QIJIS: *Qudus International Journal of Islamic Studies*
Volume 4, Issue 2, August 2016

ASPECT OF LANGUAGE ON A QUALITATIVE ANALYSIS OF STUDENT'S EVALUATION INSTRUMENT

Ismanto
STAIN Kudus , Central Java, Indonesia
ismanto_07@yahoo.co.id

Abstract

This article examined the characteristics of good student's evaluation instrument. There are at least two requirements that must be met. Those are valid and reliable. The validity of the instrument can be seen from the instrument's ability to measure what should be measured. The fact the existence of the validity of an instrument may be a grain fill, the response process, internal structure, relationship with other variables, and the consequences of the implementation of the charging instrument. Analysis of the content is then known as content validity, i.e. rational analysis of the domain to be measured to determine the representation of each item on the instrument with the ability to be measured. Content validity is submitting pieces of blue print and items of the instrument to the experts to be analyzed quantitatively and qualitatively.

Keywords: *Aspects Of Language, Student's Evaluation Instruments, Content Validity*

A. Introduction

There is a modest opinion regarding the validity of the measurement instruments used, in which the validity of an instrument is the instrument's ability to measure what should be measured (Nunnally & Bernstein, 1994: 83). And Retnawati (2016: 16) also mentions that there are also the syntheses on several expert opinions regarding the validity, stating that validity is the support of empirical facts and theoretical reasons to the interpretation of test scores or instrument, and associated with rigor. From both, these opinions can give understanding to the validity, namely the ability of items or instruments to measure what is being measured carefully.

Validity can be grouped into three types, namely the criterion validity, content validity, and construct validity (Nunnally & Bernstein, 1994, and Allen & Yen, 1979). The existence of instrument validity can be seen through the rational analysis of the content matter and empirical analysis of the test scores response data item. This rational analysis is then commonly said with a qualitative analysis of the items, and rational analysis known as quantitative analysis of the test scores. Qualitative analysis can also be applied to the non-test instrument.

One type of validity that was developed in this paper is the validity of the content, where the content validity of an instrument is the status of items in the instrument representing the components in the overall region and reflecting the contents of the object to be measured behavioral traits. By paying attention to this case, the validity of the content related to the rational analysis of the domain to be measured to determine the representation of the instrument with the ability to be measured. It means a qualitative analysis of major concern to analyze the contents of each item on both the test and non-test of an assessment instrument. On the measurement of learning outcomes, the validity of a requirement is needed in the development of the instrument.

B. Evidence Related of Contents

Model simple linear classical test X = T + E (X is observable test score, T is the true score, and E is the error of measurement) provides basis the conceptual-statistical validity, especially how the instrument can provide conformance measurement results with the purpose of the test is structured. Evidence of the content and construct validity measured by an instrument can be obtained through the rational and empirical analysis of how adequate instrument to represent the realm of content and how relevant the content domain in accordance with the intended interpretation of test scores. The contents of the test refers to the themes, the choice of words, as well as the format or the form of the items, tasks or questions used in the test.

Evidence related to the content is generally obtained through the assessment of specialists or experts on the conformity between the instrument parts with construct measured (Supratiknya, 2014: 169). Experts' agreement on matter often called the domain being measured determines the degree of validity of the content (related content). This is because the measurement instruments, for example in the form of a test or questionnaire can be proved valid if the expert believes that such instruments to measure mastery of skills defined in the domain or also the psychological constructs measured (Retnawati, 2016: 18). The presence of the experts here becomes inevitable in the assessment of the quality of item and instrument containing items.

With attention to the steps following the preparation of the instrument, the position and role of the experts will appear their crucial role in a construction instrument. The general steps in preparing testing instrument are as follows: (a) defining the test, (b) preparing spesifications test, (c) selecting a scaling method, (d) constructing the items, (e) request a review of items from a number of relevant experts, and revised as necessary; (f) to assemble the items into a form that is ready for the semi-final test piloted; (g) carried out tests on a representative sample of the population of the audience who were subjected to the test, (h) examine the characteristics of psychometric scores item by item analysis, to select items, items that can be specified as a candidate the form final scale and items that still needs to be revised or terminated, (i) examine the reliability, validity, and discrimination power of the form final test, (j) draw up a manual or guidebook test and publish test (Crocker & Algina, 2008: 66, Supratiknya, 2014: 182, Mardapi (2008), and Arifin, 2012: 88-101). The fifth step regarding review and revision of the item is part of the role of experts, both aspects of substance, construction, and language

