English Language Teaching; Vol. 15, No. 1; 2022
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Item Objective Congruence Analysis for Multidimensional Items

Content Validation of a Reading Test in Sri Lankan University

Fouzul Kareema Mohamed Ismail[1,2] & Ainol Madziah Bt Zubairi[3]

[1] Department of Curriculum and Instruction, Faculty of Education, International Islamic University Malaysia, Kuala Lumpur, Malaysia

[2] Department of English Language Teaching, Faculty of Arts and Culture, South Eastern University of Sri Lanka, Oluvil, Sri Lanka

[3] Department of Language and Literacy, Faculty of Education, International Islamic University Malaysia, Kuala Lumpur, Malaysia

Correspondence: Fouzul Kareema Mohamed Ismail, Department of Curriculum and Instruction, Faculty of Education, International Islamic University Malaysia, Kuala Lumpur, Malaysia & Department of English Language Teaching, Faculty of Arts and Culture, South Eastern University of Sri Lanka, Oluvil, Sri Lanka.

Received: December 3, 2021 Accepted: December 23, 2021 Online Published: December 24, 2021

doi: 10.5539/elt.v15n1p106 URL: https://doi.org/10.5539/elt.v15n1p106

Abstract

This paper presents the findings of a study that intended to seek the content validity (CV) evidence of an instrument to measure the reading ability of university students in Sri Lanka. The reading passages and items were adapted from CEFR aligned Learning Resource Network (LRN) materials. The items were designed based on the cognitive processing involved in completing each reading task as prescribed by Khalifa and Weir (2009). As a part of collecting evidence for content validation of the instrumentation, Item Objective Congruence (IOC) analysis is used in this study. In IOC, the congruence between the cognitive processing of reading and the test items were studied providing quantified data for CV. A pool of twelve experts examined a total of 41 test items against eight cognitive processing effectively. As the experts had chosen more than one objective for an item, the IOC formula simplified by Crocker and Aligna (1986) for multi-dimensional assessment of multiple combinations of skills was applied in the present study. The findings of the IOC indicate the experts' varying degrees of agreement in terms of what some of the items were designed to assess. 38 items had acceptable IOC indices, one item was removed from the study and two items were modified. Items having high congruence show that they test only one skill and those indicating low congruence notify that, items assess more than one cognitive processing skill. The study demonstrates the utility of the IOC method in gathering evidence for CV. Test development and validation are crucial in assessment which is the first and foremost process to evaluate educational management.

Keywords: content validation, test development, item objective congruence, assessing reading, Sri Lanka

1. Introduction

Reading is the most essential skill out of four language skills (Carrell, Devine, & Eskey, 2000; Grabe, 2009; Li & Wilhelm, 2008; Mermelstein, 2015). Whether, in the first language acquisition, or the second language learning, as knowledge is gained through reading, it is crucial in the development of the language proficiency of an individual. As the information or the knowledge is existing in text form in the world, according to Cimmiyotti (2013), reading has become a needed skill for students to master. Reading is necessary for building vocabulary and escorts to long-lasting education and improvement in the first along with second language skills. Consequently, mastery of reading ability is obligatory in schools to ensure successful learning in any subject (Nuttall, 1996). If a child is poor in reading ability, his performance in school life is inhibited (National Reading Panel, 2000). In higher education, reading is regarded to be one of the essential skills for successful academic study (Hermida, 2009). In the present virtual world, English has become the global lingua franca, thus it requires not only the native speakers but also the others to be fluent users of English. Therefore, English language teaching has become popular, and in this scenario, reading plays a vital role in learning, teaching, and assessment. For effective educational management, assessment is crucial. It brings out the mechanism to evaluate the

learning and teaching attainment of an institution. In this line, the results of this research can better shape educational management.

Reading alone is insufficient for one's successful personal and educational growth; one must also be able to correctly analyse, evaluate, reorganize, and break down ideas and facts that the writer intends to state. This process is regarded as reading comprehension (Grabe & Stoller, 2011; Mckee, 2012; Veeravagu, Muthusamy, Marimuthu, & Michael, 2010). As reading comprehension is involved in multiple processes, it is identified as a complex skill (Grabe & Stoller, 2011; Mckee, 2012; Pearson & Johnson, 1978). "A reading skill can be described roughly as a cognitive ability which a person is able to use when interacting with written texts'' (Urquhart & Weir, 1998, p. 88). Sometimes the terms reading 'skill' and 'strategy' can be interchangeably used (Grabe, 1991; Grabe & Stoller, 2011; Nuttall, 1985). However, the concept of considering reading as an underlined construct; emerged after the 2000s followed by contemplating it as a cognitive process by Pearson and Johnson (1978). They mentioned that a reader exerts different cognitive processes at each level of comprehending. As well word recognition is at the lower level of reading comprehension whereas the capacity to recognize the main ideas and to make inferences are assumed to be at higher level skills (Hudson, 2007; Khalifa & Weir, 2009; Urquhart & Weir, 1998).

In developing a reading assessment instrument, the objectives, or the constructs selected in this study are the eight cognitive processing proposed by Khalifa and Weir (2009). This study investigates to which extents these cognitive processes of reading can be identified, and to which level these processes (objectives) were agreed by experts in evaluating the items using the IOC method.

