Chabahar Maritime University
Iranian Journal of English for Academic Purposes ISSN: 2476-3187
IJEAP, (2016) vol. 5 issue. 2 (Previously Published under the title: Maritime English Journal)

Assessing Critical Thinking Ability via the Writing Process: Developing and implementing a Rating Scale

Reihaneh Sheikhy Behdani[1]

ID: IJEAP-1706-1052

Mojgan Rashtchi[2]

Received: 01/06/2016 Accepted: 04/09/2016 Available online: 2016/11/11

Abstract

The present exploratory mixed methods study aimedto develop a scale for assessing critical thinking (CT) ability of Iranian EFL learners. The study wasconducted in three phases. In phase one, the features of CT components were examined in the writingsof 120 participants. A holistic and an analytic scale were developed based on their performance in exploiting the components of CT and the features of CT components.In the second phase, the language learners' new writing samples (N=110) were analyzed to examine to what extent they could employ the CT components.Also, the reliability and validity of the newly developed scale was calculated. Factor analysis revealed that the scale measured the four constructs of clarity, fairness-breadth, depth- significance-logicalness, and accuracy-precision-relevance.In the final phase, the researchers implemented the newly developed scale on a new group of learners (N=33)to observe the degree the scale couldassess the CT ability of language learners.

Key words: Critical Thinking Ability; Components of Thinking; Rating Scale; Writing Process

1. Introduction

Critical thinking (CT) plays an importantroleinlearning(Beyer, 1987; McPeck, 1981) and puzzling out the students' CT ability seems to be crucial in this rapidly changing world (Stupple, Maratos, Elander, Hunt, Cheung, Aubeeluck, 2017). Therefore, CT assessment has witnessed the development of a wide range of generic measures. For example,Cornell Critical Thinking Test (1985), Ennis-Weir Critical Thinking Essay Test (1985), Watson-Glaser Critical Thinking Appraisal (1980) are only a few to mention. This diversity of instruments stems from different definitions of CT (Ku, 2009). Since conceptualization and assessment of CT are interdependent and the definition of

[1] Department of English, Science and Research Branch, Islamic Azad University, Tehran, Iran
Email: reihaneh.sheikhy322@gmail.com

[2] Corresponding Author: MojganRashtchi**,** TEFL Department, North Tehran Branch, Islamic Azad University, Tehran, Iran Email**:** mojgan.rashtchi@gmail.com

CT regulatesits assessment (Ku, 2009), the writersof the present study referred to various CT definitionsand selected a definition which was compatible with the newly developed scale. Then the definition of CTunderlying the newlydevelopedscalewas provided. Accordingly, the researchers re-examined the assessment of CT basedonPaul and Elder's (2014) definition and developed a scale of CT.

Earlier definitions of CT focused on the cognitive component and considered CT as a skilloraset of skills, a mental procedure, or simply as rationality (McPeck, 1981). These definitions revolved around thinking methods and rules of formal logic rather than the implication of thoughts. However, conceptualizing CT as a“reasonable reflective thinking that is focused on deciding what to believe and do”by Ennis (1987, p. 10)showed a shift towards motivational aspects of CT, namely, *disposition* (Facione, 1990; Halpern, 1998; Perkins, Jay, &Tishman, 1993). The disposition to think critically has been elucidated as the “consistent internal motivation to engage in problems and make decisions by using CT” (Facione, 2000, p. 65).Similarly,Paul and Elder (2007) explicated CT as a structured cognitive process that requires active and skillful engagement in thinking. Later, Paul (2012)described CT as a "disciplined self-directed thinking" and considered it as the "perfections of thinking appropriate to a particular mode or domain of thinking" which reveals itself as "sophistic or weak sense and fair-minded or strong sense"(p. 33). Sophistic, as Paul (2012) argued,pertains to the "interests of a particular individual or group" while excluding others; whereas, fair-minded relates to the "interests of different people or groups" (p. 33). Paul and Elder (2014) maintained that to improve the CT ability, learners need to engage in a set of intellectual processes consisting ofpurpose, concept, information, question, inference, assumption,point of view, and implication.These components of thinking, as Paul and Elder (2014) argued, require learners to progressfrom memorizing the pieces of information to the thinking process.

