# Measuring Students' Critical Thinking: A Critical Review Article

Imadin M. Zannrni

Department of Translation, Cihan University-Erbil,
Kurdistan Region, Iraq

***Abstract*—*Staring* from the increasing importance of Critical Thinking (CT), playing a pivotal role in learning English as a second language, aiming to develop the educational situation in Kurdistan Region and Iraq as well as raising the awareness of the CT, and being one of the most prominent 4Cs the researcher has conducted this paper that examines three studies; their names will be mentioned afterward, in the field of critical thinking. Each study has its methodology, instruments, and goals, so it has been necessary to examine and analyze the research methodologies, findings, discussions, and ethics. Moreover, the researcher has tried to set a compromise among the results.**

***Keywords*—Consistency, Critical thinking, Critical thinking skills, Reasoning, Reflection, Clarity.**

## I. Introduction

As indicated by Gough (1991), may be in particular in the present data age, thinking abilities are seen as critical for instructed people to adapt to a quickly evolving world. Numerous instructors accept that certain information won't be as essential to the upcoming laborers and residents as learning and new data.

In the 20th century, the capacity to take incautious, brilliant ideas has been seen differently. Deborah Gough's words cited toward the start of this report embody the current perspective in schooling about the significance of training the present understudies to think basically and imaginatively. All journalists talk about thinking abilities regarding the two related wonders of present-day innovation and quick-moving change. Robinson, for instance, states in her 1987 practicum report:

Helping youngsters to become viable scholars is progressively perceived as an immediate objective of training... Assuming understudies work effectively in a profoundly technological society, they should be outfitted with deep-rooted mastering and figuring abilities essential to obtain and handle data in a steadily evolving world.

Beyth-Marom et al. (1987) highlight this point, portraying thinking abilities as means to using sound judgment: Thinking abilities are essential instruments in a general public described by quick change, numerous choices of activities, and various individual and aggregate decisions and choices, and they mentioned the cultural variables that make a requirement for very much created thinking abilities are just essential for the story, nonetheless. Another explanation is that instructors, bosses, and others call for more and better reasoning.

Moreover, Robinson (1987) notes that while the significance of intellectual improvement has become far and wide, understudies' exhibition on proportions of higher-request thinking capacity has shown an essential requirement for understudies to foster the abilities and perspectives of successful reasoning.

Three articles have been examined in this study. The researcher has gone through the following process to conduct this study. At first, the researcher had to summarize the articles. The second step was to analyze and evaluate them, adopting a comparative-contrastive and analytical methodology. Furthermore, the research has examined the evidence that led to the three articles' conclusions. Crystal straightforward research questions and the adoption of an appropriate method for the analysis were the primary concerns. Both validity and reliability have been examined in this study.

## II. Literature Review

### *A. Critical Thinking Skills (CTS)*

Teachers are central to encourage new ideas and information in the critical thinking skills project. Therefore, it is significant to understand how to measure students critical thinking skills

Cihan University-Erbil Journal of Humanities and Social Sciences
Volume VI No. 1 (2022) 5 pages
DOI: 10.24086/cuejhss.vol6n1y2022.pp54-58
Received 05 June 2021; Accepted 22 August 2021; Regular research paper: Published 20 February 2022
*Corresponding author's e-mail: imadin.zannrni@cihanuniversity.edu.iq

(Shareef and Abbas, 2021). Depending on the following terms, terminologies, and range of vocabularies: Critical Thinking Skills (CTS), Critical Thinking Instruments, Analysis, Evaluation, Deduction, Induction, validity, reliability, Inquiry-based learning, etc. Each of the research articles above was picked from the Google Search Engine database. The researcher has chosen these keywords, because they are widely used in previously conducted journal articles in the field of Critical Thinking. The results have shown more than ten studies with the same core, so it is worth mentioning that the reviewed papers were chosen based on the following criteria and reasons: First, being a recent study, not old, is a necessity. Second, the practicality of the studies has played a key role in selecting, and it has been more superior to theory. Consequently, all of the theoretical studies were not taken into consideration. Third, relativeness has also played a key role in selecting. Fourth, the participant students are similar to Middle East students. The following Table I shows the details of the three articles.

