TPM®

CONTENT VALIDITY: DEFINITION AND PROCEDURE OF CONTENT VALIDATION IN PSYCHOLOGICAL RESEARCH

ADIYO ROEBIANTO
UNIVERSITY OF AL-AZHAR, INDONESIA

SETIAWATI INTAN SAVITRI
IRFAN AULIA
ARIE SUCIYANA
LAILATUL MUBAROKAH
UNIVERSITY OF MERCU BUANA, JAKARTA, INDONESIA

Developing research designs and instrumentation in psychological research is essential because the constructs and variables in the discipline are broad and need to be measured by specific instruments. For each instrument developed or adapted, validation such as content validation needs to be conducted. The content validation process includes a readability test determining whether the items or questions effectively represent the variables or constructs measured. This study utilized the Conjoint Community Resiliency Assessment Measure (CCRAM) which consists of 21 items and employed nine experts in psychology to provide expert judgments. Some content validity measurement methods, such as interrater reliability (IRR), Aiken's validity, content validity ratio (CVR), and content validity index (CVI), were also used. The results from all measurements of content validity indicate consistency in CCRAM instrument items. The strengths and weaknesses of each content validity measurement method are also highlighted.

Keywords: Psychological research; Reliability; Validity; Content validity; CCRAM instruments.

Correspondence concerning this article should be addressed to Adiyo Roebianto, Department of Psychology and Education, University of Al-Azhar, jl. Sisingamangaraja, 12110 South Jakarta, Indonesia. Email: adiyo.roebianto@uai.ac.id

Psychology is a branch of social science studying psychological attributes. Unlike attributes in natural science such as length, weight, and height, the attributes in psychology cannot be observed directly. This unobservable attribute is commonly referred to as a construct (Thorndike & Thorndike-Christ, 2014). Some of the constructs studied in psychology, such as community resilience, are unobservable.

Community resilience is one of the constructs that has recently been developed. The best way to hypothesize new constructs is for researchers to use structured methods to test these constructs. According to Hernandez (1991), scientific research requires a sequence of stages that must be followed and should not be changed, because it would affect its reliability and validity. Hence, testing a new construct requires a reliable and valid instrument (Rubio et al., 2003). Valid and reliable instruments will help researchers to interpret new constructs or variables (Bagozzi, 1980). Therefore, a valid instrument is an important aspect in instrument development.

Validity is a central factor in choosing an instrument for a research study. The items on the instrument need to be checked as to whether the content has been measured in accordance with the attributes to be evaluated. According to the Thorndike and Thorndike-Christ (2014), reliability is defined as the consistency of the results of repeated tests, while validity is defined as the relevancy between evidence and theories with

the score interpretation and purpose of the test. According to the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (2014), validity evidence is divided into content validity, construct validity, and criterion validity. Content validity is one of the conditions for the other validities. So it is critical during instrument development and/or instrument adaptation (Aravamudhan & Krishnaveni, 2015).

This study aims to explain the content validity concept and process in detail. Having proper content validity is the essential initial step to developing measurement tools (AERA, APA, & NCME, 2014; Rubio et al., 2003; Thorndike & Thorndike-Christ, 2014).

Content Validity

Content validity can be defined as how representative the items or tests are to measure the behavior studied (Cohen & Swerdlik, 2018; Slaney, 2017). Thorndike and Thorndike-Christ (2014) further explained content validity as the extent the test measures construct and the relevancy of the test to the aspects measured. Similarly, AERA, APA, and NCME (2014) define content validity as the correlation between the content of the test and the construct measured. In general, the content of the tests includes wording, format, and display of items. A test is considered content valid when the items relevantly measure the construct. Furthermore, attribute measurement should not use an excessive number of items; thus, representative items to measure constructs are needed.

Rubio et al. (2003) stated that researchers conducting content validation should receive some constructive feedback for developing measurement tools. That constructive feedback can be provided by panelists analyzing and evaluating the quality of measurement tools and objective criteria of the items. The feedback is the basis for revising tests or items, which upon finalization will be used for a pilot study. Thus, the test should have reasonably good psychometric properties, before being used in larger samples. The stages of content validation are explained below.

Stages of Content Validity Test

Test Blueprint/Questionnaire

Initially, to design a test, researchers should draw specification tables that contain explanations about what and how to measure, which in psychometric science terminology is commonly called a test blueprint or table of specifications (Cohen & Swerdlik, 2018; Thorndike & Thorndike-Christ, 2014). The term test blueprint, borrowed from the architectural world, is defined as action planning (Patil et al., 2015). It is used as the basis for the development of instruments/questionnaires/tests, especially those created from scratch. In scientific articles (thesis and dissertation), the test blueprint can be considered a conceptual and operational definition of the variables with existing measurement tools or tests. The inclusion of the test blueprint in writing can help readers to understand definitions and measurements of variables (Lawshe, 1975).

