JRERE 12 (2) 2023 125-133

Journal of Research and Educational Research Evaluation

http://journal.unnes.ac.id/sju/index.php/jere

The Application of Aiken's V Method for Evaluating the Content Validity of Instruments that Measure the Implementation of Formative Assessments

Siti Nurjanah[1]✉, Edi Istiyono[2], Widihastuti Widihastuti[3], Muhammad Iqbal[4], Saefudin Kamal[5]

[1,2,3,4,5]Graduate School of Education, Yogyakarta State University, Indonesia

Article Info

History Articles
Received:
10 May 2023
Accepted:
11 June 2023
Published:
30 August 2023

Keywords:
Aiken's V, formative assessment, content validity.

Abstract

This study aims to analyze the content validity of the measuring instrument used for formative assessment implementation by utilizing the Aiken method. The selection of formative assessment was based on the consensus within experts that it offers significant enhancements to the progression of classroom learning. Nevertheless, there is currently no valid instrument available to measure it. The instrument under study encompasses three key dimensions: perceptions, practices, and challenges. The instrument consists of a closed-ended questionnaire with 37 items and an open-ended questionnaire with 2 items. The content validity was evaluated by three raters who are highly experienced in their respective disciplines. After doing a quantitative study of the Aiken's V value, it was determined that all items of the instrument for implementing formative assessment were deemed valid. According to the qualitative analysis, raters provided some modest revisions for certain items. Overall, the instrument used to assess the implementation of formative assessment is deemed valid through content validity analysis relying on Aiken's method. Instructors, researchers, and policy makers can utilize the findings of this study to obtain information regarding the implementation of formative assessment through the utilization of the measuring instrument employed in this study.

✉Correspondence Address :
Department of Educational Research and Evaluation,
Graduate School of Education, Yogyakarta State University, Indonesia
55281
E-mail : siti960pasca.2023@student.uny.ac.id

p-ISSN 2252-6420
e-ISSN 2503-1732

INTRODUCTION

Utilizing valid measurement instruments is an essential cornerstone in research. Measurement instruments for research are deemed valid based on their content validity, and the validation process contributes a pivotal role in the development of these instruments (Polit & Cheryl Tatano Beck, 2006; Almanasreh et al., 2019; Yusof, 2019). Content validity serves as the primary means of connecting intangible concepts with tangible and quantifiable indicators (Wynd et al., 2003). Content validity enhances the objectivity of a measurement instrument by incorporating crucial feedback from raters, which allows instrument developers to make necessary revisions to their measures (Rubio et al., 2003). Hence, it is anticipated that instrument designers will authenticate the accuracy of the measurement instruments they provide.

Instruments that fail to meet validity requirements will have unforeseen implications in a scientific study. Instruments lacking validity may yield inaccurate measurement results (Almanasreh et al., 2019). Norbeck et al. stated that performing reliability testing on instruments with low content validity does not yield substantial advantages (Norbeck et al., 1981). If the validity of the measurement instrument is still questionable, researchers intending to utilize it must exercise utmost caution during the deployment phase (Potter & Levine-Donnerstein, 1999). The instrument developer takes measures to minimize any potential harmful impacts through rigorous validity testing.

When attempting to establish a precise definition of validity, one will encounter a multitude of meanings dispersed over different reference resources. Validity, in essence, pertains to the degree of precision exhibited by a measuring instrument in accurately assessing the specific psychological attribute it is designed to evaluate (Shi et al., 2012). Content validity, as defined by Almanasreh et al. (2019), refers to the degree to which the items of an assessment instrument are both pertinent and indicative of the intended target concept. In the study conducted by Rubio et al. (2003), it is emphasized that content validity evaluates the extent to which the items in a measure accurately depict the information they are intended to reflect. Content validity is commonly defined as the degree to which a set of items accurately represent the dimension and definitions of the concept being assessed.

