Outbreak, Surveillance, Investigation & Response (OSIR) Journal
Field Epidemiology Training Program, Division of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625903894, Fax: +6625903845, Email: osireditor@osirjournal.net, http://www.osirjournal.net

The Grammar of Science: How "Good" is Your Instrument?

Jaranit Kaewkungwal

Mahidol University, Thailand

Corresponding author email: jaranitk@biophics.org

Received: 15 Feb 2023; Revised: 14 Mar 2023; Accepted: 17 Mar 2023
https://doi.org/10.59096/osir.v16i1.262097

In order to answer a research question. the researchers have to design or select the appropriate research instrument(s) or measurement tool(s) to obtain the data related to the objective(s) of the study. Research instruments are in different formats, techniques, or methods including, but not limited to, questionnaires, clinical interviews, laboratory tests, etc. When the researchers want to measure something especially the attributes that can't be observed directly (e.g., quality of life, attitude) they have to define observable or measurable indicators to measure those attributes. Such concrete indicators are considered as constructs (aka components, dimensions, or factors) of the attribute being measured. A questionnaire is defined as an instrument for the measurement of one or more concepts or constructs which are generally resulted in aggregated item scores/scales.[1] An interview could be completely structured similar to questionnaire or an open-ended statements, questions, or stimulus word to obtain responses given by the target study population.[1] Laboratory tests involve medical procedures that examine a sample of blood, urine, or other substance from the body; the test results will be used to answer the research questions in terms of determining diagnosis, checking to see if treatment is working, or monitoring the disease over time.[2]

What Determine "Good" Quality of a Research Instrument?

The ability of a research instrument to measure the constructs under study's objectives is a vital factor in selecting or applying the instrument. The qualities of a good research instrument are (1) validity and (2) reliability. Validity and reliability are concepts used to indicate how well a method, technique or test measures something.[1,3,4] Reliability is the extent to which the measurements is dependable, self-consistent, stable and/or reproduceable when repeating under the same conditions. Validity is the extent to which the results really measure what they are supposed to measure.[3]

At this time, let's focus on the quality of a structured or unstructured questionnaire/interview developed for use in a research study. For this kind of research instrument, the researchers should consider, at the minimum, attaining "content validity" and "internal consistency" of the measurement. Content validity means the extent to which the content or topic in the instrument is truly representative of all aspects of the concept(s) or construct(s) it is designed to measure.[5] Internal consistency reliability defines the consistency of the results delivered from the measurement tool, ensuring that the various items measuring the topic(s) of interest deliver consistent responses/scores.[6]

"Content Validity"—Do You Measure the Concept(s) that You Want to Measure?

In fact, there are three common types of validity including content, construct, and criterion-related validity. Content validity is considered as a prerequisite for other validity and usually receives the highest priority in development process of a questionnaire/interview forms.[7]

Content validity is also known as definition validity or logical validity.[7] Content validity tells us how "good" the instrument is at measuring a concept. A similar idea of content validity is "Face validity" which also employs similar approach by assigning a panel of experts to use their theoretical and substantive knowledge and respond to each item whether or not it is or is not a good item.[8] However, face validity indicates that the measure appears to be valid "on its face" while content validity requires a more rigorous process, such as using a panel of experts to evaluate the logical content within the concept to be measured.[9] Steps in evaluating content validity are as follow:[7,10–12]

Step 1: Determine the Content—Identification of Concept/Construct and Item Generation

The first step is determining the content of the concept(s) or construct(s) that the instrument intends to measure. The content can be identified by literature review and/or focus group discussion or interview with the people familiar with concept. To make it manageable, a "Table of Specifications" of the content for each concept should be created. The researchers then develop items based on the table of specification. When measuring data related to behavior, feeling, or certain action that cannot be captured in a single variable or item, it is recommended to develop multiple items in consideration of measurement error and gaining more accurate responses.[12] A sample of table of specifications adopted from part of the WHOQOL instrument (measuring quality of life) composes of two constructs: physical health and psychological health.[13] Two items are developed for the three topics (content) of each construct (Table 1). In evaluating the content validity of this instrument, the researchers may assess the content validity of each construct separately and/or the entire instrument that intends to measure quality of life.

Table 1. Example of table of specification for quality of life instrument

Concepts/Constructs	Items
Physical Health	
I. Pain and discomfort	1. Do you worry about your pain or discomfort?
	2. How difficult is it for you to handle any pain or discomfort?
II. Energy and fatigue	3. How easily do you get tired?
	4. How much are you bothered by fatigue?
III. Sleep and rest	5. Do you have any difficulties with sleeping?
	6. How much do any sleep problems worry you?
Psychological Health	
I. Positive feelings	1. How much do you enjoy life?
	2. How much do you experience positive feelings in your life?
II. Self-esteem	3. How much do you value yourself?
	4. How much confidence do you have in yourself?
III. Bodily image and appearance	5. Do you feel inhibited by your looks?
	6. Is there any part of your appearance which makes you feel uncomfortable?