Thorndike and Thorndike-Christ (2014) stated that the test blueprint mainly consists of the specification of cognitive processes measured and the description of content measured by the test. They further explained that it should include methods and procedures of domain measurement that will be cross-checked by other test developers to build the framework of test development. Similarly, Cohen and Swerdlik (2018) claimed that the test blueprint should consist of 1) the definition and information of the construct measured by

the items; 2) the number of items for each aspect or dimension; and 3) the item arrangement in the test. In other words, test blueprints should draw the general picture of the contents measured by the test developed.

This study includes the following points for the test blueprint:

1. The construct measured, because a test blueprint should include the explanation of psychological constructs as well as a brief definition of the constructs.

2. Dimensions and indicators of the construct, because psychological constructs generally have aspects and dimensions with certain indicators or criteria.

3. Items measuring dimensions and aspects, because the explanation of items should include format, wording, number, favorability (based on the definition of aspects and dimensions), and unfavorability.

The three elements above should be included in the test blueprint because they are the core representations of what is measured and how the measurements should be conducted.

Expert Judgement

According to Rubio et al. (2003), experts are divided into content or domain experts and evaluator participants representing the population samples (lay experts). The authors further remarked that domain experts are professionals conducting research or publishing scientific articles on related studies. The selection of experts in the measurement field or in the topic researched should be accurate because the experts are expected to provide constructive feedback for the test design, especially in terms of fulfillment of criteria and relevancy of psychometric properties. Rubio et al. (2003) defined experts as professionals having a number of publications and experiences in related fields. Syaiful and Roebianto (2020) added that experts are individuals actively involved in the development of the disciplines.

Lay experts (participant judgments) are parts of a sample of the target population required to give perspectives on the items or tests developed. Participants' judgments on items or tests developed can determine the representativeness of the samples for the population included in the research. In this stage, participants are asked about the extent to which they understand the wording and terms of the items. Their evaluation is considered for test or item revision.

Scholars have differing opinions on the proportion of experts required in the content validation process. Lynn (1986) suggested that the minimum number of members in a panel of experts in content validation is three. Similarly, Rubio et al. (2003) mentioned that a content validation process should include at least three panels to represent content and lay experts. While other authors (see, for example, Gable & Wolf, 1993; Rubio et al., 2003; Walz et al., 1991) stated that the number can vary from two to 20. Grant and Davis (1997) believed that the number of experts should be based on the expertise needed and the scope of the research.

On the other hand, Almanasreh et al. (2018) argued that the specific number should not be determined rigidly although many studies involve 10 experts in the content validation process. Their argument is based on the difficulty of reaching a unanimous decision when the panel consists of too many experts. Nevertheless, when experts are too few, the information gathered for instrument development will be limited. Hence, the number of experts suggested is around five to 10.

A standard number of panel experts is required for the development of entirely new tests, diagnostic tests, and tests with various levels and broad coverage (high-stake testing; AERA, APA, & NCME, 2014). For instance, the comparison of achievement test results of students worldwide requires comprehensive expert judgments to evaluate content validity. The selection of the panel experts in this context should be very careful and representative. On the contrary, for tests with a smaller scope (low-stake testing) and limited use, such as a

classroom assessment or any assessment whose results do not have a significant impact, the number of panels of experts might not be as high. The test items developed for undergraduate theses (bachelor's degrees) are an example of a test that might not need a high number of panel experts.

To determine the number of experts in the panel involved in studies, researchers first should know the availability of instruments in psychological research that are used to measure a variable before asking for expert judgments. The content validation procedure is generally used as the initial step to develop instruments with limited research support (Raykov & Marcoulides, 2010; Thorndike & Thorndike-Christ, 2014). In addition, some instruments already have validity and reliability tests, which can be cited in the study as the basis for the instrument development, from the previous research. On the other hand, an expert's judgment process can be long given that it includes a back-to-back process (review-revision). The detailed and careful selection process of the panel can also be time-consuming when researchers have a pool of options. Therefore, in research such as an undergraduate thesis that is low-stakes, one expert might be sufficient to evaluate the test items developed, as long as the selection process follows the golden standard (Rubio et al., 2003), namely selecting professional experts based on their scientific publication and relevant professional experience (especially for industrial and organizational settings).