Content validity, an essential feature of instrument development, initiated at the initial phases of constructing a measuring tool. This involves the assessment of instrument elements by raters to determine their relevance to the intended content (Lynn, 1986; Almanasreh et al., 2019). Wynd et al. (2003) highlighted that content validity relies on the researcher's judgment, logic, and reasoning, which is then verified by a group of experts who possess specialized knowledge in relevant areas. Ventura-León (2022) highlighted the significance of this component and warned that relying solely on internal structure to establish validity is insufficient. Further investigation, such as examining content validity, is necessary. Furthermore, these studies emphasized the significance of evaluating the objectivity of measurement instruments through expert evaluation in order to ensure theoretical coherence and establish a foundation for researchers to analyze the content of the items employed.

Several methods have been proposed for assessing the degree of consensus among experts regarding the pertinence of the instrument's content. This study aims to evaluate the content validity of the measuring instrument by employing the method proposed by Aiken (1985). Aiken's V coefficient is a convenient and straightforward metric for evaluating the responses provided by the raters. It is not only simple to calculate but also easy to interpret (Ventura-León, 2022).

Prior research has investigated content validity using the utilization of the Aiken method (Hikmah et al., 2017; Ikhsanudin &

Subali, 2018; Maulita et al., 2019; Muliana et al., 2020; Puspitasari & Febrinita, 2021; Wati & Misbah, 2021; Yudhistira & Tomoliyus, 2020; An Nabil et al., 2022; Castro Benavides et al., 2022). Nevertheless, there has been no research performed to evaluate the validity of any instrument to measure the implementation of formative assessments using Aiken's method. Formative assessment is widely recognized by professionals as a valuable approach for enhancing progress in classroom learning (Akom, 2010; Heritage, 2007; Bulunuz & Bulunuz, 2017; Li et al., 2021). Hence, it is imperative to possess a valid measurement instrument capable of quantifying the implementation of formative assessment. This research is anticipated to yield an instrument of such nature. This study aims to investigate the validity of the measuring instrument for formative assessment by employing the Aiken method. The aims of this study will be accomplished by addressing the following research inquiries.

What is the content validity of the instrument for implementing formative assessment, as evaluated by Aiken's validity coefficient (V)?

How is the content validity of the formative assessment instrument assessed qualitatively based on the suggestions provided by raters?

Instructors, researchers, and policy makers can utilize the findings of this study to obtain information regarding the implementation of formative assessment through the utilization of the measuring instrument employed in this study.

METHODS

For the accomplishment of the research objectives, a content validity analysis was undertaken utilizing Aiken's V method. The instrument evaluated was an instrument to measure the implementation of formative assessment in the form of a questionnaire with a Likert scale of 4 answer alternates. This instrument is created as a result of modification of the instrument in research by Lajane et al. (2020). A few modifications were made to the items to suit the broader context of learning, while considering some pertinent practical considerations. The instrument consists of two questionnaires, a closed-ended questionnaire and an open-ended questionnaire. The closed-ended questionnaire consisted of 37 items, whereas the open-ended questionnaire consisted of 2 items. The open-ended questionnaire was devised so that the survey conducted with this instrument is intended to acquire broader information from the respondents. A comprehensive survey is required to acquire additional information concerning the challenges encountered by educators when implementing formative assessment. Teachers originating from various educational institutions are expected to encounter distinct limitations. The unstructured survey enabled educators to investigate novel information by offering ample room to elaborate on their current circumstances. The instrument used to gauge the execution of formative assessments comprises three dimensions: perceptions, practice, and challenges. These three dimensions are purported to encompass information regarding the execution of formative assessment. The dimensions are further dissected into numerous indications, which are thoroughly elucidated in Table 1.

Table 1. The dimensions and indicators employed to measure the implementation of formative assessment in the instruments.

