PE - 9

APPLYING CONTENT VALIDITY RATIOS (CVR) TO THE QUANTITATIVE CONTENT VALIDITY OF PHYSICS LEARNING ACHIEVEMENT TESTS

Supahar

Physics Education Department, Mathematics and Natural Science Faculty, Yogyakarta State University
pahar.fis@gmail.com

Abstract

This paper aims to provide guidelines for us in determining content validity using a quantitative approach developed by Lawshe. This approach is introduced to express content validity quantitatively. In this approach, a panel of subject-matter experts (SMEs) are required to assess whether a measurement item represents a learning continuum which is the operationalization of a theoretical construct. Inputs from this panel is then used to calculate the CVR of each item in the measuring instrument. There are three scoring alternatives to calculate the scores of all the items, namely a particular item is "[3] essential, [2] useful but not essential, or [1] not necessary for the domain being measured. The CVR score for each item can range from 1 to -1. A high score indicates that the item measured has higher content validity. Lawshe also presents a table of CVR minimum scores based on one-tailed tests of significance with p = 0.05. Because CVR values are determined by the number of panels, therefore these CVR values will depend on the number of panels employed.

Key words: Content Validity Ratio, kuantitative

INTRODUCTION

There are a number of authors who remain stating that item-total correlations are similar to the validity coefficient of a measuring instrument (for example Sukidjo Notoatmodjo, (2012: pp.164-168), and V. W. Sujarweni (2012: p.172). Reviewing further, it is revealed that validity of a measuring instrument does not merely include item-total correlation coefficients. The item-total correlation coefficients are not a validity coefficient of a measuring instrument. Rather, they are high or low item discriminating power. Thus, one cannot claim to have obtained a valid measuring instrument simply because they have a number of item-total correlation coefficients. The location of the item-total correlation is closer to reliability compared with validity. Items with a item-total correlation, if collected, will increase reliability of the measurement, and vice versa. As a result, in the process of item selection, items with a low item-total correlation are selected in order to increase reliability.

There are three types of empirical validation procedures commonly known traditionally among developers of measuring instruments, namely content validity, construct validity, and criterion-related validity. Consulting with the experts is a testing procedure for **content validity**. We know that one of the testing procedures through content validity is aimed at obtaining expert judgment. Theories can also be used to examine whether the aspects and items in the measuring instrument we have developed be in conformity with the theories on which it is based. Analyzing parts of a measuring instrument is a testing procedure for **construct validity**. Construct validity means examining the structure of the factors in the research measuring instrument. Furthermore,

making a comparison between a measuring instrument with another similar measuring instrument is the testing procedure for **criterion-related validity**. So, we need a criterion, which is another measuring instrument. We compare our measuring instrument with another one which has been valid and tested. If this valid measuring instrument measures the same thing, then it is called **concurrent validity**. If the measuring instrument measures the impact resulting from the variable being measured, then it is called **predictive validity**. Therefore, to examine validity of a measuring instrument, i.e. the extent to which the measuring instrument measures what it is supposed to measure, conventionally people will see it from three different perspectives, namely (a) based on the contents being measured, (b) based on the theoretical construct of the attributes being measured, and (c) based on the criteria of the measuring instrument.

The most essential thing in developing a measuring instrument for cognitive attributes is to fulfill the required content validity of a test. The content validity of a test indicates the extent to which the test, which consists of a set of question items, seen from its contents does measure what it purports to measure. The indicator for the extent to which it measures what it purports to measure is determined based on the degree of representativeness of the test content in relation to the subject of measurement. The content validity of a test is determined by expert judgment through an analysis process. By using established test specifications, people perform a logical analysis to determine whether the items that have been developed do measure (representative for) what they are meant to measure. This paper will describe the stages for item validation to obtain evidence for the content validity of a test using CVR according to the method proposed by Lawshe.