Following the changesinthe definition of CT, its assessment hasalsoundergone a marked change. The early tests utilizing single multiple-choice response format focused on learners' recognition or level of knowledge, and mainly tapped the cognitive components of CT. However, they were not completelyable to reveal the respondents'disposition; neither could they reflect their inclination to engage in CT (Ennis, 2003; Halpern, 2003; McMillan, 1987). In a similar vein, Halpern (2003) criticized the incomprehensiveness of multiple-choice tests arguing that such tests tendto measure CT quantitatively. The underlying reason, according to Halpern (2003), was that the test-takersdo not have the freedom to suggest their own evaluative criteria and cannot generate their personalsolutions to the problem. Similarly, the multiple-choice tests are considered unable to reveal test-takers' CT ability in unprompted contexts (Ennis

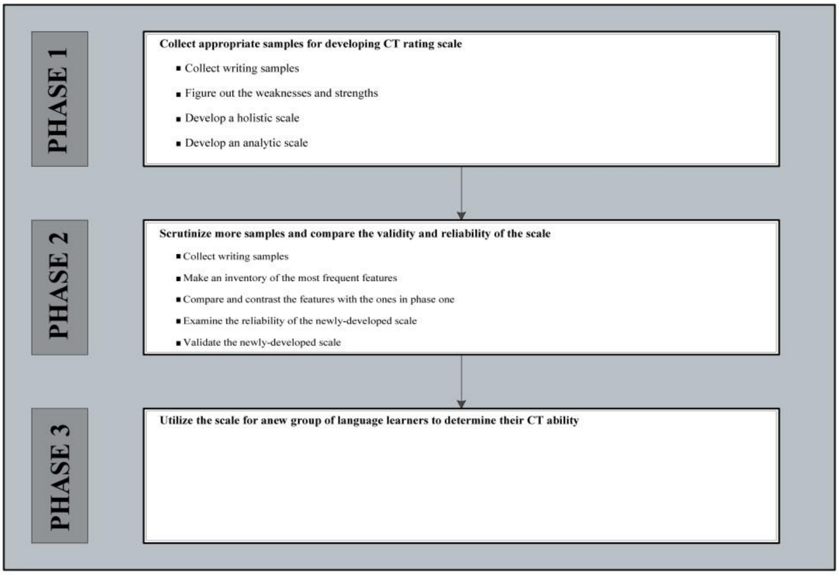
& Norris, 1990; Halpern, 2003; Norris, 2003). In response to the concerns mentioned above, Facione (1990) suggestedtheconcurrentemployment of the California CTSkills Test and the California CT Disposition Inventory (Facione&Facione, 1992) to measure CT. Nevertheless, gaging each factor using separate measures is unlikely to fill the gap between what people claim they would do (in self-reported dispositional measures) and what they actually do (in an actual test of CT skills).

Moreover, severalscholarsfavoropen-ended tests for assessing CT ability and disposition (Halpern, 2003; Norris & Ennis, 1989; Taube, 1997). Nosich(2012), also, believed that essay type formatscould assess both CT ability and disposition. Therefore, the Ennis-Weir CT Essay Test was developed as an open-ended test that measures test-takers' CT ability to examinean argument in authentic situations. However, the test does not dynamically assess the learners CT ability throughout the process, rather, it zeros in on the final product.

1.1. The Present Study

To address the deficiencystated earlier, the primary purpose of this study was to develop a rating scale,which could measureCT viatheprocess approach to writing (White & Arndt, 1991). Thereupon, the researchers followed one of the main variants of the exploratory mixed methods design called instrument development model. According to Cresswell and Plano Clark (2011), the design is a "two –phase sequential design" in which the researcher starts by exploring a topic (qualitative phase) and uses the findings of the first phase to develop the second phase (quantitative one) (p. 86). Consistent with this design, the researchers first qualitatively explored the components of CT in the scripts that led to the development of the items and scales for a quantitative survey. In the second phase, the researchers computed the reliability of the scale and validated it. Consequently, the qualitative and quantitative methods were linked together by developing the items of the instrument. In the third phase, the researchers applied the scale to examine the extent to which the scale could assess the CT ability of language learners.Non-equivalent pretest-posttest control group design was employed for the quantitative phase of the study.The following flowchart illustrates the three phases of the study:

Flowchart of the Procedures in Developing & Implementing the Scale


PHASE 1
Collect appropriate samples for developing CT rating scale
▪ Collect writing samples
▪ Figure out the weaknesses and strengths
▪ Develop a holistic scale
▪ Develop an analytic scale
PHASE 2
Scrutinize more samples and compare the validity and reliability of the scale
▪ Collect writing samples
▪ Make an inventory of the most frequent features
▪ Compare and contrast the features with the ones in phase one
▪ Examine the reliability of the newly-developed scale
▪ Validate the newly-developed scale
PHASE 3
Utilize the scale for anew group of language learners to determine their CT ability

Therefore, the following research questions were proposed:

1. Which features of CT are most frequently used in the expository writings of theIranian EFL learners?
2. Is the scale a reliable measure for estimating the Iranian EFL learners' CT ability?
3. What underlying domains of CTare measured by the variables of the newly developed scale?
4. Can the scale assess the CT ability of Iranian EFL learners?