No.	Dimensions	Indicators
1	Perceptions	Perceptions and thoughts regarding the usefulness of formative assessments
		Teacher training
2	Practices	Discuss learning focus and objectives
		Forms of formative assessment applied
		Regulations and scheduling of the execution of formative assessment
3	Challenges	Various types of perceived challenges to the implementation of formative assessment

Once the items have been compiled into a complete instrument, a validation sheet is created and disseminated to many impartial raters. Subsequently, the researchers conducted an analysis of the instrument's content validity utilising Aiken's validation method. The procedure for analysing the content validation of the instrument using Aiken's V involves the following phases:

The validation sheet was disseminated among the impartial raters. For this study, three raters were selected, all of them were experienced teachers with more than ten years of teaching experiences. Work experience is a factor taken into account when choosing raters to evaluate the validity of measuring instruments (Rubio et al., 2003). Additionally, one of the raters fulfilled the role of a *guru penggerak* who closely aligns with the independent curriculum that prioritises formative assessment, hence providing further support for this measurement domain. Additionally, two of the raters possessed knowledge and skills in the field of evaluation and measurement. The presence of this experience is raters engaged in the content validation process of a measurement instrument, instilling assurance in the precision and validity of the tool.

The raters evaluated the items of the instrument based on the indicators specified in the validation sheet. The raters are provided with 5 rating categories, which are determined by a Likert scale. The decision to include 5 rating categories instead of 4 was made due to the inherent limitations of a 4-point rating scale. This scale restricts raters from expressing uncertainty or neutrality, thereby necessitating the inclusion of an additional category. Hence, it is advisable to employ a rating categories consisting of either 5 or 3 points, particularly during the initial phases of evaluation. This will provide raters with greater flexibility to express a more precise degree of confidence and neutrality in the evaluation process (Almanasreh et al., 2019).

The rater provided the researcher with the validation results and included revisions.

The researchers modified the instrument items in accordance with feedback from all raters.

The researchers computed the validity coefficient using the Aiken method, based on the ratings provided by the raters for the instrument items.

The researcher determines the standard value for the validity coefficient that is deemed acceptable. The study involved three raters who were presented with five rating categories. The researcher established a significance level of 5% in order to determine the validity coefficient value for an item that was deemed valid, which was found to be 0.92 (Aiken, 1985). If the validity coefficient value of an item is below 0.92, the item is considered invalid. If an item is deemed invalid, the researcher has the option to either eliminate the item or make revisions to it. The determination to eliminate or modify an instrument item is made based on certain considerations. For instance, whether an indicator is still covered or not after an item is eliminated.

The researchers established the final instrument after conducting a thorough

analysis of its validity, taking into account the validity coefficient (V).

The findings of the quantitative content validation analysis conducted by three raters are given in Table 2.

RESULTS AND DISCUSSION

Table 2. Overview of the results from the content validation of the questionnaire by raters and the computation of Aiken's V value.

