Advances in Social Science, Education and Humanities Research, volume 545
Proceedings of the International Conference on Educational Assessment and Policy (ICEAP 2020)

Development of Critical Thinking Skills Measurement in Socio-Political Context

Miftachur Rohmah*, Sri Kusromaniah
Faculty of Psychology
Universitas Gadjah Mada
Yogyakarta, Indonesia
*miftachur.rohmah@mail.ugm.ac.id

***Abstract*—Critical thinking is one of the most essential goals in education nowadays. Aside from academics and career settings, the skill is considered important to be implemented in understanding socio-political issues for shaping better civil society by emphasizing consciousness and accountability. This study aims to develop measurement for critical thinking skills in Indonesia's socio-political context. Tryout involves 87 students from eight universities in Yogyakarta. The data is analyzed using both classical test theory and item response theory to obtain more comprehensive psychometric properties. It results in strong content validity, model fit and 13 good quality items from total 145 constructed items, yet a low internal consistency. Thus, the final items are not adequate to be administered in an independent test, but can be alternative questions for item bank in critical thinking skills assessment. Moreover, suggestions for further research are provided.**

Keywords—critical thinking skills, sociopolitical issues, test development, Watson Glaser Critical Thinking Appraisal (WGCTA)

I. INTRODUCTION

According to the 2019 Democracy Index, Indonesia scored 6.48 averagely on five parameters thus given a status of flawed democracy [1]. In this case, gaining a score of 6.11 shows that the quality of political participation in Indonesia needs to be improved.

Based on a joint report from the Americas Barometer 2010, the Afrobarometer 2011–2012, Meredith Weiss 2013, and Pulse Asia 2013, among 58 countries, Indonesia ranks third in terms of the percentage of money politics in elections [2]. Meanwhile, the combined data from the LSI survey, Indonesian Political Indicators and SMRC involving 210,524 subjects from 2006 to 2009 show that 18% of respondents confessed that they would accept and choose political figures who gave money or gifts [2]. These data indicate a need to improve inner quality of political participation which focuses on individual competence. Political participation cannot be optimized if society does not have sufficient abilities to contribute to positive change [3].

The urgency to increase this quality is also strengthened by the rampant political campaigns through social media [4]. The use of social media in political campaigns has several negative consequences, such as the production and distribution of fake news or hoaxes [5]. Ministry of Communication and Information Technology shows that there are more than 800,000 sites spreading fake news or hoaxes in Indonesia [6]. In a survey by MASTEL [7] involving 1,116 people, 44% of respondents stated that they received hoax news every day, while 17% stated that they received hoax news more than once a day. As many as 91% of respondents reported that the types of hoax news received focused on socio-political issues. Hoaxes have several negative impacts, such as damaging the credibility and integrity of organizers and politicians competing in general elections, causing unrest or commotion in society, and dividing national unity [8]. Poor critical thinking skills are related to a person's inability to identify fake news [9]. Thus, it is suggested that the efforts to increase public participation should be carried out by encouraging people to think critically.

Critical thinking equips a person to participate in democratic life which involves the ability to consider options, seek alternative point of view, and gather information to make right decisions [10]. It has an indirect positive effect on the orientation and implementation of political participation [11].

II. CRITICAL THINKING SKILLS

A. The Concept of Critical Thinking Skills

Cognitive abilities play an important role in processing information and making decisions in individual socio-political participation, even in collective settings [10]. This is in line with the definition of critical thinking skills, a cognitive ability to understand situations from various perspectives while separating facts and assumptions [12].

This study uses cognitive psychology approach and skills dimension in interpreting critical thinking. Critical thinking is defined as the skill to understand situations from various perspectives, recognize assumptions from facts on events, solve problems, make decisions, and learn new concepts.

B. Measurement of Critical Thinking Skills

Regarding the urgency of political participation inner quality, it is advised to increase people's critical thinking skills. A further study to gain an empirical data from standardized measurements is needed to assess the state of critical thinking skills nowadays. The measurement needs to be aimed directly at the socio-political context. However, the currently available measuring tools are considered unfitted.