Note: Examples of concepts and items adopted from WHOQOL-100 Instrument.

Step 2: Judgment by Expert Panel—Appointment and Consensus of the Experts' Opinions

To confirm the validity, a specific number of experts will be appointed to evaluate the items in the instrument developed. The panel may include lay experts who are the potential research subjects. The number of experts to be appointed is arbitrary; at least five people are generally recommended to avoid agreement by chance. Should the number of experts increase, the probability of chance agreement decreases.

The researchers may conduct a "Delphi" method with panel of experts by arranging series of iterative questionnaires or meetings.[14,15] The consensus among the experts on the items in the instrument can be obtained after intensive discussion in series of meetings or it can be based on the anonymous responses to the iterative questionnaires. The process of multiple iterations collecting both qualitative and quantitative data helps reduce the range of responses and to reach consensus based on criteria chosen a priori by the researcher.

The expert panel can provide their quantitative and qualitative viewpoints on the relevancy or representativeness, clarity and comprehensiveness of the items to measure the specified construct. The content experts may also recommend about language and scores for the items in the instrument.

Step 3: Analysis of Content Validity—Assessing Quality of the Items in the Instrument

To confirm the content validity, several statistics have been proposed and used in literature. The most common ones are based on interrater agreement: content validity ratio (CVR) and content validity index (CVI).[7,10,11,16,17]

In evaluating CVR. the experts are requested to specify whether an item is "important" or "necessary" or "essential" for the concept desired to measure or not. The experts may be asked to score each item from 1 to 3 with the degree as shown in Table 2. The formula for $CVR = (N_e – N_t/2) / (N_t/2)$, in which the N_e is the number of experts indicating "essential" and N_t is total number of expert panel. CVR is easy to compute, but not so easy to interpret; its values can range from –1.0 to +1.0,

with CVR=0 when half the experts judge an item to be relevant.[16] The decision to eliminate or keep the item is usually based on the acceptable level of significance proposed by Lawshe.[17] As shown in Table 3, with 5 experts in the panel and 6 items in the measurement tool, the numbers of N_e vary among all items. Only item#1 passes the minimum value required under Lawshe's suggestion (i.e., using 5 experts requires CVR=0.99). The not-passed items should be eliminated or may require major revision.

Table 2. Proposed scoring for each item by expert panel

Score	Relevancy	Clarity	Essential
1	Not relevant	Not clear	Not necessary
2	Must be revised	Must be revised	Useful, but not essential
3	Relevant with minor revision	Clear with minor revision	Essential
4	Very relevant	Very clear	

Table 3. Content validity ratio (CVR) based on number of experts evaluated the item essential

Item	N_e-rate "Essential"	N_t-Total	CVR	Decision
1	5	5	1	Keep
2	4	5	0.6	Eliminate
3	3	5	0.2	Eliminate
4	2	5	-0.2	Eliminate
5	1	5	-0.6	Eliminate
6	0	5	-1	Eliminate

Note: Content validity ratio (CVR) = (N_e-N_t/2) / (N_t/2), N_e: the number of experts indicating "essential", N_t: total number of expert panel
Decision on CVR based on Lawshe's Table for minimum values of CVR
Number of panelists and minimum acceptable CVR value: 5,6,7=0.99; 8=0.75; 9=0.78; 10=0.62

CVI is another commonly used in assessing content validity. It is item-level information that can also be used to make decision whether to keep, revise or eliminate the items. In evaluating CVI, the experts are asked to rate instrument items in terms of relevancy and clarity in accordance with the concept the researchers want to measure. The rating scores may range between 1 to 4 as shown in Table 2.

The simplest form of item-level content validity index (I-CVI) is based on the criteria that the researchers simply set to make decision whether to keep, revise or eliminate the items. In a typical approach the researchers would appoint odd number of members for an expert panel. The researchers then set the criteria such that, if 2 of 3, 4 of 5, or 7 of 9 experts agreed on the relevancy of an item, they will keep that item. Another way to calculate I-CVI is based on the formula: I-CVI = N_{re} /N_t in which N_{re} is the number of experts rating at level 3 or 4 to the relevancy and N_t is the total number of experts in the panel. Thus, I-CVI reflects the proportion of agreement ranging between 0 and 1 as shown in Table 4.

The CVI of all items can be summarized as the scale-level content validity index (S-CVI). It can be calculated as an average of the I-CVIs for all items on the scale, so-called scale-level content validity index, averaging calculation method (S-CVI/Ave).