### *B. A Broader Overview on the First Article*

The first study was conducted by Ghadi, who holds a BA Degree in Geography from Yarmouk University-Jordan, an MSc in Curriculum and Instruction specializing in Science Education from Muta University-Jordan, and a Ph.D. in Curriculum and Instruction at University Putra Malaysia.

This study was conducted in 2013 at the University of Putra Malaysia (UPM). It aimed at measuring the Critical Thinking Skills of undergraduates at UPM. Using unpurposive (random) sampling, 433 students were chosen to participate in a pilot study. According to Johnson (1992), pilot testing involves trying out a questionnaire before wide-scale distribution to ensure it is easy to understand and provides appropriate data. It is usually done with a small group of respondents representing the larger group. Problems will inevitably be detected. Problem items should then be revised and pilot-tested again.

The fundamental goal was to discover the suitability criteria of all the items in the CTS instruments that would be used in the actual study. The study tested the validity and reliability of four CTS, Analysis, Evaluation, Deduction, and Induction. The pilot study results demonstrated that the instruments needed to be adjusted to guarantee the exclusion of all items except for good quality items utilized in the pilot study. Furthermore, all sections of the modified CTS instruments were valid, reliable, and fit to attain information for the final research. The study adopted 22 MCQs with two alternatives, and it was a quantitative study. According to Merriam-Webster Dictionary (2020), it collects data through observation and experimentation and the formulation and testing of hypotheses.

Ross (2005) mentioned that when the research focuses on figures and statistics that might be analyzed and quantified objectively; it is recognized as quantitative research. This research aims at using mathematical methods in an attempt to provide accurate data.

From Universiti Putra Malaysia, three experts in Educational studies have checked face and content validities for the instrument. Harris (1969) defines face validity as the appearance of the test to the examinee, test administrator, and educator. Wisniewski et al. (1982) propose that face validity refers to what the test appears superficially to measure. Henning (1987) says that face validity is often determined impressionistically, for example, by asking students or teachers whether the test was appropriate to their expectations. It is a kind of an impressionistic reaction to the test. Face validity pertains to whether the test looks valid. If a test seems right to other people, such as examinees, teachers, and administrative personnel, it can be described as having test validity. Face validity has to do with the public acceptability of a test. It is sometimes known as surface validity or appearance validity. If a test appears irrelevant, silly, or childish, the result will be poor cooperation on the examinees. On the other hand, Bachman (1990) identifies a test to have content validity if its content adequately represents the language skills with which it is concerned. Content validity relates to achievement tests constructed as a sample of the syllabus materials.

### *C. A Broader Overview on the Second Article*

Al-Mahrooqi (2020) conducted the second journal article, Deputy Vice-Chancellor for Postgraduate Studies and Research at Sultan Qaboos University, Oman. The main argument in the research is the lack of investigative attention to the CTS of the Omani tertiary-level students. The study depended mainly on the Cornell Class-Reasoning Test, which was authorized by Robert H. Ennis, William L. Gardiner, Richard Morrow, Dieter Paulus, and LuciIIe Ringel in 1964. It was published by the Illinois Critical Thinking Project, Department of Educational Policy Studies, the University of Illinois at Urbana-Champaign. It is a multiple-choice deductive logic class-reasoning test. This test has 78 questions to see how well the examinees do particular thinking. The study depended mainly on this test as a source to collect data. The test was modified to cover only

Table I
Information about the Journal Article Used in this Review

| No | Article | Author(s) | Issue/vol./pp | Year | Journal |
|---|---|---|---|---|---|
| 1 | Measuring Critical Thinking Skills of Undergraduate Students in Universiti Putra Malaysia | Ghadi et al. | 3, 6, 1458-1466 | 2013 | International Journal of Asian Social Science |
| 2 | Assessing Students' Critical Thinking Skills in the Humanities and Sciences Colleges of a Middle Eastern University | Al-Mahrooqi and Denman | 13, 1, 783-796 | 2020 | International Journal of Instruction |
| 3 | Effects of using inquiry-based learning on EFL students' critical thinking skills | Wale and Bishaw | 5,9, 1-14 | 2020 | Asian-Pacific Journal of Second and Foreign Language Education |

36 questions across six-item groups associated with five CT principles. Descriptive analysis was used to calculate correct overall percentages for the entire test and each item group to determine whether participants had mastered or failed to master the critical thinking principle. 200 students (50.5% male, 49.5% female) participated in the CT test. (89.0%) They were 20 years or older, 10.0% 18 or 19 years old. The first half of the participants studied science-based colleges (50.0%), and the second half of the participants were enrolled in humanities-based colleges (50.0%).

