Teachers' Knowledge in HOTS (TKHOTS) Instrument: Content Validity Index

Nani Herlin Jamin, Shahlan Surat, Wan Muna Ruzana Wan Mohammad

The National University of Malaysia, 43600 Bangi, Selangor, Malaysia
Email: nanininja82@gmail.com

Abstract
This study examines the content validity of the Teachers' Knowledge in Higher Order Thinking Skills Instrument (TKHOTS) for Malaysian secondary teachers by calculating the Content Validity Index (CVI), which includes CVI for the item (I-CVI) and CVI for scale (S-CVI). We appointed three professional university experts to validate the instrument. The instrument involved 63 items with three main domains. The I-CVI for all the items of the three domains ranged from 0.75 to 1. The average of S-CVI for domain planning, implementation and evaluation were 1, 0.95 and 0.99, respectively. The overall S-CVI was above 0.80, which indicated high content validity of the items and proved that TKHOTS has a great potential to be promoted as the best instrument to measure teachers' knowledge in HOTS teaching and learning. However, more following analysis, including reliability, construct validity, and criterion-related validity, should be implemented to better the instrument's applicability.
Keywords: Teacher Knowledge, Validity, Higher Order Thinking Skills, Instrument, Content Validity Index

Introduction
The effectiveness of teachers in mastering knowledge relating to Higher Order Thinking Skills (HOTS) and pedagogical skills are professional knowledge and should be one of the characteristics of 21st-century teachers. Therefore, a valid instrument to measure teachers' level of knowledge in HOTS needs to be developed to review the extent of teachers' mastery of HOTS. In Malaysia, specific instruments for examining teachers' knowledge of HOTS, in general, are limited. Most studies measure teachers' knowledge in HOTS based on subject domains such as Malay Language (Isa & Mahamod, 2021) and Islamic Education (Noh et al., 2017).

Teachers' Knowledge in Higher Order Thinking Skills Instrument (TKHOTS) for Malaysian secondary teachers begins with reviewing existing standard instruments implemented in Malaysia's schools. These instruments are 1) Malaysian Education Quality Standards Second Wave and, 2) HOTS Schools Rating Standard Instrument The Ministry of Education Malaysia developed these instruments in 2017 to cultivate HOTS in teaching and learning. Based on it, we constructed 63 items with three thematic domains. In addition, we modified the items in TKHOTS regarding the context of language delivery aimed at measuring

teachers' knowledge. As a result, we believe TKHOTS can be the best instrument to measure teachers' level of knowledge in HOTS because we referred to the standard framework that supports the domain and items. Nevertheless, the development of TKHOTS needs to verify through an expert view. Hence, we intend to conduct content validation through Content Validity Index (CVI) as an essential step in the instrument development.

Literature Review
Content Validity Index
Validity is a critical aspect of quantitative research. To consider the instrument used is good, the reliability and validity of the instrument are vital (Popham, 1989). The instrument's validity is established if it provides a measure of what it measures. Validity denotes the extent to which items on a tool accurately assess the concept being measured in the research study (Masuwai et al., 2016). Even if the instrument is reported as having good one-dimensionality and reliability, poor content validity will jeopardize the instrument's psychometric utility. Furthermore, the content validation process will guarantee the instrument to have defensible, accurate, and appropriate (Furr, 2011).

Yusoff (2019) stated the five sources of validity evidence, i.e., content, response process, internal structure, relation to other variables and consequence. In educational research, Kaplan and Bush (1982) suggested four types of validity, i.e., face validity, content validity, construct validity and criterion-related validity. This study mainly addresses content validity to ensure that the questionnaire's items meet each thematic domain as defined

The instrument's measurement of content validity is crucial in instrument development (Ramli et al., 2018). Content validity is defined as the degree to which elements of an assessment instrument are relevant to and representative of the targeted construct for a particular assessment purpose (Haynes et al., 1995). The importance of content validity in the instrument psychometric and its relevance with reliability have made it an essential step in the instrument development (Zamanzadeh et al., 2015). Therefore, this study emphasized one of the critical steps in instrument development: assessing content validity using CVI. The CVI can represent the content validity evidence (Davis, 1992; Polit et al., 2007).