2. Method

2.1. Phase One

2.1.1. Participants

One hundred and twenty undergraduate students of English major from Islamic Azad University, Rasht and Lahijan Branches, drawn from a subject pool of 148 learners took part in this study. The groups were intact and were selected based on convenience sampling (Hatch &Lazaraton, 1991). They were attending an essay writing course, at Islamic Azad University.The Babel English Language Placement Test was used to examine the participants' homogeneity regarding their English Language Proficiency. The learners whose scores fell between 52 and 80 were considered as the upper-intermediate, and were selectedto participate in the study. Additionally, the researchers used Akef's (2007) rating scaleand Ennis-Weir's(1985) Essay Test to examine the consistency ofthe participants' performance in the English writing and CT. Therefore, 28non-qualified learners were eliminated from the study.

2.1.2. Instrumentation

Babel English Language Placement Test that is based on the Nelson Quick-check placement test was employed to examine the participants' homogeneity in language proficiency.It isamultiple-choice test and measures the recognition of correct responses to reading prompts, grammatical forms, and lexical choices in contexts. The time allotted for the test was 60 minutes. Three experienced university lecturers of TEFL verified itssuitability and approved the clarity of its directions. The test was piloted on 50 students who were representative of the target population. The reliability estimate computed through KR-21 displayedthat the test enjoyeda high reliability index (r=.91).

The essay writing prompts, according to Kroll and Reid (1994), are the incentives in the form of writing topics to which the students respond. They were adopted from Cambridge IELTS Test (1-10) and were in a framed format. Since there is amenace that a prompt may fail todemonstrate the participants' level of writing skill accurately, the researchers scrutinized their potential usability. Each prompt was examined and controlled with regard tothe sixvariables for evaluating the prompts introduced by Hamp-Lyons (1991), that is, contextual, content, linguistic, task, rhetorical, and evaluation variables. Subsequently, agroup of three university lecturers followed a procedure (pre-test, revise, pre-test again, accept or reject the revised form) to pilot the topics.

The next instrument was Ennis-Weir's (1985) open-ended essay test of CT, which was used to gauge an individual's ability to inspect argument. The instructions are simple and direct, and it needs 40 minutes to answer. The scoring can range from -1 to +3 for responses of the eight numbered arguments, and the scoring for the overall summary evaluation of the letter ranges from -1 to +5. The total scores can range from -9 to +29.

The Ennis-Weir test has the benefit of content validity because it proposes common situations to examine the ability to appraise and formulate arguments. The inter-rater reliability on 27 students in a college-level introductory informal logic course and 28 gifted eighth-grade students of English was reported as .86 and .82, respectively, which is a high correlation for an essay type test (Ennis & Weir, 1985).

2.1.3. Procedure

After administering the Babel English Language Placement Test, the designated participants wrote on a writing prompt to ensure that they were homogeneous on the product and process approaches to the writing skill. For the product aspect, the format of the paragraphs, the topic and the supporting sentences, coherence,unity, and transitions were inspected. The scoring scale proposed by Akef (2007) was utilizedto evaluate the process aspect. Therefore, to appraise their CT ability, the participants took the Ennis-Weir Test. Then the participants were provided with a set of writing prompts.

After collecting the scripts, the researchers analyzed them and found the indispensiblefeatures of components of CT. The featureswerelater employed to ratethescripts as *commendable*, *adequate*, *fair*, and *poor* achievements.

Table 1. Holistic Scale

Level of Performance	**Description**
Commendable Achievement	Students followed the components of reasoning based on the intellectual standards. Little or no weak points are observed.
Adequate Achievement	Students followed the componentsof reasoning based on intellectual standards, but there are minor weak points.
Fair Achievement	Students followed the componentsof thinking based on the intellectual standards in writing process, but there are noticeable weaknesses.
Poor Achievement	Students consider the componentsof reasoning, but standards are ignored. Revision is needed.

After writing the descriptions as statements that highlighted the recurrent features of the components of CT, the researchers designed an analytic rating scale for each component. One of them, as an example, is presented.

Table 2. Analytic Scale

Features of Good Achievement in Purpose (4) ➢ The purposes are formulated clearly. ➢ The purposes are classified based on the significance of the issue. ➢ The purposes are well supported with the reasons and evidence. ➢ A unity of purposes is reflected in the paragraph.
Features of Adequate Achievement in Purpose (3) ➢ The purpose is almost clear but needs more elaboration. ➢ The purposes are classified based on the most significant issues. ➢ The purpose accomplishes almost well amount of data in each paragraph. ➢ The purposes are somehow consistent but still needs more monitoring.

Features of Fair Achievement in Purpose (2)

- The purpose is somehow vague and needs clarification.
- The purpose needs to focus on matters that are more significant.
- The purpose is not well accomplished in each paragraph.
- The purposes inadvertently negate each other.

Features of Poor Achievement in Purpose (1)

- It is not clear what the purposes of reasoning are.
- The purposes are based on trivial matters.
- The purposes deal with too much in each paragraph.
- The purposes are inconsistent.