Questionnaire Administration

Questionnaires presented to experts for evaluation mainly should contain an introduction, research scope (such as description and questionnaire instructions), format response, and space for notes/comments from experts (Rubio et al., 2003). The introduction should cover the aim of the research, criteria of expert judgment for content validation, instrument and scoring descriptions, and items. These elements, especially the instrumentation, are essential because they significantly influence the judgment given by experts. The clearer the description and analysis process, the more detailed the expert judgment on the items will be. However, for the lay expert judgment from the samples, Rubio et al. (2003) suggested that the wording be as simple as possible to ensure participant comprehension.

Practically, Almanasreh et al. (2018) observed that the administrative information needed to conduct expert judgment is (1) a cover letter explaining the purpose of the research, detailed description of the instruments, and the analysis process, (2) content validation assessment sheets, and (3) a copy of instrument development forms. The availability of these documents can help experts to understand and evaluate the instrument thoroughly.

In terms of item evaluation, Rubio et al. (2003) suggested that there are four criteria of content validity, namely 1) representativeness, 2) clarity, 3) factor structure, and 4) comprehensiveness. Representativeness refers to whether the items represent the content domain described in the conceptual definition, while clarity relates to how clear the sentences and wording of the items are. Each item has four scale anchors from 1 to 4 for representativeness and clarity aspects. Items with a value of 1 are not representative and clear at all, while items with a value of 4 are very representative and clear in terms of wording.

Furthermore, Yaghmaie (2003) suggested that the criteria for content validity are relevancy, clarity, simplicity, and ambiguity. Overall, the relevancy and clarity aspects in Yaghmaie (2003) are similar to the representativeness and clarity aspects that Rubio et al. (2003) proposed. However, Yaghmaie included the simplicity aspect which relates to how straightforward and easy to understand the wording is, and the ambiguity aspect which describes how the item may have ambiguous meanings. This study

includes comprehensiveness/relevancy and clarity aspects for the analysis process because experts tend to agree these criteria are the most important ones.

Different criteria were proposed by Sireci (1998; Sireci & Faulkner-Bond, 2014), categorizing the important elements in content validity into domain definition, domain representation, domain relevancy, and feasibility test of the test development procedure. Domain definition refers to how a construct can be measured by the test developed. Domain representation explains the level of representativeness and measurement level of a test to measure the defined domain. Domain relevancy describes how relevant each item is to the measured domain. Finally, the feasibility test of the test development procedure refers to the evaluation of the entire processes conducted to develop the tests, to ensure that each item is proper and representative of the construct measured, not of other unrelated constructs.

Rubio et al. (2003) suggested that, in terms of factor structure, experts are required to categorize items into a certain factor. Factors can be defined as dimensions and aspects; thus, test developers need to define each factor/dimension included in the test development. However, this stage can be omitted if only one factor/dimension is included. Another important stage is to categorize items based on factors/dimensions prior to the expert judgment on the accuracy of the categorization. A correction of which dimension/factor is measured by the items will be provided when the categorization is considered inaccurate.

The final stage of item evaluation is an analysis of comprehensiveness, or whether the items have represented the constructs measured. This process will determine whether items should be added or removed from the test instrument. The example of expert judgment adapted from Rubio et al. (2003), Yaghmaie (2003), Zamanzadeh et al. (2015), is shown in Table 1.

TABLE 1
Expert judgment format

Instruction — The questionnaire aims to evaluate the content validity of the items developed. Please provide an analysis based on the following descriptions:

- Assess the relevancy level of each item with a 1-4 scale, where 1 indicates that the item is *not at all relevant*, and 4 indicates that the item is *very relevant*. The space for comments on revision (if necessary) is provided in the sheet.
- The analysis of clarity level follows the same 1-4 scale procedure.
- Decide the categorization of each item based on the factor. Definition and description of each factor have been provided. If items do not belong to the factors described, a separate note explaining which factors are measured by the items can be provided.
- Lastly, assess the comprehensiveness of all items and determine whether items should be revised or removed.

Thank you for your participation.

Theoretical definition	*Relevancy*	*Clarity*	*Factor*
Explaining the construct measured by the questionnaire, the conceptual definition, and the operational definition of the constructs.	1. The item is not relevant. 2. The item needs major revision. 3. The item needs minor revision. 4. The item is relevant.	1. The message of the item is not clear. 2. The item needs major revision. 3. The item needs minor revision. 4. The message of the item is clear.	Provide lists and definitions of the factors. 1 = factor 2 = factor 3 = factor 4 = other (write the factor)
Item	Relevancy score	Clarity score	Factor
Item 1 …	1 / 2 / 3 / 4	1 / 2 / 3 / 4	1
Item 2 …	1 / 2 / 3 / 4	1 / 2 / 3 / 4	2
Item 3 …	1 / 2 / 3 / 4	1 / 2 / 3 / 4	3

