International Journal of Quantitative and Qualitative Research Methods
Vol.6, No.1, pp.46-57, February 2018
Published by European Centre for Research Training and Development UK (www.eajournals.org)

STATISTICAL ANALYSIS: INTERNAL-CONSISTENCY RELIABILITY AND CONSTRUCT VALIDITY

Said Taan EL Hajjar
Ahlia University

ABSTRACT: *In testing a proposed model of any academic field, huge extents of data are collected and analyzed through a proposed or adopted scale by researchers. It is very significant to test the reliability and the construct validity of this scale. This study aims to explore the important strategies or tools that should be followed to achieve a reliable and valid scale. It also aims to study the two concepts in association with scale testing as well as to afford a snapshot of the recent understandings of the reliability and validity of scale tests as expressed in previous research studies. This paper emphasis mainly on the theoretical and practical parts of reliability, and the affiliation between the two perceptions. Seeking to synthesize research results about the construct validity and internal-consistency reliability of items (factors), we focus on the four main measurements: Cronbach's alpha, inter- item correlation, corrected item – total correlation, and Cronbach's alpha if item deleted. These measurements would provide researchers with a clear knowledge about purifying the survey's scale before using it in collecting the data set. This kind of research is required so that a complete understanding of the construct validity and internal – consistency reliability of items is realized.*

KEYWORDS: Scale, Cronbach's Alpha**,** Internal-Consistency**,** Reliability**,** Construct, Validity

INTRODUCTION

In most of the academic research studies, especially in the field of business, researchers employ a survey study to test a developed model or testing hypotheses to support the theory it has to be measured. For the statistical expert functioning with researchers the estimation of reliability and validity is a mission often encountered. The measurement of aspects differ particularly in the field of business as they are associated with the quantification of abstract and invisible factors. Then in several cases the meaning of quantities is only incidental.

Most perceptions in the behavioral businesses have significance within the setting of the theory that they belong to. Each perception, then, has a functioning definition which is ruled by the principal theory. If a perception is elaborated in the hypothesis testing to satisfy the theory it should be measured. So the primary decision that the study is tackled with is how the perception shall be measured (e.g. the type of measure). In fact, there are many types of measurement. It can be observational, interview, self-report, etc. These types eventually get customized with a precise form such as observing an event on a video-tape, observing a continuous activity, self-report measures (e.g. surveys that can be open-ended or close-ended), Likert-type scales etc. There is no doubt that each type of measure possesses definite types of concerns that require to be exposed in order to make the measurement significant, precise, and efficient.

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

One more important aspect is the population along which it is intended to be measured. This choice is not completely reliant on the theoretical pattern but further to the availability of the instant research question.

Another important issue is the aim of the scale or measure. That means," What is it that the researcher wants to do with the measure?" "Is it settled for a precise study or is it settled with the expectation of general procedure with similar populations?"

At the time these perceptions are satisfied and a measure is established, which is a sensitive process, then the following questions are to be raised "how can be sure that we are certainly measuring what we need to measure?" since the concept that we are measuring is intangible, and "can we be certain that if we retest again the measurement we will have similar results?" The first question is associated to validity, while the second one is related to reliability.

Young researchers frequently ask the following question: "Can we adopt a questionnaire developed by a certain scientist to be used in collecting our data?" In fact, there is more than one response. Firstly, if this adopted questionnaire satisfies the researcher's research needs, then it is possible to be applied in his/her research study. Secondly, the researcher can omit or add some factors related to this adopted survey before collecting the data. However, in both cases it is extremely recommended to test the reliability and validity of this survey. The reason is many features may affect the validity and the reliability of these adopted questionnaires such as culture, population, sample size, etc. In addition, we should keep in mind that the old scientist purified his questionnaire through a sample selected from a certain population. Some researchers start to develop their questionnaire from zero. That is they create the factors of their latent variables. In this case, a factor analysis test is needed in which reliability and validity are parts of it. Therefore, for a non- statistician researchers, it is preferable that they adopt a previous reliable - valid scale, make little modification on its constructs to fit their topic, and, then, test the reliability and validity of this new modified scale.