In terms of construction, Watson-Glaser critical thinking construct has been widely accepted and used [13,14]. Watson and Glaser refer critical thinking as the ability to look at a situation and clearly understand it from multiple perspectives while separating facts from opinions and assumptions [15]. Watson – Glaser Critical Thinking Appraisal (WGCTA) has five subscales: inference, recognizing assumptions, deduction, interpretation, and evaluation of arguments [16]. This construct is used in this study because of its two distinguishing indicators (recognizing assumptions and analysing arguments) are considered relevant to the urgency of applying critical thinking skills in socio-political studies.

C. Present Study

This study aims to develop a measuring tool for critical thinking skills as an evaluation of individual critical thinking skills in their application in socio-political studies, as well as a process of collecting accurate data to make analyses related to the role of cognitive abilities in socio-political participation. Samples are taken from college students as critical thinking is highly related to education [17,18]. This research is expected to be one of the foundations for a cognitive approach in improving the quality of democracy.

III. METHODS

This study equips non-experimental quantitative methods. The process of developing instrument is carried out in accordance with the guidelines from Hambleton and Jones [19] and Terwee et al. [20] by utilizing statistical analysis of classical test theory (CTT) and item response theory (IRT). CTT is used to represent whether a measuring tool is good and reliable as assessed from its validity and reliability index [21]. Meanwhile, IRT is used to overcome the shortcomings of CTT with measurement precision that depends on latent variables so that it is considered to be able to predict the level of difficulty of questions and the actual ability of respondents [22]. This combination strategy has been widely used in many studies [23,24] and is recommended to gain more comprehensive psychometric properties [25,26].

A. Item Construction

The construction stage starts from preparing a test specification that refers to the construct of Goodwin Watson and Edward Glaser's critical thinking skills. The ability to think critically is directly related to one's knowledge of the item topics [27]. One of the criteria for consideration in question construction is the purpose, domain, and context of each item [28]. Thus, the preparation of grain content begins with a study of popular socio-political issues in Indonesia through newspapers, government publications, YouTube discussion channels, social media, and educational publications which results in 25 issues.

Fig. 1 illustrates the item design. The presentation of each subtest consists of 3 parts: the issue, items, and answer choices. Instructions and examples of how to perform each subtest are shown on the page preceding the three sections. The instructions and samples used are the translation of Pearson's Watson - Glaser Critical Thinking Appraisal Practice Test [29]. Generally, each issue is followed by six items each followed by two answer choices.

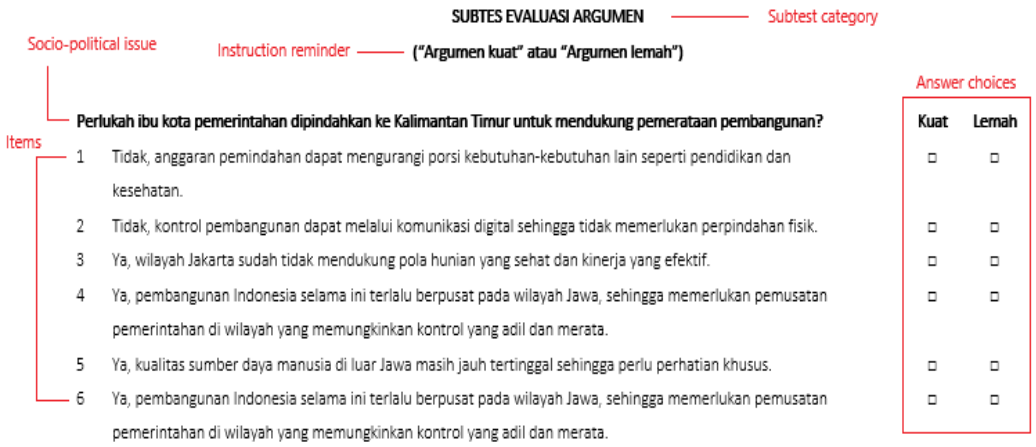

Fig. 1. Illustration of item design.

B. Face Validity

Face validity test is carried out through an assessment by an expert in the field of cognitive psychology. Considering the number of questions is too many for empirical testing, an expert suggested for eliminating several lengthy main scenarios. Thus, three issues followed by 17 items were aborted, leaving 128 items in 5 subtests.