Results show that participant students did not master or fail all five assessed principles. However, they recorded significantly higher scores on four of the six-item groups than foundation students in the earlier study. Furthermore, the researcher concludes that female participants received higher overall test scores than their male counterparts, although there was no difference based on the college of study. However, the researchers believe that the research had few limitations. The first was the small sample ($N = 200$) compared to the number of students at SQU (7300). Second, the modified version of the Cornell Class-Reasoning Test, Form X. the small research sample has limited the study's outcomes.

### *D. A Broader Overview on the Third Article*

The article was conducted by Wale, a lecturer at Woldia University, Ethiopia. The second researcher is Kassie Shifere Bishaw, a lecturer at the English Language and Literature Department, Bahir Dar University, Ethiopia. Both authors contributed notably in conception and design, data acquisition, analysis and interpretation of data, and revising the manuscript critically, taking public responsibility for the entire content.

The fundamental contention in the examination was to check the request put together learning impacts concerning understudies' CTS. A semi-test plan which utilized a time-series program with single gathering members was used. 20 EFL college understudies partook and took progressed composing abilities course were chosen using the particular testing technique. Tests, center gathering conversation and understudy intelligent diary were utilized to accumulate information on the understudies' basic reasoning abilities. While the quantitative data were analyzed using One-Way Repeated Measures ANOVA, the personal data were examined through portrayal. The discoveries of the review uncovered that utilizing request-based factious composing guidance upgrades understudies' basic reasoning abilities. Consequently, request-based guidance is proposed to further develop understudies' basic reasoning abilities. The technique upgrades understudies' translation, investigation, assessment, deduction, clarification, and self-guideline to the center of basic reasoning abilities.

## III. Research Methodology

### *A. Reliability and Validity*

Before evaluating the validity and reliability of the reviewed journal articles, it is necessary to define these two terms. The validity and reliability of each piece are analyzed based on the literature.

Alderson and Banerjee (2001) explain reliability as the degree to which a measurement tool produces stable and consistent results. A test is supposed to be dependable if it gives similar outcomes over and over when it is shown on various events. By and large, if individuals get comparable scores on equal types of a test, for example utilizing multiple types of a trial which attempt to quantify similar abilities and capacities using similar strategies for testing, equivalent length, and level of trouble, this demonstrates that the test is solid.

Reliability is defined again by Crocker and Algina (1986) "Whenever a test is administered, the test user would like some assurance that the results could be replicated if the same individuals were tested again under similar circumstances. This desired consistency (or reproducibility) of test scores is called reliability". In other words, the sum of the parts should be reproducible so that a score can be meaningful and interpretable.

According to Alderson and Banerjee (2001), the validity or invalidity of those consistent scores is another question. However, an instrument must be reliable first, and then it can be valid. It is a kind of an argument that the two terms are used interchangeably because a test has to be reliable to be good. However, the reverse might not be accurate. The definition is the most noticeable difference between reliability and validity. The measurement consistency is estimated by reliability; in other words, instrument measures are calculated every time, using the same subjects and conditions. On the other hand, the accuracy of measurement is estimated by validity; in other words, it assesses the degree to which it is supposed to measure.

As mentioned before, the first research journal article measures the undergraduates' CT. A famous journal publishes the International Journal of Asian Social Science for its high-quality publications. For evaluating the validity and reliability of this study, we have to ascertain the method of data collecting, introducing, examining the content, and to what degree the findings could be generalized (Cohen et al., 2011). It is essential to say that the study is over-generalized because a small portion of students only participated as samples. This might be understood as a shortcoming since UPM students are 7300.

