International Journal of Assessment and Evaluation in Education
Vol 8, Dec 2018 (7-22) ISSN 2232-1926 / eISSN 2600-870X

Elicitation of L2 Learners' Reading Comprehension Skills and Strategies through Cognitive Diagnostic Assessment

Karim Shabani
Allameh Mohaddes Nouri University Iran.
shabanikarim@gmail.com

Received: 20 February 2018; Accepted: 23 June 2018; Published: 21 December 2018

Abstract

Cognitive diagnostic assessment (CDA) as an innovative criterion-based approach to assessment has been the center of increasing attention in recent decades due to its inherent merits over its previous psychometric counterpart. Due to the lack of research on the post hoc analyses of classroom assessments to diagnose L2 learners' strengths and weaknesses in the multidivisible skill of reading comprehension, and also its significance in high-stakes tests, this article offers the implementation of samples of high-stakes tests in the classroom environment due to its beneficial consequences and washback for the learners to increase their educational chances. Therefore, the present article is intended to shed light on the mental processes which the examinees went through in responding to the multiple choice questions of four reading comprehension passages, elicited from the recent Iranian University Entrance Exam 2015. Time-series design was conducted to forty female students aged between 16 and 17 years old who made the control and experimental groups in an EFL context at a high school to help them enhance their reading comprehension skills and strategies by giving proper diagnostic feedback and intervention. The analysis of data collected through triangulation of methods, namely interview, think-aloud, and self-assessment confirmed the usefulness of consistent cognitive diagnostic assessment (CDA) for both teacher and students. Moreover, the CDA analysis led to the development of a number of online reading comprehension strategies which are normally belittled in non-interactive assessments. Finally, the article suggests the use of consistent CDA as a dependable and trustworthy procedure to diagnose and enhance L2 learners' reading comprehension processes.

***Keywords**: Cognitive diagnostic assessment, interactive, washback, L2 reading.*

INTRODUCTION

Assessment as a core feature of L2 education is significant in classroom because "it sends a clear message to students about what is worth learning, how it should be learned, and how well they are expected to learn" (Moss, 2013, p. 235). Moreover, language tests can be used as valuable sources of information to reveal if L2 learning and teaching has been effective. Furthermore, language tests can be used as sources of feedback on the processes of learning and teaching. Hence, there is a reciprocal association between language testing and teaching (Gu, 2014).

Success in language teaching and learning is highly contingent upon valid diagnosis of L2 learners' strengths and weaknesses in a specific domain. It seems that among the language skills, reading comprehension (RC) as a receptive skill needs more tools to diagnose the strengths and weaknesses of students in second or foreign language context because of three main reasons. The first reason includes the general view that RC skill is divided into both micro and macro skills which are under the direct influence of RC subcomponents such as vocabulary, structure, and word

recognition and functions. Harding, Alderson, and Brunfaut (2015, p. 8) asserted that "it is likely that several different abilities are involved in successfully completing the task, including word and morpheme recognition, syntax, knowledge of vocabulary and possibly sentence structure". Furthermore, Harding et al. (2015) stated that the extent to which test items are diagnostic is unclear; Reading comprehension sub-skills do not exist in isolation and are interrelated. They argue that "one cannot make inferences before understanding specific detail, or understanding the main idea, or evaluating a text" (Harding et al., ibid, p. 7). Hence, reading is regarded as a complex skill which carries many sources of difficulty (Nation, 2009). The second reason according to Alderson (1984, cited in Harding et al., 2015, p. 4) is that "SFL [second or foreign language] readers are typically reading in a language that they have not mastered, and therefore L2 reading problems will be at least as much as language-related as reading-related". The next reason relates to the challenges that have been made between two main beliefs on the dimensions of reading comprehension: The first belief is that reading is a unidimensional skill, consisting of a single global construct, or a bidimensional construct involving general reading and vocabulary abilities (Rost, 1993), and on the contrary, the second takes a multidivisible view towards reading comprehension (Alderson, 2005; Badrasawi, Kassim, & Daud, 2017).

REVIEW OF LITERATURE

The Paradigm Shift in Assessment

The paradigm shift in assessment from psychometric behaviorist perspective towards a broader model of educational assessment occurred due to the need for dominance of 'assessment *for* learning' over 'assessment *of* learning'. According to Hernández (2012, p. 492), "traditional assessment practices are usually good at evaluation but they are often lacking in description and fail to provide students with advice and support to improve their learning". However, the paradigm shift in assessment is beyond the traditional functions of testing as proficiency, placement, achievement, and diagnosis described by many authors like Alderson (1990), Baird et al. (2014), Brown (2005) and Farhady, Ja'farpur, and Birjandi (2009). In recent decades, the paradigm shift is directed towards a new approach to assessment namely cognitive diagnostic assessment (CDA) which calls for a "more descriptive test information and detailed score reporting for improving instructional designs and guiding students' learning" (Jang, 2008, p. 118). Likewise, CDA aims and needs to provide learners descriptive and interpretable diagnostic feedback to help them "take actions to close the gap between their current competency level and their desired learning goals" (Jang, 2008, p. 118).

The new trends in assessment lay stress on the continuous nature of assessments and the need to include interaction between teachers and learners in classroom (Alavi, Kaivanpanah & Shabani, 2012). As Hancock (1994) stated, assessment is usually an ongoing strategy which monitors student learning and involves them in decision making about the extent to which their performance corresponds to their ability. Therefore, "assessment … should be viewed as an interactive process that engages both teacher and student in monitoring the student's performance" (Hancock, 1994, p. 3).

