A CEFR-Informed English Test: How Content Validity Index and Content Validity Ratio are used for Content Validation

Kamarul Ariffin Ahmad[1*] & Izazol Idris[2]

[1]Fakulti Bahasa dan Komunikasi, Universiti Pendidikan Sultan Idris,
35900 Tanjong Malim, Malaysia
[2]Fakulti Pembangunan Manusia, Universiti Pendidikan Sultan Idris,
35900 Tanjong Malim, Malaysia
*Email: kamarul.a@fbk.upsi.edu.my

ABSTRACT

Current education roadmap in Malaysia has seen enormous initiatives taken by the government to ensure the education system is at par with the international standard where the English syllabus and assessments are tailored according to the Common European Framework of Reference (CEFR). This includes higher learning institutions, thus this research aimed to develop the English entrance test that is CEFR-informed. 11 CEFR experts were involved in the study where they reviewed 60 multiple choice questions for reading, grammar and reading comprehension. The results showed that all items were accepted and valid based on the computation of Content Validity Index (CVI) and Content Validity Ratio (CVR). The lowest CVI recorded was *.82* with S-CVI of *.95* and the lowest CVR was *.64* which were above the suggested cut-off value by experts and research. The study concluded that all the items were carefully developed based on the framework and the needs of the test.

Keywords: CEFR, Content Validity Index (CVI), Content Validity Ratio (CVR), English Entrance Examination

INTRODUCTION

The current aspiration of the education system in the country is to prepare the future generation of Malaysia for the economic growth and the national development. In accordance with the notions, the rapid changes and development happening globally, Malaysia Education Blueprint 2013 - 2025 was introduced to address the public concern of the preparedness of the future generation and the capability of the education system in producing generation who is competitive as well as ready to face the 21st century challenges (Kementerian Pendidikan Malaysia, 2013). The Blueprint aims to achieve three key objectives which are (1) to ensure the quality, at the same time improving equity and to enforce unity, (2) to ensure a holistic approach to the students in the 13 years of education, and (3) to ensure that the products (i.e. the students) of the education system meet the demand and the expectations of the industry.

The three key objectives are then divided into 11 shifts to transform and reform the education system in the country. The shifts are to uphold the objectives and to outline in detail what needs to be done to ensure all key objectives are successfully achieved. Shift 2 explains and outlines the needs of ensuring that the students are bilingual – proficient in both Bahasa Malaysia and English language, while encouraging them to learn a third language. With the shift, the English Language Education Reform in Malaysia: The Roadmap 2015 – 2025 was developed and introduced by the English Language Standards and Quality Council Malaysia. The roadmap aims to align the English language learning ecosystem in Malaysia that consists of the curriculum, teaching methodology and the

assessments, to an international standard that is the Common European Framework of Reference (abbreviated as CEFR) for English language (Zuraidah Mohd Don & Mardziah Hayati Abdullah, 2019). It also encompasses all levels of education in Malaysia, from pre-school education to tertiary education.

In line with the aims of the education system in Malaysia, revisiting the documents including the curriculum and the assessment for English learning was done to guarantee its adherence to the roadmap. With a particular interest towards the assessment, this study found that the English entrance examination instrument for the Diploma in English programme needed to be re-developed as reported by a study by Kamarul Ariffin Ahmad and Zahari Suppian (2022). Their study found that the existing instrument did not have the records showing the instrument's empirical statistics in terms of separation index, reliability and found that there were several items with problematic distractors. An interview with then Head of Department of the English language revealed that there was no evidence suggesting the abidance of the instrument to any item development guidelines and the CEFR (Hasimah Ja'afar, 2020) which motivated the re-development of the instrument.