A script received fourwhen it represented clarity, significance, precision, and consistency. In fact, the scale did not include the scripts that followed one feature and ignored the others. To avoid this sort of unwanted dependency of features on each other and too much detail, the researchers suggested a Likert scale format for the rating scale (Table 3).

Table 3. Features of Purpose

The extent to which the purpose is clearly stated.	0	1	2	3	4
The extent to which the purpose is worthwhile and well-chosen.	0	1	2	3	4
The extent to which the adopted purposes are precise.	0	1	2	3	4
The extent to which the purposes are consistent with each other.	0	1	2	3	4

The students' scripts received four when they communicated to the readerclearly, three when they needed more elaboration, two when they had an unclear purpose that could be improved with some changes, one when the purpose was vague and required a fundamental revision, andzerowhenthe component of thinking was lacking.

2.2. Phase Two

In phase two, the researchers attemptedto make an inventory of the CT features that were repeatedlyused in the 110 scripts written by the participants. Then they compared and contrasted the features with the ones discovered in phase one. Also, the reliability and validity of the newly developedscalewerecomputed.

2.2.1. Participants

Initially, a total number of 85 undergraduate students of English Translation and TEFL at Islamic Azad University, Rasht and Lahijan Branches, participated in Babel English Language Placement Test. The results of the test led to the deletion of 35 participants whose scores surpassed one standard deviation above and below the mean. Afterward, the participants took a writing test to corroborate their homogeneity on thewriting skill. Subsequently, Ennis-Weir Essay Test (1985) was administered to gauge the participants' CT before the treatment. The selected learners, then, were required to write abouta set of prompts.

2.2.2. Instrumentation

Babel English Language Placement Test, a set of essay writing prompts, and Ennis-Weir Essay Test (1985) were used for data gathering.

2.2.3. Procedure

Followinghomogenization, the learners were expected to write abouta set of prompts. The researchers analyzed the writingsconcerningthe components of CT and figured out the potential problems on each component of CT, and discussed them with the learners. As the next step, theresearchers estimatedthereliability and validity of the scale.

2.3. Phase Three

In phase three, the established scale was employed to inspect whether it could assess the CT ability of EFL learners. In fact, the researchers aimed to examine the extent the newly developed scale could determine the CT ability of a group of EFL learners.

2.3.1. Participants

The participants were 95 students majoring in English Translation and TEFL from Islamic Azad University, Rasht and Lahijan Branches. They had taken an essay writing course in the fifth semester of their education. The results of Babel English Language Placement Test, Ennis-Weir Essay Test (1985), and the newly established scale revealed that the groups were homogeneous with regard to their general language proficiency and CT ability. Akef's (2007) rating scale was employedto rate the writing performance of the participants. Accordingly, 63 students were selected and randomly assigned to experimental (n=33) and control groups (n=30). The purpose of this phase was to explore the extent to which the scale could assess the CT ability of EFL learners via writing.

2.3.2. Instrumentation

The same instruments used in phases one and two were utilized. In addition, the researchers used the newly developed CT rating scale. The scale took a process approach to the assessment of CT ability of Iranian EFL learners. The researchers aimed to appraise the CT ability of students via writing process. They established 32 items derived from the performance of the experimental group, thediscussionswith them, and theliterature on CT. Considering the results of factor analysis, the researchers developed four latent constructs for the eight components of CT including clarity, fairness-breadth, depth-significance-logicalness, and accuracy-precision-relevance. The results revealed a high reliability index (r=.97) via Cronbach's alpha.

2.3.3. Procedure

Sixty-three learners who were selected to participate in the study sat for a writing exam. The writings were rated following the product and process aspects. Also, the script utilizing the developed scale was analyzed, and the mean of the two raters' scores was considered as the final score for each individual. Moreover, the Ennis-Weir Test was administered to ensure the homogeneity of the participants regarding CT. The participants’ writingswere blind corrected by a raterwith reference to the newly developed scale to control bias. Then the collected data were analyzed to explore whether there was any discrepancy between the CT ability of the learners.

3. Results

3.1. ResultsPhase 1

To answer the first research question, the researchers analyzed the students' scripts. The following were themost frequent features:statingthe truth, expressing ideas in a simple way, presenting adequate data for each item, not exaggerating, considering different aspects of the topic, being specific, understandingthe relationship between the evidence given and the conclusion made, giving examples to support the reasons, relevance, brevity, providing justifications, using vivid language, comprehensibility,coherence, providing counter arguments,giving enough details,and consistency.

Subsequently,as shown in Table 4,the scale comprised eight components of purpose, concept, information, question, inference, assumption, implication, and point of view, eachconsisting of four constructs. Additionally, the quality of alearner's performance about each component wasincorporated(0 to 3).