The measurements of performances and constructs are issued to error that may influence the measurements of reliability and validity. In fact, a measuring instrument that is reliable provides the same measurements when a researcher frequently measure the similar unchanged items or events. The detected value on an instrument can be measured by:

$$\text{Detected value} = \text{true value} + \text{error} \quad (1)$$

Knowing that an instrument is considered reliable if it precisely reflects the true value, and then minimizes the residual (error) factor. The reliability coefficient is computed by the following formula:

$$\text{Reliability Coefficient} = (\text{True variability}) / (\text{Total detected variability}) \quad (2)$$

That is, if the value of a reliability coefficient is equal to 0.75, this indicates that 75% of the variability in detected values is assumed to describe true individual differences and 25% of the variability is due to random residual.

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

In this paper, there will be brief literature review about the reliability and validity concepts, and a procedure used to express reliability and validity tests.

LITERATURE REVIEW

In order for a survey to be valuable and of use, it must be both reliable and valid. If a scale is valid, then it must be reliable. For instance, students have different marks on a given exam every time they take it, the exam is not probably to foretell anything. However, when an exam is reliable, that does not imply that it is necessary valid. For instance, we may measure strength of grip very reliably, but that does not make it a valid measure of intelligence. So reliability is a necessary, but not sufficient, condition for validity.

Validity analysis requires to emphasize on the aim and use of the results achieved by a certain test, and needs further than a quantitative analysis of the obtained results. Specifying the significance of reliability and validity measurements on the part of instructors, Weigle (2007) states "Teachers should not hesitate to ask questions about the reliability and validity of the tests that their students are required to take and how test results will be used, and teachers should be proactive in bringing issues of questionable testing practices to the attention of administrators." The progression of test validation requires gathering information about the reliability and validity of a test in terms of its satisfaction of the aim it is intended for, and scores' consistency on the basis of evidence resulted from the scores. In addition, O'Neill (2003) states "Validity research involves a dynamic process that requires an examination of procedures and results, use of this information to revise and improve assessment practices, and an examination of revised practices in a never-ending feedback loop."

Types of Reliabilities

Reliability is one of the most significant components of test quality. It is involved with the reproducibility, consistency, or an examinee's performance on the test. Reliability is the total consistency of a certain measure. A measure is considered to have a high reliability when it yields the same results under consistent conditions (Neil, 2009). There are several types of reliability such as Test-Retest Reliability, Alternate - Forms Reliability, Split –Half reliability, and Internal Consistency Reliability.

Test-Retest Reliability

Test-Retest reliability or simply Stability Test is computed through correlation. It is the degree to which values are consistent through any repeated test. The most direct way of estimating reliability is to manage or administer the test two times to the identical set of themes and then correlate the two measurements at each time (Karl, 2012). The correlation coefficient (r) between the two sets of values designates the degree of reliability. For perceptive field tests, correlations more than or equal to 0.90 are good. For subjective tests, correlations more than or equal to 0.50 are acceptable. For attitude measures, correlations more than or equal to 0.70 are significant (Robert, 2004). This test also indicates score variation that arises from a testing session to another one as a consequence of residuals of measurement. Robert (2004) states that the procedures for a test-retest reliability

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

encounters some problems. That is, the responses in the second test may essentially change as a result of the previous test. A second problem is that several constructs of interest essentially change over time independent of the stability of the measure. A third problem is that the interval range between the two administrations is considered too long and the factor in which the researcher is trying to measure may have altered. Finally, the interval range may be too short so that reliability is inflated due to memory.