In instrument development, according to Shrotryia and Dhanda (2019), content validity is the prerequisite measure before the other steps in validation take place. This crucial step should foresee if any of the items would behave peculiarly either not measuring the intended construct or containing ambiguity that might hamper respondents' comprehension while attempting the instrument. This step is further detailed and can be observed in Ahorsu et al (2020) in which they explained that the content validity stage would also involve omitting items based on the judgements made by the panels of experts. Above all, the focal point of instrument development would be the item; on how it is worded, the length and the complexity of the structure as well as the vocabulary used to form the stem.

Almanasreh et al (2019) outlined four key elements in determining content validity which include (1) the definition of the construct, (2) the representation of the construct, (3) the relevance of the item to measure the domain, and (4) the pertinence of the construction stages or procedure. With the introduction and the implementation of CEFR in the Malaysia English language environment, a new instrument that is used for university entrance which defines the construct is called for. All four key elements mentioned by Almanasreh et al (2019) should be taken into consideration throughout the process of the development. This is to ensure that the instrument developed would facilitate the selection of the future candidates for the programme.

In the educational setting, the usual practice and widely used method in testing would begin with the preparation and development of the table of specification (Adom et al, 2020). Test developers utilise this table to guarantee that items used to test the knowledge or skills concur with the learning objective. Adom et al added that this is the crucial step in ensuring that the test is designed to measure what it intends to measure i.e the validity of the content. This is also mentioned by Shrotryia and Dhanda (2019) where they added the second step which is the judgement phase where it involves a group of experts of the field to go through the process of item vetting.

The development of instrument testing English language proficiency based on the Common European Framework of Reference (CEFR) is abstract thus it could be complex to construct (Almanasreh et al, 2019; & Nor Ashiquin C. Alih et al, 2020). This is also seen in Aryadoust (2020) where McNamara stated that to ensure an English proficiency test is only to measure one single construct was deemed complex. With the paucity of research pertaining to CEFR for English language and assessment in Malaysian setting, a careful construction of the item was deployed with the reference to Cohen-Swerdlik's model (2009) to identify and define the test, and also to Haladyna's Taxonomy of Developing Multiple Choice Questions (2004) to construct each of the items.

This study aimed to develop an English entrance examination to be used and facilitate the selection of future students for Diploma in English programme for the institution. The focus of this particular study was to identify and obtain the quantified content validity of the items developed with the reference to CEFR thus working towards making the instrument a CEFR-informed test. Therefore, this study intended to find out Content Validity Indices (CVI) and the Content Validity Ratio (CVR) of the instrument as the first two stages mentioned by Shrotryia and Dhanda (2019).

MATERIALS AND METHODS

This research adopted a quantitative approach as the nature of the procedure which was to quantify the expert judgement and to represent these judgements in numbers. On the other hand, the instrument in this study was developed in three stages based on the syntheses made from literature namely pre-judgement stage, judgement stage and post judgement stage. Pre-judgement stage which was prior to the development of the instrument and its content, the process began with the test developer and the head of department discussing the instrument with the reference to Cohen-Swerdlik's 14 Questions to define a test. Upon the conclusion of the responses to the 14 questions, a table of specification was developed to ensure that the instrument for the English entrance examination adhered to the curriculum set by the institution. Several meetings with the faculty members were conducted to discuss and to obtain the approval of the content of the instrument - the table of specification was vetted by the head of department and other members of the faculty. Then the items were developed based on the accepted and approved table of specification. 60 multiple choice items were developed for the entrance test which doubled the original number of items in the real test paper.