Moreover, L2 reading comprehension researchers (e.g., Anderson, Bachman, Perkins, & Cohen, 1991) have suggested the use of triangulation of methods such as interview, self-assessment, and think-aloud verbal protocols in assessments. Bachman (2000, p. 7) stated that "because the results of language tests are still most frequently reported quantitatively- as scores or numbers- it is essential that we continue to develop ways in which to utilize quantitative and qualitative approaches in a complementary fashion".

Due to the significance and dominance of RC in high school and university textbooks, and also consideration of RC as a fundamental and value-laden subject in most university entrance examinations, there is a need to implement high-stake tests in the classroom setting in order to support the students in increasing their educational chances.

Hence, the present study was designed by the concerns as to whether CDA can have any significant effects on the students' reading comprehension abilities and to what extent CDA is able to diagnose the most effective skills and strategies for the enhancement of reading comprehension performance.

Cognitive Diagnostic Assessment

Cognitive diagnostic assessment (CDA) is a new approach to language testing designed on the criterion-referenced principles to assess the learners' specific strengths and weaknesses in a given domain, provide them timely and fine-grained diagnostic feedback and intervention in an ongoing process, and 'remediate learners' weaknesses to close the gap between their performances and actual abilities" (Jang, 2008, p. 119). Furthermore, Jang (2008) stated that the CDA approach serves assessment *for* learning instead of assessment *of* learning. It is a sort of assessment as a learning tool which provides teachers to modify their classroom instruction due to the students' needs indifferent multidivisible skills, and also inform examinees of their cognitive strengths and weaknesses in assessed skills, and guide their learning. In fact, CDA is increasingly used in the past few decades to align assessment with learning (Liu, 2014) and benefit both teachers and students. CDA approach is intended to raise the students' consciousness in implementing diverse skills and subskills in responding to the test items through involving them in self-assessment, and verbal think-aloud protocols as well.

Diverse studies have been conducted in recent years based on cognitive diagnostic assessment (e.g., Doe, 2013; Gu, 2011; Huff & Goodman, 2007; Jang, 2005, 2008; Kim, 2011; Lee &Sawaki, 2009; Liu, 2014; Ravand, Barati, Widhiarso, 2013). Huff and Goodman (2007) asserted that CDA is a demand of psychometricians and cognitive psychologists for having a more cognitively informed test design, scoring, and reporting in order to better inform teaching and learning. They also stated that CDA is an integrated system of curriculum, instruction, and assessment which focuses on the process of building knowledge by students based on curriculum objectives, facilitating 'knowledge building and active learning' (p. 22) through instructors, and providing feedback to teachers due to assessment results for remediation and designing a new curriculum.

For instance, Jang (2009) conducted a study to elicit 11 ESL students' use of reading skills in three different passages including 37 items written in three different modes through think-aloud verbal protocols. The result of the study was elicitation of 16 text-related skills which were then reduced into nine categories of reading processing skills.

The main ambition of the present study is to take a step toward filling the gap in the literature on exploring the effectiveness of CDA design implementation and diagnosing the examinees' strengths and weaknesses in RC skills and subskills through triangulation of data. Relatively few studies have been conducted on eliciting the high school students' skills and strategies before they take their first high-stake university entrance examination in order to provide them fine-grained and descriptive diagnostic feedback through post hoc analyses of their performances; a technique which might help them close the gap between their competence and performance and increase their educational chances. Though diverse studies have been conducted with respect to reading strategies (e.g. Rymniak & Shanks, Lougheed, 2003; Gallagher, 2000; Sullivan, Brenner, & Zhong, 2004; Rupp, Ferne & Choi, 2006) relatively few studies have been reported to elicit reading skills and strategies through CDA design in a continuous trend to examine its effect on improving the L2 learners' RC abilities. Moreover, there is a paucity of research in Iranian context on the implementation of time-series design along with administration of four samples of public university entrance examinations; hence, the impetus to conduct the present study.

Finding response to the following research questions was in the agenda of the present study:

RQ1: Do the students who receive cognitive diagnostic assessment (CDA) in reading comprehension activities outperform the ones who don't?

RQ 2: To what extent can consistent cognitive diagnostic assessment (CDA) diagnose L2 learners' reading comprehension strategies?

RQ 3: What is the learners' perception towards the effectiveness of CDA in enhancing L2 learners' reading comprehension?

METHODOLOGY

Participants

The participants of the study were recruited from 60 female high school students in Babolsar, Iran who studied English as a foreign language in Shahed high school, aged between 16-17 years. After conducting the McGraw-Hill Comprehension Placement Test to homogenize the sample, 40 students who were placed into Comprehension B1 relying on the placement test results were selected. They were randomly assigned to two groups of control and experimental, each consisting of 20 students. It should be mentioned that the participants were taken from intact classes for which the researcher herself was the teacher. Therefore, they can be considered an intact group selected through convenience sampling method.

Instrumentation

Several instruments were utilized in this study including McGraw-Hill Comprehension Placement Test (Test 3) (r=0.83) to homogenize the samples, high school RC-based textbooks, four samples of the tasks from The Iranian *University Entrance Exam* 2015, a self-assessment questionnaire with a set of qualitative questions developed and checked by two experts to ensure the relevance of its items and trustworthiness (McCann & Clark, 2003), the data garnered through think-aloud protocols and a structured interview that was designed by the researcher and checked for accuracy and relevance by two colleagues to ensure its objectivity (Berg, 2004), the teacher's immediate diagnostic feedback and intervention, as well as the numerical scores of each conducted test. SPSS software (v.22) was also used for descriptive analysis of the collected numerical data for the experimental and control groups.