2. Literature Review

The theoretical background for this study is gained through the evidence supporting three stages of instrument development. The first part of the literature seeks information in developing the test instrument plus assessment of reading, and the second part focuses on content validation (CV). The last stage includes the evidence for evaluation of the CV using multi-dimensional objectives of IOC analysis. Figure 1 illustrates these stages involved in reviewing the literature for the present study.

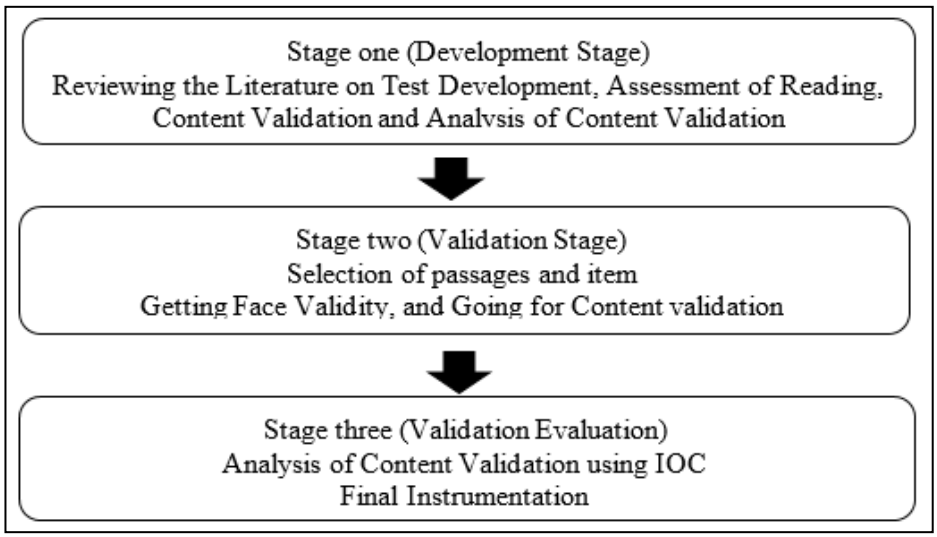

Figure 1. Theoretical Framework for Gathering Evidence for Content validation

2.1 Theoretical Framework

2.1.1 Test Development and Validation

Testing is compared to a type of architectural process, Fulcher's (2010, p. 94) "The test design cycle" is a good illustration of the endless procedure of test design. Assessment which is a continuous process plays an enormous role in the teaching-learning process (Popham, 1999, 2000). It helps teachers and learners to improve teaching and learning. The goal of a test is a major factor in how a test is developed and validated (Bachman & Palmer, 1996; Chapelle, Jamieson, & Hegelheimer, 2003). Chapelle et al. (2003) distinguished three such distinct objectives: use, infer, and effect. Focusing on these three purposes, and teaching, learning, and assessing conditions prevailing among the target population, the current study intends to develop a reading instrument.

Understanding what the tests are going to measure is a crucial factor in the development of a test design (Alderson, 2000). In this regard, the intended objective of the test is to determine reading proficiency. The

assessment of reading skills is primarily examined through the findings of the earlier works of Munby (1978), Grabe (2009; 1991), Grabe and Stoller (2011), Koda (2005), Hudson (2007), Khalifa and Weir (2009), Urquhart and Weir (1998), Carrell et al. (2000), and the like.

The table of specifications recommended by Alderson (2000) and Fulcher (2010) is pivotal in the construction of test items. The test designers define the content or ability domain through the format of the "table of specifications" from which they decide on items and test tasks to develop tests. Therefore, factors influencing test content are thus an important part of both test development and test use. Hence, an analysis of the test content is significantly meaningful for validation (Bachman, 1990, p. 244). The table of specifications developed for this study was designed focusing on the parameters of cognitive processing skills, types of reading, and item format (test method). The test items were developed on the construct of assessing the socio-cognitive process of reading powered by Khalifa and Weir (2009).

2.1.2 Assessing Reading Comprehension

A better understanding of reading facilitates its assessment (Alderson, 2000). Reading comprehension is a complex multifaceted process, and assessing it is also a challenging effort. Measuring reading has been a problem to test developers, material designers, and teachers (Mckee, 2012). According to Alderson (2000), reading test development is affected by the reader variable and the text variable. Reader's background knowledge, linguistic competence, cultural awareness, physical, psychological characteristics, strategies they use, as well as the reader motivation are the factors affecting the reader variable. Content and title of the text, text type, and item format, text readability index including word and sentence count, and grammatical level, etc., influence the text variable (Alderson, 2000; Khalifa & Weir, 2009; Kobayashi, 2002; Urquhart & Weir, 1998).

Generally, based on the structure, style, purpose, and discourse mode, texts are broadly divided into the narrative and expository types (Weaver & Kintsch, 1991). Expository texts provide information about a particular topic, or information that requires deep reading rather than reading for pleasure, or entertainment that needs shallow comprehension (Zhang & Duke, 2008). Narrative texts belonging to the nature of reading for pleasure; are written using the past tense, in the form of stories employing temporal sequence and applying common ordinary vocabulary (Medina & Pilonieta, 2006). However, there is no specified structure ensued in expository writing.