Figure 1. Model of Item Development adopted from Gregory (2014) and Cohen-Swerdlik (2009)
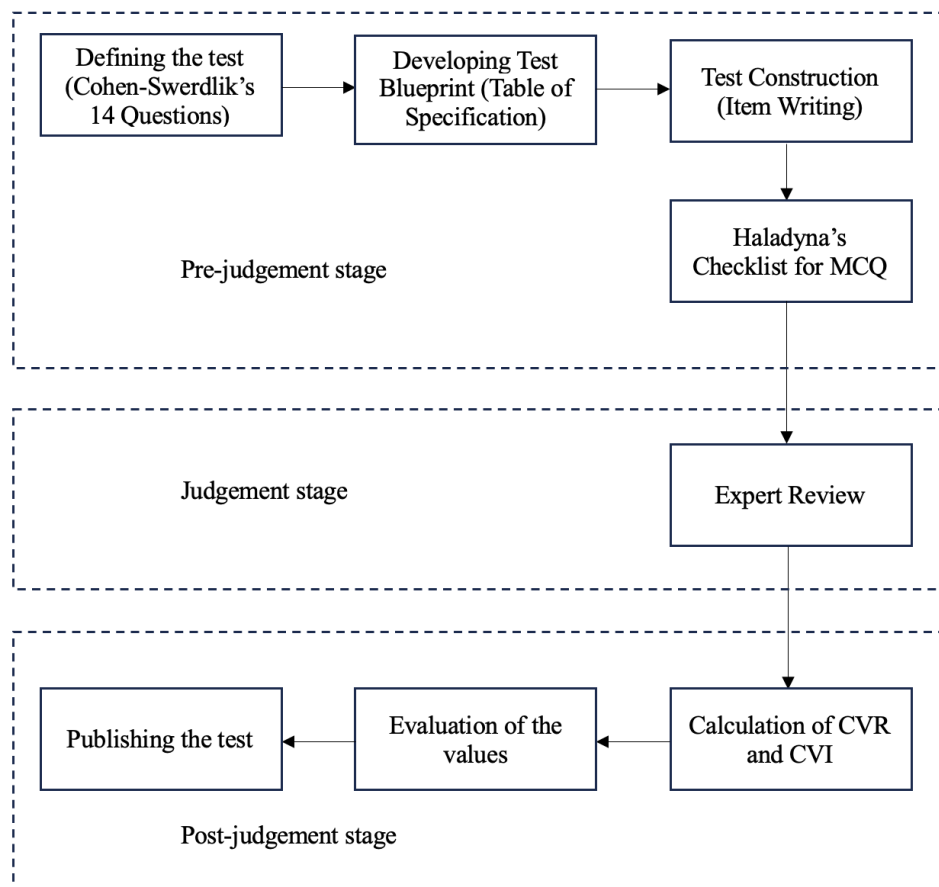
Multiple choice items type was selected to be used in the examination as it was feasible. As big number of test takers played a role in deciding the type of items used in an instrument, the English entrance test which would have an average of 2,000 test takers per seating was developed with multiple choice items to facilitate easy scoring, objective marking and could foster prudence amongst the test takers when answering the questions (Kamarul Ariffin Ahmad, 2018; Rachmat & Arfiandhani, 2019; & Babo et al, 2020). Since this is an English instrument, passages were included in the instrument as stimuli to test candidates' Grammar, Vocabulary and Reading Comprehension skills. The passages used were analysed with Text Inspector (i.e. https://textinspector.com) to guarantee their CEFR levels were up to what was desired.

In the Grammar and Vocabulary sections, words or phrases were lifted from the passages and four options were presented to the test takers to choose from. The words or phrases taken out from the passages were selected based on their word class (i.e. for Grammar items) and their CEFR levels - this was conducted based on the analyses obtained from the Text Inspector. The options prepared for each of the blanks were also carefully selected to suit the context to avoid give-aways and according to the CEFR levels. Whereby, the items developed for the Reading Comprehension sections were based on Bloom Taxonomy – the practice outlined for the institution.