The topic of the second research journal article is assessing students' CTS. The research is published by the International Journal of Instruction, well-known for being a prestigious publisher. By comparing the sample to the population, one quickly recognizes this study's over-generalization. Another weak point is that the study was included only the public universities and neglected the private ones. Furthermore, the study used a modified Cornell Class-Reasoning Test with 38 questions and 40 questions. As a result, the participants have a solid chance to achieve a better score in answering these 40 questions.

"Effects of using inquiry-based learning on EFL students' critical thinking skills" is the third study, which stands out because it has different sampling with a different methodology

and over-generalized results. Since the researcher adopted a quasi-experimental design, the research will be qualitative because it deeply examines a small sample.

*B. Population and Sample*

Johnson (1992) defined a population as an entire group of subjects (persons) on whom results will be applied. The population may vary according to different factors such as the purpose and research questions. Because it is challenging to survey the entire group of interest (the population), researchers select a subgroup (convenience sample). This sample has elements or persons because of their accessibility. The model must be similar to the population of interest, while the volunteer sample consists of persons who volunteer to participate in a study. We could make judgments about how the results might generalize to the population. Any generalization must be made based on a reasonable sample compared with the population. Simple random sampling involves selecting a piece to know the probability of each element is determined. He also mentioned that leveling the population and selecting models from these levels is called stratified sampling.

The population of the first research is all students from UPM, and the sample was 433 students. They were selected randomly. The population of the second research was all students from SQU, and the model was 200 students from SQU. 50.5% were males, and 49.5% were females. 50% were in the Humanities colleges, and 50% were in the Sciences Colleges. The population of the third research was all students from Woldia University, and the sample was 20 second-year students.

*C. Research Instrument*

According to Johnson, 1992, after deciding the population (sample), it is essential to determine what collecting information will be the most beneficial and effective. Questionnaires, interviews, direct observation of language use are the most common methods. In questionnaires, the researcher asks questions and gets answers from the participants. It can range from short 5-item instruments to long documents. They can be administered by mail, in person, or by phone. It is widely used because it takes less time and is less expensive. A questionnaire can be open-ended, in which the respondents answer with their own words as they want. This format is helpful for qualitative information because you get different answers. A questionnaire can be closed-ended, in which the respondents can choose one from among a limited or specific number of responses like multiple choice. This format is beneficial for gaining quantitative information and is easier to analyze. The last step in a questionnaire is a pilot test: the questionnaire should be tried out with respondents similar to those who will respond in the study before doing the survey. The feedback from those respondents can help the researcher modify to adjust the questionnaire (Ali et al. 2017).

The first research used a pilot study, using 22 *MCQs*. The second research also used a Cornell Class-Reasoning Test (MCQs) questionnaire, Form X (78 questions). The researchers adopted a modified version with 34 queries.

On the other hand, the third research is classified as experimental. Johnson (1992) defines it as a quantitative approach designed to investigate the effects of suggested reasons. In an experiment, the researcher mainly aims at establishing a two-phenomena cause-effect relationship. The researcher seeks to establish two variables, the independent one (IV), causes that change in a variable (DV); in other words, it affects the dependent one. He also mentioned many variations of the experiment and many forms of quasi-experimentations. A quasi-experimented from active investigation by purposive selection for subjects assigned to experimental groups. This kind of experiment is used if the researcher is interested in independent variables (IV) that cannot be randomly assigned. And this occurs when question IV is an originated characteristic.

## IV. Findings and Discussion

After summarizing the studies discussing the authors, aims, research methodologies, instruments, participants, validity, reliability, and results, the researcher discusses the following points.

Regarding titles, the second article has a broad label that needs to be narrowed down because the Asian students' CTS in Oman may differ from their counterparts in India and Malaysia. Furthermore, this study has studied only the public sector. However, it mentioned the faculties. The title of the third article was too general. The first article has the most specific title by saying the core of the research and the place.

In terms of the instruments, all researchers use one tool. The first and the second articles used the quantitative method and adopted a close-ended. The 22 and 36 MCQs were used, respectively, leaving a good chance to have different results. Furthermore, the second study used a form of a test that has never been reviewed in The Mental Measurements Yearbook. On the other hand, the third study used a quasi-experiment, a qualitative approach with more reliable and valid results.

In terms of sampling, participants, and population, the first study used a small number of participants compared with the total number of students at UPM and the methodology. It neglected the details of the participants—for example, the age, gender, and the faculty. The second study was more professional because it introduced an informative background.