The main purpose of reading is comprehension. Comprehension is assessed through the student's ability to recollect the facts of what they have read (Allington, 2001). As comprehension is an unobserved behaviour, assessing reading comprehension can be made through the careful application of testing techniques. Alderson (2000, p. 202) mentions that this "test technique" should be fair with all students. Selected response (SR) and constructed response (CR) are the two classifications of the test method. Multiple choice questions (MCQ), true/false or right/wrong/ not given items, and gapped test are among the formats in the SR method. The CR includes "short answer questions (SAQ)", "cloze" and "gap filling", "information transfer", and "reading into writing" types (Khalifa & Weir, 2009, pp. 87-91).

Table 1. Cognitive Processing Skills of Reading in Khalifa and Weir (2009)

Socio-Cognitive Processing	**Description**
Word Recognition (WR)	The reader classifies the same word in question or determines a word meaning independently and matches it in the text. This occurs at the word level.
Lexis Access (LA)	The reader uses morphological knowledge of a word to identify synonyms, antonyms, hypernyms, or other related words and matches it in the text. This occurs at the word level.
Syntactic Parsing (SP)	The reader employs syntactic (grammatical) knowledge to establish comprehension to identify answers without logical problems. This can occur at the clause or sentence level.
Establishing Propositional (core) Meaning (EPM)	The reader expeditiously uses morphological and syntactic to determine the meaning of a sentence at the local level. It is a literal understanding of what is on the page. This occurs at the sentence or clause level.
Inferencing (I)	The reader goes beyond literal or explicitly stated meaning to infer a further significance. The reader can selectively read the paragraphs for main ideas and implicitly expressed ideas in the text. This can occur at the sentence level, paragraph level, or text level.
Building a Mental Model (BMM)	In a comparative and contrastive text type, the reader recognizes important distinctions and uses numerous features of the text to form a wider mental model. This occurs at a whole text level.

Creating a Text Level Structure (CTLS)	The reader applies genre expertise to determine the text's structure and purpose by analysing and distinguishing major ideas from supporting details. A trained reader decides how the various sections of the text work together, and which parts of the text are vital to the intent of the author or the audience. This occurs at the text level.
Creating an Inter-Textual Representation (CITR)	Understanding the text and comparing it across other texts. This occurs beyond the text level.

Furthermore, to develop the parameters for a reading test, a test designer must first determine the constructs or subskills of reading that the test intends to measure with the understanding of aforesaid text and reader variables. Understanding the reading constructs is a must for item writing. The eight socio-cognitive skills informed by Khalifa and Weir (2009) have been widely used to guide item writing when designing the items. Starting from word recognition (WR) to creating an inter-textual representation (CITR) the hierarchical representation of cognitive processing skills goes higher. In other words, WR is the easiest skill whereas CITR is the hardest among them all. According to Khalifa and Weir (2009), the first four skills (WR, LA, SP, and EPM) belong to low order thinking (LOT) skills, and the rest fit to high order thinking (HOT). Table 1 elaborates the cognitive processing skills proposed by Khalifa and Weir (2009).

2.1.3 Content Validation

Content validity (CV) is the minimum quality requirement for an instrument development at the item development stage (Halek, Holle, & Bartholomeyczik, 2017). CV means “the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose” Haynes, Richard & Kubany, 1995, p. 238). In simple, it can be referred that a test should be able to measure what it intends to measure as it has been highlighted by Turner and Carlson that “An important component in test development is providing evidence that the items created are measuring the content or construct they are defined to measure” (2003, p. 164).

CV is done through expert judgment, in other words, it can be obtained through the involvement of a group of subject matter experts (SME) thinking about the significance of individual items within an instrument (Creswell, 2012; Crocker & Algina, 1986). Moreover, CV is crucial for an assessment tool when the results are utilized as evidence in the selection of the examinee for the enrolment of an educational or occupational prospect, or promotion (Wilson, Pan, & Schumsky, 2012). The judgment of the experts is utilized to confirm the appropriateness and accountancy of the item. Furthermore, to define the construct, it is essential “to engage the expertise of a subject matter specialist in the design and the development of the language test” (Bachman & Palmer, 1996, p. 96).

Lawshe's technique proposed in 1975, has been frequently used to set up and evaluate content legitimacy in different fields (Ayre & Scally, 2014). Hence, Crocker and Algina (1986) added three more steps to Lawshe’s method suggesting four main steps to identify the content validity of an instrument. The CV processes consist of defining the performance domain of interest, selecting a trained panel of experts in the subject area, providing a systematic framework for rating or matching items to the performance domain, and obtaining and summarizing data from the rating or matching process (ibid).

Though CV is crucial in measurement, still its procedures are not well discussed in many research studies. Content validation cannot be written within a paragraph referring that the items are good enough according to expert judgment as it was discussed in many research surveys (Crocker, 2003). Therefore, the present study applied a thorough system to validate the content.

