Journal of Studies in Learning
and Teaching English
Volume. 6, Issue. 2, Ser. 12, (2017), 55-78

A Comparative Study on the Effect of the Formative Use of Confidence-Based Scoring and Conventional Scoring on Iranian EFL Learners' Grammar Improvement

Firooz Sadighi
Professor of English Language and Linguistics
English Department, Faculty of Humanities
Islamic Azad University, Shiraz Branch
Shiraz, Iran
Email: firoozsadighi@yahoo.com

Mohammad Sadegh Bagheri
Assistant Professor of TEFL
English Department, Faculty of Humanities
Islamic Azad University, Shiraz Branch
Shiraz, Iran
Email: bagheries@gmail.com

Masoomeh Salehi*
Ph.D Candidate, Faculty of Humanities
Department of English Language
Islamic Azad University, Shiraz Branch
Shiraz, Iran
Email: mas.salehi84@gmail.com

Abstract. This study aimed at comparing the effect of the formative use of confidence-based scoring and standard scoring on the Iranian EFL learners' grammar improvement. Participants of this study were 72 freshman students majoring in English translation. The study was conducted in three English grammar classes. These classes were divided into two groups: one class was designated as Group 1 of the study, which was the control group, and the other two classes formed Group 2, which was the experimental group. In order to homogenize

Received: March 2017; Accepted: June 2017

*Corresponding author

the two groups, the grammar section of the January 2004 TOEFL test was given to all classes at the beginning of the semester, and the two groups were matched based on the students' scores on this test. In other words, the participants were selected in a way that the mean scores of the two groups on the TOEFL test were equal. Both groups received the same formative tests during the semester based on the content of the course. The formative tests in Group 1 were scored conventionally, and the formative tests in Group 2 were scored in a confidence-based manner. To compare the effect of the formative use of these two scoring methods on the Iranian EFL learners' grammar improvement, at the end of the semester, the final exam scores of the two groups were compared by an independent-samples t-test. Since the mean of the final exam scores of Group 2 was significantly more than that of Group 1, it was concluded that the formative confidence-based assessment was better than the regular formative assessment in improving learning English grammar.

Keywords: Confidence-based assessment, formative assessment, partial knowledge scoring

1. Introduction

In recent years, a great deal of literature related to education has dealt with innovative methods of evaluation. The search for innovative assessment methods has been motivated by different reasons such as enthusiasm of expert teachers, changes in the context of assessment (Gibbs, 2006), and external influences by governments, employers, and professional groups (Bryan & Clegg, 2006).

As a result of these changes, alternatives to traditional assessment practices have been proposed, and students are more actively involved not only in teaching and learning activities but also in assessment itself (Bryan & Clegg, 2006). One of the recent innovations in assessment practices is a confidence-based assessment. This type of assessment is not an alternative method of assessment. Instead, it is an alternative method of scoring or marking traditional tests. In other words, an alternative assessment refers to alternatives to traditional testing practices, especially standardized multiple-choice tests. However, a confidence-based assessment is a new way of marking the so-called objective tests, such as real/false, multiple-choice, matching, and even short answer tests. As its supporters suggest, the confidence-based assessment is a useful method for both formative and summative assessments (Gardner-Medwin & Gahan, 2003; Gardner-Medwin, 2006).

In a confidence-based assessment, the students are asked not only about the correct answer but also about their level of confidence in the answer they give. In the most common type of confidence-based assessments, there are three confidence levels: low, mid, and high. There are a couple of confidence-based marking schemes. Two of these schemes are used more commonly than others. In a scheme for true/false questions, the following marks will be given to each answer: correct with high confidence = 3, correct with mid confidence = 2, correct with low confidence = 1, wrong with low confidence = 0, wrong with mid confidence = -2, and wrong with high confidence = -6. In another scheme for multiple-choice questions, most of the given marks are the same as the above scheme except for the wrong answer with mid confidence, which receives -1 point, and wrong answer with high confidence, which receives -4 points. The reason for the difference between the penalties in these two marking schemes is that in true/false questions, the probability of answering an item by pure chance and getting the answer right is 50 percent. Therefore, the penalties must be higher in order to make students report their real level of confidence. It goes without saying that in both schemes if the question is left unanswered, the mark will be 0, but students usually do not leave questions unanswered (Gardner-Medwin, 2006).