DISCUSSION

Development of a measuring instrument for cognitive attributes (including achievement test) involves a series of sequential activities carried out in stages, which more or less has been standardized, which must be carried out consecutively in order to produce the expected specifications of the measuring instrument with sufficient quality. The stages taken have to be comprehensive, detailed, and specific representing the overall quality and characteristics which the measuring instrument to be developed should have. An ideal test should have complete, clear, and detailed specifications so that two or more developers of measuring instruments of equivalent qualifications who use those specifications separately will produce equivalent and interchangeable measuring instruments, in which the instruments differ only in terms of sampling (or tasks, or statements) included in each of these instruments.

The development of measuring instrument specifications is essentially a decision-making process. Each decision must be made based on considerations of various things, such as cognitive attributes to be measured, theoretical bases, the subject to be measured, objectives of the measuring activity, methods to use the measurement results, the effects of various alternatives on the reliability and validity of the measuring instrument, and so on. According Sumadi Suryabrata (2002: p. 48), things to be considered are generally include the following: specifying the area to be measured, conceptual or theoretical bases, the subject to be measured, the objective of the measuring activity, materials included in the measuring instrument, the type of question used, the total number of questions, the difficulty and distribution indexes, test blueprints, the tasks of the item writers, question development plans, the schedule for the distribution of the measuring instrument.

In general, development of an achievement test is done following the stages proposed by Oriondo & Dallo-Antonio (1984: p. 34) who state that test development should follow a number of stages, namely. (a) planning the test, (b) trying out the test, (c) determining the validity, (d) determining the reliability, and (e) determining and interpreting the test scores. The activity of planning a test includes: setting the goal, preparing the table of specifications, selecting the appropriate item format, writing items, and revising the items based on the input/

expert judgment. The activity of trying out a test includes: trying out a test and analyzing the result to obtain evidence in terms of empirical validity and reliability estimation of the measuring instrument, as well as preparing the measuring instrument format based on the specifications of the expected items. The measuring stage is intended to collect data as the basis for the determination and interpretation of the obtained test scores. Figure 1 below briefly summarizes the stages to develop a test.

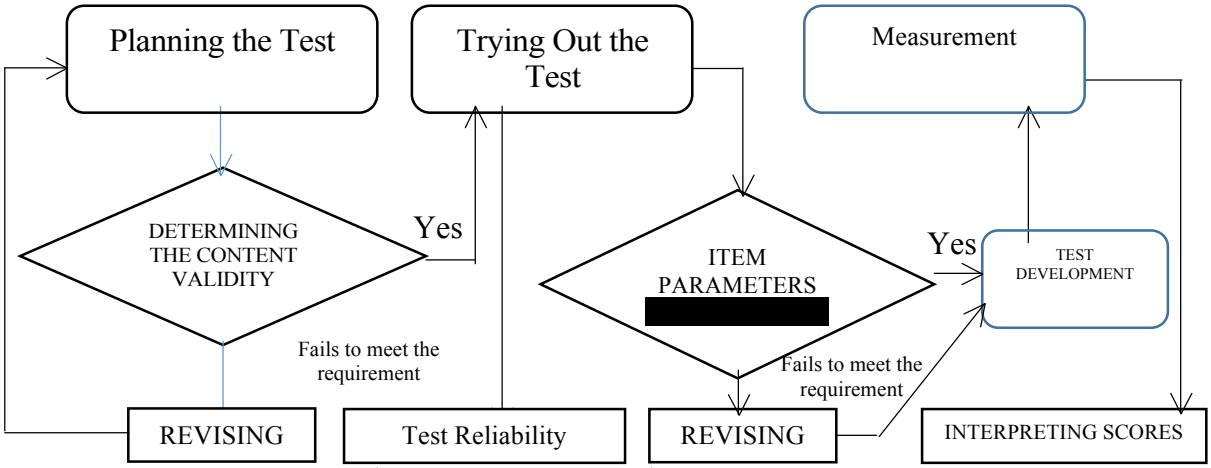

Figure 1. Stages to Develop a Test

Achievement test specifications should at least include the following: the area to be measured, the subject to be measured, the objective of the measuring activity, materials included in the measuring instrument, the type of question used, the total number of questions, and test blueprints. Test blueprint development aims to formulate as precisely as possible the scope and the emphasis of the test as well as its parts, so that the formulation can provide effective guidelines for test developers. In this test blueprint, indicators are formulated for the items that have been formulated in the area of measurement, the objective of the testing and materials to be tested. Table 1 below presents a sample test blueprint for achievement tests.