2.1.4 Evaluation Through Item Objective Congruence

An evaluation of the efficacy of items in measuring one or more objectives, accomplished by an unbiased expert panel provides evidence of CV. The most significant assessment at this stage, according to Berk (1984), is determining if items and objectives are congruent. The remaining item analyses are meaningless if there is insufficient proof that the items are measuring what they are supposed to measure. The index of item-objective congruence (IOC) introduced by Rovinelli & Hambleton (1977) is one method to quantitatively measure content experts' judgments of items to evaluate the fit between test items and the table of specifications (Berk, 1984; Turner, Mulvenon, Thomas & Balkin, 2002; Turner & Carlson, 2003). Further, Turner and Carlson (2003) cite Berk (1984) “that an evaluation of the match between items and objectives is the most important assessment during the content validation stage” (p. 164).

IOC is a process in which SMEs rate individual items on the degree to which they agree or do not agree with the specific objectives listed by the test developer (Turner et al., 2002). Accordingly, an expert evaluates each item

by giving a rating of 1 for clearly measuring objective, -1 for not clearly measuring, or 0 for the unclear objective. After the experts rate the items, the results are calculated to create the indices of IOC for each item on each objective. Rovinelli and Hambleton's (1977) formula is used under the assumption that an item measures only one objective.

In a condition where an item measures more than one objective, the multidimensional item formula simplified by Crocker and Aligna is utilized to evaluate the similarity between an item and a set of objectives (Turner et al., 2002; Turner & Carlson, 2003).

3. Methodology

Concerning time frame and cost effect, reading test materials from Learning Resource Network (LRN) were adapted instead of developing a new instrument. The LRN is a globally recognized awarding organization accredited by Ofqual in England and by the UK British Council since 2011. Firstly, a preliminary investigation on the selection of passages was carried out before the construction of the test instrument. Although the passages were selected according to Common European Frameworks of Reference (CEFR) levels, the Flesch Kincaid reading ease analysis using Text Inspector software for readability analysis of the passages; was conducted to confirm the levels of reading difficulty.

Table 2. Readability index according to Text Inspector Analysis

Passage No	Text Title	LRN CEFR Level	Text Inspector CEFR level	Flesch Reading Ease	Token count	Sentence count
1	Ice Cream	B1	B1+	71.93	150	8
2	The Tradition of Coffee Drinking	B2	C2	52.26	405	16
3	Supersonic Flight	C1	C1+	50.13	396	18
4	The Shard	IELCA(M/L)	C1	56.65	659	35

Table 2 presents a summary of the readability indices of the selected passages. Later, the passages along with the items were reviewed by researchers along with two other experts in reading to make sure whether the items measure the intended cognitive processing skills of reading, for instance, whether the items check for WR, LA, SP, and the rest. Finally, four passages ranging from B1 to C1 CEFR levels of LRN passages were adopted along with 38 items. Three items were constructed by the researchers. All passages have a maximum of nine, ten, or eleven questions. Almost all the passages selected in the tests belong to the purpose of reading for orientation (expository passages). Since the target population samples are ESL adult learners, they require reading for orientation to be motivated towards academic achievement.

Since collecting evidence for CV is crucial for test development (Alderson, 2000), it was decided to go for expert judgment. The request for being the expert judgment was sent around thirty experts who have ample experience in English language teaching, language testing, assessment of reading, and development of test materials around the world considering the recommendations suggested by Crocker, Llabre, and Miller (1988) in appointing the expert panel. Among them, twelve experts accepted to participate in the validation process. A summary of the research, a letter from the Postgraduate (P.G) office of the Faculty of Education elaborating the appointment of an expert panel, testlet with the key, and the IOC rating sheet were attached at the first email correspondence to the experts. (See Table 3 for example of the IOC rating sheet for passage 1, there were three similar rating sheets employed for the remaining 3 passages).

Table 3. IOC rating sheet for Passage 1 in Test 1

Text & title	level	Item No	Items	*Please assign a rating from 1 to 3 **Socio-Cognitive Reading Skills							*Please assign a rating from 1 to 3 ***Type of Reading				Remarks
				WR	LA	SP	EPM	I	BMM	CTLS	CITR	Ex/Loc	Ex/Glob	Ca/Loc	Ca/Glob	
CEFR B1 Ice-Cream		1	For questions **1-10**, choose the best answer (A, B, or C) and **fill in the gaps**. …..……. a hot summer day A. on B. in C. at													
		2	……. where does ice cream come from? A. wondered B. wondering C. wonder													
		3	……. the Chinese A. from B. of C. by													
		4	……was originally made A. He B. It C. What													
		5	…… of the freezer in the 20th century A. inventing B. invention C. inventor													
		6	……. was made from milk A. what B. whose C. which													
		7	……. no way to store A. is B. was C. will be													
		8	……. in great quantities A. sold B. cleared C. done													
		9	……one of the most popular desserts A. became B. becoming C. become													
		10	……. flavours and colours A. acceptable B. accurate C. many													

Accordingly, a passage was validated by a minimum of three experts, as feedback from at least three judges for each task is recommended for better rater performances (Crocker et al., 1988; Fulcher, 1997). Therefore, according to the above endorsement, a pool of twelve experts was honestly involved in the validation process. Since the present research has been carried out during the COVID-19 pandemic, the researcher utilized the maximum facilities provided by the virtual learning environment. In between the validation process, a minimum of fifteen to twenty email communications was conducted between the researcher and some of the raters. Alderson, Clapham and Wall (1995, p. 63) mention that "they must take each item as if they were the students", thus, it took them a considerable amount of time to give their judgment. With some raters, the researcher had online meetings to clarify the doubts that arose while validating. The researcher is very much grateful for the services provided by the expert panels.