In detail, aspects of the tests content of that need to be evaluated according to Supratiknya (2014: 169-170) include: (a) the sufficiency, whether the content of the test are insufficient or inadequate in terms of representing domain specific content to be measured; (b) clarity, whether the content clearly reflects the realm of specific content to be measured in terms such as not to confuse with other specific content domain; (c) relevance, i.e. whether the contents of such tests have compatibility with the specific content domain to be measured; (d) the conformity between the items and tasks are used as stimuli in these tests with the definition of the construct measured; (e) the presence or not of bias in the form of gender bias in the test content, culture, age or the other of social grouping factors; (f) the possibility of a "construct irrelevant variance" (variances that are irrelevant to the construct measured) and "construct underrepresentation" (lack of adequate representation of the construct measured), which shows the extent of the possibility of such tests measure exceeds (construct irrelevance variance) or less (construct underrepresentation) of that measure. This type of evidence is associated with content validity with regard to first construct the realm of content and targeted measurement.

C. Review and Revision Item

After each item is composed based on the table of specifications and guidelines for writing other relevant items,

items should then requested a review from the experts. Those should be relevant to revision and should be completed by advice from the reviewer before those are assembled into the shape of a semi-final that is ready to be tested. According to Crocker and Algina (2008: 81), study or review of that item will include an examination of the matters as follows: (a) accuracy, the accuracy of the formulation of conceptual constructs or psychological attributes are measured along with the formulation of the operations to the indicators behavior, even down to the items selection format; (b) the appropriateness or relevance to test specifications, its relevance to a table of specifications, particularly related to the conformity between content items with both components of content and process, as well as the number of items corresponding distributions are planned in the table of specifications, (c) technical item-construction flaws, namely the presence or not of various errors of technical preparation items, such as the presence of more than one idea or problem in a item, use the negative form or words that might give a clue towards an answer such as “always”, “never”, and the like, (d) grammar, (e) offensiveness or appearance of “bias”, namely the choice of words can give the impression offend or discriminate against certain groups, and (f) the level of readability, the level of difficulty of the languages spoken in comparison to group audiences will be subjected to the test.

Aspects of the substance and the construction can be checked on the basis of accuracy, appropriateness or relevance to the test specifications and technical item-construction flaws, while the quantity should receive more attention on aspects of language examined on the basis of grammar, offensiveness or appearance of “bias”, and the level of readability. Aspects of language into a specific role in review of item by experts, given in the context of Indonesian with the emergence of numerous ethnic groups with their own language, linguists’ presence should have an important role in each review of item in each school subject.

Instrument development can also occur by making the instrument; by adapting to the instrument that comes from abroad. The instrument cannot be directly used in Indonesia because of differences in language and culture. Therefore, the adaptation of instrument is needed to develop instruments to translate it into Indonesian and adapt them to the culture of Indonesia. According to Hambleton and Patsula (1998), the action to adapt or translate tests into languages or other cultures, mainly due to the following reasons: (a) often adapting or translating the test is cheaper and easier than making a new test in the local language, (b) if the purpose of testing is to measure the psychological aspects of community cross-cultural or cross-country, adapting the test is the most effective way to create tests in the local language, (c) at least the experts in the country who are able to make a test, (d) there is a sense of safe to use on tests that have been adapted rather than test the newly created, especially when adapted tests are tests that are already well known, and (6) usually the similarity or belief persists with the measurement results, even though the test was a different language. Instrument development through translation of a foreign language at least involves two linguists; the master of foreign languages and Indonesian itself. The two are Indonesian.

The procedure is done by translating the instrument adaptation of the instrument using the techniques of backward-translation or translation back and forth (Suharsono & Istiqomah, 2014). Instrumental translation into Indonesian is done by foreign language experts, and then it is followed by consulting the results of the translation to an Indonesian. Then it is translated back into the original language (foreign) like languages such authentic instrument originated from and back to an expert consulted a foreign language. After consulting to a linguist, it is back translated into Indonesian. The goal is to avoid the mistakes of meaning contents of these items as well as testing the content validity, for example through professional judgment or review by experts.