DATA ANALYSIS

Aiken's Validity

According to Aiken (1985) the procedure for determining whether the items are valid or not begins with the judgment (ratings) of an item by *n* raters (judges). The judgments are based on how accurately the items represent the constructs being measured. The maximum number of raters is 25 and the minimum is two; whereas the maximum number of rating categories is seven and the minimum is two (Aiken, 1985). The formula of Aiken's validity is as follows:

$$V = \frac{\Sigma r - lo}{n(c-1)} \quad (1)$$

In the formula above, V is Aiken's validity; ∑ is the sum of *r*-lo (*r* = raters' judgment of the items; lo = the lowest category score); *n* is the number of raters; *c* is the number of categories of the items. The range of V coefficients is 0 to 1, a high value indicating that an item has high content validity, or a set of items has high content validity in the judgment of a single rater (Aiken, 1985).

Kappa Interrater Reliability (IRR)

Kappa interrater reliability was initially proposed by Jacob Cohen in 1960 (McHugh, 2012). The Kappa coefficient (κ) is a robust statistic useful for either interrater or intrarater reliability testing. The formula of Kappa IRR is:

$$\kappa = \frac{Pr(\alpha) - Pr(e)}{1 - Pr(e)} \quad (2)$$

where P*r*(α) represents the actual observed agreement and P*r*(e) represents chance agreement. Just like in correlation, the Kappa coefficient can range from –1 to 1, where 0 represents the amount of agreement that can be expected from random chance, and 1 represents perfect agreement between the raters. The Kappa coefficient in standardized scores is thus interpreted in the same way. Cohen (1960) suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating *no agreement*, 0.01-0.20 as *none to slight*, 0.21-0.40 as *fair*, 0.41-0.60 as *moderate*, 0.61-0.80 as *substantial*, 0.81-1.00 as *almost perfect agreement*.

Content Validity Ratio (CVR)

The CVR method requires experts to determine whether an item is needed to measure the construct measured by instruments (Ayre & Scally, 2014; Lawshe, 1975; Zamanzadeh et al., 2015). A panel of experts is asked to give a score of 1 to 3 on a scale, where the value of 1 means the item is *not necessary*, 2 means the item is *useful but not essential*, and 3 means the item is *essential*. The value of CVR varies from –1 to 1. The closer the value to 1, the higher the agreement among experts is, and the more the items should be included in the instrument, while the closer the value to –1, the lower the agreement among experts; thus, the items need to be removed or revised. The formula to determine CVR is as follows:

$$CVR = (N_e - N/2) / (N/2) \quad (3)$$

Coefficient N_e is the number of experts or panelists giving a score of 3 or *essential*, while N is the number of experts or panelists. Lawshe (1975) suggested that the minimum criterion for CVR value to be considered acceptable is based on the number of experts. For example, if the number of experts is 10, the minimum number of CVR considered acceptable is .62.

Content Validity Index (CVI)

McCoach et al. (2003) described the CVI as a method used to summarize item relevancy score from a panel of experts. In general, the CVI method is similar to the CVR, except in the analyzed aspects and calculation formula. To calculate content validity, CVI calculation includes relevancy and clarity, which then are scored with a 1-4 scale, as explained in Table 1.

To obtain the CVI in each item (I-CVI), the number of experts/panelists giving a score of 3 or 4 to each item is calculated and divided by the number of experts involved in the assessment (Rubio et al., 2003; Zamanzadeh et al., 2015). For example, if five of seven experts give a score of 3 or 4 to an item, the I-CVI of the item is 5:7 = .714 (I-CVI = .714).

According to Zamanzadeh et al. (2015), items with I-CVI below .7 (I-CVI < .70) should be removed, while items with I-CVI between .70 and .90 (.70 ≤ I-CVI ≤ .90) should be revised, and items with I-CVI above .90 (I-CVI > .90) should remain.

The CVI in instrumentation level or Scale CVI/S-CVI can be measured by two approaches. The first approach requires universal agreement (UA) among experts (Zamanzadeh et al., 2015) where test developers calculate the number of items that scored 3 or 4, and the total score is divided by the total number of items. For example, if three of nine items are scored 3 or 4 by experts, the score of S-CVI/UA is 3:9 = .333 (SCVI/UA = .333). The second approach is relatively more straightforward because it only requires experts to calculate the average score of I-CVI in the instrument (I-CVI) (Lawshe, 1975; Rubio et al., 2003; Zamanzadeh et al., 2015). According to Davis (1992), the newly developed instruments should have a CVI minimum value of .80 out of 1.00.