C. Pilot Study

In the next stage, a pilot study was carried out to find 60 items for tryout. Purposive sampling method was used for easy access. Data collection involved 134 undergraduate students from the Department of Nutrition and Psychology, Universitas Gadjah Mada. The test duration was 10-15 minutes, with an additional 5 minutes for filling out the informed consent form and giving instructions, administered in a classical paper-and-pencil format. Through the analysis of the ITEMAN version 3.00 program, out of 128 total items, 12 items in each subtest were taken with consideration of the discrimination power and difficulty index.

D. Tryout

Tryout is an empirical test to test the validity and reliability of the instrument. The administered scale consisted of 60 items from the analysis of the pilot study which were divided equally into 5 subtests. The criteria for participants were active students, willing to spend 1 hour, and being present at the test location. As many as 87 students from eight universities in Yogyakarta participated in the tryout. The paper-and-pencil test is carried out in classrooms, either collectively or individually.

E. Content Validity Ratio (CVR)

A total of 60 items from the pilot study were tested for content validity using CVR which involved five experts in the field of psychology (two master students, one doctoral student, and two lecturers in the field of cognitive psychology). The CVR form that has been collected from the experts is processed using Excel software. With 5 raters, the validity value of the CVR approved is equal to or more than 0.99 [30].

F. Contruct Validity and Reliability

The statistical data processing procedure from the tryout was carried out after obtaining the content validity results. Thus, some data obtained from the tryout is not processed in statistical analysis if the item does not have good content validity based on CVR. Model fit and construct validity analysis was carried out through the one-factor confirmatory factor analysis (CFA) model using the demo version of MPLUS software. Those items then went through reliability tests with Cronbach's alpha.

G. Item Response Theory

Before executing IRT analysis, assumption tests of unidimensionality and local independence were carried out [31,32] using one-factor CFA model for categorical variables with WLS/WLSMV estimator. Then, the item analysis is carried out using robust maximum likelihood (MLR) estimator. The 2PL IRT analysis describes two technical properties: the level of item difficulty and the power of discrimination. Theoretically, the item difficulty index ranges from -4 to 4, but generally accepted values are between -2.80 and 2.80 [32]. Meanwhile, in the difference power parameter, a positive value means the large probability of answering correctly along with the high ability of the participants. Conversely, a negative score indicates that the likelihood of answering correctly decreases as the participant's ability increases [33].

IV. Results and Discussion

A. Participant

A total of 221 college students were involved in this research. This number is a combination of two stages data collection. The first stage was a pilot study involving 134 respondents from Universitas Gadjah Mada. The second stage was a tryout procedure which involved 87 students from eight universities in Yogyakarta.

B. Pilot Study Analysis

This part shows 12 items in each subtest which are selected from pilot study statistical analysis using ITEMAN. The main parameters for selecting items are the discrimination index and the difficulty level of the items.

TABLE I. Discrimination Index

Category	Subtests				
	DED	***EVA***	***INF***	***INT***	***REC***
Very good	1, 11, 15, 17	1, 2, 3, 8, 10, 11, 13, 14, 15, 18	2, 3, 11, 12, 14, 18, 20	2, 5, 7, 8, 9, 13, 18	4, 6, 9, 11, 16
Good	3, 5, 7, 8, 16	6	16, 17	3, 10, 16	10, 14, 15
Moderate	12	-	1, 5, 13	14	1, 3, 7, 17
Poor	4, 14	9	-	1	-
Total	12 items	12 items	12 items	12 items	12 items

Note: DED = Deduction, EVA = Evaluation of Argument, INF = Inference, INT = Interpretation, REC = Recognizing Assumption

The numbers represent item identities in each subtest. In Table I, items with 'very good' discrimination index become priority. However, these items are not necessarily selected because of the consideration from the balance of difficulty level and number of items in each issue.

TABLE II. DIFFICULTY LEVEL

Category	Subtest				
	DED	***EVA***	***INF***	***INT***	***ASU***
Moderate	1, 3, 7, 11, 12, 14, 17	1, 2, 3, 8, 9, 11, 13, 14, 15	2, 5, 14, 16, 17	1, 2, 7, 8, 9, 13, 14, 16	1, 3, 4, 6, 11, 15, 16
Very easy	4, 5, 8, 16	6, 10, 18	1, 3, 13, 18	3, 5, 10, 18	7, 9, 10, 14, 17
Very difficult	15	-	11, 12, 20	-	-
No one answered correctly	-	-	-	-	-
Total	12 items	12 items	12 items	12 items	12 items

Note: DED = Deduction, EVA = Evaluation of Argument, INF = Inference, INT = Interpretation, REC = Recognizing Assumption

In Table II, the proportion of difficulty level in each main scenario is preferred. Items that fall into the 'no one answered correctly' category are eliminated immediately. Meanwhile, items in the 'very easy' and 'very high' categories can be preferred if necessary, to equalize the difficulty level of other items in an issue.