Interview

In order to enhance the reliability and validity of the results of the study, the whole participants in experimental group (N=20) were included in the structured interview. The L2 learners were asked questions about the usefulness of the conducted design of CDA in improving their reading comprehension skill, and identifying their strengths and weaknesses in reading comprehension text digestion and responding to its items. The L2 learners were also asked about how they would predict their performance in the approaching high-stakes Iranian University Entrance Exam 2016.

Design

A quasi-experimental design and more precisely time-series design with teach-test-teach-test format was implemented in the present study along with the manipulation of Cauley and McMillan's (2010, p. 2) formative assessment cycle with the following tenets:

1. Ongoing assessment allows both for fine-tuning of instruction and student focus on progress.
2. Immediate assessment helps ensure meaningful feedback.
3. Specific, rather than global, assessments allow students to see concretely how they can improve.

In order to diagnose the students' strengths and weaknesses in a specific domain, and deliver proper and fine-grained feedback to the learners, we need to successively involve them in a sort of task related to the content and curriculum of the course. As McNamara (2006) stated, "tasks are the conditions under which this evidence [evidence for drawing the inferences about test-takers based on the test] might be sought" (p.47); that is, engaging the students in the relevant tasks would provide such evidence. Since the tests used in the study were multiple choice items, various sources of

evidence were needed to diagnose the students' strengths and weaknesses in each subskill, and comment on their abilities in RC thereof. The reason is that "multiple choice questions can result in test performances that may not accurately reflect students' ability to construct meaning from texts" (Ozuru, Rowe, O'Reilly & McNamara, 2008, p. 1002). Anderson et al. (1991, p.44) also stressed the use of think-aloud protocols for RC construct presentation:

> With respect to the construct of reading comprehension it is known that direct assessment of the …trait is impossible since it is a mental operation which is unobservable. Using think-aloud protocols is a way of getting at the unobservable behavior of reading comprehension.

The think-aloud protocol along with self-assessment and interview are included in the design as triangulation of data sources.

Procedure

The present study was conducted within three months during the course. As the first step, the aims and goals of the assessment administration with CDA design, its method and procedures were shared with the experimental group verbally in a separate session held before the initial test administration. In the first month, the participants were under guided instruction of diverse reading skills and strategies, vocabulary and grammar based on their RC-based textbook at the high school. The themes in the book were practiced by the L2 learners in a dynamic question and answer format offered through the teacher and the textbook content. Next, the learners were given the necessary feedbacks explicitly or implicitly depending on the situation during 70 minute sessions being held each week. Therefore, the situation in the first month was quite identical for both groups. The differentiation between the two groups was initiated in the second month:

Throughout the CDA design, the control group participated in taking four RC samples of the recent Iranian University Entrance Exam 2015 offered with an interval of two weeks, the same as the experimental group, but received no reading strategy and test-taking strategy intervention during the test administrations; however, due to ethics new vocabularies were written on the board and the texts were translated with the learners' collaboration; moreover through taking consistent tests they practiced test-taking process and benefited the test administrations somehow. Conversely, the experimental group was involved in a different design developed by the researcher and received reading strategy and test-taking strategy interventions to surmount their weaknesses and do the test-taking process more consciously in subsequent tests. Meanwhile, the tests had no time limitation and were actually a kind of power tests rather than speedy tests. In fact, the purpose was to provide a low-anxiety condition in which lets the examinees have the chance to complete all items. In addition, the students were informed of the fact that their received scores on the tests do not influence their summative assessment grades. The goals and procedures of test administration, scoring method, and the way the scores used to judge on the students' abilities were clearly and precisely described in a sheet called general consideration sheet and was appended to the initial page of the test booklet to motivate the students to try a different condition comparing the previous tests that were normally administered in preceding courses.

The last page of the test booklet included a self-assessment sheet (see appendix A). The self-assessment sheet was designed on the basis of multidivisible feature of RC including questions which were targeted at exploring the process in which the students deal with a RC passage, its subcomponents and items. Moreover, the self-assessment sheet dealt with eliciting the students' mental processes of reading strategies and test-taking strategies to contribute the diagnosis of the strengths and weaknesses of the examinees, and hence benefit both students and teachers as two major agents of change for the process of teaching and learning. Therefore, each test booklet included a general consideration sheet, one sample of RC, selected from the recent (2015) Iranian University Entrance Exam (with diverse readability levels) including four multiple choice questions, and a self-assessment sheet respectively.

After one month, one prepared test booklet as mentioned before was delivered to the participants in the experimental group every other two weeks till the end of the term. So, during a term, four samples of RC were conducted. Once the students accomplished the test items, they were asked to complete the self-assessment sheets. After the whole participants had accomplished the task, they were requested to go through verbal think-aloud protocols and describe the way they dealt with the passage and test items to the teacher, and their peers. This process would both raise the examinees' consciousness in how they performed the test, and their attempt to enhance their reading and test-taking strategies in subsequent tests (Jang & Wagner, 2014; Doe, 2015). Moreover, this provides an opportunity for the classroom teacher to deliver proper and timely, fine-grained intervention and feedback to students, and modify her instruction due to the learners' strengths and weaknesses (Lee, 2015). After the final administration of the tests, the teacher conducted an interview with the whole participants in a separate session in order to elicit their opinions about the effectiveness of the conducted design on diagnosing their strengths and weaknesses in RC skill, and the strategies they used in reading comprehension passages as well. Thus, the procedures in the study followed time-series design with *teach-test-teach-test format.*