Table 4.The Designed Rating Scale & Components

Components of Reasoning and Intellectual Standards		**Levels of Performance**				
Purpose	The extent to which the purposes are clear	0	1	2	3	4
	The extent to which the purposes are significant	0	1	2	3	4
	The extent to which the purposes are precise	0	1	2	3	4
	The extent to which the purposes are relevant to each other	0	1	2	3	4
Concepts	The extent to which the concepts are clear	0	1	2	3	4
	The extent to which the concepts are accurate	0	1	2	3	4
	The extent to which the concepts are precise	0	1	2	3	4
	The extent to which the concepts are relevant	0	1	2	3	4
Information	The extent to which the pieces of information are relevant to each other	0	1	2	3	4
	The extent to which the pieces of information are clear	0	1	2	3	4
	The extent to which the pieces of information are accurate	0	1	2	3	4
	The extent to which the pieces of information are fair	0	1	2	3	4
Question	The extent to which the questions are logical	0	1	2	3	4
	The extent to which the questions are clear	0	1	2	3	4
	The extent to which the questions are significant	0	1	2	3	4
	The extent to which the questions are deep	0	1	2	3	4
Inference	The extent to which the inferences are logical	0	1	2	3	4
	The extent to which the inferences are broad	0	1	2	3	4
	The extent to which the inferences are deep	0	1	2	3	4
	The extent to which the inferences are fair	0	1	2	3	4
Assumption	The extent to which the assumptions are clear	0	1	2	3	4
	The extent to which the assumptions are precise	0	1	2	3	4
	The extent to which the assumptions are fair	0	1	2	3	4
	The extent to which the assumptions are broad	0	1	2	3	4
Point of view	The extent to which the points of view are clear	0	1	2	3	4
	The extent to which the points of view are broad	0	1	2	3	4
	The extent to which the points of view are fair	0	1	2	3	4
	The extent to which the points of view are relevant to each other	0	1	2	3	4
Implication	The extent to which the implications are fair	0	1	2	3	4
	The extent to which the implications are broad	0	1	2	3	4

	The extent to which the implications are precise	0	1	2	3	4
	The extent to which the implications are significant	0	1	2	3	4

3.2. Results Phase 2

To answer the second research question, the reliability of the scale was calculated via Cronbach’s alpha. The result showed a high reliability index(r=.97).

As shown in Table 5, the KMO measure is 0.96, which is “meritorious” according to Kaiser's (1974) classification of measure values. Also, as the table indicates,Bartlett's test of sphericity is statistically significant (*p*< .05).

Table 5.The KMO & Bartlett's Test of Sphericityfor the CT Scale

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.961
Bartlett's Test of Sphericity	Approx. Chi-Square	13669.157
	df	496
	Sig.	.000

Table 6.The Initial Eigenvalues & Total Variance for the CT Scale Items

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	19.184	59.951	59.951	19.184	59.951	59.951	8.151	25.473	25.473
2	2.136	6.674	66.625	2.136	6.674	66.625	5.892	18.412	43.886
3	1.086	3.392	70.017	1.086	3.392	70.017	5.761	18.004	61.889
4	1.027	3.211	73.228	1.027	3.211	73.228	3.628	11.339	73.228

5	.910	2.843	76.071
6	.807	2.522	78.593
7	.616	1.926	80.519
8	.580	1.812	82.331
9	.550	1.718	84.049
10	.477	1.490	85.539
11	.423	1.322	86.862
12	.374	1.170	88.032
13	.342	1.068	89.100
14	.322	1.005	90.105
15	.281	.877	90.982
16	.270	.843	91.826
17	.261	.817	92.642
18	.233	.729	93.371
19	.227	.710	94.081
20	.218	.682	94.762
21	.201	.629	95.392
22	.180	.563	95.954
23	.179	.560	96.514
24	.166	.520	97.034
25	.154	.482	97.516
26	.143	.447	97.963
27	.140	.439	98.401
28	.127	.396	98.797
29	.110	.345	99.142
30	.105	.330	99.471
31	.093	.291	99.762
32	.076	.238	100.000

The criterion for choosing the components to retain was based on aneigenvalue-one criterion. As Table 6 shows, four constituents were extracted accounting for 73.22% of the variance.

Table 7. Rotated Components Matrix for the Inclusion of Items in the Components

Rotated Component Matrix[a]

	Component			
	1	2	3	4
q9	.800			
q4	.774			
q6	.774			.343
q3	.758			.301
q14	.658	.307	.332	
q7	.655		.395	.385
q8	.642	.356		.411
q22	.605	.437	.300	
q31	.590	.528	.333	
q28	.565	.563	.350	
q10	.552	.475	.343	.373
q11	. 465	.447	.438	.381
q23		.808		
q24	.303	. 734		
q12		.693		.411
q19		.680		
q29	.390	. 670		
q26	.386	. 663		.323
q27	.449	. 660		
q30	.401	. 629		.344
q18		.547		

q17			.739	
q2	.443		.707	
q20	.402		.685	
q16			.683	.387
q15	.393		.568	.496
q32	.528	.339	.542	
q13	.312	.341	.426	.405
q5				.848
q21	.301			.742
q1			.463	.727
q25		.531	.321	.576

To answer the third research question, therotated component matrix was used. As Table 7 shows, 12 items inconstruct 1, 9 items inconstruct 2, 7 items inconstruct3, and 4 items construct 4were established. The constructs were clarity, fairness-breadth, depth-significance-logicalness, accuracy-precision-relevance, respectively.