C. Content Validity Ratio (CVR) Lawshe

Based on Lawshe's [30] recommendation, item validity is only accepted if the item's CVR is ≥ 0.99 for five raters. With the statistical data analysis procedure using Excel program, as many as 36 items had an index of 1 and are considered valid.

D. Tryout Analysis

After 60 items from the pilot study went through the content validity test, the tryout data from 36 items with good validity were processed using classical test theory and item response theory analysis.

1) Construct validity: One-factor model of confirmatory factor analysis is used to assume the relationship between 5 subtests and the variable critical thinking skills. The Kolmogorov Smirnov test showed that the data were not normally distributed (p <0.05). Thus, CFA analysis is carried out with a robust maximum likelihood (MLR) estimator because it is resistant to a small number of subjects and an abnormal data distribution [34,35].

TABLE III. ONE-FACTOR MODEL CFA FIT MODEL TEST

Cut-off Good Fit	Result
Chi-Square (p-value > 0,05)	X^2 3,198 (p-value 0,6695)
CFI ≥ 0,90 TLI ≥ 0,90	CFI 1 TLI 1
RMSEA ≤ 0,05	RMSEA 0
SRMR ≤ 0,08	SRMR 0,035
Conclusion	**Good Fit**

Table III shows the model fit test result compared with several good fit parameters according to Zhao and UCLA: Statistical Consulting Group. Thus, there is an agreement between the model from empirical data and the theoretical model [36].

The next step is measuring the construct validity of the tested model. Fig. 2 shows the factor loadings and p-values to make conclusions regarding the validity of the constructs following the CFA test.

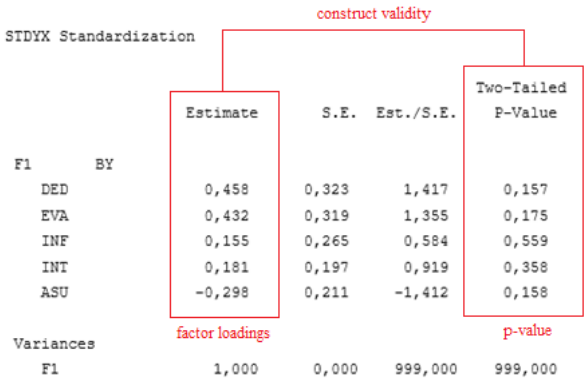

Fig. 2. Construct validity statistics.

Based on the statistical results in Fig. 2, the factor load value on all subtests is below 0.6, which indicates that all subtests are not part of the construct of the critical thinking skills variable [37]. However, the p-value for all factors is more than 0.05 which means that the conclusion of construct validity cannot be generalized to the population since the data obtained from the sample does not provide sufficient evidence.

2) Reliability: Cronbach's alpha analysis was performed on each subscale which results in a score of 0.037 for the deduction subtest; 0.374 for the evaluation of argument subtest; 0.436 for inference subtest; score of 0.214 for interpretation subtest; a score of 0.150 for the subtest recognizing assumptions; and a score of 0.344 for the five subtests combined. These values are under the criteria for a good score, which is ≥ 0.7 [38]. Thus, this data can be interpreted as the scale's reliability in this study has low internal consistency.

3) Unidimensionality and local independence assumptions: As illustrated in Table IV, four subtests, except for the deduction subtest, meet the assumptions of unidimensionality and local independence. Due to the limited data processing capacity and considering items which form a fit model, 14 variables were eliminated in this process.