The following table demonstrated the table of test specification as approved by the head of the department and no items were developed for A1 and A2 CEFR levels as the candidates, who were the Malaysian high school leavers, were expected to master and to be equipped with at least CEFR band B1 (Zuraidah Mohd Don & Mardziah Hayati Abdullah, 2019):

Table 1. Table of Test Specification for the CEFR-Informed English Entrance Test

Sections	Number of Items					Total Items
	CEFR Bands					
	A1	A2	B1	B2	C1+	
Grammar			2	6	2	10
Vocabulary			2	6	2	10
Total			4	12	3	20
	Bloom Taxonomy					
	C1	C2	C3	C4	C5	C6	
Reading Comprehension	1	1	3	3	1	1	10
Total	1	1	3	3	1	1	30

Out of the ten items developed for Grammar, three items tested test takers' knowledge about Tenses, and one item was developed to test each of the following word classes: Possessive Pronouns, Preposition, Modals, Determiner, Adverbs, Conjunctions and Phrasal Verbs.

The final phase of the development would be the calculation of the Content Validity Ratio (CVR) and the Content Validity Index (CVI) for each of the items and the instrument based on the responses from the experts. As could be observed from Norshahira Osman et al (2021) and Villagomez et al (2022), a minimum score of *.63* for CVR would signify the relevance of the inclusion of the items and according to Lawshe (1975, as cited in Zeraati & Alavi, 2014) claimed that the cut-off value of CVR with 11 experts would be *.59*. On the other hand, the cut-off value for I-CVI was set tat *.78* (Lynn, 1986) and the cut-off value for S-CVI was set at *.80* (David, as cited in Mohd Effendi at al, 2017; & Ahmad Tarmizi Abu et al, 2021). I-CVI is the item level of validity index where each expert would decide whether the item is measuring the construct.

The calculation of I-CVI*;*

$$y = \frac{\left[n1 \left(\frac{x}{3} \right) + n2 \left(\frac{x}{3} \right) + \cdots n10 \left(\frac{x}{3} \right) \right]}{10}$$

where
y	=	Item level Content Validity Index (I-CVI)
n1..n10	=	Coding for experts (n1 = Expert no.1.. n10 = Expert no.10)
x	=	Total score from each expert; binary system (1 = yes, 0 = no)

The calculation of S-CVI

$$S\text{-}CVI = \frac{CVI^1 + CVI^2 + \ldots}{N^{item}}$$

where
CVI^l	=	CVI value for Item 1
N^{item}	=	Total number of items

The calculation of CVR;

$$CVR = \frac{Ne - (\frac{N}{2})}{N/2}$$

where

Ne	=	The number of experts who declared the item relevant
N	=	The total number of experts involved

The experts involved in this study were the Common European Framework of Reference for English language experts in Malaysia. They were trained by the Ministry of Education Malaysia and involved with either the development of the CEFR or the training of English language teachers nationwide. A total of 11 experts were involved in this study which were identified through the snowballing method - each of the experts was asked to nominate another expert which was then invited to participate. Each of the participated experts was given the hardcopy of the instrument consisted of 60 multiple choice items (20 questions each for Grammar, Vocabulary and Reading Comprehension) where they needed to evaluate in terms of the mapping accuracy to the CEFR level and their relevance to be included in the instrument to measure proficiency. Each of the experts were given two weeks to complete the evaluation.

RESULTS AND DISCUSSION

The findings of this study would be presented with the presentation of the demographic findings of the experts and followed by the calculation of the content validity indices and the content validity ratios of the items.

Table 2. CEFR Experts Fractions Based on Position and Years of Experience

Position	Years of Experience	
	≤ 10 years	> 10 years
Senior Lecturer	1	5
English Teachers	0	5
Total	11	

The 11 experts consisted of three males and eight females in which all three male experts were English teachers and the remaining eight female experts were a combination of both teachers and the senior lecturers. The experts were gathered through snowballing method, and this may expose the results to biasness thus recruiting only CEFR experts who are appointed by the Ministry of Education of Malaysia was seen as a means to reduce the noise from biasness. In addition, these experts were scattered throughout the county where they are actively involved in teaching the new English language syllabus that is aligned with the CEFR. The recruitment of these experts with various backgrounds in teaching would further strengthen the results of this study. Given that the teachers consisted of those teaching in primary and secondary schools, and the lecturers teaching in the universities would provide an extensive review to the instrument that may cater various levels of the CEFR. This is in accordance with the English language syllabus in Malaysia where each level of education is designed to achieve a band or a different band in CEFR (i.e. A2 for primary school leavers, B1/B2 for secondary school leavers and B2/C1 for graduates).