There are two practical advices needed to prepare a review of item (Supratiknya, 2014: 199). The first item is the presentation of the draft item pool to request a review. To facilitate the experts conducting a review, otherwise the draft item pool presented systematically following the grid distribution both content and process. In addition, a complete draft design of tests ranging from definitions to table of specifications should also be included; so that the experts can perform their duties optimally. That is examining the item. Secondly, experts need to be involved differently, including experts in various fields as well as laymen. The expert group could include specialists in a particular matter or discipline as well as teachers or lecturers subjects or certain subjects. They especially can be asked to do review in aspects of accuracy or timeliness and relevance of the draft formulation construct about the table of specifications. The expert group also needs to include psychometric experts in order to give a particular review of the accuracy of format choice items as well as the presence or not of various technical errors of preparation of items. It is referred to lay the real experts also unnecessary expert status, but who knows in depth the characteristics of the group to be subjected to the test. In particular, this common resource can be held to examine related word choices that will not cause a certain bias as well as the level of appropriate language difficulty for the target audience group. The need for a review of item becomes a necessity in terms of review and revision of items as needed, to repair the item for the future.

Every review that is collected from various experts critically needs to be processed and used as a basis to make improvements or revisions to the draft item pool. One possible unfavorable tendency of making up anything includes preparation of the matter, it is defensive to feel more out of other people and not easy to accept input. Every critical note from the experts has to be accepted as a clue about the possibility of something that has not been settled in the draft pool item, and it must be observed and followed up with revised as necessary

(Supratiknya, 2014: 200). Strengthening of experts in terms of review and revision of the items is an insight of the rater in addition to the items, as well as an understanding of the groups involved in the construction and filling of the instrument, making it more comprehensively to the overall understanding of the components involved in the evaluation process of learning.

D. Proof of Content Validity

Evidence of regarding the content matter or the evaluation instrument of learning above is also confirmed by the statement of Kumaidi (2014), that "... The approach that should be avoided is meant proving the validity of which is based on an analysis of item (item analysis), especially the use of the correlation coefficient score of item s and total score test (r_{ix}). "That means proving validity by calculating the correlation item with total need to be avoided, in other words that the user inaccuracy product-moment correlation (r_{ix}) as the index validity as a mistake and should be avoided.

Proof of content validity can be done by reviewing the items, which covers aspects of the material or substance, construction, and language (Arifin, 2012: 144-145, Mardapi, 2012: 182, and Retnawati, 2016: 42), as well as emphasized by Permendikbud number 23 year 2016 article 14th, paragraph 2nd states that the assessment instruments with terms of three aspects can be used by educational units up to the government. Material aspects such as the suitability of the questions with indicators, limit the question and a clear answer, the distracters are functioning properly, the material content in accordance with the purpose of the test, the material content in accordance with the level, type of school, and grade. Construction aspects such as the formulation of the sentence problems or questions should use the question word or command that demands an answer unraveled; the subject matter did not give instructions to the answer key; the subject matter is free of the question is a double negative; there are clear instructions about how to do the problems, contains guidelines for scoring, images, graphics, tables, diagrams, discourse and the like contained in the matter to be presented clearly and function; length of the answer choices are relatively equal; the answer choices does not use the expression “all the answers to the above are false” or “all of the answer choices in the top are right” and like; no clue leads to the correct answer; the answer choices are homogeneous; the answer choices in the form of numbers or time listed in order of the size of these figures or chronological; and items do not depend on the answers to the previous item.

While aspects of language (also Basuki, 2010: 186) in the form of the formulation of communicative sentences is to use simple language and words that are already known to the student; item use of the Indonesian language is good and true; the formulation of the items do not use the word or phrase created an interpretation double or misunderstanding, do not use the language of the local or regional; the formulation of questions does not contain words that can offend the students; and the answer choices are not repeat word or phrase of the same word. Three aspects mentioned above involve some experts, such as experts on the matter, measurement, and language. Agreements review outcomes of each expert as a form of proof of the validity of the content, i.e. using validity index to determine the quality of the item, which can then be followed by a revision as necessary to point to the validity below the standard required.