3.3. Results Phase 3

Independent-samples-t-test was run to answer the fourth question. ANCOVA was used to adjust posttest scores for any probable differences in the pretest. The assumptions of linearity, multicollinearity, normality of residuals, homogeneity of variances, and homogeneity of regression were met before running ANCOVA.

Table 9.Independent-Samples T-Test for the CT Gain Scores

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Gain_CT_Non.CT	Equal variances assumed	.22	.64	36.62	63	.000	-28.26061	.771

As shown in Table 9, there was a statistically significant difference between the CT scores of the participants, supporting that the scale shows the difference in CT ability of the learners, *t* (61) = 36.62, *p*< .05.

4. Discussion

4.1. Phase 1

The present study aimed to constructa rating scale for assessing the CT components. The scale comprises eight components that form CT (Paul & Elder, 2014) and four constructs as the benchmarks for assessing the componentsof the scale. According to Nosich (2012), the components of thinkingare "important touchstone or point of assessment in critical analysis and in the assessment of critical thinking" (p. 98). ToNosich, components of thinking are the fundamental structures of thinking which can be characterized as "micro-skills out of which larger-domained critical thinking abilities" are built (p. 98). The first component, (*purpose*) assessed the extent to which the students avoid wandering topics from paragraph to paragraph. The second (*concept*)evaluated the degree to which students identify the key concepts in the writing prompt. The third (*evidence*)measured the extent to which the learners utilized reliable information. The next (*question*)gauged the magnitude to which students ask a question regarding the accuracy of the data. The fifth (*inferences*)evaluated learners' ability in making logical and justified conclusion from the available data. *Assumption*, the sixth component, measured whatever the students take for granted. *Point of view*, the seventh component, estimated the degree to which the learners' point of view is insightful and justified,also the extent to which they consider other or even opposite perspectives. The eighth component, (*implication*), assessed the extent to which students discriminate between what is implied by a statement from what is inferred by people carelessly.

The researchers compared the students' scripts with the CTcomponents and realized that the level of performance for each component differed regarding four constructs including clarity, farness-breadth, depth-significance-logicalness, and accuracy. In fact, these constructs provided the researchers with the criteria for CT assessment. From a philosophical perspective, some criteria are needed to make a judgment or support decisions (Case, 2005; Lipman, 1988).Similarly,Bailin, Case, Coombs, and Daniels (1999) contended that a set of criteria isrequired for weighing the arguments and positions of others, for assessing evidence and one's own thoughts. To Bailin et al., these criteria can be in the form of "*standards* for judging the adequacy of claims about meaning; the *credibility* of statements made by authorities; the *strength* of inductive arguments; and the *adequacy* of moral, legal, and aesthetic reasons" (p. 291). Furthermore, Paul (2012) referred to the criteria as the qualities of thought that should be communicated to students. Thus, Paul (2012) suggested being explicit about the criteria for evaluating students' performance. Similarly, Bailin et al. (1999) and Case (2005) discussed the application of the knowledge of criteria as a resource for judging whether the students could think critically.

4.2.Phase 2

The newly developed scale bears some similarities and differences with the available models on CT. Unlike Ennis' (1987) model that overlooked the quality of categories, the scale developed in the present study included some standards to evaluate the components of CT. However, it is consistent with Ennis's model in that both considered clarity as the essential component and a priority throughout the whole scale.

Compared to Henri and Rigault's (1996) whose considered the cognitive functions in online conferencing as CT skills and evaluated CT in two levels as surface and in-depth, the present scale assessed CT in a formal setting (classroom). Similar to Henri and Rigualt's model, the scale has some indices to evaluate CT categories. The index of the developed scale is collected based on the students' performance on each component of CT and student-teacher interaction throughout a semester. The only difference is that the evaluation index in this scale was calculated for each component of CT in four levels; whereas, in Henri and Rigualt’s, a two-level index was developed.

In the same vein, McLean (2005) who contended that a CTassessment tool requires both categories and quality, the present scale included componentsas the categories of CT and intellectual standards as the quality index of the CT components.