Alternate - Forms Reliability

Alternate - Forms Reliability test is computed through correlation. "It is the degree of relatedness of different forms of the same test" (Ralph & Robert, 2007). This test requires using in a different way word questions in order to measure the same factor or construct. That is, two identical tests in every method except for the real factors included. Factors shall emphasis on the same exact item of conduct with the same terminology and difficulty levels (e.g. factors shall differ only in their wording). According to Robert (2004), Alternate- Forms reliability is put to evade the practical influences that may inflate Stability reliability (i.e. respondents may remember how they replied on the identical factor during the administration of the first test). This test correlates the two values. Two tests that are identical in every way except for the actual items included. The aim of this test is that the researcher want to identify if there exist two or more forms of a test, then the two forms are equivalent (on σ's, means, and correlations with other measures) and highly correlated. The correlation coefficient, *r*, between alternate forms may be used as an estimator of the reliability tests (Karl, 2012). So it correlates the two scores. The attained coefficient is named coefficient of stability. Problems that encounter the procedures of this test are the difficulty of creating two forms that are fundamentally equivalent, and that they involve two administrations.

Split –Half Reliability

It can be difficult to administer a test two times to estimate its reliability. Practice other changes between two times: Time 1 and Time 2 might undermine stability estimates of reliability. Another method is to divide the factors into two sets, compute each subject's value on the each half, and correlate the two groups of values (Karl, 2012). This test needs only one administration, and it is very appropriate for a very long test. The most used way to divide the test into two parts is using the approach: odd-even. As longer tests appear to be more reliable, and as Split-Half Reliability test characterizes the reliability of a test only half providing the actual test, a modification formula shall be applied to the coefficient- Spearman-Brown correction formula.

$$r_{sb} = \frac{2\,r_{hh}}{1 + r_{hh}} \tag{3}$$

Note that Split-Half Reliability test is a procedure of internal consistency reliability, and r_{hh} is the Half-Test Reliability Coefficient. A problem that appears of using the split-half process is that the reliability estimate achieved using one pair of random halves of the items is likely to differ from that achieved using another pair of random halves of the factors. Here appears the question: Which random half is the one we should use? This problem can be solved by computing the Spearman-Brown corrected split-half reliability coefficient for every one of the possible split-halves and then

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

find the Cronbach's alpha. Cronbach's alpha is a minor bound to the true reliability of a measuring instrument, and that it possibly will underestimate the true reliability (Osburn, 2000).

Internal Consistency Reliability

Internal consistency evaluates the consistency of results across factors within a test. Cronbach's alpha is the most used internal consistency measure, which is generally founded as the mean of all possible split-half coefficients (Cortina, 1993). It is an overall of an earlier procedure of estimating internal consistency.

Internal Consistency Reliability test determine how all factors on the test relate to all other factors.

It is applied to sets of factors proposed to measure different features of the same concept. Its process works as a single component hits only one feature of a concept. If many different factors are employed to gain information about a specific construct, then the data set is more reliable.

Cronbach's alpha shows degree of internal consistency. It is a meaning of the number of factors in the scale and the degree of their inter-correlations. It ranges from zero to one exclusively, and it measures the proportion of variability that is shared among factors (in another words it is the covariance among factors). Moreover, if all factors dispose to measure the same entity, then they are highly related and the value of alpha must be high. On the other hand, if all factors dispose to measure different entities, then the correlation among each other is very low, and the value of alpha is low, too. Note that the main cause of measurement error is because of the sampling's content.

The conceptual formula of Cronbach's Alpha is defined by:

$$\alpha = \frac{K\,\bar{r}}{[1+(K-1)\bar{r}\,]} \qquad (4)$$

Where K is number of factors and $\bar{r}$ is the average correlation among all factors (the mean of the $K\,(K\text{-}1)\,/2$ non-redundant *correlation coefficients* (i.e., the mean of an *upper triangular*, or *lower triangular*, correlation matrix).