The strict agreement rate can be calculated based on absolute agreement on relevancy among experts, so-called scale-level content validity index, universal agreement calculation method (S-CVI/UA) as shown in Table 4.

Table 4. Item content validity index based on number of items considered relevant by 5 panelists

Item	Relevant (rating 3/4)	Not relevant (rating 1/2)	I-CVI	Decision
1	5	0	1	Excellent
2	4	1	0.8	Appropriate
3	3	2	0.6	Eliminate
4	2	3	0.4	Eliminate
5	1	4	0.2	Eliminate
6	0	5	0	Eliminate

Note: *Scale-level content validity index, averaging calculation method (S-CVI/Ave) = 3/6 = 0.50*
Scale-level content validity index, universal agreement calculation method (S-CVI/UA) = 1/6 = 0.17
Decision on I-CVI: >78% "Appropriate"; 70–79% "Need revision"; <70% "Eliminate"

Step 4: Pre-testing the Instrument—Ensuring the Questions and Answers are Comprehensible.

As recommended in literature, the draft of the instrument should be administered to 5–15 people representing targeted research participants in 2–3 rounds.[12] The responses can be in the format of verbalizing the mental process of what they understand while providing responses to the items in the instrument. This step may be repeated with revised version of the instrument should the items are still unclear or misunderstood by the respondents.

"Internal Consistency"—Is Your Research Instrument Reliable?

Quality of research instrument should be assessed not only for its validity but also its reliability. One of the most common and simplest statistics for reliability of data collected via questionnaire/interview forms is Cronbach's Alpha. Cronbach's alpha reflects internal consistency of a set of scores/scales of the items. It refers to the extent to which the instrument is a consistent measure of a concept.[4,8] The researchers should calculate reliability of the research instrument before they actually use that tool. However, researchers can also calculate it after they get the data for their research project to confirm the quality of their tool.

In order to have the reliable and reproducible evidence showing the consistency or stability of an instrument, the researchers must try out the instrument with a sufficient sample size. It is typically recommended to test the instrument with 30 subjects. However, for a single coefficient alpha test, assuming the null hypothesis of Cronbach's alpha coefficient=0, the sample size may be less than 30 to achieve a minimum desired effect size of 0.7 but if the null hypothesis set the coefficient larger than zero, a larger sample size is needed.[18]

Cronbach's alpha is computed by comparing the variances for all individual item scores with the variance of the total score: The formula is $\alpha = (k/(k-1))(1-\sum\sigma_{yi}^2/\sigma_x^2)$ in which k refers to the number of items, σ_{yi}^2 is the variance of item i, and σ_x^2 is the variance of the observed total scores (Table 5). Another way to calculate Cronbach's alpha is based on correlations among the items. The formula based on correlations for standardized $\alpha = k\, r_{avg} / (1 + ((k-1)\, r_{avg}))$ in which k refers to the number of items and r_{avg} is the average of all correlations among all pairs of items (Table 6).[19,20]

When the item is measured as binary (yes-no) rather than scores/scales (e.g., Likert's scale or others), Cronbach's alpha formula will be adjusted to the so-called Kuder-Richardson 20 formula. The use and interpretation of the Kuder-Richardson 20 formula is similar to the Cronbach's alpha.

Table 5. Cronbach's alpha based on item variance and total variance

Item	1	2	3	4	5	6	Total
Subject	y	y	y	y	y	y	x
1	y	y	y	y	y	y	x
2	y	y	y	y	y	y	x
3	y	y	y	y	y	y	x
⋮	y	y	y	y	y	y	x
n	y	y	y	y	y	y	x
	σ_{y1}^2	σ_{y2}^2	σ_{y3}^2	σ_{y4}^2	σ_{y5}^2	σ_{y6}^2	σ_x^2

Table 6. Cronbach's alpha based on interitem correlations

Item	1	2	3	4	5	6
1	1	r_{12}	r_{13}	r_{14}	r_{15}	r_{16}
2		1	r_{23}	r_{24}	r_{25}	r_{26}
3			1	r_{34}	r_{35}	r_{36}
4				1	r_{45}	r_{46}
5					1	r_{56}
6						1

Note: r_{avg} = average of r's

Coefficient of reliability ranges from 0 to 1. As a rule of thumb, the meaningful level of reliability is typically set as an acceptable threshold; for Cronbach's alpha reliability it is usually set at the minimum value of 0.7 or preferable 0.8.[19]

It is important to note that the value of alpha is dependent to the number of items; with high number of items, the alpha coefficient could be somewhat high even when the average interitem correlation is low. For example, the alpha coefficient is 0.71 when the average

interitem correlation of 10 items is 0.2.[20] It is thus recommended that a minimum alpha coefficient should be set at higher level when making important decisions from the use of a particular instrument; some suggest to use reliability at the level of 0.9 or above.[8,21] It should also be noted that Cronbach's alpha is not a measure of unidimensionality. The instrument may have high alpha even when there are multiple underlying dimensions or constructs. Test of constructs is a part of "construct validity" which can be analyzed by other kind of statistics, i.e., factor analysis.