4.3. Discussion phase 3

After applying the scale for a new group of learners, the researchers figured out some differences in learners' performance.First, some participants diverged from the topic of the writing, whereas some participants followed the primary purpose of the prompt and discussed it in details. Moreover, all participants could recognize the major concepts; however, the interpretations ofsomeof themwere anchoredin the social or personal understanding of the topic whiletangiblymisusing the concepts. Some learners, on the other hand, looked for every item in a dictionary. They supported their ideas with sufficient evidence. To this aim, they referred to the information both for and against theirownstance. They were also concerned about the reliability of the sources from which the data were taken. Similarly, theyverified the accuracy of the stated information. Each paragraph of their compositions addressed one of the major questions. On the other hand, some learners did not support their claims with enough evidence nor did theyanswerany significant question. The inferencesofsomeof the learnerswere not well reinforced by evidence and were rooted in questionable assumptions; whereas, the other participants recognized the assumptions leading to inferences. Considering each topic as multidimensional and paying attention to different aspects of it was a conspicuous feature of the writings. Nevertheless, the writings of the otherstudents were limited in scope. Unlike some learners, the others anticipated the ramifications (both positive and negative) of their reasoning and they were able to explicate each consequence. Furthermore, they traced out the significant implications of their rationale and could clarify why they imply so.

5. Conclusion

Assessing CT is possible in a variety of formats. The multiple-choice format is the most widely used to save time and toreduce expense.However, multiple-choice format, as Ennis (1993) put forth, is not comprehensive.According to Nosich (2012), essay format, as an alternative, iscomprehensiveand it assesses both CT ability and disposition; yet,itisexpensive and time-consuming.Questionnaires that are also utilized to assess CTare usually culture-bound, and their appropriateness for EFL learners may be questionable. To compensate for the drawbacks, the researchers developeda rating scale that utilizes the writing process to assess the CT ability among language learners.First, the researchers collected 290 writing samples from 120 language learners. After investigating the features of CT among the scripts in phase one, the researchers went through the second phase of the study. In phase two, the researchers inspected the newly collected scripts (N=110) to identify the learners' weaknesses and strengths in using the components of CT. The results of phase one and two were compared and contrasted, subsequently.After computing the reliability, the researchers corroborated thenewlydeveloped scale. At the final phase, the scale was put into practice to examine to what extent it could efficiently assess CT ability of the participants.The results indicated a significant difference between the experimental and control groups.

The findings of this study may assist language learners, teachers, and syllabus designers.Language learners can use the scale to realize their strengths and weaknesses in CT skill. The scale also can be used by language learners as a guideline for the development of CT. Also, the scale can be employed by individuals who intend to assess their level of CT ability.

The scale can be utilized by teachers, too. In fact, it can assistteachers to discover learners' reasoning ability(Lai, 2011; Norris, 1989through writing processes, locate discursiveness in their writingand help them avoid it. More prominently, the rating scalehelps teachers to explorethe thinking process of their students. Additionally, novice teachers can benefit from the scale indesigninga course plan for writing classes to promote both CT skills and the writing ability. Finally, the ratingscaleallows consistent and effective assessmentof CT ability.The scale also aids syllabus designers to take a new perspective on the selection of materials for writing courses and include the skills of CT in the table of contents.

REFERENCES

Akef, K. (2007). *Assessing the steps adopted by Iranian student writers in their writing process: A model for developing rating scale descriptors* (Unpublished doctoral dissertation), Islamic Azad University, Tehran, Iran.

Angelo, T.A., & Cross, K. P. (1993).*Classroom assessment techniques: A handbook for college teachers*. San Francisco, CA:Jossey-Bass.

Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking.*Journal of Curriculum Studies, 31*(3), 285-302.http://dx.doi.org/10.1080/002202799183133

Beyer, B. (1987).*Practical strategies for the teaching of thinking*. Boston, MA: Allyn& Bacon.

Case, R. (2005).Moving CT to the main stage.*Education Canada, 45* (2), 45-49.

Creswell, J. W., & Plano Clark, V. L. (2011).*Designing and conducting mixed methods research*. London: Sage.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9-26). New York, NY: W. H. Freeman.

Ennis, R. H. (1993). Critical thinkingassessment.*Theory into Practice, 33* (3), 179-186.http://dx.doi.org/10.1080/00405849309543594

Ennis, R. H. (2003). Critical thinking assessment.In D. Fasko (Ed.), *Critical thinking and reasoning* (pp. 293-310). Cresskill, NJ: Hampton Press.

Ennis, R. H., &Norris, S. P. (1990). Critical thinking evaluation: Status, issues, needs. In J. Algina& S. M. Legg (Eds.), *Cognitive assessment of language and math outcomes* (pp. 1-42). Norwood, MA: Ablex.

Ennis, R. H., & Weir, E. (1985).*The Ennis-Weir critical thinking essay test*. Pacific Grove, CA: Midwest.

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction.* Millbrae, CA: The California Academic Press.

Facione, P. A. (2000). The disposition toward critical thinking: Its character, measurement, and relation to critical thinking skill. *Informal Logic, 20* (1), 61-84.

Facione, P. A., &Facione, N. C. (1992).*The California critical thinking dispositions inventory (CCTDI) and CCTDI test manual*. Millbrae, CA: California Academic Press.