In general, Cronbach's alpha increases as the inter-correlations among test factors increase, and then it is recognized as an internal consistency estimate of reliability of test values. Since inter-correlations among test factors are maximized when all factors measure the same construct, Cronbach's alpha is considered to indirectly indicate the degree to which a group of factors measures a particular unidimensional latent construct. However, it is easy to demonstrate that tests with the same test range and variance, but different fundamental factorial structures may result in the same values of Cronbach's alpha. In fact, a lot of researchers have proven that α may undertake fairly high values even though the set of factors measures many unrelated latent constructs (Trippi, 1976, Knapp, 1991; Schmitt, 1996; McDonald, 1999; Zinbarg et al, 2006). As a result, Cronbach's alpha is mostly employed when the factors measure different practical extents within a single construct. Indeed, reliability is directly related to the validity of the measure.

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

Types of Validity

Validity is questionably the most vital measures for test quality. The word validity tends to whether or not the test measures what it claims to measure. Validity is the amount of which an idea, deduction or measurement is well-originated and relates precisely to the real world (Brains et al, 2011). The term "valid" is resulting from the Latin word “validus”, which means “strong”. The validity of a measurement instrument is reflected to be the degree to which the instrument measures what it asserts to measure; in this situation, the validity is said to be equivalent to accuracy.

Validity is employed in the context of the following three concepts: form of the test, purpose of the test, and the population for whom it is proposed. Therefore, avoid to ask the question “Is this a valid test?” Preferable to be replaced by “how valid is this test for the decision that I need to make?” In fact, there are different types of validity such as construct validity, convergent validity, content validity, and discriminant validity.

Construct Validity

Construct validity is the degree in which a test measures a theoretical construct that is intended to be measured. Simply states that the construct validity of a measurement or an operation is the amount to which it really measures what it claims to measure. Many times scientists evaluate or measure abstract constructs. When the aspect being measured is known as an abstract construct that is estimated from straight apparent events, then it may be known as “construct validity.”(Karl, 2012).

The method of validating the clarifications about that construct as designated by the test score is construct validation. This may be achieved experimentally, e.g. when a researcher wants to validate a measure of nervousness. There is a hypothesis states that nervousness rises when subjects are under the threat of an electric shock, then the threat of an electric shock should raise nervousness scores (Nunnally, 1978).

Convergent Validity

Convergent validity is a tool frequently employed in behavioral sciences. It involves the degree to which two measures of constructs that theoretically should be related, are indeed related (William, 2006). In another words, an effective assessment of convergent validity indicates that a test of a perception is highly correlated with other tests intended to measure theoretically similar perceptions. For example, to demonstrate the convergent validity of a test of statistics skills, the grades on the test may be correlated with grades on other tests that are also planned to measure elementary statistics ability; high correlations among the test grades would be proof of convergent validity. Hence, convergent validity refers to the highly theoretical correlation between the instrument employed and the measures of other constructs.

Content Validity

Content validity is a decision of how fitting the factors look like to a panel of assessors who have good knowledge about the subject matter (Robert, 2004). It speaks of how precisely a measurement

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

tool taps into the various features of the particular construct in question (Chris, 2016). The aim of content validity test is to find out how the questions really assess the construct. In another words, this test will discover which questions must be included in the survey and which should be not. For example, if we want to test knowledge on Lebanon geography it is not reasonable to have several questions limited to the geography of Mexico. Content validity is a subjective decision; there is no correlation coefficient. Indeed, this method of testing is a logical method rather than empirical one due to the fact that if we want to realize whether or not the whole content of the construct is characterized in the test, we compare the test task with the content of the construct. A survey has content validity if its factors are randomly selected from the sample space of all possible factors, and to measure this validity it is highly recommended to exert enough effort to describe the population interest and then request experts to judge how the selected sample describe well that population (Karl, 2012).

Discriminant Validity

Discriminant validity is used to test whether measurements which are expected to be related are, in fact, unrelated (John et al, 2000). This test could refer to the question: "Is the instrument employed not well correlated with measures of other constructs to which it should not be related?" For instance, we might estimate grades on our test not to have a high correlation with other different tests. So Convergent validity can be estimated using correlation coefficients.

Convergent validity, along with discriminant validity, is a certain kind of construct validity. When assessing their test validity, Campbell and Fiske (1959) have announced the perception of discriminant validity. They assured the significance of using both convergent and discriminant validation methods when measuring new tests. An effective assessment of discriminant validity indicates that a test of a perception is not highly correlated with other tests intended to measure theoretically different perceptions.