TABLE IV. DISCRIMINATION INDEX

Cut-off Good Fit	Subtests				
	DED	***EVA***	***INF***	***INT***	***REC***
Chi Square p-value > 0.05	X^2 4.869 p-value 0.0876	X^2 10.690 p-value 0.2976	X^2 6.532 p-value 0.6857	X^2 5.139 p-value 0.8220	X^2 0.179 p-value 0.9144
CFI ≥ 0.90 TLI ≥ 0,90	CFI 0.600 TLI 0.000	CFI 0.938 TLI 0.897	CFI 1.000 TLI 1.000	CFI 1.000 TLI 1.000	CFI 1.000 TLI 1.000
RMSEA ≤ 0.05	RMSEA 0.128	RMSEA 0.046	RMSEA 0.000	RMSEA 0.000	RMSEA 0.000
Conclusion	Poor Fit	Good Fit	Good Fit	Good Fit	Good Fit

Based on these two logistical parameters (2PL) illustrated in Table V, Table VI, Table VII, and Table VIII, out of the 22 items in 4 subtests, there are five items with negative discrimination index and two of them have an abnormal level of difficulty, thus should be eliminated. In addition, there are four other items with positive discrimination index, yet outside the normal threshold, therefore those items should be eliminated to increase the precision of instrument. Thus, based on the item response theory analysis, there are 13 items that have good item characteristics.

TABLE V. EVALUATION OF ARGUMENT SUBTEST 2PL ANALYSIS

Item	Discrimination Index	Difficulty Level	Conclusion
EVA1	1.044	-0.569	Good
EVA2	0.910	-2.438	Good
EVA3	0.028	-30.044	Poor
EVA5	0.950	-1.492	Good
EVA6	1.404	-2.016	Good
EVA7	4.268	-0.809	Good

TABLE VI. INFERENCE SUBTEST 2PL ANALYSIS

Item	Discrimination Index	Difficulty Level	Conclusion
INF1	1.457	-0.201	Good
INF2	1.823	-1.846	Good
INF3	0.477	4.995	Poor
INF7	0.877	-0.032	Good
INF8	1.763	-1.034	Good
INF9	0.281	4.145	Poor

TABLE VII. INTERPRETATION 2PL SUBTEST ANALYSIS

Item	Discrimination Index	Difficulty Level	Conclusion
INT1	0.401	1.243	Good
INT2	0.080	-10.770	Poor
INT3	-0.344	2.044	Poor
INT6	-0.312	1.554	Poor
INT7	-0.061	12.054	Poor
INT8	-0.917	0.279	Poor

TABLE VIII. RECOGNIZING ASSUMPTION 2PL SUBTEST ANALYSIS

Item	Discrimination Index	Difficulty Level	Conclusion
ASU2	1.000	0.894	Good
ASU3	1.179	-1.737	Good
ASU4	-0.148	5.434	Poor
ASU7	1.177	-1.426	Good

From total 145 constructed items, there are 36 items with strong content validity and 13 good quality items, yet a low internal consistency. There are several suggestions for improvement, including involvement of item revision process following subject matter experts' suggestion [39] and empirical testing with a larger number of samples [40].

V. CONCLUSION AND IMPLICATIONS

The final scale is not adequate for assessment process, but the items can be included as item bank in critical thinking skills assessment. Therefore, this psychometric study still needs to be developed to support the quality of socio-political participation. In the future, the scale can be used to assess the ability of people who will be involved in certain political positions, such as political parties or governments.

ACKNOWLEDGMENT

Gratitude goes to Mrs. Sri Kusromaniah as the supervisor of this study as well as a rater in content validity test; members of Keluarga Aktivis Mahasiswa Muslim Indonesia Yogyakarta who have supported and provided information in the preliminary study; Mr. Thomas Dicky Hastjarjo as an expert in testing the face validity and research reviewer; Tengku Nurfa, Muhammad Azka Hifni, Mrs. Heru Astikasari, and Mrs. Maria Nugraheni as raters in testing the content validity; Mrs. Sutarimah Ampuni as research reviewer; and ultimately 221 college students who were involved in the process of collecting empirical data.

REFERENCES

[1] The Economist Intelligence Unit, Democracy Index 2019, 2020. Accessed on: Jul. 2, 2020. [Online]. Available: EIU, http://www.eiu.com/Handlers/WhitepaperHandler.ashx?fi=Democracy-Index-2019.pdf&mode=wp&campaignid=democracyindex2019

[2] B. Muhtadi, "Politik Uang dan New Normal dalam Pemilu Paska-Orde Baru," Jurnal Antikorupsi INTEGRITAS, vol. 5, no. 1, pp. 55-74, 2019.