Item No.	[Dimension, Indicator]	Rater 1	Rater 2	Rater 3	S1	S2	S3	∑S	V	Judgement
Closed-ended questionnaire										
1	[1.a]	5	5	5	4	4	4	12	1.00	Valid
2		5	5	5	4	4	4	12	1.00	Valid
3		5	5	5	4	4	4	12	1.00	Valid
4		5	5	5	4	4	4	12	1.00	Valid
5	[1.b]	5	5	5	4	4	4	12	1.00	Valid
6		5	5	5	4	4	4	12	1.00	Valid
7		5	5	5	4	4	4	12	1.00	Valid
8		5	5	5	4	4	4	12	1.00	Valid
9		5	5	5	4	4	4	12	1.00	Valid
10		5	5	5	4	4	4	12	1.00	Valid
11	[2.a]	5	5	5	4	4	4	12	1.00	Valid
12		5	5	5	4	4	4	12	1.00	Valid
13		5	5	5	4	4	4	12	1.00	Valid
14	[2.b]	5	5	5	4	4	4	12	1.00	Valid
15		5	5	5	4	4	4	12	1.00	Valid
16		5	5	4	4	4	3	11	0.92	Valid
17		5	5	5	4	4	4	12	1.00	Valid
18		5	5	5	4	4	4	12	1.00	Valid
19		5	5	4	4	4	3	11	0.92	Valid
20		5	5	4	4	4	3	11	0.92	Valid
21		5	5	5	4	4	4	12	1.00	Valid
22		5	5	5	4	4	4	12	1.00	Valid
23		5	5	5	4	4	4	12	1.00	Valid
24		5	5	5	4	4	4	12	1.00	Valid
25	[2.c]	5	5	5	4	4	4	12	1.00	Valid
26		5	5	4	4	4	3	11	0.92	Valid
27		5	5	5	4	4	4	12	1.00	Valid
28		5	5	5	4	4	4	12	1.00	Valid
29		5	5	5	4	4	4	12	1.00	Valid
30		5	5	5	4	4	4	12	1.00	Valid
31		5	5	5	4	4	4	12	1.00	Valid
32	[3.a]	5	5	5	4	4	4	12	1.00	Valid
33		5	5	5	4	4	4	12	1.00	Valid
34		5	5	5	4	4	4	12	1.00	Valid
35		5	5	5	4	4	4	12	1.00	Valid
36		5	5	5	4	4	4	12	1.00	Valid
37		5	5	5	4	4	4	12	1.00	Valid
Open-ended questionnaire										

Item No.	[Dimension, Indicator]	Rater 1	Rater 2	Rater 3	S1	S2	S3	∑S	V	Judgement
1	[3.a]	5	5	4	4	4	3	1	0.92	Valid
2		5	5	4	4	4	3	2	0.92	Valid

Based on the dimensions, in the perception dimension, all items have a validity coefficient (V) value of 1.00. In the practice dimension, the maximum V value is 1.00 and the lowest is 0.92. There are 3 items that have V 0.92. In the challenge dimension, all items have a V value of 1.00. All items in the open-ended questionnaire received a validity coefficient (V) of 0.92.

All items were classified as valid with a validity coefficient value ≥ 0.92. The acknowledged validity coefficient value of ≥ 0.92 adjusts the characteristics of the rater and the rating categories that can be chosen by the rater (3 raters and 5 rating categories) (Aiken, 1985). A total of 33 items out of 37 items in the closed-ended questionnaire got a validity coefficient value of 1.00. A validity coefficient value closer to 1 indicates a more perfect agreement between raters (Shi et al., 2012; Ventura-León, 2022; Yusof, 2019). Consensus among raters is crucial for coefficient V since it uses averages in its analysis, which requires adequate concentration of data to be evaluated effectively (Merino-Soto, 2018). This attempt can be improved by discussing revised instrument items with the raters. Although this review is indeed not necessary to conduct because the items are classified as valid, the instrument developer may initiate it with the hope that the instrument items developed can be even better. Although Aiken's V is acquired high in scale development, instrument developers can still undertake a thorough review of content validity to assure the accuracy of the instrument developed.

The four remaining items on the closed-ended questionnaire and two items on the open-ended questionnaire revealed a validity coefficient of 0.92. This outcome indicates that some raters selected response ratings other than 4 (which implies relevance) and 5 (which implies high relevance) for the indicators listed in the validation sheet. The validity coefficient serves as a criterion for determining if an item should undergo revision, deletion, or replacement (Polit & Cheryl Tatano Beck, 2006). The items have undergone re-evaluation in order to generate more refined assertions than previously. The enhancement of instrument items was conducted by evaluating quantitative data in the form of values provided by raters. The instrument developer reviewed the items again, taking into consideration the indicators on the validation sheet that received a score lower than 4 or 5 from the rater. Content validity can be quantitatively evaluated by calculating various indicators to measure their strengths and flaws (Shi et al., 2012).