Several previous studies have shown the use of CVI to support the validity of the instruments. The instruments, such as 'Anatomy Education Environment Measurement inventory' (Hadie et al., 2017), 'Assessment Instrument for Employee Engagement' (Shrotryia & Dhanda, 2019), 'Instrument for measuring Patient-Centered Communication (Zamanzadeh et al., 2015) and 'Teaching and Learning Guiding Principles Instrument' (Masuwai et al., 2016) used CVI to establish the content validity. We proposed this method to the TKHOTS instrument to expand the CVI-related studies.

TKHOTS Instrument
TKHOTS was developed by performing three steps, as Zamanzadeh et al. (2015) stated – identifying the content domain, generating the sample items, and constructing the instrument. The content domain is the content area related to the measured variables (Beck & Gable, 2000). First, we conducted an extensive review of the literature to determine the content domain for teachers' knowledge in HOTS. A clear relationship was seen between teachers' knowledge and student development (Isnon & Badusah, 2017), teaching practices in the classroom (Mariani & Ismail, 2013) and teachers' beliefs (Soliman, 2017). Then, relating to the context of the study, we did a review of educational policy on HOTS worldwide. Based on reviews and a brief discussion with Malaysian policy researchers, we can conclude that

HOTS in teaching and learning involves three domains, namely planning, implementation, and evaluation. Each stage contains actions and qualities that need to be implemented and measured.

In the second step, we generated the sample items by analysing existing standard instruments implemented in Malaysia's schools, namely 1) Malaysian Education Quality Standards Second Wave and 2) HOTS Schools Rating Standard Instrument). We classified the items into three domains, i.e., planning, implementation and evaluation. We also modified the items in sentence structure to align with the research purpose, i.e., teachers' knowledge. To determine the content domain in HOTS based on policy and the standard HOTS assessment, we use a specifications table, as suggested by Lynn (1986). The table of specifications reviews the alignment of a set of items (placed in rows) with the concepts forming the construct under study (placed in columns) through collecting the evidence and by analysing data (Zamanzadeh et al., 2015). Lynn (1986) mentioned that the instrument construction is the third step in which the items are refined and organised in a suitable format and sequence. Finally, we collected the finalised items in a usable form. Therefore, we constructed the 63-items, which were eight items in planning, 38 items in implementation and 17 items in the evaluation domain. We organised these items in the form of a questionnaire with 5 points Likert scale, starting from (1) Strongly disagree, (2) Disagree, (3) Somewhat Agree, (4) Agree and (5) Strongly Agree.

Validating the items by experts
The gist of the study was validating the items by a specific number of experts. The expertise panels must possess extensive knowledge and demonstrate a good grasp of the related subject (Masuwai et al., 2016). Among the selection criteria for the experts are having a background in the research area, possessing related working experience, being diverse in giving an opinion, and having up-to-date knowledge (Powell, 2003). Their function is to deeply review each of the items proposed before deciding whether to retain or remove it (Ramli et al., 2018). However, the recommended number of experts has always been inconsistent. The minimum acceptable expert number is two (Davis, 1992; Yusoff, 2019), three (Shrotryia & Dhanda, 2019), and five (Zamanzadeh et al., 2015). The maximum number of experts has not been confirmed yet, thus involving more than ten experts is irrelevant. As Polit and Beck (2021) stated, increasing the number of experts decreases the chances of agreement. Yusoff (2019) suggested the number of experts for content validation should be at least six and does not exceed 10.

The viewpoints of the experts, both quantitative and qualitative, can prove the content validity. These views include the instrument's relevance, necessity, representativeness, and comprehensiveness. There are several methods, such as CVI (Zamanzadeh et al. 2015), Content Validation Ratio CVR (Ramli et al., 2018), and Kappa statistic (Shrotryia & Dhanda, 2019), can be used as quantification. Therefore, the adequacy of the final content of the instrument would be based on the collective opinion of these experts based on their professional assurance (Masuwai et al., 2016).