Data collection and analysis

The data of the study were collected from the examinees' responses to the test and self-assessment questionnaire, think-aloud protocols, the teacher's immediate diagnostic feedback and intervention, and the final structured interview. The main reason behind incorporating both quantitative and qualitative data collection and analysis methods was that in this way one could gain a more profound insight into the cognitive processes involved in the reading comprehension tasks. To find an accountable response to the first research question of the study, the data collected from each test were recorded and analyzed by SPSS software (v.22), through running ANOVA to appraise the performance of both control and experimental group in the conducted design, and use the results. Next, the post hoc test was run to both find the interrelations of tests in experimental group and thus appraise the efficacy or inefficacy of the consistency of CDA administrations. The qualitative data accumulated through self-assessment sheets, and also verbal think-aloud protocols were analyzed and the blind spots were immediately instructed by the teacher. The interventions relevant to reading skills and strategies were also presented verbally on the spot. Interview data were transcribed verbatim, and analyzed using qualitative content analysis (Kiani, Alibakhshi & Akbari, 2009, p. 107). Finally, the whole qualitative data garnered were recorded for further analyses.

RESULTS

Analysis of the First Research Question

In order to appraise the effectiveness or ineffectiveness of the conducted CDA design, and also the kind of interventions which the experimental group received in a consistent way during four test administrations, a repeated measures ANOVA was run to test the significance of differences among four sample means. Before running the repeated measures ANOVA test, we checked the data for normality and the results are as follows:

Table 1 Test of Normal distribution

	Shapiro-Wilk		
	Statistic	df	Sig.
Test 1 Control	.701	20	.000
Test 2 Control	.825	20	.002
Test 3 Control	.847	20	.005
Test 4 Control	.749	20	.000

Test 1 Experimental	.841	20	.004
Test 2 Experimental	.874	20	.014
Test 3 Experimental	.770	20	.000
Test 4 Experimental	.853	20	.006

Shapiro-Wilk goodness-of-fit test conducted to examine the normal distribution of the data showed significant results for all the tests denoting that the distribution was not normal; hence, the decision was made to use a non-parametric test for the statistical analysis.

Experimental Group's progress over time:

Table 2 Descriptive Statistics for the experimental group

			Std.			Percentiles		
	N	Mean	Deviation	Minimum	Maximum	25th	50th (Median)	75th
T1	20	7.2500	4.12789	.00	15.00	5.0000	10.0000	10.0000
T2	20	7.0000	5.47723	.00	15.00	1.2500	5.0000	10.0000
T3	20	11.5000	5.87143	.00	20.00	5.0000	15.0000	15.0000
T4	20	14.0000	5.52506	5.00	20.00	10.0000	15.0000	20.0000

The descriptive statistics run on the experimental group's scores showed a very little decline from Test 1 (M= 7.25, SD= 4.12) to Test 2 (M= 7.00, SD= 5.47). However, the results showed a higher mean score on Test 3 (M= 11.50, SD= 5.87) and Test 4 (M= 14.00, SD= 5.52). The mean development is clear in Figure 1 as well.

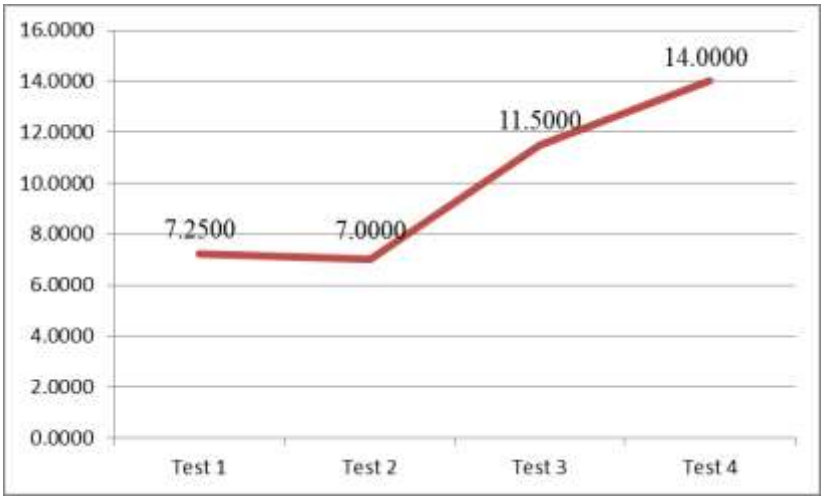

Figure 1 Mean development for the experimental group

Table 3 Results of Friedman Test

N	20
Chi-Square	14.030
df	3
Asymp. Sig.	.003

a. Friedman Test

The results of the Friedman Test showed that there was a statistically significant difference in the test scores across the four time points ($\chi 2$ (3, n = 20) = 14.03, p= .003). Inspection of the median values showed a decrease from Test 1 (*Md*= 10.00) to Test 2 (*Md*= 5.00). However, median values showed an increase to Test 3 and 4 (*Md*= 15.00).

Table 4 Wilcoxon Signed Ranks Test as multiple comparison

	T2 - T1	T3 - T1	T4 - T1	T3 - T2	T4 - T2	T4 - T3
Z	-.263[b]	-2.173[c]	-3.256[c]	-2.218[c]	-2.934[c]	-1.398[c]
Asymp. Sig. (2-tailed)	.793	.030	.001	.027	.003	.162

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

c. Based on negative ranks.