[3] M. Okome, "Quality Vs. Quantity - Global and Local Responses Towards Increasing Women's Political Participation: Nigeria's 2015 Elections and Women's Political Participation," in Proceedings of the

Thematic Dialogue On 2015 General Elections and the Future of Women Political Participation, Jul. 29, 2020 Ikeja.

[4] I.A. Ratnamulyani and B.I. Maksudi, "Peran Media Sosial dalam Peningkatan Partisipasi Pemilih Pemula di Kalangan Pelajar di Kabupaten Bogor," Sosiohumaniora, vol. 20, no. 2, pp. 154–161, 2018.

[5] I. Syahputra, "Demokrasi Virtual dan Perang Siber di Media Sosial: Perspektif Netizen Indonesia," Jurnal ASPIKOM - Jurnal Ilmu Komunikasi, vol 3, no. 3, pp. 457-475, 2017.

[6] A. Pratama, "Ada 800 Ribu Situs Penyebar Hoax di Indonesia", 2020. Accessed on: Jul. 2, 2020. [Online]. Available: CNN Indonesia, https://www.cnnindonesia.com/teknologi/20161229170130-185-182956/ada-800-ribu-situs-penyebar-hoax-di-indonesia

[7] Masyarakat Telematika Indonesia, "Hasil Survey Mastel tentang Wabah Hoax Nasional," 2017. Accessed on: Jul. 2, 2020. [Online]. Available: Mastel, https://mastel.id/press-release-infografis-hasil-survey-mastel-tentang-wabah-hoax-nasional/

[8] A. Elcaputera and A.W. Dinata, "Penegakan Hukum Penyebaran Berita Bohong (Hoax) dalam Penyelenggaraan Pemilu 2019 Ditinjau dari Konsep Keadilan Pemilu," 2020. Accessed on: Jul. 2, 2020. [Online]. Available: KPU, https://journal.kpu.go.id/index.php/ERE/article/view/154

[9] P. Machete and M. Turpin, "The Use of Critical Thinking to Identify Fake News: A Systematic Literature Review," 2020. Accessed on: Jul. 20, 2020. [Online]. Available: doi: 10.1007/978-3-030-45002-1_20

[10] L. Pinto and J. Portelli, "The role and impact of critical thinking in democratic education: Challenges and possibilities," In Critical thinking education and assessment: Can higher order thinking be tested?, J. Sobocan, L. Groarke, R.H. Johnson & F. Ellett, Ed. London: Althouse Press, 2009, pp. 299-320.

[11] E.M. Guyton, "Critical Thinking and Political Participation: Development and Assessment of a Causal Model," Theory & Research in Social Education, vol. 16, no. 1, pp. 23-49, 1988.

[12] R.J. Sternberg and D.F. Halpern, "Conclusion: How to Think Critically about Politics … and Anything Else!" Critical Thinking in Psychology, pp. 354–376, 2020.

[13] J. H.McMillan, "Enhancing college students' critical thinking: A review of studies," Res High Educ 26, pp. 3–29, 1987.

[14] Oermann, H. Marilyn, Gaberson, and B. Kathleen, Evaluation and Testing in Nursing Education, 3rd ed. New York: Springer Publishing Company, 2009, pp. 131.

[15] Pearson Education, "WatsonGlaser™ III (W-G III) Frequently Asked Questions," 2018. Accessed on: Jul. 22, 2020. [Online]. Available: Talent Lens, https://us.talentlens.com/content/dam/school/global/TalentLens/us/watson-glaser-iii-faqs.pdf

[16] R. Zulmaulida, W. Sanusi, and J. Dahlan, "Watson-Glaser's Critical Thinking Skills," Journal of Physics: Conference Series, vol. 1028, no. 012094, 2018.

[17] J. Delors, Learning: The Treasure Within: Report to UNESCO of the International Commission on Education for the Twenty-first Century. Paris: UNESCO, 1996. Accessed on: Jul. 2, 2020. [Online]. Available: UNESCO, http://unesdoc.unesco.org/images/0010/001095/109590eo.pdf

[18] Kementrian Pendidikan dan Kebudayaan RI, "Konsep dan Implementasi Kurikulum 2013," 2014. Accessed on: Jul. 2, 2020. [Online]. Available: Kemdikbud, https://www.kemdikbud.go.id/kemdikbud/dokumen/Paparan/Paparan%20Wamendik.pdf

[19] R.K. Hambleton and R.W. Jones, "Comparison of classical test theory and item response theory and their applications to test development," Educational Measurement: Issues and Practice, vol. 12, no. 3, pp. 38–47, 1993.