Proof of the validity index can be done by various methods. First, the method of Nieveen (1999: 126) with steps, i.e. (a) determining the average for each criterion of validator or rater (K): $\mathrm{K} = \frac{\Sigma \mathrm{v}}{\mathrm{n}}$, (b) finding the average three aspects (A): $\mathrm{A} = \frac{\Sigma \mathrm{K}}{\mathrm{m}}$, and (c) finding the average total validity of all aspects (RTV): $\mathrm{RTV} = \frac{\Sigma \mathrm{A}}{\mathrm{p}}$. With v is the assessment score rater (e.g. 1, 2, 3, or 4), n is the amount of rater, m is the number of criteria in every aspect (e.g. 4 to v above), and p is the amount of aspect, each step can be seen in the following illustration:

Table 1 Rater assessment results:

No	Aspect and Criteria	Assessment			K	A	RTV
		Rater 1	Rater 2	Rater 3			
	Aspect Format						1.89
1.	Aaa	2	2	3	2.33	1.25	
2.	Bbb	3	2	3	2.67		
	Aspect Substance						
1.	Ccc	3	4	4	3.67	1.92	
2.	Ddd	4	4	4	4		
	Aspect Language						
1.	Eee	2	3	3	2.67	1.5	
2.	Fff	4	3	3	3.33		

Note: data above is manipulated

Using the criterion of validity (Khabibah, 2006: 76) may have one of the following categories: 3.25 ≤ RTV≤ 4 means that it is valid, 2.5 ≤ RTV<3.25 means that a valid, 1.75 ≤ RTV<2.5 means less valid, and 1≤ RTV<1.75 means invalid. So from the data sample assessment of the points above, obtained RTV = 1.89 in the category of less valid.

Second, the method of Gregory (2007) is an index to show the results of expert judgments agreement on the content validity of an instrument. This index ranges from 0 to 1. By making contingency table for a minimum of two experts, with the first category is not relevant (NR) and less relevant (LR) into categories of weak relevance (or can be categorized 0), and a second category for which sufficient relevant (SR) and very relevant (VR) into categories of strong relevance (or can be categorized as 1). An Index deal with the experts of content validity is a comparison of the number of item of the two experts with the category of strong relevance to whole items. The following contingency table is given to each expert as assessor to provide a check mark (V):

Table 2. Contingency rater assessment:

No	Item descripsion	Assessment			
		NR	LR	SR	VR
1	Aaaa		V		
2	Bbbb			V	
3	Cccc				V

Note: data above is manipulated

And the results table of rater assessment as follows:

Table 3. Rater assessment results:

Number of item	Rater 1	Rater 2
1	1	1
2	1	0
3	0	1

Note: data above is manipulated

If there are two experts involved in the assessment instrument, it indexes the content validity using Gregory method can be calculated by a formula,

$$VI = \frac{D}{(A+B+C+D)} \text{ or } VI = \frac{D}{k}$$

where VI is a content validity index D is the amount of item of the two experts with strong relevance categories, k is the amount of whole item, while the A, B, and C is the amount of item of both experts and/or one of them with a weak relevance. Gregory's formula above can be developed for the assessment of three or more experts as raters. For three raters, Gregory's formula can be written as follows:

$$VI = \frac{H}{(A+B+C+D+E+F+G+H)} \text{ or } VI = \frac{H}{k}$$

Where H is the amount of item of three expert categories strong relevance, as well as the A, B, C, D, E, F, and G is the amount of item of the three experts, and/or two and/or one of them by weak relevance. The amount of capital letter stating variance assessment of the raters against the item can be calculated with $2^n 2^n$, where n is the amount of rater.

Third, the method Aiken (1980, 1985) in Retnawati (2016: 18) and Azwar (2012) is an index V-Aiken as an index of agreement rater regarding validity, meaning that the index V-Aiken is an index of agreement raters of the appropriateness of item with indicators to be measured using a particular item. V-Aiken index ranges from 0 to 1. By making contingency table for a minimum of three experts, the categories are not relevant (NR, was given a score of 1), less relevant (LR, was given a score of 2), is sufficient relevant (SR, was given a score of 3), and very relevant (VR, were given a score of 4). Contingency table given to each expert as rater like table 2, while the results of the assessment of rater written as follows:

Table 4. Rater assessment results:

Number of item	Rater 1	Rater 2	Rater 3
1	2	1	2
2	3	4	3
3	4	4	3

Note: data above is manipulated

The V-Aiken index can be calculated as follows:

$$V = \frac{\sum s}{n(c-1)}$$

Where V is the index of an agreement raters regarding item validity, s is score given to each rater minus the lowest score in the category of scoring used (s = r - l_o, with r is score category of rater selection and l_o is lowest score in the category of scoring), n is the amount of rater; and c is the amount of categories that can be selected rater.