METHODS

Instruments

This study adapts the Conjoint Community Resiliency Assessment Measure (CCRAM) developed by Leykin et al. (2013). The instrument consists of 21 items and five responses with a Likert scale of 1-5. A score of 1 indicates *strongly disagree*, while a score of 5 indicates *strongly agree*. The CCRAM instrument has five dimensions (Cohen et al., 2013; Leykin et al., 2013):

1. *Leadership* represents the capability of leaders to direct communities or groups when facing crisis or disruption. This dimension contains six items and is labeled with "L." The indicators in this aspect are:
    a. Trust in leaders or decision-makers
    b. Conviction in leadership's perspectives toward justice and proper service
    c. Function in community.

2. *Collective efficacy* represents the social cohesion among neighbors combined with the willingness to cooperate to achieve shared objectives. This dimension contains five items and is labeled with "E." The indicators in this aspect are:

a. Collective efficacy
b. Supports in community
c. Involvement to cooperate.

3. *Readiness* represents the manifestation of social learning through a feedback process to build resilience. This dimension contains four items and is labeled with "R." The indicators in this aspect are:

a. Relatives and acquaintances in the community for emergency conditions
b. Perspectives toward community readiness for emergencies.

4. *Attachment to place* represents a phenomenon that combines various aspects related to the relationship between people and place. This aspect also includes the influence of emotion, knowledge, belief, and attitude related to a place. This dimension contains four items and is labeled with "A." The indicators are:

a. Emotional attachment to the community
b. Sense of belonging
c. Pride in the community
d. Ideological identification with the community.

5. *Social trust* represents a trust that people are dependable and available to act according to the policy. This dimension contains two items and is labeled with "B." The indicators are:

a. Trust in relationships among community members
b. Quality relationships among community members.

The item content validity index or I-CVI of each item will be calculated and the calculation of the scale content validity index or S-CVI will be conducted for the test.

Adapting the Instruments

According to Beaton et al. (2000), the instrument adaptation process includes the following stages.

Stage 1 — Instrument translation. The first stage involves the translation of instruments from English to Bahasa Indonesia and back translation into English. The translation process was conducted by four translators having a background in psychology and expertise in English. After the translation of the 21 items was conducted, the proofreading was done by several participants. The proofreading process was conducted to check participants' understanding of the items. Once the participants' comprehension of the wording was ensured, the experts' judgment followed. In other words, the first stage employs a qualitative approach to analyze the instrument.

Stage 2 — Synthesis. In the second stage, the experts and/or translators sit together to synthesize the results of translations. The translation results from experts and/or translators were discussed to ensure all the issues were addressed. In the end, the experts and/or translators completed the questionnaire resulting in a new version from their discussion.

Stage 3 — Back translation. Based on the new version of the questionnaire according to the experts and/or translators, this new instrument is back-translated into the original language. This process makes sure that the translated version has a similar meaning or item content as the original version.

Stage 4 — Expert review/expert judgment. In the fourth stage, experts are required to analyze the 21 items, especially in terms of relevancy and clarity. This study involves nine experts to analyze the

CCRAM instrument. An online discussion was held to overview the CCRAM instrument and the aim of the research. Then, an assessment form containing columns for scoring and comments was provided for the analysis process. The forms were then returned after the scoring was completed.

Participants (Panel Experts)

To determine the content validity of the instrument developed, experts are required to provide judgment. This research involves nine panel experts in psychology in various majors, such as social psychology, clinical psychology, and psychometry. Of the nine panel experts, seven have doctoral degrees in psychology and two hold master's degrees in psychology; in terms of profession, three experts work as psychologists and lecturers, while the rest (six) are lecturers or academicians. Of the nine experts involved, three are males, and six females.

An online briefing was conducted to explain the instruments developed. An explanation of the job description, that is analyzing items of CCRAM in terms of relevancy and clarity, was also given. The experts were provided with assessment forms, and the scoring given was kept confidential from other experts. The result of the scoring was then returned to the researchers as it was, without intervention and discussion.

Results

The calculation of CVI in item and scale levels was conducted. The CVR of items and Aiken's validity coefficients were calculated, as well as Kappa interrater reliability. The results of the calculations are shown in Table 2.