1.1 Background of formative confidence-based assessment

Multiple-choice tests are among the most common test formats throughout the world. The traditional method of scoring multiple-choice tests, called number-right scoring, is to give one point for each correct answer and zero points for a wrong answer or no answer. However, the problem with this scoring method is that if the student does not know the answer at all, he or she has a 25-percent chance of getting the answer right through choosing one of the options by mere chance. To solve this problem, testing specialists proposed formula scoring, in which a wrong answer will receive a penalty which is a function of the number of choices (for example in four-choice questions, the penalty is -0.33) (Frary, 1988).

The shortcoming of both of the scoring methods mentioned above is that they do not take partial knowledge into account. They assume that a student either knows or does not know the answer to a question, and there is nothing in between. However, there are occasions when a student thinks that one of the options is correct but he or she is not sure about it (Burton, 2002). In reaction to this shortcoming of number-right scoring and formula scoring, many scholars proposed methods of scoring partial knowledge. These methods include: "(1) confidence weighting (or confidence-based assessment), (2) answer-until-correct scoring, (3) option weighting, (4) elimination and inclusion scoring, and (5) multiple answer scoring" (Kurz, 1999; p. 1).

The confidence-based assessment has a history of more than 80 years. According to Echternacht (1972), the first article published about confidence-based assessment was Henver (1932). The participants in this study were given two true/false tests about aesthetics and music. To score these tests, four different scoring methods were used: (1) the number right, (2) the number right minus the number wrong, (3) a weighted right-answer score, and (4) a weighted-right minus a weighted-wrong score. Using reliability estimate by Spearman-Brown formula, Henver (ibid) concluded that the best method of scoring was the weighted-right method.

Another early study on the confidence-based assessment was Soderquist (1936). This study was conducted with a true/false test given to university students in their extracurricular classes. Two different methods were used to score the tests: the weighted-right minus weighted-wrong, and the number right minus the number wrong. Reliability estimates using the Spearman-Brown formula showed that the weighted-right minus the weighted-wrong scores had a higher reliability. This finding is not precisely consistent with Henver's (1932) findings which showed that the reliability of weighted-right scores was more than that of weighted-right minus weighted-wrong scores. Echternacht (1972) attributed this difference in the results of these two studies to the fact that Soderquist's participants were aware of the scoring method while Hevner's participants were not.

In the 1940s, there was a decline in the number of confidence-based assessment studies, but after that time, studies with this topic continued to be conducted. Of course, at that time, the name "confidence-based assessment" was not used. The more common names for this method were "confidence weighting of scores" or simply "confidence testing". The designation "confidence-based assessment" began to be used in the 1990s.

For years, the main focus of such studies was on reliability and validity of confidence-based scores. A few studies (e.g., Dressel & Schmid, 1953) also tried to show that confidence-based tests were better able at discriminating students of diverse abilities. In the 1970s, the confidence-based assessment began to be criticized because it might favor examinees with specific personality traits (e.g., Jacobs, 1971). Therefore, after that time, some studies investigated the relationship between confidence-based assessment and specific personality traits.

The studies investigating the reliability, validity, and personality bias of confidence-based assessment looked at confidence-based assessment as an alternative scoring method for summative tests. It was only in the 1980s when the idea of using confidence-based assessment for formative tests was proposed. Testing and education specialists realized that the level of a student's confidence in his or her answer could be used as a basis for improving learning and teaching. Pioneers of formative confidence-based assessment were Darwin Hunt and James Bruno. The preliminary studies of these scholars were done in the 1980s, but their leading papers were published in the 1990s.

Hunt (1993) referred to "usable" and "unusable" knowledge. According to him, knowledge is usable when a person is sufficiently sure of the knowledge so that it will be used to make decisions and take actions. Knowledge is unusable when a person is not confident enough to use it for deciding and taking actions. Usable knowledge may be either correct or incorrect. Hunt believes that the goal of education and training should be to help learners acquire and retain knowledge which is both correct and uesable and to identify and remedy useable knowledge which is incorrect. The way to achieve this goal is incorporating learners' self-assessment about their confidence level into our assessment practices.

Bruno (1987) also referred to a similar notion and proposed that there is a link between knowledge, confidence, and behavior. He stated that knowledge alone is necessary but not enough to create action. Instead, it is the combination of knowledge and confidence which(leads to behavioral outcomes and empowers people to act. People who are confidently correct will take actions that are productive. The reverse is also true in that) people who are confident about misinformation will also take action, but the results of this effort is consistently negative and potentially dangerous. Based on this belief, Bruno proposed his two-dimensional assessment process, which was initially called Information Reference Testing (IRT). These two dimensions are knowledge and confidence, the combination of which makes the knowledge-behavior quadrants (Figure 1). The learner's response will fall into one of these quadrants, and based on this, it is decided how much more teaching and practice is needed until the learner achieves mastery.