The evaluation of content validity in this study extended beyond only assessing Aiken's V coefficient value. The evaluation of the instrument was also conducted by taking into account the recommendations provided by the raters on the instrument items. This review is a crucial evaluation of the content validity of an instrument. This perspective can mitigate the limitations of quantitative analysis, which solely concentrates on items that are already present in scale development (Shi et al., 2012). Furthermore, the active involvement of raters in offering suggestions, both orally and in written form, serves as a significant catalyst for enhancing the appropriateness of items within the specific domain of focus. Each suggestion is deemed significant in the endeavor to enhance the domain and its items (Yusof, 2019). During the evaluation phase, raters play a crucial role by offering suggestions on whether to add or remove items, carefully evaluating each word in the item, and providing valuable input to enhance its development. The involvement of raters is essential to ensure that the domains and items accurately represent the concepts to be measured, leading to a more valid

instrument. The examination of both quantitative and qualitative data by raters serves as the foundation for instrument developers to make decisions regarding the disposition of items: whether they should be kept, altered, removed, or included (Almanasreh et al., 2019). The qualitative evaluation allows for a more complete improvement of items by not confining them to the values and indicators specified on the validation sheet.

The raters suggested some minor adjustments, which included adjusting the wording to align with contemporary educational needs and situations, enhancing sentence clarity and communicativeness, and correcting any typographical errors. An example of a suggestion made by the raters pertained to item 7 of the closed-ended questionnaire. The rater proposed the revision of the word "theoretical" as the teacher's focus is more inclined towards the tangible advantages for students rather than the hypothetical ones. These guidelines are crucial for instrument developers to ensure the accuracy of the instruments in measuring their intended quantities.

Since all the Aiken's V values were met, no items were removed from the instrument. Therefore, it can be concluded that the instrument indicators can be accurately represented by the instrument items, allowing for the measurement of formative assessment implementation. On the other hand, a low V coefficient value indicates that an item needs to be modified, whilst a very low V value shows the potential for removing the item (Shi et al., 2012). A low degree of agreement may indicate a lack of consistency in the comprehension of the concept being measured (Polit et al., 2007). If these values are maintained in a haphazard manner, the usage of unrelated items or the exclusion of significant items in the process of constructing the concept could potentially weaken the internal coherence of the concept. If the desired level of content validity has not been attained, it is necessary to carry out an additional round involving either the same raters or different raters. This will enable further enhancements in validation and the creation of a more precise instrument (Almanasreh et al., 2019).

The limitations of this study prevented the instrument from undergoing further analysis in terms of construct validity and reliability tests. The researchers were unable to carry out these examination due to constraints in terms of limited resources, including energy, cost, and time, which hindered their ability to gather a substantial number of teachers as participants for the instrument trial. In future research, it is recommended to conduct these tests in order to obtain instruments that possess both validity and reliability.

CONCLUSION

The content validity of the instrument for implementing formative assessment was valid, as indicated by the Aiken's validity coefficient (*V*). Based on a qualitative review using the feedback given by raters, there were a small number of minor suggestions regarding specific items. Overall, the instrument designed to measure formative assessment implementation was found to be valid through content validity analysis using the Aiken methods. To enhance the quality of future research, it is advisable to do reliability tests to ensure that the instruments used are both valid and reliable. Future research can employ this instrument to measure the implementation of formative assessment with teachers as respondents.

REFERENCES

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings, educational and psychological measurument. *Educational and Psychological Measurement*, *45*(1), 131–142.

Akom, G. V. (2010). *Using formative assessment despite the constraints of high stakes testing and limited resources: A case study of chemistry teachers in Anglophone Cameroon*. Western Michigan University.

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214–221.

An Nabil, N. R., Wulandari, I., Yamtinah, S., Ariani, S. R. D., & Ulfa, M. (2022). Analisis indeks Aiken untuk mengetahui validitas isi instrumen asesmen kompetensi minimum berbasis konteks sains kimia. *Paedagogia*, *25*(2), 184.