Convergent validity can be initiated when two similar constructs relate to one another, while discriminant validity relates to two unrelated constructs that are simply discriminated. To summarize, in order to establish construct validity, researchers have to validate both convergence and discrimination (Henseler et al, 2014).

Reliability and Validity Analysis

As we expressed in the literature there are different tests to achieve reliability and validity analysis. One of these tests is to accomplish an internal consistency reliability in which Cronbach's alpha is calculated for good estimation of this consistency (Cronbach 1951; Nunnally 1979). A value of Cronbach's alpha between 0.6 and 0.8 is acceptable (Wim et al, 2008). To measure the construct validity in an empirical approach the rule of thumb item-to-total correlations, and the inter-item correlations are adequate. Cohen (1988) states that if inter-item correlation lies within 0.10 and 0.29, then there is a weak correlation for both positive and negative values, and when inter-item correlation lies within 0.30 and 0.49 a medium correlation, and lastly if inter-item correlation is between 0.50 and 1.00 a strong correlation. Moreover, Robinson (1991) recommends that, in an empirical approach and as a rule of thumb, if the score of the item-to-total correlations is more than 0.50 and the inter-item correlations exceeds 0.30, the construct validity is satisfied. Hence,

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

the process to establish internal consistency reliability and validity of constructs proposed to describe a given scale of a certain model requires to determine the following measures:

1- Cronbach's alpha: Measures the degree of internal consistency.

2- Inter-item Correlation: An acceptable value should be > 0.30

3- Item-Total correlation: An acceptable value should be > 0.50

4- Cronbach's alpha if item deleted: Measures the value of Cronbach's alpha coefficient after the removal of the corresponding item.

These tests are statistical tests which could be conducted using SPSS.

The interpretation of these tests is represented below.

Consider a construct "A "which is associated with three proposed items as indicated in Table 1.

Table 1. Proposed items to construct "A "

Construct	Item
A	X1 X2 X3

X1, X2, and X3 are the proposed items that correspond to questions in which they are included in the questionnaire under the construct "A". Internal consistency reliability and convergent validity tests would indicate which of the X's items shall be remain and which ones shall be omitted.

Suppose that through SPSS software we got the results indicated in Table 2, Table 3, and Table 4:

Table 2 Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.806	.886	3

Table 2 shows that Cronbach's alpha is equal to 0.806 > 0.7, then it **can** be concluded that the scale used to measure the three items measuring the construct "A" is reliable, and then we move to study " Inter-item correlation". Note that if **supposedly** the value of Cronbach's alpha lies within 0.6 and 0.7, then it can be deduced that the scale employed to measure the three items measuring "A" needs to be tested for low value of Cronbach's alpha. In this case, it is preferable to keep these items for further analysis by examining the validity measures through the tests Inter-item correlation, Item-Total correlation, and Cronbach's alpha if item deleted. Along with the Cronbach's alpha values, the internal consistency among the items can be tested using the item-to- item and item-to-total correlation measures which are also part of the reliability tests.

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

Table 3 Inter-Item Correlation Matrix

	X1	X2	X3
X1	1.000	.423	.168
X2	.423	1.000	.534
X3	.168	.534	1.000

Table 3 indicates that internal consistency values with respect to item-to-item correlation were more than 0.3, except for the inter- item correlation between X1 and X3 (correlation = 0.168 < 0.300). This result indicates that there is doubt to eliminate of X1 and X3, but it is preferable to keep them for further examinations.

Table 4 Item-Total Statistics

	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
X1	.402	.630
X2	.765	.523
X3	.323	.856

(1) Overall alpha: As the results in Table 2 show, overall alpha is 0.806, which is very high and indicates strong internal consistency among the three items. Essentially this means that respondents who tended to select high scores for one item also tended to select high scores for the others; similarly, respondents who selected a low scores for one item tended to select low scores for the other 'A" items. Thus, in general, knowing the score for one "A" item would enable one to predict with some accuracy the possible scores for the other two "A" items.