Regarding the research ethics, neither the first study nor the second-mentioned any second-mentioned third study said some points regarding the institution, funding, competing interests, and resources available and neglected the other ethical principles.

The first article was a pilot study to conduct more extensive research, and the results showed that the instruments should be modified. The second study's results were vague and did not answer the research questions. The third study had satisfying results which harmonized with the previous studies and the research questions.

In terms of resources, previous studies, and literature reviews, the first and the second articles used 28 different resources. On the other hand, the third study was the most resourceful, using 32 resources. It is worth mentioning that the online and offline resources were a mix of books, journals, and research. The second article used the most recent resources.

## V. Conclusion

Overall, the third study is the most comprehensive, reliable, and valid for several reasons. First, the title presents the topic explicitly and simply. Second, using many new (up-to-date) references and resources make the research fruitful and informative. The third is the use of a variety of data-gathering instruments. Fourth, using the qualitative tool and the one-way repeated measures ANOVA gives the research a good amount of validity and reliability through pre-tests and post-tests. Fifth, the results perfectly answered the research question. Sixth is the unique way of listing and presenting the research. Finally, it is the only study to mention some research ethics.

## References


Alderson, J., & Banerjee, J. (2001). Language testing and assessment (Part I). *Language Teaching*, 34, 213-236.

Ali, O.A., Al-Salami, Q.H., Kamar, S.H. (2017). *Financial Statistics* (Arabic Edition). 1st ed. Baghdad, Iraq: Alwan-Library.

Al-Mahrooqi, R., & Denman, C.J. (2020). Assessing students' critical thinking skills in the humanities and sciences colleges of a Middle Eastern university. *International Journal of Instruction*, 13, 783-796.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing*. Oxford, United Kingdom: Oxford University Press.

Banerjee, S.C., Greene, K., Magsamen-Conrad, K., Elek, E., & Hecht, M.L. (2015). Interpersonal communication outcomes of a media literacy alcohol prevention curriculum. *Translational Behavioral Medicine*, 5(4), 425-432.

Beyth-Marom, R., Novik, R., & Sloan, M. (1987). Enhancing children's thinking skills: An instructional model for decision-making under certainty. *Instructional Science*, 16(3), 215-231.

Cohen, L., Manion, L., & Morrison, K. (2002). *Research Methods in Education.* United Kingdom: Routledge.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Orlando, FL: Holt, Rinehart, and Winston.

Ghadi, I.N., Bakar, K.A., Alwi, N.H., & Talib, O. (2013). Measuring critical thinking skills of undergraduate students in universiti Putra Malaysia. *International Journal of Asian Social Science*, 3, 1458-1466.

Gough, N. (1991). Narrative and nature: Unsustainable fictions in environmental education. *Australian Journal of Environmental Education*, 7, 31-42.

Harris, D. (1969). *Testing English as a Second Language*. New York: McGraw-Hill.

Henning, G. (1987). *A Guide to Language Testing: Development, Evaluation, Research*. South Carolina: Create Space Independent Publishing Platform.

Johnson, D.M. (1992). *Approaches to Research in Second Language Learning*. New York, USA: Longman. p. 16-253.

Robinson, I.S. (1987). *A Program to Incorporate High-Order Thinking Skills into Teaching and Learning for Grades K-3*. Washington DC: Institute for Education Sciences.

Ross, K.N. (1978). *Sample Design for Educational Survey Research*. Oxford: Pergamon Press.

Shareef, L.B., & Abbas, N.J. (2021). A study of the teaching-learning challenges of the 21st century at university ELT classroom. *Cihan University-Erbil Journal of Humanities and Social Sciences*, 5, 16-24.

Wale, B.D., & Bishaw, K.S. (2020). Effects of using inquiry-based learning on EFL students' critical thinking skills. *Asian-Pacific Journal of Second and Foreign Language Education*, 5, 1-14.

Wisniewski, J.J., Genshaft, J.L., Mulick, J.A., & Coury, D.L. (1987). Test-retest reliability of the revised children's manifest anxiety scale. *Perceptual and Motor Skills*, 65, 67-70.