Methodology
The content validity was based on six steps recommended by Yusoff (2019); (1) Preparing content validation form, (2) Selecting a review panel of experts, (3) Conducting content validation, (4) Reviewing domain and items, (5) Providing score on each item, and (6) Calculating CVI. The first step, the content validation form (refer to Appendix A), was prepared

to ensure the expert panels have clear expectations and understanding about the task. The rating scale of relevance, as recommended by Yusoff (2019), Davis (1992), Polit et al. (2007) and Lynn (1986), has been used for scoring individual items. They were asked to rate each scale item at a four-point scale along the item rating continuum, 1 = not relevant, 2 = somewhat relevant, 3 = quite relevant, and 4 = highly relevant. A four-point scale avoided a neutral point (Shrotryia & Dhanda, 2019). It is also recommended to define each domain to facilitate the scoring process.

In the second step, the expert review panels were selected based on the individual expertise with the topic to be studied. In this research, we selected four professional experts from four universities in Malaysia based on the mentioned criteria, i.e., three in Educational Psychology and one in Evaluation and Measurement. They are still active in research and publication in their respective fields. Table 1 shows the list of expert panels information.

Table 1
List of Expert Panels

No	Initial and Code	Expertise	Gender	University
1.	Assoc Porf NA (P1)	Educational Psychology	Female	Mara University of Technology, Selangor
2.	Dr. SYL (P2)	Educational Psychology	Female	University of Malaya, Kuala Lumpur
3.	Dr. SN (P3)	Educational Psychology	Female	Islamic Science University of Malaysia, Negeri Sembilan
4.	Prof ZA (P4)	Evaluation and Measurement	Male	Sultan Zainal Abidin University, Terengganu

In the third step, the content validation was conducted, which can be done by the face-to-face or non-face-to-face approach. In this study, we decided to run by non-face-to-face by sending the content validation form to the experts through e-mail. We gave explicit instruction to facilitate the content validation process and did a systematic follow-up to improve the response rate and time. We took almost two months to collect the data (from July to September 2020). Yusoff (2019) stated that the response rate and time might be the challenging factor for the non-face-to-face approach because of difficulty to get a response on time and at risk of not getting a response at all from the expert. However, cost-saving is the most significant advantage.

In the fourth step, the expert panels reviewed the domain and items represented through the form. They were given a clear definition of the domain and requested to examine its items before critically scoring each of them. They also evaluated the items through written and verbal comments. The written comments can be stated in the last column in the form (see Figure 1), and the verbal was brought by telephone conversation (for P1) and face-to-face meeting (for P3). All comments are taken into consideration to refine the domain and its items. Next, the experts were requested to independently provide their viewpoints on each item based on the relevant scale. Finally, the experts submitted the complete content validation form by e-mail (for P1 and P2), post (for P4), and by hand (for P3).

Figure 1
Example of written comment given by the panel

The last step was calculating the CVI. CVI for the item (I-CVI) and CVI for scale (S-CVI) are two forms of CVI. I-CVI (item-level content validity index) is the proportion of content experts giving items a relevance rating of 3 or 4. The relevance rating must be coded as 1 (relevance scale of 3 or 4) or 0 (relevance scale of 1 or 2). The formula stated as,

I-CVI = (agreed item)/ (number of expert)

Two methods for calculating S-CVI are S-CVI/Ave (scale-level content validity index based on the average method) and S-CVI/UA (scale-level content validity index based on the universal agreement method). S-CVI/Ave is the average I-CVI scores for all items on the scale or the average proportion relevance judged by all experts. The formula stated as,

S-CVI/Ave = (sum of I-CVI scores)/ (number of item)

S-CVI/UA is the proportion of items that all experts achieve a relevance scale of 3 or 4. Universal agreement (UA) score is given as 1 when the item reached 100% experts in agreement; otherwise, the UA score is given as 0. The formula stated as,

S-CVI/UA = (sum of UA scores)/ (number of item)

In this study, the cut-off score of CVI values should be 1 as recommended by Polit and Beck (2021) and Polit et al. (2007). In addition, the definition and formula were based on the recommendations by Lynn (1986), Davis (1992), Polit and Beck (2021) and Polit et al. (2007)

Results
The I-CVI for all the items of the three domains ranged from 0.75 to 1. The S-CVI/Ave for domain planning, implementation and evaluation were 1, 0.95 and 0.99, respectively. The value of I-CVI, S-CVI/Ave, and S-CVI/UA for all the three domains can be referred to in Tables 2, 3 and 4. Based on the results, we can conclude that overall S-CVI was above 0.80, which indicated high content validity of the items for the domain of TKHOTS. Therefore, it is recommended that a minimum S-CVI should be 0.8 for reflecting content validity (Lynn, 1986; Polit & Beck, 2021).