(2) Corrected Item – Total Correlation: Table 4 highlights the column containing the "Corrected Item – Total Correlation" for each of the items. This column displays the correlation between a given "A "item and the sum score of the other two items. The correlation between X2 and the sum of X1 and X3(i.e. item2 + item 3) is r = 0.765 > 0.500. This means that there is a strong, positive correlation between the scores on the item 2(X2) and the combined score of the other two items (X1 and X3). This result assesses how well X2 score is internally consistent with composite scores from all other items that remain. However, the correlation between X1 and the sum of X2 and X3 is r = 0.402 < 0.500 which is approximately a low fair correlation for item – analysis purposes, then that item should be removed and not used to form a composite score for the variable in question. Similarly, X3 is not internally consistent with the other two items (r = 0.323 < 0.500). So when we create the composite (overall) score for the construct "A" (the step taken after reliability analysis) we **may create** the composite using only items X2 and we would simply ignore scores from item X1 and X3 because they were not internally consistent with the other items. Note that for any remained construct in the scale it must be expressed by at least three reliable and valid items.

(3) Cronbach's Alpha if item deleted: Table 4 displays Cronbach's alpha that would result if a given item were deleted. This column of information is valuable for determining which items from

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

among a set of items contributes to the total alpha. The value presented in this column represents the alpha value if the given item were not included. For X2, the Cronbach's alpha if item 2 was deleted would drop from the overall total of 0.806 to 0.523. This item appears to be useful and contribute to the overall reliability of the construct "A". Then this item (X2) could be remained.

For X1, the Cronbach's alpha if item 1 was deleted would drop from the overall total of 0.806 to 0.630. Since alpha would drop with the removal of X1, this item appears to be useful and contribute to the overall reliability of the construct "A". Then this item (X1) can be remained although it has very low inter – item correlation and item-total correlation scores. However, Cronbach's alpha would increase from 0.806 to 0.856 if item 3 were deleted or not used for computing an overall "A" score. So should this item be removed and should the overall construct "A" composite be created only from items 1 and 2? In this case the answer is yes because firstly, alpha is increased by a large degree from deleting item 3, secondly, item 3 does not correlate very well with the composite score from item 1 and 2 (the item – total correlation for item 3 is $0.323 < 0.50$), thirdly, it has very low inter – item correlation with X1 (inter-item correlation score between X3 and X1 $= 0.168 < 0.300$). Hence, since deletion of item 3 results in large change, and since item 3 does not correlate well with the composite of items X1 and X2, then there is a statistical reason to drop item X3.

CONCLUSION AND RECOMMENDATIONS

Internal- consistency reliability and construct validity tests determine how all factors on the test relate to all other factors and which factors (items) should be remained in the scale. They are applied to sets of factors proposed to measure different features of the same concept. Their processes work as a single component hits only one feature of a concept. They can be measured through the scores of Cronbach's alpha coefficient, inter-item correlation, item-total correlations, and Cronbach's alpha if item deleted. However, the **actual** validity of the scale will be determined only after testing the content validity, construct reliability and discriminant validity of the instrument in which confirmatory factor analysis is employed.

Reliability also may be determined as a correlation between two events, frequent use of the scale (stability), and Correspondence of factors (internal consistency). Validity and reliability are two significant characteristics of behavioral measure and are referred to as psychometric properties.

It is important to keep in mind that validity and reliability are not an all or none issue but a matter of degree.

Finally, it is recommended for the scientists to adopt and apply a reliable and valid instrument related to their subject area with some modifications in the constructs and items. In this case, all what they need is to test the reliability and the validity of this instrument; otherwise, an exploratory factor analysis, confirmatory factor analysis, and path analysis are to be employed by using a structural equation model. Scientists have also bear in mind that Cronbach's alpha coefficient likely needs at least three items to give a good score. Indeed, when the number of items is increased will expand this estimator and will probably lead to a significant score while the items are unrelated among themselves. It is also recommended for the scientist to avoid quick judgement on dropping

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

any item in doubt during reliability and validity analysis; further analysis for these items will make the study significantly pure.