Conclusion

So, you now know how to confirm the quality of your research instrument regarding the content validity and internal consistency. But…are you sure that your instrument has captured all essential theoretical constructs, or it will be reproducible when the entire research process is conducted again? That will need further steps.

Suggested Citation

Kaewkungwal J. The grammar of science: how "good" is your instrument? OSIR. 2023 Mar;16(1):40–5. doi:10.59096/osir.v16i1.262097

References

1. Oosterveld P, Vorst HCM, Smits N. Methods for questionnaire design: a taxonomy linking procedures to test goals. Qual Life Res. 2019 Sep;28(9):2501–12.
2. National Cancer Institute. NCI's dictionary of cancer terms: laboratory study [Internet]. Bethesda (MA): National Institutes of Health, U.S. Department of Health and Human Services; [cited 2023 Feb 15]. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/laboratory-test>
3. Middleton F. Reliability vs. validity in research: difference, types and examples. Amsterdam: Scribbr; 2019 Jul 3 [updated 2023 Jan 30, cited cited 2023 Feb 15]. <https://www.scribbr.com/methodology/reliability-vs-validity>
4. Tang W, Cui Y, Babenko O. Internal consistency: do we really know what it is and how to assess it? Journal of Psychology and Behavioral Science. 2014 Jun;2(2):205–20.
5. Frost J. Content validity: definition, examples & measuring [Internet]. [place unknown]: Jim Frost; c2023 [cited 2023 Feb 15]. <https://statisticsbyjim.com/basics/content-validity/>
6. Shuttleworth M. Internal consistency reliability [Internet]. [place unknown]: Explorable.com; 2009 Apr 26 [cited 2023 Feb 15]. <https://explorable.com/internal-consistency-reliability>
7. Zamanzadeh V, Ghahramanian A, Rassouli M, Abbaszadeh A, Alavi-Majd H, Nikanfar AR. Design and implementation content validity study: development of an instrument for measuring patient-centered communication. J Caring Sci. 2015 Jun 1;4(2):165–178. doi:10.15171/jcs.2015.017.
8. Goforth C. Using and interpreting Cronbach's alpha [Internet]. Charlottesville (VA): Statistical Consulting Associate, University of Virginia Library; 2015 Nov 16 [cited 2023 Feb 15]. <https://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/>
9. Rubio DM, Berg-Weger M, Tebb SS, Lee ES, Rauch S. Objectifying content validity: Conducting a content validity study in social work research. Social Work Research. 2003 June:27(2);94–104. doi:10.1093/swr/27.2.94.
10. Zeraati M, Alavi NM. Designing and validity evaluation of quality of nursing care scale in intensive care units. J Nurs Meas. 2014;22(3): 461–71. doi:10.1891/1061-3749.22.3.461.
11. Yusoff MSB. ABC of content validation and content validity index calculation. Education in Medicine Journal. 2019;11(2):49–54. <https://doi.org/10.21315/eimj2019.11.2.6>
12. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quinonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front. Public Health. 2018 Jun 11; 6:149. doi:10.3389/fpubh.2018.00149.
13. Division of Mental Health and Prevention of Substance Abuse, World Health Organization. Programme on mental health: WHOQOL user manual [Internet]. Geneva: World Health Organization; 1998 [cited 2023 Feb 15]. 106 p. <https://apps.who.int/iris/handle/10665/77932>
14. Koller I, Levenson MR, Gluck J. What do you think you are measuring? a mixed-methods procedure for assessing the content validity of test items and theory-based scaling. Front. Psychol. 2017 Feb 21; 8:126. doi:10.3389/fpsyg.2017.00126.

15. Barrett D, Heale R. What are Delphi studies? Evidence-Based Nursing. 2020;23(3):68–9. doi:10.1136/ebnurs-2020-103303.

16. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? appraisal and recommendations. Res Nurs Health. 2007 Aug;30(4):459–67. doi:10.1002/nur.20199.

17. Lawshe CH. A quantitative approach to content validity. Personnel Psychology. 1975; 28(4):563–75. doi:10.1111/j.1744-6570. 1975.tb 01393.x.

18. Bujang MA, Omar ED, Baharum NA. A review on sample size determination for Cronbach's alpha test: a simple guide for researchers. Malays J Med Sci. 2018;25(6):85–99. doi:10.21315/mjms2018.25.6.9.

19. Nunnally JC. Psychometric theory. 2nd ed. New York: McGraw-Hill; 1967.

20. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334.

21. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.