Table 2
I-CVI and S-CVI for 'Planning' domain

Domain: Planning

ITEM	P1 (NBA)	P2 (SYL)	P3 (SNH)	P4 (PZA)	EXPERTS IN AGREEMENT	I-CVI	UA
Pn1	1	1	1	1	4	1	1
Pn2	1	1	1	1	4	1	1
Pn3	1	1	1	1	4	1	1
Pn4	1	1	1	1	4	1	1
Pn5	1	1	1	1	4	1	1
Pn6	1	1	1	1	4	1	1
Pn7	1	1	1	1	4	1	1
Pn8	1	1	1	1	4	1	1
					S-CVI/AVE	1	
					S-CVI/UA		1

Table 3
I-CVI and S-CVI for 'Implementation' domain

Domain: Implementation

ITEM	E1 (NBA)	E2 (SYL)	E3 (SNH)	E4 (PZA)	EXPERTS IN AGREEMENT	I-CVI	UA
Pl1	1	1	1	1	4	1	1
Pl2	1	1	1	1	4	1	1
Pl3	1	1	1	1	4	1	1
Pl4	1	1	1	1	4	1	1
Pl5	1	1	1	1	4	1	1
Pl6	1	1	1	1	4	1	1
Pl7	1	0	1	1	3	0.75	0
Pl8	1	1	1	1	4	1	1
Pl9	1	1	1	1	4	1	1
Pl10	1	1	1	1	4	1	1
Pl11	1	0	1	1	3	0.75	0
Pl12	1	0	1	1	3	0.75	0
Pl13	1	1	1	1	4	1	1
Pl14	1	1	1	1	4	1	1
Pl15	1	1	1	1	4	1	1
Pl16	1	1	1	1	4	1	1
Pl17	1	0	1	1	3	0.75	0
Pl18	1	0	1	1	3	0.75	0
Pl19	1	1	1	1	4	1	1
Pl20	1	1	1	1	4	1	1
Pl21	1	1	1	1	4	1	1
Pl22	1	1	1	1	4	1	1

Pl23	1	0	1	1	3	0.75	0
Pl24	1	1	1	1	4	1	1
Pl25	1	1	1	1	4	1	1
Pl26	1	1	1	1	4	1	1
Pl27	1	1	1	1	4	1	1
Pl28	1	1	1	1	4	1	1
Pl29	1	1	1	1	4	1	1
Pl30	1	0	1	1	3	0.75	0
Pl31	1	1	1	1	4	1	1
Pl32	1	1	1	1	4	1	1
Pl33	1	1	1	1	4	1	1
Pl34	1	1	1	1	4	1	1
Pl35	1	1	1	1	4	1	1
Pl36	1	1	1	1	4	1	1
Pl37	1	1	1	1	4	1	1
Pl38	1	1	1	1	4	1	1
					S-CVI/AVE	**0.95**	
					S-CVI/UA		**0.82**

Table 4
I-CVI and S-CVI for 'Evaluation' domain

Domain: Evaluation

ITEM	E1 (NBA)	E2 (SYL)	E3 (SNH)	E4 (PZA)	EXPERTS IN AGREEMENT	I-CVI	UA
Pn1	1	1	1	1	4	1	1
Pn2	1	1	1	1	4	1	1
Pn3	1	1	1	1	4	1	1
Pn4	1	1	1	1	4	1	1
Pn5	1	1	1	1	4	1	1
Pn6	1	1	1	1	4	1	1
Pn7	1	1	1	0	3	0.75	0
Pn8	1	1	1	1	4	1	1
Pn9	1	1	1	1	4	1	1
Pn10	1	1	1	1	4	1	1
Pn11	1	1	1	1	4	1	1
Pn12	1	1	1	1	4	1	1
Pn13	1	1	1	1	4	1	1
Pn14	1	1	1	1	4	1	1
Pn15	1	1	1	1	4	1	1
Pn16	1	1	1	1	4	1	1
Pn17	1	1	1	1	4	1	1
					S-CVI/AVE	**0.99**	
					S-CVI/UA		**0.94**