Table 1. A Sample Test Blueprint for Physics Achievement Tests on Heat

Test Objective : .....
Test Subject : .....

Teaching Materials	Sub-Materials	Indicator	Cognitive Domain Being Measured				Type of Question	Total	
			Factual (C1-C6)	Conceptual (C1-C6)	Procedural (C1-C6)	Metacognitive (C1-C6)		f	%
KALOR	A…..	A1……			1(C1)		Multiple choice		
		A2……	4 (C3)						
		A3……							
		etc.		8(C4)					
	B…..	B1……							
		B2……			12 (C2)				
		B3……							
		etc.							
	C…	C1……				30(C5)			

	…								
		C2……							
		C3……				40(C6)			
		etc.							
SUM								40	100

There are three main aspects to decide whether a physics achievement test is good or not, namely: its substance, construct, and language. Viewed from the substance aspect, it must represent the competencies to be assessed and from the construction aspect, it must meet the technical requirements in accordance with the type of instrument used. Viewed from the language aspect, the language it uses must be good, correct and communicative following the level of development of the students. A sample format to analyze test items is presented in Table 2.

Content validity actually does not have a qualitative value. The emphasis of the content validation approach is through expert or professional judgment. The term *professionals* here refers to experts in the domain being measured. If we develop a measuring instrument to assess achievement in physics, we may consult with physics teachers, physicists, experts in physics education and experts in psychometrics. Examination of a test's content validity relies on the accuracy in determining the test domain. Indicators/ items in the measuring instrument must consist of a representative sample of the indicators/ items of the domain to be measured. Although the statistical and psychometric coefficients of the correlation cannot be used to assess this content validity, several approaches have been proposed by experts to measure it, for example is the approach developed by Lawshe (1975: pp. 563-575) which proposes content validity ratios/ CVR as the statistics and V Aiken's (1985: pp.131-142).

In the approach developed by Lawshe, a panel of subject-matter experts are asked to indicate whether an item of measurement is "[3] essential, [2] useful but not essential, or [1] not necessary" as a form of the operationalization of the theoretical construct. The inputs provided by this panel are then used to calculate the CVR of each item in the measuring instrument. To measure the CVR, a number of experts (panel) are asked to review each item in the measuring instrument. There are three scoring alternatives, namely a particular item is "[3] essential, [2] useful but not essential, or [1] not necessary" compared with the domain being measured. This scoring is done on all items.

An item's CVR score ranges from 1 to -1. A high score indicates a high content validity. An item with a CVR = 0 means that half of the panel indicate that the item is relevant to the domain being measured. Thus, a positive value indicates that more than half of the panel indicate an item is good enough to be involved in the measuring instrument. Items with a very low CVR will not be used in the pilot test/ try out. Items with a low CVR value suggest that those items do not represent the domain to be measured. The content validation approach investigates the extent to which the items comprising a test represent the theoretical content this instrument is intended to assess.

The general verification of content validity is indeed still unable to define accurately the domain to be measured because there are quite a lot of human behavior samples to be used to comprise the items. The accuracy of the content validity can be achieved if the instrument development defines the domain to be measured well and the instrument items are correctly written. During the content validation process, experts use the definition of the domain to be measured we have arranged as a basis to assess the extent to which a measurement item represents the intended measured domain.

$$CVR = \frac{\left(N_e - \frac{N}{2}\right)}{(N-1)}$$

CVR = content validity ratio,
Ne = total SME panelists indicating "essential" (score = 3),
N = total number of SME panelists

Lawshe in his paper presents a table of CVR critical values to examine the significance of an item's CVR. Unfortunately, this table is less practical as to be indicated as satisfying with a 5% level of significance, an item should have a score of 0.78 which requires 8 panelists. If there are only three panelists, the minimum CVR score necessary is 0.99. The biggest problem in this case is that gathering a great number of panelists in order that the critical value required is not too high is not realistic. Therefore, according to Saifuddin Azwar (2014: p.115), a CVR should be interpreted relatively within the range of -1.0 to 1.0. All the items with a positive CVR are defined as having content validity, while those items with a negative CVR should be eliminated.