As the results of Friedman Test were significant, the Wilcoxon Signed Rank Test was run as a post-hoc to compare the sets of scores. Having considered the adjustment to the p-value as a result of multiple comparisons, i.e. six comparisons in all, the results showed statistically significant results from Test 1 to Test 4 (z= -3.25, p= .001) with a large effect size (r = -.51) and from test 2 to Test 4 (z= -2.93, p= .003) with a medium effect size which tends to be large (r = -.46).

Control Group's progress over time:

Table 5 Descriptive Statistics for the control group

						Percentiles		
	N	Mean	Std. Deviation	Minimum	Maximum	25th	50th (Median)	75th
T1	20	7.2500	3.02403	5.00	15.00	5.0000	5.0000	10.0000
T2	20	6.0000	4.75727	.00	20.00	5.0000	5.0000	10.0000
T3	20	8.0000	5.93828	.00	15.00	1.2500	10.0000	15.0000
T4	20	7.0000	5.23148	.00	20.00	5.0000	5.0000	10.0000

The descriptive statistics run on the control group's scores showed a decrease from Test 1 (M= 7.25, SD= 3.02) to Test 2 (M= 6.00, SD= 4.75). However, the results showed a higher mean score on Test 3 (M= 8.00, SD= 5.93) and a decrease on Test 4 (M= 7.00, SD= 5.23). The mean development is clear in Figure 2 as well.

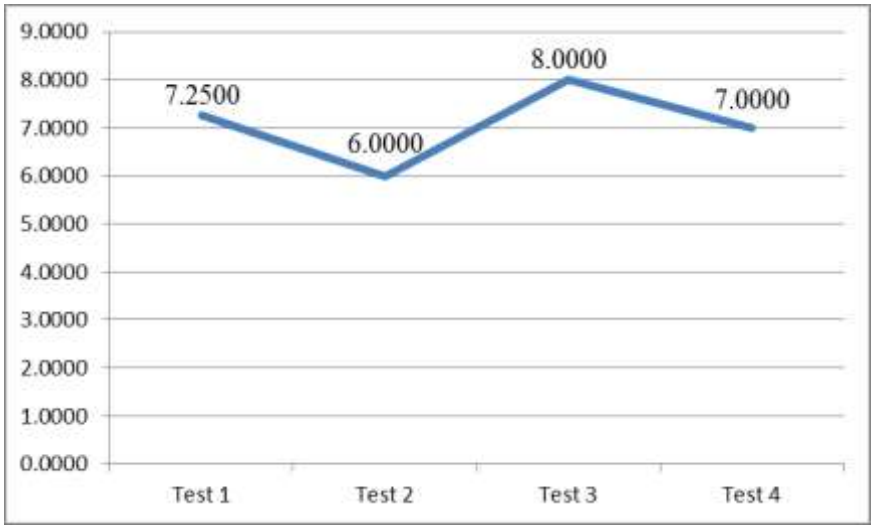

Figure 2 Mean development for the control group

Table 6 Friedman Test

N	20
Chi-Square	3.448
df	3
Asymp. Sig.	.328

a. Friedman Test

The results of the Friedman Test showed that there was a statistically non-significant difference in the test scores across the four time points ($\chi 2$ (3, n = 20) = 3.44, p= .328). Inspection of the median values showed no difference from Test 1 (*Md*= 5.00) to Test 2 (*Md*= 5.00). However, median values showed an increase to Test 3 (*Md*= 10.00) while a decrease to Test 4 (*Md*= 5.00).

To compare the performances of the two groups on each test, the non-parametric Mann-Whitney U test was used, which yielded the following results.

Table 7 Mann-Whitney U test to compare the performance of the groups on each test

	T1	T2	T3	T4
Mann-Whitney U	189.500	177.500	131.500	73.000
Wilcoxon W	399.500	387.500	341.500	283.000
Z	-.311	-.639	-1.961	-3.576
Asymp. Sig. (2-tailed)	.756	.523	.050	.000
Exact Sig. [2*(1-tailed Sig.)]	.779[b]	.547[b]	.063[b]	.000[b]

a. Grouping Variable: Groups

b. Not corrected for ties.

The results of the Mann-Whitney U Test showed a medium significant difference on Test 3 (U = 131.50, z = –1.96, p = .050, r = -.31) and a large significant difference on Test 4(U = 73.00, z = –3.57, p = .000,r = -.56). Therefore, the groups had similar performance on Test 1 and Test 2 while the experimental group outperformed the control one on Test 3 and Test 4.

Taken together, these results suggest that utilizing cognitive diagnostic assessment (CDA) design had significant effects on eliciting and improving the L2 learners' reading comprehension skills and strategies. Specifically, our results suggest that the more consistent is the CDA, the more

significant-level difference is observed between the terminal and initial tests. It means that any assessments in an applied setting should be implemented continuously to diagnose the students' strengths and weaknesses, and lay claim on investigating their construct in a specific domain.

Analysis of the Second Research Question

Before implementing CDA design, the students were under guided instruction on some reading strategies for one month, according to their high school textbooks which were based on reading comprehension. The cornerstone of the book has been founded on three main phases namely pre-reading, while-reading, and post-reading strategies. In fact, the authors of the book aimed at teaching the students how to code and specifically decode the data in the articles to become efficient writers and readers. Hence, through reading scientific articles of the average word count of 625, the students had to learn, practice and master the following strategies while responding to the textbook's exercises:

Skimming, scanning, using imagination, heeding the punctuation marks, ways of defining or restating the words, using surrounding words to conjecture the new words, using cues, word formation, finding the references, finding the main ideas, and using the contrastive words to relate the meanings of the sentences or paragraphs.