Table 4. Descriptions of the Subject Matter Experts (SME)

Demographic information (Variables)		N	%
Affiliation	IIUM	3	25.0
	UniMaS	1	8.3
	UniSZA	1	8.3
	UTM	1	8.3
	University of Bedfordshire - CRELLA-UK	1	8.3
	Uni of Kelaniya -SL	2	16.7
	Uni of Colombo- SL	1	8.3
	SEUSL-SL	2	16.7
Qualification	Post-Doctoral	1	8.3
	PhD.	9	75
	PhD reading	1	8.3
	M.A reading	1	8.3
Teaching & Language Testing experience	>=30 years	2	16.7
	20-29 years	3	25
	10-19 years	3	25
	0-9 years	4	33.3
Gender	Male	1	8.3
	Female	11	91.7

Table 4 shows brief information of the experts participating in the CV. Through this content validation of the above expert panel, it is admissible to the researcher to answer the question raised by Haynes et al. (1995) that "does the test measure what it intends to measure?"

3.1 Data Analysis

The ratings of the SME for the present research indicate that there are items that measure more than one objective. Therefore, the multidimensional item formula simplified by Crocker and Aligna (1986) was handled to calculate the indices of IOC. The adjusted formula for the multidimensional item is as follows:

$$I'_{ik} = \frac{(N)\,\mu_k - (N-p)\,\mu_l}{2N - p}$$

In this formula, where $\boldsymbol{I'_{ik}}$ is the "index of item-objective congruence for item $\boldsymbol{I}$" on a set of objectives $\boldsymbol{k}$. Here $\boldsymbol{N}$ represents the number of objectives, $\boldsymbol{p}$ symbolises the number of valid objectives, $\boldsymbol{\mu_k}$ indicates the judges' mean rating of item $\boldsymbol{i}$ on the valid objectives $\boldsymbol{k}$. Further, $\boldsymbol{\mu_l}$ is the judges' average rating of item $\boldsymbol{i}$ on the invalid objectives $\boldsymbol{l}$ (Crocker & Aligna (1986) as cited in Turner & Carlson, 2003, p. 169).

Table 5. Example of IOC indices

Item No	*Index of IOC*	Objectives (average ratings of three experts)							
		1(WP)	2(LA)	3(SP)	4(EPM)	5(I)	6(BMM)	7(CTLS)	8(CITR)
11	0.978	**1.00**	-0.67	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00
12	0.578	-0.67	-0.33	-1.00	**0.33**	-1.00	-1.00	-1.00	-1.00
13	1.00	-1.00	-1.00	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00

Using this formula, the items of this survey were analysed to receive the accepted congruence as shown in Table 5. This is an example of comparing interpretations for the first three items of the second passage of the test. This table includes the serial number of the item, the index value for IOC (valid objective), as well as the mean ratings of the three experts on each objective.

In the current research, raters were briefed on the procedure of rating and assigned a rating from 1 to 3 for each item for its objectives (based on socio-cognitive skill and type of reading of Khalifa and Weir (2009) as follow:

1) refers to that the item definitely measures the objective
2) refers to uncertainty whether the item measures the objective
3) refers to that the item does not measure the objective

They were not informed "which item is meant to be matched with which objective" as recommended by Osterfind (1997, p. 259). Hence, the judges freely measured the items. Once they completed the task, their ratings were then coded by the researchers according to the criteria suggested by Turner and Carlson (2003). For example, in Table 5. Item 11 indicates that all experts approved that the item is evaluating Objective 1 and not measuring Objectives 3, 4, 5, 6, 7, and 8, as well as the item, has a high IOC value indicating .978. In this study, Objective 1 is Word Recognition constructed as WR as well as Objective 2 refers to LA. Similarly, SP, EPM, I, BMM, CTLS, and CITR are recognized by Objective 3 to 8 respectively. However, two of three experts believed that the item is a measure of Objective 2. Comparably, Item 12 has a valid IOC reporting at .578 which is an accepted value according to Brown (2005), Supparerkchaisakul, Mohan and Fansler (2017), and Takwin, Pansri, Parnichparinchai and Vibulrangson, (2018), although Pengruck, Boonphak and Sisan, (2019) used IOC indices of 0.60 to 1.00 as accepted values. However, Brown (2005) mentions that if the index of the IOC is between 0.5 and 1.00, it suggests that the item is acceptable, but if IOC falls below 0.5, it means that the item is not fitting and must be removed or reviewed.

