading ability of students by using the Reading Achievement Test. Researchers compare the Reading Achievement Test scores of students with the teachers' rating scores on students' reading abilities. High correlation between the two scores indicates that there is high concurrent validity in the test.

For the Predictive validity, it correlates test scores with criterion scores which obtained in the future. It means that the scores or data are obtained in different time. For example, Employment Test used to measure the performances of employee in a company or organization. At first, researchers give the test to employee and after six months, the supervisors asked to give evaluation to the performances of employee. Then, researchers compare the test scores and supervisors' rating scores to see the level of validity. The difference between Concurrent and Predictive validity is the time frame used to obtain the data and scores.

For Construct validity, construct is a theoretical, intangible quality or trait in which individuals differ. It is abstract and hard to be measured. Thus, it needs some indicators or signs to represent it. A construct is a collection of

related behaviors that can represent the things that researcher want to measure. Construct validity is evidence that an idea or concept is being measured by a test (Neukrug & Fawcett, 2006).

For example, depression is a construct and it manifested by some behaviors such as lethargy, difficulty concentrate and loss of appetite. Homogeneity refers to a test measure a single construct. Homogeneous refers to the single component or subtest in a Homogeneity test. The purpose of homogeneity is selecting items which potential to form a homogeneous scale.

Convergent validitydefined as a test highly correlates with other variables which have same or overlap constructs. For example, researchers would like to take the Beck Depression Inventory-II (BDI-II) to compare with others tests which have same variables as well. The result shows that, BDI-II has high correlation with Scale for Suicide Ideation (r =. 37); Beck Hopelessness Scale (r = . 68); Hamilton Psychiatric Rating Scale for Depression (r =. 71); and Hamilton Rating Scale for Anxiety (r =. 47). Lastly, for the Discriminant validity, it means that a test does not correlate with the variables or test which are not measure the different variables or constructs.

- Relationship between Reliability and Validity

A good validity need to have good reliability established first. However, a good reliability does not lead to a good validity. A good reliability only reflex that the scores in a measurement is appeared consistently.

A good validity may leads to reliability. When the measurement or test tends to measure what researchers tend to measure, the validity occurred and

thus the reliability occurred also. In a test, reliability is necessary but not sufficient for validity. In other words, measure can be reliable but not valid; valid measures must be reliable, however.