In addition to the aforementioned strategies, the list of strategies the students reported along with the teacher's feedback during the CDA administration is as follows:

Table 8 List of Elicited Strategies during CDA Administration

No.	Strategies
1	Reading the first line, or the first two lines of the passage, or the whole passage to find the main idea.
2	Reading the first and last line of each paragraph
3	Paraphrasing the sentences
4	Highlighting the cues
5	Underlining the cues
6	Translation
7	Writing notes like meanings and references on the words
8	Rereading the passage
9	Rereading the difficult sentences
10	Reading the passage chunk by chunk
11	Skipping unknown words
12	Heeding to capitalizations
13	Matching the item stem or alternatives with the related paragraphs in the text
14	Skipping the difficult alternative to check the other alternatives

15	Rejecting the wrong options in the test item
16	Skimming before scanning
17	Scanning the test items before skimming the text
18	Stopping reading the options when the correct option is reached
19	Heeding the negative markers
20	Recognizing suggestion, obligation, command, warning, deduction, and possibility cues in the text

Analysis of the third research question

The self-made assessment sheet of reading comprehension (see appendix B) was prepared so that it could help the researcher diagnose the factors which might hinder the students' grasping of a reading comprehension text either semantically or syntactically. As a matter of fact, the researcher intended to figure out if the main problem triggers responsible for crippling the candidates' reading comprehension were the unknown vocabularies or unfamiliar structures. Moreover, the researcher needed to be cognizant if the students used a specific strategy to fathom the text and respond to the relevant questions or not. Thus, if the students neglected some reading comprehension strategies, the instructor could scaffold them and fill the gaps. It is notable that though the tests were administered with no time limit, the students had to write the start and end time of the tests at the top of the test sheet. This was done in order to help the students better assess their capabilities and improvement while assuming that it would help them to relieve their anxiety.

However, the analysis of the contents of the structured interview conducted with the 20 participants in experimental group has confirmed the effectiveness of CDA design for L2 learners. The strategy instruction and providing the L2 learners timely diagnostic feedback and intervention succeeded in raising their consciousness of reading strategies. The frequency of the gleaned qualitative data through structured interview was analyzed through SPSS software. On the whole, the data analysis indicated that ninety percent of students could successfully recognize their strengths and weaknesses. During the CDA design implementation they were acquainted with reading strategies so as to better analyze and comprehend the passages. The time allocated to tests decreased in terminal tests (Test#3, Test #4) comparing the initial tests (Test#1, Test #2) administered during the study. The results of the interview also revealed that not knowing a couple of words in the reading comprehension passage was not so cloggy for students in Tests # 3 and # 4 any longer. Utilizing the taught reading comprehension skills and strategies, the dominant participants expressed their self-reliance and improvement in using the same skills and strategies in pilot tests administered by other organizations. However, the dominant weakness among the students was the inability to link the sentences (30%) in the reading comprehension passages and also hesitancy in choosing the correct option between 2 item alternatives out of four alternatives (20%). Regarding the negative effects of CDA on the participants, there was somehow a trade-off between negative and positive remarks. Thirty five percent of the students believed CDA had no negative effects, and the other thirty five percent of participants stated that irrelevant to the efficacy of the instructor's intervention and diagnostic feedback, it was boring and time-consuming since they had to fill in the self-assessment sheets during the assessment. More than 85 percent of the students predicted that they will be able to use the same skills and strategies in the approaching high-stakes Iranian University Entrance Exam 2016.

By way of summary, the overall attitude towards CDA indicated that the learners benefited from the key elements of CDA and perceived its role quite effective for their learning.

DISCUSSION AND CONCLUSION

The current quasi-experimental study investigated the impact of cognitive diagnostic assessment on the way learners' reading skills and strategies could be elicited through the interactive process occurred between the teacher and students in a continuous trend. The results of the study indicated that the participants in the experimental group benefited from the CDA instruction, and hence outperformed the control group. Thanks to the implemented design for experimental group, the teacher could raise the students' consciousness of taking reading skills and strategies in reading comprehension passages, and help them diagnose their strengths and weaknesses in the specific subskills. Likewise, through giving proper immediate diagnostic feedback to students, the teacher succeeded in motivating the students to attempt more to close the gap between their competence and performance. According to Hernández (2012, p. 491), "some of the limitations of a traditional approach to assessment may be overcome when students become actively engaged with the feedback and have to act upon it to improve their work on their learning" (Gibbs & Simpson, 2004; Hernández, 2008).

The post task individual and group conferences between the teacher and students led to identifying new test-taking strategies as well as other strategies to understand the texts. In other words, modification in instruction and intervention to meet learners' academic needs, as well as the change in students' attitude towards RC was strictly contingent upon introduction of the tests in an ongoing process. This is what was defined as washback validity by Messick (1996). Messick (1996) defined washback as "a good or bad practice that is *evidentially linked* to the introduction and use of the test" (cited in Baily, 1999, p. 5).