Further, this view was proved by the designers of the IOC, Rovinelli and Hambleton (1977, pp. 15-16) that "if an item to be a perfect match to an objective, while the others were not able to make a decision, the computed value of the index would be 0.50". It means that Item 12 was agreed by two experts that it assesses Objectives 4. Furthermore, it was disagreed by two experts that it is not clearly measuring Objective 1, but an expert was uncertain whether this item measures Objective 1. If any of the experts are uncertain of the objective whether the item measures it or not, that time the congruence value indicates 0.0. They all mentioned that this item is not measuring Objectives 3, 5, 6, 7, and 8. If an objective was disagreed by two experts, and agreed by one, this objective has a congruence of -0.33 as it was attained by Item 12 for Objective 2. Consequently, Item 13 has a high IOC index reporting at 1.00. In this case, all experts agreed this item targets Objective 5 and does not measure Objectives 1, 2, 3, 4, 6, 7, or 8.

3.2 Limitations and Directions for Future Research

In developing a test, the three types of validity, for instance, content validity, construct, and criterion-related validity are considered, however, in the present study only the evidence for content validity is achieved. Evidence for other validity factors can be gained in future research. Despite the limitations, this study provides ample information on CV using IOC for multiple objectives.

4. Results and Discussion

As shown in Table 6 the highlighted number indicates that the item measures the particular cognitive processing skill. The values of columns three (WR) to ten (CITR) indicate the average rating of all three experts for each cognitive processing skill (objective or the underlined construct of the study). Out of 41 items, only 3 items were identified with low IOC indices of less than 0.50. These items were further investigated. This resulted from a scenario in which judges either rated more than one objective or were not certain of the valid socio-cognitive skills for those items. Nevertheless, this is normal as some questions may be categorised as measuring more than one cognitive category while there are more objectives as there are eight in the present study (Kim, 1996).

Table 6. Sample IOC Indices According to the Adjusted Formular for the Present Study

NO		WR	LA	SP	EPM	I	BMM	CTLS	CITR	P	μk	μl	N	IOC	Result
1	1	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
2	2	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
3	3	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
4	4	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
5	5	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
6	6	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
7	7	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
8	8	**0.33**	-1.00	-0.33	-1.00	-1.00	-1.00	-1.00	-1.00	1	0.33	-0.905	8	0.6	Accepted
9	9	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
10	10	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
11	1	**1.00**	-0.67	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1	1	-0.952	8	0.978	Accepted
12	2	-0.67	-0.33	-1.00	**0.33**	-1.00	-1.00	-1.00	-1.00	1	0.33	-0.857	8	0.578	Accepted
13	3	-1.00	-1.00	-1.00	-1.00	**1.00**	-1.00	-1.00	-1.00	1	1	-1	8	1	Accepted
30	1	-1.00	-1.00	-1.00	-1.00	-0.67	**0.33**	-0.33	-0.33	1	0.333	-0.762	8	0.533	Accepted
31	2	-1.00	-1.00	-1.00	-1.00	-0.67	**0.33**	-0.33	-0.33	1	0.333	-0.762	8	0.533	Accepted
32	3	-1.00	-1.00	-1.00	-1.00	-0.67	**0.33**	-0.33	-0.33	1	0.333	-0.762	8	0.533	Accepted
33	4	-1.00	-1.00	-1.00	-1.00	-0.67	**0.33**	-0.33	-0.33	1	0.333	-0.762	8	0.533	Accepted
34	5	-1.00	-1.00	-1.00	-1.00	-0.67	**0.33**	-0.33	-0.33	1	0.333	-0.762	8	0.533	Accepted
35	6	-1.00	-1.00	-1.00	-1.00	-1.00	0.00	**1.00**	-0.67	2	0.5	-0.944	8	0.69	Accepted
36	7	-1.00	-1.00	-1.00	-1.00	-0.67	**1.00**	-0.33	-1.00	1	1	-0.857	8	0.933	Accepted
37	8	-1.00	-1.00	-1.00	**0.33**	-1.00	-0.33	-0.33	-1.00	1	0.333	-0.810	8	0.556	Accepted
38	9	-0.67	-0.67	-0.67	**0.00**	-1.00	-1.00	-0.33	-1.00	1	0	-0.762	8	0.356	Unaccepted
39	10	-1.00	-1.00	-1.00	-1.00	**0.33**	-0.33	-0.33	-1.00	1	0.333	-0.810	8	0.556	Accepted
40	11	-1.00	-1.00	-1.00	-0.33	**0.33**	-0.33	-0.33	-1.00	1	0	-0.713	8	0.333	Unaccepted
41	12	-1.00	-1.00	**0.33**	-1.00	-1.00	-1.00	-0.67	-0.33	1	0	-0.857	8	0.4	Unaccepted

Table 6 illustrates the IOC indices of the experts relating to items 1-13, and 30-41 measuring the cognitive processing skills of reading which is a little different from the way how Jusoh (2018) applied IOC for a reading test instrument for evaluating five objectives. In this analysis, as can be seen from Table 6, three items (Item 38, 40, and 41) had invalid IOC indices less than 0.50. These items were further investigated with the consultation of the experts. Finally, Items 40 and 41 were removed from the test, and Item 38 was modified before the test was piloted.

Items with unacceptable IOC values test the cognitive skills like EPM as in Item 38, I as in Item 40, and SP in Item 41 as per the current judgement of the experts for these items. From the congruence between these unacceptable items and objectives, researchers were unable to give a watertight implication that there is a correlation between the objectives (cognitive processing skills) and the unacceptable IOC indices.