From the calculation of the index VI or V, an item or instrument can be categorized based on the index. According Retnawati (2016: 19), if the index is less than or equal to 0.4 is said to be less valid, from 0.4 to 0.8 is said to be sufficient valid, and if it is greater than 0.8 is said to be very valid.

E. Conclusion

Validity is the ability of items or instruments to measure what is being measured. Content validity of an instrument is the status of items in the instrument representing the components in the overall region of the object and reflects the behavioral traits which are measured. The validity of an item or instrument can be seen through the rational analysis of the content matter and empirical analysis of the test scores of item response data. This rational analysis is then commonly said with a qualitative analysis of the items (test and non-test) by experts involving content, measurement, and language. While rational analysis is known as quantitative analysis of the test scores. While rational analysis is known as quantitative analysis of the test scores. Proof of the content validity can be done by reviewing the item(s), which include aspects of material, construction, and language, which are then expressed in content validity index. Method of proving the content validity can use several options, tailored to the needs and conditions, the method of Nieveen, Gregory, and/or Aiken.

REFERENCES

Allen, M.J. & Yen, W.M. 1979. *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole Publishing Company.

Arifin, Z. 2012. *Evaluasi Pembelajaran.* Jakarta: Direktorat Jenderal Pendidikan Islam Kementerian Agama RI.

Azwar, S. 2013. *Validitas dan Reliabilitas.* Yogyakarta: Pustaka Pelajar.

Basuki, I.A. (Ed.). 2010. *Dasar-dasar Evaluasi Pembelajaran Bahasa Indonesia*. Malang: Universitas Negeri Malang.

Crocker, L. & Algina, J. 2008. *Introduction to Classical and Modern Test Theory*. OH: Cengage Learning.

Gregory, R.J. 2007. *Psychological Testing: History, Principles, and Application*. Boston: Pearson.

Hambleton, R. K. & Patsula, L. 1999. Increasing the Validity of Adapted Test: Myth to be Avoided and Guidelines for Improving Test Adaptation Practices. *Journal of Applied Testing Psychology*, August 1999. Acossiation of Test Publishers (ATP).

Khabibah, S. 2006. *Pengembangan Model Pembelajaran Matematika dengan Soal Terbuka untuk Meningkatkan Kreativitas Siswa Sekolah Dasar*. Disertasi doktor, tidak diterbitkan, Universitas Negeri Surabaya, Surabaya.

Kumaidi. 2014, 24 Mei. *Validitas dan Pemvalidasian Instrumen Penilaian Karakter*. Makalah disajikan dalam seminar Nasional Pengembangan Instrumen Penilaian Pendidikan Karakter yang Valid, di Fakultas Psikologi, Universitas Muhammadiyah Surakarta.

Mardapi, D. 2008. *Teknik Penyusunan Instrumen Tes dan Non Tes*. Yogyakarta: Mitra Cendikia Press.

Mardapi, D. 2012. *Pengukuran, Penilaian & Evaluasi Pendidikan*. Yogyakarta: Nuha Litera.

Nieveen, Nienke. 1999. Prototyping to Reach Product Quality. In Van den Akker, Jan., et. al. *Design Approaches and Tools in Education Training* (pp 125 – 135). Dordrecht: The Netherland Kluwer Academic Publisher.

Nunnally, J.C., & Bernstein, I.H. 1994. *Psichometric Theory* (3rd edition). New York: McGraw-Hill.

Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 23 Tahun 2016 Tentang Standar Penilaian Pendidikan

Retnawati, H. 2016. *Validitas, Reliabilitas & Karakteristik Butir*. Yogyakarta: Parama Publishing.

Suharsono, Y. & Istiqomah. 2014. Validitas dan Reliabilitas Skala Self-efficacy. *Jurnal Ilmiah Psikologi Terapan*. Vol. 02, No.01, 144-151.

Supratiknya, A. 2014. *Pengukuran Psikologis.* Yogyakarta: Penerbit Universitas Sanata Dharma.