[20] C. Terwee, S. Bot, M. Boer, D. van der Windt, D. Knol, J. Dekker, L. Bouter, and H. De Vet, "Quality criteria were proposed for measurement properties of health status questionnaires," Journal of Clinical Epidemiology, vol. 60, pp. 34-42, 2007.

[21] L.J. Cronbach, Essentials of Psychological Testing. New York: Harper Collins Publishers, 1990.

[22] R. Jabrayilov, W. Emons, and K. Sijtsma, "Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment," Applied psychological measurement, vol. 40, no. 8, pp. 559–572, 2016.

[23] S.P. Reise, K.F. Widaman, and R.H. Pugh, "Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance," Psychological Bulletin, vol. 114, no. 3, pp. 552–566, 1993.

[24] S. Bourion-Bédès, R. Schwan, J. Epstein, V. Laprevote, A. Bédès, J.L. Bonnet, and C. Baumann, "Combination of classical test theory (CTT) and item response theory (IRT) analysis to study the psychometric properties of the French version of the Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form (Q-LES-Q-SF)," Quality of Life Research, vol. 24, no. 2, pp. 287–293, 2014.

[25] P. Irwing, T. Booth, and D. Hughes, The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development. New Jersey: John Wiley & Sons Ltd, 2018.

[26] W.H. Finch, and B.F. French, Educational and Psychological Measurement, First publ. New York: Taylor & Francis, 2019.

[27] G. Bester and G.E. Pienaar, "Measuring critical thinking in a political context," South African Journal of Education, vol. 22, pp. 286, 2002.

[28] R. Rodríguez, R. Martínez, and J. Muñiz, "New Guidelines for Developing Multiple-Choice Items," Methodology: European Journal of Research Methods for The Behavioral and Social Sciences, vol. 2, pp. 65-72, 2006.

[29] The Psychological Corporation, Watson-Glaser Critical Thinking Appraisal–UK Edition Practice Test. London: Pearson Assessmen, 2002.

[30] C.H. Lawshe, "A Quantitative Approach to Content Validity," Personnel Psychology, vol. 28, no. 4, pp. 563-575, 1975.

[31] R.K. Hambleton, H. Swaminathan, and H.J. Rogers, Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publications, 1991.

[32] F.B. Baker, The Basics of Item Response Theory, 2nd ed. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation, 2001.

[33] M. Ul Hassan and F. Miller, "Discrimination with unidimensional and multidimensional item response theory models for educational data," Communications in Statistics - Simulation and Computation, pp. 1-21, 2019.

[34] B. Muthen, L. Muthen, and T. Asparouhov, "Estimator choices with categorical outcomes," 2015. Accessed on: Jul. 29, 2020. [Online]. Available: StatModel, https://www.statmodel .com/download/Estimator Choices.pdf

[35] Y. Rosseel, "Mplus estimators: MLM and MLR. First Mplus User meeting – October 27th 2010. Utrecht University, the Netherlands,", 2017. Accessed on: Jul. 29, 2020. [Online]. Available: UGent, https://users.ugent.be/~yrosseel/lavaan/utrecht2010.pdf

[36] B. Wesolowski, "Model–Data Fit," In The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, 2018. Accessed on: Jul. 29, 2020. [Online]. Available doi: 10.4135/9781506326139.n439

[37] J.F. Hair, W.C. Black, B.J. Babin, and R. E. Anderson, Multivariate Data Analysis, 7th ed. New Jersey: Prentice Hall, 2009.

[38] J.C. Nunnally and I.H. Bernstein, Psychometric theory, 3rd ed. New York: McGraw-Hill, 1994.

[39] L. Sternod and B. French, "Test Review: Watson, G., & Glaser, E.M. (2010). Watson-GlaserTM II Critical Thinking Appraisal," Journal of Psychoeducational Assessment, vol. 34, no. 6, pp. 607–611.

[40] D.L. Bandalos, "Relative Performance of Categorical Diagonally Weighted Least Squares and Robust Maximum Likelihood Estimation," Structural Equation Modeling: A Multidisciplinary Journal, vol. 21, no. 1, pp. 102–116, 2014.