Discussion

87 percent of all items in TKHOTS achieved the I-CVI of 1.0, which met the cut-off value as suggested by Polit and Beck (2021) and Polit et al. (2007). 13 percent of items achieved the I-CVI of 0.75, and we modified it according to the recommendation of panel experts. Item P7 was suggested to drop even though the I-CVI was 0.75 because it consists of an unclear and confusing statement. In conclusion, 62- items of TKHOTS retained from the original version were deemed relevant to the thematic domains, i.e., planning, implementation and evaluation based on the high CVI.

Conclusion

Content validity is vital to ensure an instrument's overall validity; therefore, content validation needs a systematic approach based on the evidence and best practice (Yusoff, 2019). Understanding content validity is essential for the researchers because they should realize if the instruments, they use for their studies are suitable for the construct, population under study, and socio-cultural background in which the study is carried out (Zamanzadeh et al., 2015). More following analyses, including reliability, construct validity, and criterion-related validity, should be implemented to better the instrument's applicability.

This study examines the content validity of the TKHOTS instrument for Malaysian secondary teachers using CVI. As HOTS becomes essential in education, the extent of the reliability and validity of the TKHOTS has critical implications for the various stakeholders of education institutions. As a result, TKHOTS has a great potential to be promoted as a better instrument to measure teachers' knowledge in HOTS teaching and learning. The need to look at the content validity of a scale is important. Therefore, this study is seen to contribute to the quality of instruments that measure teachers' level of knowledge in HOTS. Furthermore, through this quality instrument, it is hoped that the measurements carried out will produce valid data findings and can be generalized to the actual population of the study.

References

Beck, C. T., & Gable, R. K. (2000). Postpartum Depression Screening Scale: Development and Psychometric Testing. *Nursing Research*, *49*(5), 272–282.

Che Noh, M. A., Ajmain, M. T., & Engku Abdul Rahman, E. Z. (2017). Teachers' Practice of Higher Order Thinking Skills in the Lesson of Islamic Education. *Tinta Artikulasi Membina Ummah*, *3*(2), 1–13.

Davis, L. L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, *5*(4), 194–197.

Furr, R. M. (2011). *Scale Construction and Psychometrics for Social and Personality Psychology*. SAGE Publications.

Hadie, S. N. H., Hassan, A., Ismail, Z. I. M., Asari, M. A., Khan, A. A., Kasim, F., Yusof, N. A. M., Manan Sulong, H. A., Tg Muda, T. F. M., Arifin, W. N., & Yusoff, M. S. B. (2017). Anatomy education environment measurement inventory: A valid tool to measure the anatomy learning environment. *Anatomical Sciences Education*, *10*(5), 423–432.

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.

Isnon, H., & Badusah, J. (2017). Kompetensi Guru Bahasa Melayu Dalam Menerapkan Kemahiran Berfikir Aras Tinggi Dalam Pengajaran Dan Pembelajaran. *Jurnal Pendidikan Bahasa Melayu*, *7*(1), 56–65.

Kaplan, R. M., & Bush, J. W. (1982). Health-related quality of life measurement for evaluation research and policy analysis. *Health Psychology*, *1*(1), 61–80.

Lynn, M. R. (1986). Determination and Quantification of Content Validity. *Nursing Research*, *35*(6).

Mariani, A., & Ismail, Z., (2013). Pengaruh Kompetensi Guru Matematik Ke Atas Amalan Pengajaran Kreatif, *2nd International Seminar on Quality and Affordable Education (ISQAE 2013)*

Masuwai, A., Tajudin, M. N., & Saad, N. S. (2016). Evaluating the face and content validity of a Teaching and Learning Guiding Principles Instrument (TLGPI): A perspective study of Malaysian teacher educators. *GEOGRAFIA OnlineTM Malaysian Journal of Society and Space*, *12*(3), 11–21.