Table 3. CEFR Experts Highest Level of Education

Level of Education	Number of Experts
Bachelor Degree	3
Master Degree	2
Doctoral Degree	6

Based on the expert level of education, this research has encapsulated various levels from bachelor degree up to doctorate level. This indicates that the experts involved in this study not only have extensive knowledge in the CEFR deployment in the country but also have the first-hand experience with teaching the English language at various levels and the experience in teaching students from various walks of life. This would further strengthen the dependability of the outcome of this study.

Table 4. Content Validity Index for the Instrument

Items	Expert Agree	CVI	Items	Expert Agree	CVI
1	11	1.00	31	11	1.00
2	11	1.00	32	11	1.00
3	11	1.00	33	11	1.00
4	10	.91	34	11	1.00
5	11	1.00	35	11	1.00
6	11	1.00	36	11	1.00
7	11	1.00	37	11	1.00
8	10	.91	38	9	.82
9	10	.91	39	10	.91
10	10	.91	40	11	1.00
11	10	.91	41	10	.91
12	9	.82	42	9	.82
13	10	.91	43	9	.82
14	9	.82	44	9	.82
15	9	.82	45	10	.91
16	10	.91	46	10	.91
17	10	.91	47	10	.91
18	10	.91	48	10	.91
19	10	.91	49	10	.91
20	10	.91	50	10	.91
21	11	1.00	51	11	1.00
22	11	1.00	52	11	1.00
23	11	1.00	53	11	1.00
24	11	1.00	54	10	.91
25	10	.91	55	11	1.00
26	11	1.00	56	11	1.00
27	11	1.00	57	11	1.00
28	11	1.00	58	11	1.00
29	11	1.00	59	11	1.00
30	11	1.00	60	11	1.00
			S-CVI = .95		

Overall, the I-CVI values for items that were developed for this instrument achieved the desired readings as suggested by the literature. The lowest value recorded was *.82* with the highest value of *1.00* indicating that all experts agree to the inclusion of the item to measure the intended construct. The S-CVI was also high at *.95* which indicates that the universal agreement of the experts towards all of the items in the instrument was almost perfect.

Table 5. Content Validity Ratio of the Instrument

Items	Ne (Rated essential)	CVR	Items	Ne (Rated essential)	CVR
1	11	1.00	31	11	1.00
2	11	1.00	32	11	1.00
3	11	1.00	33	11	1.00
4	10	.82	34	11	1.00
5	11	1.00	35	11	1.00
6	11	1.00	36	11	1.00
7	11	1.00	37	11	1.00
8	10	.82	38	9	.64
9	10	.82	39	10	.82
10	10	.82	40	11	1.00
11	10	.82	41	10	.82
12	9	.64	42	9	.64
13	10	.82	43	9	.64
14	9	.64	44	9	.64
15	9	.64	45	10	.82
16	10	.82	46	10	.82
17	10	.82	47	10	.82
18	10	.82	48	10	.82
19	10	.82	49	10	.82
20	10	.82	50	10	.82
21	11	1.00	51	11	1.00
22	11	1.00	52	11	1.00
23	11	1.00	53	11	1.00
24	11	1.00	54	10	.82
25	10	.82	55	11	1.00
26	11	1.00	56	11	1.00
27	11	1.00	57	11	1.00
28	11	1.00	58	11	1.00
29	11	1.00	59	11	1.00
30	11	1.00	60	11	1.00

The overall CVR values for the items were all above the baseline or the cutoff as suggested by literature. Unlike CVI where the experts rate the items based on their judgment of the item's ability to measure the construct, CVR calculates the expert's agreement on the relevance of the inclusion of the items to measure the construct. With the lowest value of *.64* and the highest value of *1.00*, this indicates that the experts agreed with the inclusion of the items to measure English language proficiency. Nevertheless, it is wise to note that there are a few items that scored low, even though accepted, that needed to be used with precautions. This means that the items were found to be almost at the grey area where it could be or could not be measuring the intended CEFR level.