Almost all the items were able to reach an agreement that they clearly measure or somewhat measure the intended socio-cognitive skills of reading. SP was identified to be the cognitive processing found to have the most agreement. 13 items measured SP out of which 11 items scored an IOC value of 1 which is the highest value for agreement. Moreover, BMM, WR, LA, and EPM scored above 0.9 IOC index, however, in many cases there was confusion among raters in agreeing whether an item tested I, BMM, or CTLS as can be predicted from Items 30 to 34. As well in some cases, some of them faced difficulties in differentiating between the cognitive processing skills like SP and WR as can be seen in Item 8. Two raters agreed this item tested WR whereas one of them agreed that it tested SP.

To summer up, compared to the items testing LOT (low cognitive processing), items measuring HOT (high cognitive processing) seem to be the most difficult to be agreed on among raters. Therefore, this item analysis has implications on item writing and training of item writers as the high cognitive processing was found to be the most difficult to be agreed on. Further, this study has consequences on the validity of these types of items when they are tested.

Table 7. Summary of Cognitive Processing Skill of the Test According to IOC Indices

WR	LA	SP	EPM	I	BMM	CTLS	CITR	Total	LOT	HOT
2	1	13	9	5	10	1	0	41	25	16

Table 7 summarises the number of items under each cognitive processing skill employed in the test passages. It shows that the test maintains a balance between the LOT and HOT skills.

5. Conclusion

The index of IOC has shown to be a valuable tool for test developers in determining item validity prior to the pilot test administration. Nonetheless, Rovinelli and Hambleton's (1977) formula is appropriate only for the unidimensional objective. The innovations in multidimensionality assessment methods, plus multiple constructs for creating multidimensional items, have geared the demand for content validity evaluation procedures for items measuring multiple objectives. The adjusted IOC index formula of Croker and Aligna (1986) can be used for both unidimensional and multidimensional objectives.

The use of experts' judgment in the IOC method for validity evidence ensures that test items assess the construct that they were designed to measure. Validity in assessment is of utmost importance as test setters are accountable for the test they set. In the other sense, assessment of reading facilitates the learning and teaching process to take decisive directions by evaluating the course of an institution, which in return should certainly enhance educational management better.

References

Alderson, C. J., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Alderson, J. C. (2000). *Assessing reading*. Cambridge Assessment English. https://doi.org/10.1017/CBO9780511732935

Allington, R. (2001). *What really matters for struggling readers: Designing research-based programs*. Longman.

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

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford university press.

Bachman, L. F., & Palmer, a. S. A. S. (1996a). *Language Testing in Practice*. Oxford University Press. https://doi.org/10.2307/328718

Bachman, L. F., & Palmer, A. S. (1996b). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Berk, R. (1984). Conducting the item analysis. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 97-143). Johns Hopkins University Press.

Brown, J. D. (2005). *Testing in language programs: a comprehensive guide to English language assessment*. McGraw-Hill College.

Carrell, P. L., Devine, J., & Eskey, D. E. (2000). *Interactive approaches to second language reading*. Cambridge University Press.

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20*(4), 409-439. https://doi.org/10.1191/0265532203lt266oa

Cimmiyotti, C. B. (2013). *Impact of reading ability on academic performance at the primary level* [Dominican University of California]. https://doi.org/10.33015/dominican.edu/2013.edu.18

Creswell, J. W. (2012). *Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research*. Pearson.

Crocker, L. (2003). Teaching for the Test: Validity, Fairness, and Moral Action. *Educational Measurement: Issues and Practice, 22*(3), 5-11. https://doi.org/10.1111/j.1745-3992.2003.tb00132.x

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. ERIC.

Crocker, L., Llabre, M., & Miller, M. D. (1988). The Generalizability of Content Validity Ratings. *Journal of Educational Measurement, 25*(4), 287-299. https://doi.org/10.1111/j.1745-3984.1988.tb00309.x

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing, 14*(2), 113-139. https://doi.org/10.1177/026553229701400201

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

Grabe, W. (2009). *Teaching and Testing Reading*. In Michael H. Long & Catherine J. Doughty (Eds.), *The Handbook of Language Teaching*. https://doi.org/10.1002/9781444315783.ch24

Grabe William, & Stoller, F. L. (2011). *Teaching and researching reading* (2nd ed.). Routledge. https://doi.org/10.1002/9781405198431.wbeal1174

Grabe, William. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*(3), 375-406. https://doi.org/10.2307/3586977

Grabe, William. (2009). *Reading in a second language: Moving from theory to practice*. Ernst Klett Sprachen. https://doi.org/10.1017/CBO9781139150484

Halek, M., Holle, D., & Bartholomeyczik, S. (2017). Development and evaluation of the content validity, practicability and feasibility of the Innovative dementia-oriented Assessment system for challenging behaviour in residents with dementia. *BMC Health Services Research, 17*(1). https://doi.org/10.1186/s12913-017-2469-8

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment, 7*(3), 238-247. https://doi.org/10.1037/1040-3590.7.3.238

Hermida, J. (2009). The Importance of Teaching Academic Reading Skills. *The International Journal of Research and Review, 3*, 20-30. Retrieved from https://julianhermida.com/teachingshowcasereading.pdf

Hudson, T. (2007). *Teaching second language reading*. Oxford University Press.