Mat Isa, N. S., & Mahamod, Z. (2021). Tahap Pengetahuan, Sikap Dan Masalah Guru Bahasa Melayu Terhadap Penerapan Kemahiran Berfikir Aras Tinggi Dalam Pengajaran Dan Pembelajaran Komsas. *Asian People Journal (APJ)*, *4*(1), 93–107.

Polit, D., & Beck, C. (2021). *Essentials of Nursing Research: Appraising Evidence for Nursing Practice* (Tenth, North American ed.). LWW.

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

Popham, J. W. (1989). *Modern Education Measurement: A Practitioner's Perspective (Second Edition)* (Facsimile, Subsequent ed.). Prentice Hall.

Powell, C. (2003). The Delphi Technique: myths and realities. *Methodological Issues in Nursing Research*, 41(4), 376–382.

Ramli, N. F. B., Talib, O. B., Manaf, U. K. B. A., & Hassan, S. A. B. (2018). Content Validity of STEMTIP Using CVR Method. *International Journal of Academic Research in Business and Social Sciences*, *8*(7).

Shrotryia, V. K., & Dhanda, U. (2019). Content Validity of Assessment Instrument for Employee Engagement. *SAGE Open*, *9*(1).

Soliman, M. S. S. (2017). Primary Teachers' Beliefs and Knowledge about Self-regulated Learning in the Kingdom of Saudi Arabia. *International Journal of Educational Sciences*, *18*(2017).

Yusoff, M. S. B. (2019). ABC of Content Validation and Content Validity Index Calculation. *Education in Medicine Journal*, *11*(2), 49–54.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. R. (2015). Design and Implementation Content Validity Study: Development of an instrument for measuring Patient-Centered Communication. *Journal of Caring Sciences*, *4*(2), 165–178.

A: Example of Content Validation Form

KESAHAN INSTRUMEN PENGETAHUAN GURU DALAM PEMBELAJARAN DAN PEMUDAHCARAAN KEMAHIRAN BERFIKIR ARAS TINGGI

> **Instruction to panel experts**

YBhg Dr,

Instrumen ini mengandungi tiga sub konstruk dan 63 item yang berkaitan dengan pengetahuan guru dalam Pembelajaran dan Pemudahcaraan (PdPc) Kemahiran Berfikir Aras Tinggi (KBAT). Penilaian pakar pada setiap item dalam sub konstruk tersebut adalah diperlukan. Penilaian yang dibuat adalah berdasarkan definisi dan terminologi yang dikhususkan pada setiap sub konstruk. Pakar boleh menggunakan skala pengukuran yang berikut bagi membuat penilaian terhadap item:

> **Rating scale of relevance**

Darjah kerelevenan:

1 = item tidak relevan dengan sub konstruk yang diukur.

2 = item agak relevan dengan sub konstruk yang diukur.

3 = item hampir relevan dengan sub konstruk yang diukur.

4 = item sangat relevan dengan sub konstruk yang diukur.

> **Definition of each domain**

Sebarang komen dan cadangan hendaklah diisi dalam ruangan catatan.

Sub konstruk: Perancangan dalam PdPc KBAT

Definisi: Persepsi terhadap informasi atau maklumat berkaitan perancangan PdPc KBAT PdPc dari segi tindakan dan kualiti merangkumi penyediaan Rancangan Pengajaran Harian (RPH) dan bahan sumber pendidikan.

Item Cadangan	Darjah Kerelevenan				Catatan
Saya tahu perancangan PdPc KBAT melibatkan perkara seperti yang berikut.					
1. PdPc KBAT dirancang berdasarkan kepelbagaian aras keupayaan murid.	1	2	3	4	
2. PdPc KBAT dirancang berdasarkan peruntukan masa yang ditetapkan.	1	2	3	4	
3. PdPc KBAT dirancang berdasarkan ketetapan kurikulum/arahan yang berkuat kuasa.	1	2	3	4	
4. Peranan guru sebagai perancang bagi memastikan kesediaan dan persediaan yang rapi dalam melaksanakan PdPc.	1	2	3	4	