At last but not least, it is recommended to have a large sample when doing any test. For instance, SPSS tool will run to give significant results only if the sample size is more than 200. It is preferable to be at least 240.

REFERENCES

Brains, Willnat, Manheim, Rich, 2011. *Empirical Political Analysis 8th edition*. Boston, MA: Longman p. 105

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Chris Clause, (2016). Content Validity: Definition, Index & Examples. Research Methods in Psychology: Help and Review. Chapter 11, lesson 8. URL: study.com/academy/.../content-validity-definition-index-examples.html

Cohen L., (2013). Research Methods in Education, Sixth Edition - STIBA Malang.URL: www.stiba-malang.ac.id/.../RESEARCH%20METHOD%20COHEN%20ok... by L Cohen - Cited by 12126.

Cortina, J.M., (1993). What Is Coefficient Alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, 78(1), 98–104.

Henseler, J., Ringle, C.M., Sarstedt, M., (2014). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43 (1), 115–135.

John, O.P., & Benet-Martinez, V. (2000). Measurement: Reliability, Construct Validation, and Scale Construction. In Reis, H.T., & Judd, C.M. (Eds.). *Handbook of Research Methods in Social and Personality Psychology*, pp 339-370. Cambridge, UK: Cambridge University Press

Karl L. Wuensch, (2012). A Brief Introduction to Reliability, Validity, and Scaling. *Reliability-Validity- Scaling. docx.*URL: http://www.core.ecu.edu/psyc/.../Reliability-Validity-Scaling.docx

Knapp, T. R., (1991). Coefficient alpha: Conceptualizations and anomalies. *Research in Nursing & Health*, 14, 457-480.

McDonald, R. P.(1999).Test theory: A unified treatment. *Psychology Press*. ISBN 0-8058-3075-8

Neil R. Carlson (2009). Psychology: the science of behavior (4th Canadian Ed.). *Toronto: Pearson*. ISBN 978-0-205-64524-4

Nunnally, J., 1978. Psychometric Theory. *McGraw-Hill, New York.*

O'Neill, P., (2003). Moving Beyond Holistic Scoring Through Validity Inquiry. *Journal of Writing Assessment*, Vol.1, No.1, pp.47-65

Osburn H. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5, 343-355.

ISSN 2056-3620(Print), ISSN 2056-3639(Online)

Ralph Rosnow & Robert Rosenthal (2008). Essentials of Behavioral Research: Methods and Data Analysis. *McGraw-Hill Humanities/Social Sciences/Languages*; 3rd edition, April 4, 2007.

Robert S. Michael, 2004. Measurement: Reliability and Validity. *Y520 Strategies for Educational Inquiry.* URL: http://www.indiana.edu/.../week.../reliability_validity_2up

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8, 350–353.

Trippi, R. & Settle, R. (1976). A Nonparametric Coefficient of Internal Consistency. *Multivariate Behavioral Research*, 4, 419-424. URL: http://www.sigma-research.com/misc/

Weigle, S.C., (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, 16, 194–209.

William M.K. Trochim, (2006). Convergent & Discriminant Validity. *Research Methods: Knowledge Base.* URL: http://www.socialresearchmethods.net/kb/convdisc.php

Wim J., Katrien W., Patrick D. P., and Patrick V. K., (2008). Marketing Research with SPSS. *Prentice Hall*; *Pearson Education*. ISBN: 978-0-273-70383-9. 274-275.

Zinbarg, R., Yovel, I., Revelle, W. & McDonald, R. (2006). Estimating generalizability to a universe of indicators that all have an attribute in common: A comparison of estimators for α. *Applied Psychological Measurement*, 30, 121–144.

ISSN 2056-3620(Print), ISSN 2056-3639(Online)