Table 2
I-CVI and S-CVI instrument CCRAM

Items	Relevant (rating 3 or 4)	Irrelevant (rating 1 or 2)	I-CVI	Clarity (rating 3 or 4)	Clarity (rating 1 or 2)	I-CVI	Interpretation
L_1	9	0	1.00	8	1	.89	Appropriate
L_2	8	1	.89	6	3	.67	Removed
L_3	9	0	1.00	9	0	1.00	Appropriate
L_4	9	0	1.00	9	0	1.00	Appropriate
L_5	9	0	1.00	8	1	.89	Appropriate
L_6	7	2	.78	7	2	.78	Revised/Removed
E_1	9	0	1.00	9	0	1.00	Appropriate
E_2	9	0	1.00	9	0	1.00	Appropriate
E_3	9	0	1.00	9	0	1.00	Appropriate
E_4	9	0	1.00	9	0	1.00	Appropriate
E_5	9	0	1.00	9	0	1.00	Appropriate
R_1	9	0	1.00	9	0	1.00	Appropriate
R_2	8	1	.89	8	1	.89	Appropriate
R_3	9	0	1.00	9	0	1.00	Appropriate

(table 2 continues)

Table 2 (continued)

Items	Relevant (rating 3 or 4)	Irrelevant (rating 1 or 2)	I-CVI	Clarity (rating 3 or 4)	Clarity (rating 1 or 2)	I-CVI	Interpretation
R_4	9	0	1.00	9	0	1.00	Appropriate
A_1	9	0	1.00	9	0	1.00	Appropriate
A_2	7	2	.78	7	2	.78	Revised/Removed
A_3	9	0	1.00	9	0	1.00	Appropriate
A_4	9	0	1.00	9	0	1.00	Appropriate
B_1	8	1	.89	8	1	.89	Appropriate
B_2	7	2	.78	7	2	.78	Revised/Removed

Note. L = leadership; E = collective efficacy; R = readiness; A = attachment to place; B = social trust.

The calculation of I-CVI indicates that in terms of relevancy and clarity, one item needs to be removed, three items need to be revised/removed, and the rest are appropriate. The lowest value of I-CVI is .67, for Item L_2, in the clarity aspect. The item to be removed from the CCRAM instrument is Item L_2 ("I believe that the elected or candidate mayor can lead and implement the city regulation properly in emergency situations"). The items to be revised are Items L_6, A_2, and B_2.

The average value of I-CVI of relevancy is .95 and I-CVI of clarity is .93. And the value of S-CVI/UA (Scale-content validity item/universal agreement) of relevancy is 15/21 = .71 and S-CVI/UA of clarity is 13/21 = .62. Interpretation of I-CVIs: if the I-CVI is higher than 79%, the item is considered appropriate; if it is between 70 and 79%, it needs revision; if it is less than 70%, it is removed. Conclusions were made based on the two I-CVI (Zamanzadeh et al., 2015).

According to the average of I-CVI in terms of relevancy and clarity, the CCRAM instrument is considered satisfactory because the values of both aspects are above .90 (S-CVI = .95 and .93, respectively). While for S-CVI/UA, the value of relevancy and clarity are .71 and .62, respectively. The finding indicates that there are items that need to be revised in terms of clarity (Table 3).

TABLE 3
CVR instrument CCRAM

Items	N_e	CVR-relevancy	N_e	CVR-clarity	Interpretation
L_1	9	1.00	8	.78	Remaining
L_2	8	.78	6	.33	Revised/Removed
L_3	9	1.00	9	1.00	Remaining
L_4	9	1.00	9	1.00	Remaining
L_5	9	1.00	8	.78	Remaining
L_6	7	.56	7	.56	Revised/Removed
E_1	9	1.00	9	1.00	Remaining
E_2	9	1.00	9	1.00	Remaining
E_3	9	1.00	9	1.00	Remaining
E_4	9	1.00	9	1.00	Remaining
E_5	9	1.00	9	1.00	Remaining
R_1	9	1.00	9	1.00	Remaining
R_2	8	.78	8	.78	Remaining
R_3	9	1.00	9	1.00	Remaining

(table 3 continues)

Table 3 (continued)

Items	N_e	CVR-relevancy	N_e	CVR-clarity	Interpretation
R_4	9	1.00	9	1.00	Remaining
A_1	9	1.00	9	1.00	Remaining
A_2	7	.56	7	.56	Revised/Removed
A_3	9	1.00	9	1.00	Remaining
A_4	9	1.00	9	1.00	Remaining
B_1	8	.78	8	.78	Remaining
B_2	7	.56	7	.56	Revised/Removed

Note. N_e = the number of panelists deciding that the items are essential; CVR or content validity ratio = (N_e – N/2) / (N/2). With the number of experts = 9, items with CVR below .75 will be revised or removed (according to Lawshe ,1975); if the number of experts is 9, the minimum value of CVR is .75. L = leadership; E = collective efficacy; R = readiness; A = attachment to place; B = social trust.