As mentioned earlier, the new cognitive approaches assume learning as reward for students rather than grades (see table 1). As Cauley and McMillan (2010, p. 2) stated "if students believe learning is important, they will exert greater effort". Several studies conducted on continuous assessment (McDowell, Sambell, Bazin, Penlington, Wakelin, Wickes & Smailes, 2005; Trotter, 2006) supported the students' learning and inspiring motivation in learners as Hernández (2012, p.499) explains "students, in particular, associate continuous assessment with motivation to learn on an ongoing basis and believe it provides opportunities to get feedback on their learning". Moreover, continuous feedback help students to "understand what they are learning, set goals, and self-assess…through self-assessment students can judge their own work, identify discrepancies between current and desired performance, and implement further learning activities to enhance their understanding or skills" (Cauley& McMillan, 2010, pp. 4, 5).

It seems that time-series design provides a condition for *ipsative assessment* which is defined by Gipps (1994) as an "assessment in which the pupil evaluates his/her performance against his/ her previous performance" (p. 1). Haugh (2011) stated that ipsative assessments focus on assessing how far the learners have improved comparing their previous performance, and appraising how effective was the feedback for their development. Thus, "ipsative feedback focuses on learner progress rather than a performance gap" (p. 2). Therefore, the time-series design with a '*teach-test-teach-test format'* can be an appropriate substitution for ipsative assessments through meeting the learners' language needs, closing their performance gap as well as considering their development comparing their previous performance.

The findings of the present study relying on the positive influence of CDA in engaging the students in learning process are aligned with Jang (2005, 2009) and Doe (2013). Most of the studies have been conducted on the pre- and post-test basis in assessment field, but the current study is the first time-series design which implemented four consistent tests, and thus could practically support the indispensability of conducting consistent CDAs, due to the significance of systematic gathering of evidence, and also due to the intrinsic nature of validation as an ongoing process in applied setting. Nonetheless, the present article hopes to be considered as an admissible offering to classroom teachers to conduct consistent CDA through time-series design. However, further studies are needed to be conducted to explore the effectiveness of CDA design in diverse contexts and settings, with learners of different genders and diverse language proficiency levels. Moreover, the findings of the study have been directed to the significance of detailed post-test reports and washback offered by

Bailey (1999) and Spolsky (1990). It is also needed to conduct feasible designs of CDA on value-laden subjects which search for the stakeholders' beneficial consequences. Furthermore, more studies are needed to appraise the efficacy of consistent CDA design in different language skills like listening, speaking, and writing. Finally, more studies are needed to integrate consistent CDA design with technology.

REFERENCES

Alavi, S. M., Kaivanpanah, S. & Shabani, K. (2012). Group dynamic assessment: An inventory of mediational strategies for teaching listening. *Journal of Teaching Language Skills*, *3*(4), 27-58.

Alderson J. C. (1990). Testing reading comprehension skills. *Reading in a Foreign Language*, *6*(2). 425-438.

Alderson, J. C. (2000). *Assessing reading*.UK: Cambridge University Press.

Alderson, J.C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London/New York Press.

Alderson, J. C., Brunfaut, T., & Harding, L. (2014). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics* 1-25. doi: 10.1093/applin/amt046.

Anderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, *8*(1), 41-66.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language testing*, *17*(1), 1-42.

Badrasawi, K. J. I., Kassim, N. L. A. & Daud, N. M. (2017). The effects of test characteristics on the hierarchical order of reading skills. *Malaysian Journal of Learning & Instruction*, *14*(1), 63-82.

Baird, J., Hopfenbeck, T. N., Newton, P., Stobart, G. & Steen-Utheim, A.T. (2014). *Assessment and learning: State of the field review*. Oslo: Knowledge Centre for Education.

Baily, K.M. (1999). *Washback in language testing*. Educational Testing Service, Princeton, New Jersey.

Berg, B.L. (2004). *Qualitative research methods*. Boston: Pearson Press.

Brown, J.D. (2005). *Testing in language programs*. Prentice Hall.

Cauley, K. M., & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *83*(1), 1-6.

Doe, C. D. (2013). *Validating the Canadian Academic English Language Assessment for diagnostic purposes from three perspectives: Scoring, teaching, and learning* (Unpublished Doctoral dissertation). Queen's University:Kingston, Ontario, Canada.

Doe, C. (2015). Student interpretations of diagnostic feedback. *Language Assessment Quarterly*, *12*(1), 110-135.

Farhady, H., Jafarpour, A., & Birjandi, P. (2009). *Testing language skills*. Tehran: *SAMT* Publishers.

Gallagher, N. (2000). *DELTA's key to the TOEFL test.* McHenry, IL: Delta.

Gibbs, G. & D. Simpson, C. (2004-05). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, *1*, 3-31.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. Psychology Press.

Gu, Z. (2011). *Maximizing the potential of multiple-choice items for cognitive diagnostic assessment* (Unpublished doctoral dissertation). University of Toronto.

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing*, *31*(1), 111-133.

Hancock, C. R. (1994). *Alternative assessment and second language study: What and why.* Eric Documents: ED376695.

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, *32*(3) 317-336.

Herna´ndez, R. (2008). *The challenges of engaging students with feedback*. Paper presented at the 4th Biennial EARLI/Northumbria Assessment Conference, 27–29. August, postdam/Berlin.

Hernández, R. (2012). Does continuous assessment in higher education support student learning? *Higher Education*, *64*(4), 489-502.

Huff, K., & Goodman, D. P. (2007).The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education*. 19-60. Cambridge: Cambridge University Press.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hughes, G. (2011). Towards a personal best: A case for introducing ipsative assessment in higher education. *Studies in Higher Education*, *36*(3), 353-367.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG-TOEFL* (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign, Urbana, IL.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chappelle, Y.-R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117-131). Ames, IA: Iowa State University.