Bulunuz, N., & Bulunuz, M. (2017). Effect of formative assessment-based instruction on high school students' conceptual understanding of balance and torque. *Journal of Inquiry Based Activities (JIBA)*, *7*(1), 21–33.

Castro Benavides, L. M., Tamayo Arias, J. A., Burgos, D., & Martens, A. (2022). Measuring digital transformation in higher education institutions – content validity instrument. *Applied Computing and Informatics*.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? In *Phi Delta Kappan* (Vol. 89, Issue 2).

Hikmah, N., Yamtinah, S., Ashadi, & Indriyanti, N. Y. (2017). Analisis validitas isi instrumen computerized two-tier keterampilan proses sains pada materi termokimia. *Prosiding Seminar Nasional Pendidikan Sains (SNPS)*, 40–45.

Ikhsanudin, & Subali, B. (2018). Content validity analysis of first semester formative test on biology subject for senior high school. *Journal of Physics: Conference Series*, *1097*(1), 1–9.

Lajane, H., Gouifrane, R., Qaisar, R., Chemsi, G., & Radid, M. (2020). Perceptions, practices, and challenges of formative assessment in initial nursing education. *The Open Nursing Journal*, *14*(1), 180–189.

Li, T., Yeung, M., Li, E., & Leung, B. (2021). How formative are assessments for learning activities towards summative assessment? *International Journal of Teaching and Education*, *9*(2), 42–57.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*(6), 382–386.

Maulita, S. R., Sukarmin, S., & Marzuki, A. (2019). The content validity: Two-tier multiple choices instrument to measure higher-order thinking skills. *Journal of Physics: Conference Series*, *1155*(1).

Merino-Soto, C. (2018). Confidence interval for difference between coefficients of content validity (Aiken's V): A SPSS syntax. *Anales de Psicologia*, *34*(3), 587–590.

Muliana, N., Pada, A. U. T., & Nurmaliah, C. (2020). Content validity of conation assessment. *Journal of Physics: Conference Series*, *1460*(1).

Norbeck, J. S., Lindsey, A. M., & Carrieri, V. L. (1981). The development of an instrument to measure social support. *Nursing Research*, *30*(5), 264–269.

Polit, D. F., Beck, C. T., & Owen, S. V. (2007). Focus on research methods: Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing and Health*, *30*(4), 459–467.

Polit, D. F., & Cheryl Tatano Beck. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing and Health*, *29*(5), 489–497.

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*(3), 258–284.

Puspitasari, W. D., & Febrinita, F. (2021). Pengujian validasi isi (content validity) angket persepsi mahasiswa terhadap pembelajaran daring matakuliah matematika komputasi. *Journal Focus Action of Research Mathematic (Factor M)*, *4*(1), 77–90.

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Encyclopedia of Food Sciences and Nutrition*, *27*(2), 94–104.

Shi, J., Mo, X., & Sun, Z. (2012). Content validity index in scale development. *Journal of Central South University (Medical Sciences)*, *37*(2), 152–155.

Ventura-León, J. (2022). Back to content-based validity. *Adicciones*, *34*(4), 323–326. https://doi.org/10.20882/adicciones.1213

Wati, M., & Misbah, M. (2021). The content validity of the assessment instrument on the characters of wasaka in wetland environment physics learning. *Journal of Physics: Conference Series*, *1760*(1).

Wynd, C. A., Schmidt, B., & Schaefer, M. A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research*, *25*(5), 508–518.

Yudhistira, D., & Tomoliyus. (2020). Content validity of agility test in karate kumite category. *International Journal of Human Movement and Sports Sciences*, *8*(5), 211–216.

Yusof, M. S. B. (2019). ABC of content validation and content validity index calculation. *Education in Medicine Journal*, *11*(2), 49–54.