These items are Item 12, 14, 15 and 38. Items 12, 14 and 15 were grammar items and Item 38 was vocabulary items. Even though the CEFR with the use of English Grammar Profile has outlined the use of different word classes for different CEFR levels, the item developers have to always bear in mind that at times the item may be developed wrongly and could be testing a different topic or subject. This is common when the item developers intended to test grammar but in reality the item is testing vocabulary and vice versa. For future use and to ensure more rigorous results, these items were suggested to be revised to ensure a clearer objective of the test; whether to test grammar or vocabulary. Nevertheless, the decision to retain the items is fair as long as the CVI value falls within the accepted range.

Figure 2. Example of Confusion between Grammar and Vocabulary Item

---

They _____ their claim by adding that as people lose weight, their BMR (basal metabolic rate) drops and the calorie is harder to maintain.

    A. certify
    B. concur
    C. extend
    D. endorse

The item (Figure 2) was developed to test the students' ability at B2 level (Present Simple - can use simple present tense for reporting verbs) but it may also be testing the vocabulary where the test takers have to understand the meaning of each word rather than using the reporting verb accurately. Secondly, the options given also were not developed according to the intended CEFR level (i.e. B2) with 'certify' and 'concur' are unlisted in the English Vocabulary Profile and 'endorse' is listed as C2 - 'extend' was appropriate for B2 level. Therefore, as item developers for CEFR-informed instruments, they have to be meticulous when writing items for grammar and vocabulary.

CONCLUSION

This research highlights the importance of content validation with a straightforward method with the use of content validity index and content validity ratio. As mentioned and supported by Massitah Kipli and Ahmad Zamri Khairani (2020), this approach helps item developers to quantify experts' agreement towards developing a valid instrument. The indices that serve as solid and objective proofs should also be used widely in the development of new instruments. This supports the idea of utilising this model by teachers in school who may not have the needed statistical knowledge to conduct a more complicated method for content validation provided that they can at least locate one CEFR expert within the jurisdiction of the education district office.

With the results of this study, it can be concluded that content validity analysis could be done with the help of the experts and simple analyses. As mentioned in the discussion, developing items without stringent vetting process by the experts may resort to the use of items that may not measure the intended CEFR level or could be irrelevant for the assessment. Some items, though accepted by the majority of the experts, may possess threats to the validity if they were not carefully developed. This provides invaluable insights to future item developers to be careful and mindful when testing grammar and vocabulary. As literature iterated that language assessment could be complex and the validating process could be manifold, this could be a steppingstone for item developers of CEFR-informed tests to ensure the validity thus moving towards better assessment with teacher-made tests.

References

Adom, D., Adu-Mensah, J., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education (IJERE)*, *9*(1), 109. https://doi.org/10.11591/ijere.v9i1.20457

Ahmad Tamizi Abu, Rabiatul Adawiah Ahmad Rashid, & Salmiza Saleh. (2021). Al-Alaq Teaching Module: The Process for Evaluating Content Validity and Its Feasibility. *Journal of Contemporary Social Science and Education Studies (JOCSSES)*, *1*(2), 175–184. Retrieved from https://jocss.com/index.php/multidiscipline/article/view/7

Ahorsu, D. K., Lin, C.-Y., Imani, V., Saffari, M., Griffiths, M. D., & Pakpour, A. H. (2020). The fear of covid-19 scale: Development and initial validation. *International Journal of Mental Health and Addiction*, *20*(3), 1537–1545. https://doi.org/10.1007/s11469-020-00270-8 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. https://doi.org/10.1016/j.sapharm.2018.03.066