Table 3 shows that of the 21 items, four need to be revised/removed. Those items are L_2 ("I believe that the elected or candidate mayor can lead and implement the city regulation properly in emergency situations"), L_6 ("In my city, children get proper attention"), A_2 ("I have a sense of belonging to the place where I live"), and B_2 ("Relationships among communities in my city are good"). The finding is in line with the results of CVI shown in Table 2 given that the items removed and revised are the same. Next, the Aiken's validity coefficient of each item (Aiken's validity) and Kappa interrater reliability of each CCRAM dimension are shown in Table 4.

The Kappa interrater reliability index of the CCRAM instrument in relevancy and clarity aspects indicate satisfactory scores, which are .90 and .87. In general, the minimum criterion for interrater reliability is .80 (Everitt & Skrondal, 2010; Klein, 2018; McHugh, 2012). Hence, in terms of reliability, the CCRAM instrument is considered reliable.

In Aiken's validity, items with a value below .50 are removed (Aiken, 1985). In this study, the items which are divided into four categories should have a minimum value of .70. Thus, based on the criterion, two items — L_1 and L_2 — with Aiken's validity value of .67 and .59 need to be revised or removed, while other items are considered valid and should remain.

Discussion

Content validity tests with I-CVI and CVR consistently indicate that four items need to be removed or revised. From the leadership dimension, the items removed are L_2 and L_6, while from the attachment to a place aspect, the item removed is A_2, and from the social trust dimension, the item removed is B_2. Those items have index values below .70 or .70-.79 for CVI (Zamanzadeh et al., 2015) and below .75 for CVR (Lawshe, 1975). In general, items removed or revised indicate that the situations or conditions are considered irrelevant in Indonesian contexts. For instance, Item L_6, stating "In my city, children get proper attention," is considered irrelevant and unclear by the experts.

In Aiken's validity, there are two items — L_1 and L_2 — with Aiken's values of .67 and .59, respectively. The items scored low especially in clarity, indicating that the message or meaning of the items is considered unclear by the experts. In terms of interrater reliability, the CCRAM instrument indicates satisfactory reliability in relevancy and clarity, because the values, .90 and .87, respectively, are above .80. Therefore, the CCRAM instrument is considered reliable.

TABLE 4
Coefficient of interrater reliability and Aiken's validity

Items	Kappa interrater reliability of relevancy	Aiken's relevancy	Kappa interrater reliability of clarity	Aiken's clarity	Interpretation
L_1		.89		.67	Revised/Removed
L_2		.81		.59	Revised/Removed
L_3		.96		.81	Remaining
L_4		.89		.74	Remaining
L_5		.93		.85	Remaining
L_6		.74		.70	Remaining
E_1		.93		.89	Remaining
E_2		.96		.96	Remaining
E_3		.96		.93	Remaining
E_4		.89		.81	Remaining
E_5	.90	.96	.87	.93	Remaining
R_1		.96		.93	Remaining
R_2		.85		.89	Remaining
R_3		1.00		.93	Remaining
R_4		.93		.96	Remaining
A_1		.93		.89	Remaining
A_2		.74		.70	Remaining
A_3		.85		.81	Remaining
A_4		.85		.81	Remaining
B_1		.85		.85	Remaining
B_2		.78		.74	Remaining

Note. Kappa interrater reliability measures the scale level, while Aiken's validity measures the item level. L = leadership; E = collective efficacy; R = readiness; A = attachment to place; B = social trust.

CONCLUSION

The CVI measurement methods, I-CVI, S-CVI, and CVR, are straightforward content validation methods that only count the proportion of items or relevant items to be divided by the total number of items. The closer the value to 1, the higher the content validity of the items. Nevertheless, CVI and CVR methods are not sensitive to varied expert judgments. Despite a more complex measurement, the Aiken's validity index is a more sensitive measurement that can measure the agreement of raters in scoring.

In terms of reliability, the calculation of S-CVI/UA or S-CVI Ave is relatively easier to use because it is based on the average scores and proportion. Generally, values of I-CVI or CVR are in line with the value of S-CVI (Rubio et al., 2003; Zamanzadeh et al., 2015). Moreover, considering the scoring agreement among experts, the interrater reliability (IRR) measurement is considered precise because it correlates the scoring of one expert with the others. The value of IRR is high when the scoring of items is consistent among experts, indicating satisfactory consistency or reliability.

In the CCRAM instrument used as an example in this study, the content validity value is satisfactory based on CVI, CVR, and Aiken indexes. Although four items need to be revised in terms of clarity, the S-CVI/UA, S-CVI/Ave, and IRR indicate that the reliability of the items is satisfactory.

REFERENCES

Aiken L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*, 131-141.