Gellin, A. (2003). The effect of undergraduate student involvement on critical thinking: A meta-analysis of the literature 1991-2000. *Journal of College Student Development, 4* (6), 746-762. http://dx.doi.org/10.1353/csd.2003.0066

Halpern, D. F. (1998).Teaching critical thinking for transfer across domains: Dispositions, skills, structure, training, and metacognitive monitoring.*American Psychologist, 53* (4), 449-455.http://dx.doi.org/10.1037/0003-066X.53.4.449

Halpern, D. F. (2003).*Thought and knowledge: An introduction to critical thinking*. Mahwah, NJ: Laurence Erlbaum.

Hamp-Lyons, L. (1991). *Assessing second language writing in academic contexts*. Norwood, MA: Ablex.

Hatch, E., &Lazaraton, A. (1991).*The research manual: Design and statistics for applied linguistics*. Boston, MA: Heinle&Heinle.

Henri, F., &Rigault, C. R. (1996).Collaborative distance learning and computer conferencing. In T. T. Liao (Ed.), *Advancededucational technology: Research issues and future potential* (pp. 45-74). New York, NY: Springer-Verlag.

Kaiser, H. F. (1974). An index of factorial analysis simplicity.*Psychometrika, 39*, 31-36. http://dx.doi.org/10.1007/bf02291575

Kroll, B., & Reid, J. (1994). Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing, 3* (3), 231-255.http://dx.doi.org/10.1016/1060-3743(94)90018-3

Ku, K. Y. L. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. *Thinking Skills and Creativity, 4*(1), 70-76. http://dx.doi.org/10.1016/j.tsc.2009.02.001

Kuhn, D. (1999). A developmental model of critical thinking.*Educational Researcher, 28* (2), 16-26.http://dx.doi.org/10.3102/0013189X028002016

Lipman, M. (1988). Critical thinking: What can it be? *Educational Leadership, 46* (1), 38-41.

McLean, C. L. (2005). Evaluating critical thinking skills: Two conceptualization. *Journal of Distance Education, 20* (2), 1-20.

McMillan, J. (1987). Enhancing college student's critical thinking: A review of studies.*Research in Higher Education, 26*, 3-29.

McPeck, J. E. (1981*). Critical thinking and education*. New York, NY: St Martin's Press.

Norris, S. P. (2003). The meaning of critical thinking test performance: The effects of abilities of abilities and dispositions on scores. In D. Fasko (Ed.), *Critical thinking and reasoning: Current research, theory and practice*. Cresskill, NJ: Hampton Press.

Norris, S. P. & Ennis, R. H. (1989).*Evaluating critical thinking.* Pacific Grove, CA: Midwest.

Norris, S. P., Ennis, R. H. (1989). *Evaluating critical thinking*.Pacific Grove: Critical thinking Books and Software.

Nosich, G. (2012). A model for the national assessment of higher order thinking. In R. Paul (Ed.), *Critical thinking: What every student needs to survive in a rapidly changing world* (pp. 78-123). Dillon Beach, CA: Foundation for Critical Thinking.

O'Hare, L. O., &McGuinness, C. (2009).Measuring critical thinking, intelligence, and academic performance in psychology undergraduates.*The Irish Journal of Psychology, 30* (3-4), 123-131. http://dx.doi.org/10.1080/03033910.2009.10446304

Paul, R. (2012). *Critical thinking: What every person needs to survive in a rapidly changing world*. Dillon Beach, CA: Foundation for Critical Thinking.

Paul, R., & Elder, L. (2003).*The thinker's guide to how to write a paragraph: The art of substantive writing*. Dillon Beach, CA: Foundation for Critical Thinking.

Paul, R., & Elder, L. (2006).*Critical thinking: Concepts and tools.*The Foundation of Critical Thinking.Retrieved fromhttps://www.criticalthinking.org/files/Concepts_Tools.pdf

Paul, R., & Elder, L. (2007).*Consequential validity: Using assessment to drive instruction*. Retrieved from: www.criticalthinking .org

Paul, R., & Elder, L. (2014).*Critical thinking: Tools for taking charge of your professional and personal life*. Cranbury, NJ:Pearson.

Perkins, D. N., Jay, E., &Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *Merrill Palmer Quarterly, 39*, 1.

Stupple, E. N. J., Maratos, F.A., Elander, J., Hunt, T. E., Cheung, K. Y. F., Aubeeluck, A. V. (2017). Development of critical thinking toolkit (CriTT): A measure of student attitudes and beliefs about critical thinking. *Thinking Skills & Creativity, 23*, 91-100.http://dx.doi.org/10.1016/j.tsc.2016.11.007

Taube, K. T. (1997). Critical thinking ability and disposition as features of performance on a written critical thinking test. *Journal of General Education*, *46*, 129-164.

White, R., & Arndt, V. (1991).*Process writing*. London: Longman.