Aryadoust, V., Ng, L. Y., & Sayama, H. (2020). A comprehensive review of rasch measurement in Language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Babo, R., V. Babo, L., T Suhonen, J., & Tukiainen, M. (2020). E- assessment with multiple-choice questions: A 5 year study of students' opinions and experience. *Journal of Information Technology Education: Innovations in Practice*, *19*, 001–029. https://doi.org/10.28945/4491

Cohen-Swerdlik (2009). *Psychological Testing and Assessment: An Introduction to Tests and Measurement (7ᵗʰ ed.)*. The United States of America: McGraw-Hill.

Haladyna, T.M., (2004). *Developing and Validating Multiple-Choice Test Items (3ʳᵈ ed.).* London: Lawrence Erlbaum.

Hasimah Ja'afar. (2020, February). Discussion on the Current Instrument for Entrance Examination. Personal Interview.

Kamarul Ariffin Ahmad, (2018). *Item Analysis and Score Dependability of English Proficiency 1 Examination in a Public University* (Master's Dissertation). University of Malaya, Kuala Lumpur.

Kamarul Ariffin Ahmad, & Zahari Suppian. (2022). International and National Teacher Education Conference (INTEC 2022). In *Teacher Education in The Next Normal* (pp. 117–126). Hat Yai; Prince of Songkla University.

Kementerian Pendidikan Malaysia, Malaysia Education Blueprint, 2013-2025: Preschool to post-secondary education (2013). Putrajaya.

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

Massitah Kipli, & Ahmad Zamri Khairani. (2020). Content Validity Index: An application of validating CIPP instrument for programme evaluation. *Int Multidiscip Res J*, *2*(4), 31-40.

Mohd Effendi, Ahmad Zamri, Hisyamsani Idris & Normawati Rahman. (2017). Kesahan Kandungan Pakar Instrumen IKBAR Bagi Pengukuran AQ Menggunakan Nisbah Kesahan Kandungan.

Nor Ashiquin C. Alih, Masdinah Alauyah Md Yusoff, & Abdul Halim Abdul Raof. (2020). Teachers' Knowledge and Belief on the CEFR Implementation in Malaysian ESL Classroom. International Journal of Multidisciplinary and Current Educational Research (IJMCER), 2(5), 126–134.

Norshahira Osman, Norashida, S. R., & Lukman, Z. M. (2021). Content validity determination of the suicidal ideation behaviour assessment instruments using content validity ratio (CVR) formula. *Journal of Drug Delivery and Therapeutics*, *11*(4-S), 81–85. https://doi.org/10.22270/jddt.v11i4-s.4968

Rachmat, N. A., & Arfiandhani, P. (2019). "I use multiple-choice question in most assessment I prepared": EFL teachers' voice on summative assessment. *ETERNAL (English, Teaching, Learning, and Research Journal)*, *5*(1), 163. https://doi.org/10.24252/eternal.v51.2019.a13

Shrotryia, V. K., & Dhanda, U. (2019). Content validity of assessment instrument for employee engagement. *SAGE Open*, *9*(1), 215824401882175. https://doi.org/10.1177/2158244018821751

Zeraati, M., & Alavi, N. M. (2014). Designing and validity evaluation of quality of nursing care scale in Intensive Care Units. *Journal of Nursing Measurement*, *22*(3), 461–471. https://doi.org/10.1891/1061-3749.22.3.461

Zuraidah Mohd Don, & Mardziah Hayati Abdullah. (2019). *What the CEFR is and isn't*. Kementerian Pendidikan Malaysia. https://www.moe.gov.my/en/menumedia/printed-media/newspaper-clippings/what-the-cefr-is-and-isn-t-free-malaysia-today-27-mei-2019