Almanasreh, E., Moles, R., & Chen, T. F. (2018). Research in social and administrative pharmacy evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214-221. https://doi.org/10.1016/j.sapharm.2018.03.066

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). Standards for educational and psychological testing (Vol. 148). American Educational Research Association.

Aravamudhan, N. R., & Krishnaveni, R. (2015). Establishing and reporting content validity evidence of Training and Development Capacity Building Scale (TDCBS). *Management*, *20*(1), 131-158. https://hrcak.srce.hr/141598

Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe's Content Validity Ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*(1), 79-86. https://doi.org/10.1177/0748175613513808

Bagozzi, R. P. (1980). *Causal models in marketing*. John Wiley.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, *25*(24), 3186-3191. https://doi.org/10.1097/00007632-200012150-00014

Chadha, N. K. (2009). *Applied psychometry*. SAGE Publications.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104.

Cohen, O., Leykin, D., Lahad, M., Goldberg, A., & Aharonson-Daniel, L. (2013). The Conjoint Community Resiliency Assessment Measure as a baseline for profiling and predicting community resilience for emergencies. *Technological Forecasting and Social Change*, *80*(9), 1732-1741. https://doi.org/10.1016/j.techfore.2012.12.009

Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and measurement* (9th ed.). McGraw Hill.

Davis, L. (1992). Instrument review: Getting the most from your panel of experts. *Applied Nursing Research*, *5*, 194-197. https://doi.org/10.1016/S0897-1897(05)80008-4

Everitt, B. S., & Skrondal, A. (2010). *The Cambridge dictionary of statistics* (4th ed.). Cambridge University Press.

Gable, R. K., & Wolf, J. W. (1993). *Instrument development in the affective domain: Measuring attitudes and vales in corporate and school settings*. Kluwer Academic.

Grant, J. S., & Davis, L. L. (1997). Selection and use of content experts for instrument development. *Research in Nursing & Health*, *20*, 269-274. https://doi.org/10.1002/(sici)1098-240x(199706)20:3<269::aid-nur9>3.0.co;2-g

Hernandez, C. A. (1995). The experience of living with insulin-dependent diabetes: Lessons for the diabetes educator. *The Diabetes Educator*, *21*(1), 33-37. https://doi.org/10.1177/014572179502100106

Klein, D. (2018). Implementing a general framework for assessing interrater agreement in stata. *Stata Journal*, *18*(4), 871-901. https://doi.org/10.1177/1536867x1801800408

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563-575.

Leykin, D., Lahad, M., Cohen, O., Goldberg, A., & Aharonson-Daniel, L. (2013). Conjoint Community Resiliency Assessment Measure-28/10 items (CCRAM28 and CCRAM10): A self-report tool for assessing community resilience. *American Journal of Community Psychology*, *52*(3-4), 313-323. https://doi.org/10.1007/s10464-013-9596-0

Lynn, M. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382-385.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain* (3rd ed.). Springer.

McHugh, M. L. (2012). Lessons in biostatistics interrater reliability: The kappa statistic. *Biochemica Medica*, *22*(3), 276-282. https://hrcak.srce.hr/89395

Patil, S., Gosavi, M., Bannur, H., & Ratnakar, A. (2015). Blueprinting in assessment: A tool to increase the validity of undergraduate written examinations in pathology. *International Journal of Applied and Basic Medical Research*, *5*(4), S76-S79. https://doi.org/10.4103/2229-516x.162286

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. Routledge. https://doi.org/10.4324/9780203841624

Rubio, D. M. G., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94-104. https://doi.org/10.1093/swr/27.2.94

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment*, *5*(4), 299-321.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psichotema*, *26*(1), 100-107. https://doi.org/10.7334/psicothema2013.256

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical and practical dimensions*. Macmillan Publishers Ltd. https://doi.org/10.1057/978-1-137-38523-9

Syaiful, I. A., & Roebianto, A. (2020). Adapting and examining the factor structure of the Self-Compassion Scale in Indonesian version. *Jurnal Psikologi*, *47*(3), 175-205. https://doi.org/10.22146/jpsi.57608

Thorndike, R. M., & Thorndike-Christ, T. (2014). Measurement and evaluation in psychology and education. *Journal of the American Statistical Association*, *56*(296), 1029. https://doi.org/10.2307/2282039

Walz, C. F., Strickland, O., & Lenz, E. (1991). *Measurement in nursing research* (2nd ed). F. A. Davis.

Yaghmaie, F. (2003). Content validity and its estimation. *Journal of Medical Education Spring*, *3*(1), 25-27.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A.-R. (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, *4*(2), 165-178. https://doi.org/10.15171/jcs.2015.017