Jusoh, Z. (2018). *A Rasch Analysis Of Reading Skill Across Text Type And Item Format* (Doctoral dissertation).

Khalifa, H., & Weir, C. J. (2009). *Examining reading: research and practice in assessing second language reading* (Vol. 29). Cambridge University Press.

Kim, H. (1996). Assessing attainment of Bloom's cognitive levels using testlets and multi categorical IRT. *ERA-AARE Joint Conference*.

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*(2), 193-220. https://doi.org/10.1191/0265532202lt227oa

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524841

Li, H., & Wilhelm, K. H. (2008). Exploring Pedagogical Reasoning: Reading Strategy Instruction From Two Teachers' Perspectives. *The Reading Matrix, 8*(1), 96-110.

Mckee, S. (2012). Reading Comprehension, What We Know: A Review of Research 1995 to 2011. *Language Testing in Asia, 2*(1), 45-58. https://doi.org/10.1186/2229-0443-2-1-45

Medina, A. L., & Pilonieta, P. (2006). *Once upon a Time: Comprehending Narrative Text*.

Mermelstein, A. D. (2015). Reading Level Placement and Assessment for ESL/EFL Learners: The Reading Level Measurement Method. *ORTESOL Journal, 32*, 44-55.

Munby, J. (1978). *1978: Communicative syllabus design*. Cambridge: Cambridge University Press.

National Reading Panel. (2000). *Report of the National Reading Panel*. National Institute of Child Health and Human Development.

Nuttall, C. (1985). Survey Reviews: Recent materials for the teaching of reading. *English Language Teaching Journal, 39*. https://doi.org/10.1093/elt/39.3.198

Nuttall, C. (1996). *Teaching reading skills in a foreign language* (new ed.). Heinemann.

Osterfind, S. J. (1997). *Constructing test items: Multiple-choice, constructed-response, performance and other formats*. Kluwer Academic Publishers.

Pearson, P., & Johnson, D. (1978). *Teaching reading comprehension*. New York: Holt, Rinehart & Winston.

Pengruck, L., Boonphak, K., & Sisan, B. (2019). Early childhood education: A confirmatory factor analysis concerning Thai administrators' creative administration. *Asia-Pacific Social Science Review, 19*(1), 17-32. Retrieved from https://apssr.com/wp-content/uploads/2019/03/RA-2.pdf

Popham, W. J. (1999). *Classroom assessment: What teachers need to know*. Allyn & Bacon.

Popham, W. J. (2000). Assessing mastery of wish-list content standards. *NASSP Bulletin, 84*(620), 30-36. https://doi.org/10.1177/019263650008462004

Rovinelli, R. J., & Hambleton, R. K. (1977). On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity. *Dutch Journal of Educational Research, 2*, 49-60.

Supparerkchaisakul, N., Mohan, K. P., & Fansler, K. (2017). Developing a Scale for University Citizenship Behavior: Thai and US Academic Contexts. *International Journal of Behavioral Science, 12*(2), 71-89. Retrieved from https://so06.tci-thaijo.org/index.php/IJBS/article/view/80221

Takwin, M., Pansri, O., Parnichparinchai, T., & Vibulrangson, S. (2018). Developing a Self-Assessment Instrument for Analysis of the Social and Personal Competencies of Teachers in Senior High Schools in Indonesia. *Journal of Physics: Conference Series, 128, 2nd International Conference on Statistics, Mathematics, Teaching, and Research* 2017 9-10 October 2017, Makassar, Indonesia. https://doi.org/10.1088/1742-6596/1028/1/012088

Turner, R. C., Mulvenon, S. W., Thomas, S. P., & Balkin, R. S. (2002). Computing indices of item congruence for test development validity assessments. *27th Annual SAS Users' Group International Conference*, Miami, United States.

Turner, Ronna, C., & Carlson, L. (2003). Indexes of Item-Objective Congruence for Multidimensional Items. *International Journal of Testing, 3*(2), 163-171. https://doi.org/10.1207/S15327574IJT0302_5

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a Second Language: Process, Product and Practice* (G.N. Candlin (ed.)). Longman.

Veeravagu, J., Muthusamy, C., Marimuthu, R., & Michael, A. S. (2010). Using Bloom's taxonomy to gauge students' reading comprehension performance. *Canadian Social Science, 6*(3), 205-212. https://doi.org/10.3968/j.css.1923669720100603.023

Weaver, C. A., & Kintsch, W. (1991). Expository text. In R. Barr et al. (Eds.), *Handbook of Reading Research*. Vol. 2. New York: Longman.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development, 45*(3), 197-210. https://doi.org/10.1177/0748175612440286

Zhang, S., & Duke, N. K. (2008). Strategies for Internet reading with different reading purposes: A descriptive study of twelve good Internet readers. *Journal of Literacy Research, 40*(1), 128-162. https://doi.org/10.1080/10862960802070491