In addition to CVRs as an item's content validity statistics, the statistics of the Content Validity Index (CVI) can also be calculater, which indicates the content validity of a test. This CVI is the mean of the total CVRs of all the items. CVI Computation is performed only on selected items, i.e. items which have been declared to have a satisfactory CVR. Polit and Beck (2006: pp. 248-497) recommend that CVI reporting should also be coupled with the report on the range of selected items' CVR values. Please note that even for selected items based on the CVR criteria, it does not mean that it is not necessary for these items to be analyzed in terms of internal consistency, especially to improve the reliability of the measuring instrument.

$$CVI = \frac{\sum CVR}{k} \quad \text{where} \quad k = \text{the number of the items}$$

Table 2. A Sample Format to Analyze Test Items

No.	Aspect Analyzed	Item Number							
		1	2	3	4	5	6	...	N
A	Materials								
	1. Test items suit the indicators								
	2. The materials tested suit the competency								
	3. The answer key provided suit the questions/ statements								
	4. Ect.								
B	Construct								
	1. Using question word/ instructions which require answers								
	2. The instruction to do the test is clearly described.								
	3. The scoring method is provided.								
	4. Tables, figures, maps and the like in the question items are clearly stated and illustrated								
	5. Ect.								
C	Language								
	1. Items are arranged communicatively								

	2. Standardized language is used								
	3. The phrases used are not ambigous.								
	4. Etc.								
Expert Judgment [1] Not necessary [2] Useful but not essential [3] Essential		[1] [2] [3]	[1] [2] [3]	[1] [2] [3]	[1] [2] [3]	[1] [2] [3]	[1] [2] [3]	[1] [2] [3]	[1] [2] [3]

CONCLUSIONS AND SUGGESTIONS

A. Conclusions

1. Content validity actually does not have a quantitative value, therefore it does not belong to the category of empirical validity. It can be assessed using a coupleof methods, for example is the agreement of a panel of experts in assessing whether items we have developed are in conformity with the measuring construct or not.
2. Although the statistical and psychometric coefficients of the correlation cannot be used to assess this content validity, several approaches have been proposed by experts to measure it, for example is the approach developed by Lawshe which proposes content validity ratios/ CVR.
3. To measure the CVR, a number of experts (panel) are asked to review each item in the measuring instrument. There are three scoring alternatives, namely a particular item is essential, useful but not essential, or not necessary compared with the domain being measured. This scoring is done on all items.
4. An item's CVR score ranges from 1 to -1. A high score indicates a high content validity. A positive value indicates that more than half of the panel indicate an item is good enough to be involved in the measuring instrument.

B. Suggestions

1. Accuracy of the content validity can be achieved if at the stage of the instrument development, the domain to be measured is well defined and the instrument items are correctly written.
2. The tight recommendation given by Lawshe will require a large number of Subject Matter Experts (SME) to make the critical value required not too high.

REFERENCES

Aiken, L.R. (1985).Three coeffisients for analyzing the reliability and validity of Rating. Educational and Psicological Measurement, 45, 131-142.

Lawshe, C.H. (1975). A Quantitative approach to content validity. Personnel Phychology,28, 563-575.

Polit, D.F. & Beck,C.T. (2006). The content validity index. Are you sure you know what's being reported? Critique and recommendations. Resarch in nursing and helth, 29, 489-497.

Saiffudin Azwar. (2014). Reliability and validity.Yogyakarta. Pustaka Pelajar

Soekidjo Notoatmodjo. (2012). Metodologi penelitian kesehatan. Jakarta. PT. Rineka Cipta.

Sumadi Suryabrata. (2002). Pengembangan alat ukur psikologi. Yogyakarta. Andi Offset.

V. Wiratna Sujarweni. (2012). SPSS untuk paramedis. Yogyakarta.Gava Media.