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly*, *6*(3), 210-238.

Jang, E. E., & Wagner, M. (2014). Diagnostic feedback in language classroom. In A. Kunnan (Ed.), *Companion to language assessment* (pp. 693-711). Wiley-Blackwell.

Kiani, G. R., Alibakhshi, G., Akbari, R. (2009). On the consequential validity of ESP tests: A qualitative study in Iran. *The Journal of Applied Linguistics, 2*(1), 103-126

Kim, H. S. J. (2011). *Diagnosing examinees' attributes-mastery using the Bayesian inference for binomial proportion: A new method for cognitive diagnostic assessment* (Unpublished doctoral dissertation). Georgia Tech University.

Lee, Y. W., & Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly*, *6*(3), 239-263.

Lee, Y. W. (2015). Diagnosing diagnostic language assessment. *Language Testing, 32*(3), 299-316.

Liu, H. H. T. (2014). The conceptualization and operationalization of diagnostic testing in second and foreign language assessment. Teachers College. *Columbia University Working Papers in TESOL & Applied Linguistics*, *14*(1), 1-12.

Lougheed, L. (2003). *How to prepare for the TOEFL test*. Hauppauge, NY: Barron.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly: An International Journal*, *3*(1), 31-51.

McCann, T., & Clark, E. (2003). Grounded theory in nursing research: Part 2—Critique. Nurse Researcher, 11(2), 19-28.

McDowell, L., K. Sambell, V. Bazin, R. Penlington, D. Wakelin, H. Wickes, & J. Smailes (2005). *Assessment for learning: Current practice exemplars from the Centre for Excellence in Learning and Teaching.* Newcastle: Centre for Excellence in Teaching and Learning. University of Northumbria.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13, 241-256.*

Moss, C. M. (2013). Research on classroom summative assessment. In J. H. McMillan (Ed.), *Handbook of research on classroom assessment* (pp. 235–255). Los Angeles: Sage.

Nation, P. (2009). *Teaching ESL/EFL reading and writing*. New York, NY: Routledge.

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods*, *40*(4), 1001-1015.

Ravand, H., Barati, H., &Widhiarso, W. (2013). Exploring diagnostic capacity of a high Stakes reading comprehension test: A pedagogical demonstration. *Iranian Journal of Language Testing*, *3*(1), 12-37.

Rost, D. (1993). Assessing the different components of reading comprehension: Fact or fiction? *Language Testing*, *10*(1), 79-82.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language testing*, *23*(4), 441-474.

Rymniak, M. & Shanks, J. (2002). *TOEFL CBT exam*. New York: Kaplan.

Spolsky, B. (1990). Social Aspects of Individual Assessment. In de Jong, John H.A.L., & Stevenson, D.K. (Eds.), Individualizing the Assessment of Language Abilities, 3-15. Clevedon, Philadelphia: Multilingual Matters Ltd.

Sullivan, P.N., Brenner, G.A. & Zhong, G.L.Q. (2004). *Master the TOEFL 2005*. Lawrenceville, NJ: Thomson/Peterson.

Trotter, E. (2006). Student perceptions of continuous summative assessment. *Assessment and Evaluation in Higher Education, 31*(5), 505-521.

Appendix A

Self-assessment Sheet of Reading Comprehension

1. I knew the separate words, but couldn't relate them. Yes ☐ No ☐
2. I had difficulties with linking the sentences. Yes ☐ No ☐
3. I had problems with some structures that impeded the understanding the text.

   Yes ☐ No ☐ * If yes, name. …… .
4. I had problems with some lexicons that impeded the understanding the text.

   Yes ☐ No ☐ * If yes, name. …… .
5. Please write down the unknown vocabularies.

…….. .
6. Please write down the unknown structures.

…….. .
7. Explain the strategies you have used to understand the passage and respond the items.

Appendix B: Interview Questions

1. Did holding consistent cognitive diagnostic tests (CDA) help you identify your strengths and weaknesses in reading comprehension?
2. Talk about your strengths in reading comprehension.
3. Talk about your weaknesses in reading comprehension.
4. How far do you think consistent tests(CDA) helped you improve in reading comprehension?
5. What were the positive effects of CDA?
6. What were the negative effects of CDA?
7. Have you identified your actual abilities in reading comprehension?
8. Which skills and strategies helped you have good performance on reading comprehension?
9. Did appending 'general considerations' to your tests influence your performance in reading comprehension?
10. Did appending 'general considerations' to your tests decrease your anxiety during the assessment?
11. Did appending 'general considerations' to your tests help you have a better performance on reading comprehension?
12. Did holding untimed (with no time limit) tests affect your performance on reading comprehension?
13. How did holding untimed tests affect your performance on reading comprehension?
14. Did your teacher's feedback help you improve in reading comprehension?
15. Did your teacher's immediate feedback help you diagnose your strengths and weaknesses in reading comprehension?
16. Did your teacher's immediate feedback help you improve your performances in next tests?
17. How far peer-learning helped you improve in reading comprehension?

18. Now, what's your attitude toward reading comprehension after taking consistent tests followed by teachers' immediate feedbacks and self-assessment papers?
19. What's your prediction of your performance in reading comprehension part of University Entrance Examination or other reading comprehension tests held by other organizations?
20. Do you think you could use the same skills and strategies in reading comprehension tests of University Entrance Examination or other reading comprehension tests held by other organizations?
21. Please add anything else you would like about the design of CDA and the way it was implemented.