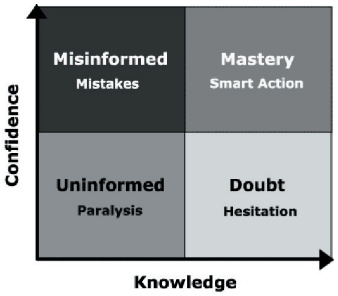
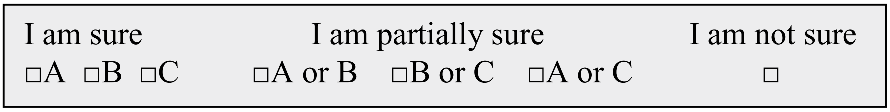


Figure 1. Bruno's (1987) knowledge quadrants

The most comprehensive source explaining Information Referenced Testing is Bruno (1995). Unlike other confidence-based marking schemes, in Bruno's model, the identification of correctness and confidence is made simultaneously. This method, which is used with three-choice questions, is shown in Figure 2. This process is not suitable for questions with more than three choices because in that case, there will be a lot of choices in "I am partially sure" part



I am sure	I am partially sure	I am not sure
□A □B □C	□A or B □B or C □A or C	□

Figure 2. Bruno's (1995) assessment model

In 2005, Bruno made a piece of software based on his IRT model. Since then, the name IRT was replaced by Confidence-Based Learning.

In a more recent study, Novacek (2013) provided a review of studies conducted on the confidence-based assessment. He stated that the confidence-based assessment offers “a middle ground between the traditional multiple-choice answer and a lengthy essay response” and concluded that it “increases the retainability of learned material and identifies topics in which people are misinformed” (p. 403).

Confidence-based learning and formative confidence-based assessment are not just theoretical issues. A number of empirical studies have been carried out to examine their usefulness. These studies are reviewed in the next section.

1.2 Empirical studies on formative confidence-based assessment

The confidence-based assessment is not just a summative testing method. Rather, it is a tool for improving learning when it is used throughout a course. There are many studies in the literature which refer to the benefits of using the confidence-based assessment to improve learning.

One of the earliest studies on formative confidence-based assessment was conducted by Issroff and Gardner-Medwin (1998) who offered free online confidence-based tests to medical students at University College London. Results of their study at the end of the first-year medical course showed a high level of voluntary use, particularly towards exam time. In the questionnaires given to the participants of this study, most students considered the confidence assessment useful to them. Many students found that they were helped in identifying their strengths and weaknesses and in distinguishing between knowledge, misconception, and guesswork.

In another study, Khan, Davies, and Gupta (2001) used a web-based formative confidence-based assessment system with undergraduate clinical students. This study was not of experimental design, and no comparison has been made between the students working with this system and those who did not. However, the authors believe that this system, which allowed measurement of confidently possessed the correct knowledge and identified misinformation, was helpful both to the students and the teachers.

Another study with a different design was conducted by Davies (2002) on undergraduate students of the computer. He gave four tests to his participants during a semester and scored the tests based on the students' confidence level. The same tests had been given to another group of students the previous year, but they were scored conventionally. The results showed that the group of participants whose tests were scored based on their confidence levels gained better results than the previous-year students.

Hunt (2003) also states that self-assessment of responses by test takers improves learning. Moreover, Hunt (1982) showed that the order of executing the answer and the self-assessment response affects learning. Hunt's (1982) study demonstrated that people who first answered followed by the self-assessment response were more accurate, i.e., better able to discriminate between knowing and not knowing the correct answer than people who first gave the self-assessment response followed by the answer.

In a correlational study, Gvozdenko and Chambers (2007) incorporated self-assessed confidence into online formative and summative assessment in basic mathematics skills administered to tertiary students. Running a correlation between the self-assessed confidence and the accuracy data obtained through another computerized test, they showed the usefulness of self-assessment of confidence to the learners.

Serradell-Lopez, Lara-Navarra, Castillo-Merino, and Gonzalez- Gonzalez (2010) also conducted a study on the formative confidence-based assessment. They gave two confidence-based tests to 200 undergraduate students of business administration and management at two different times during a course. They found that in the second test, both the knowledge and the confidence of the students increased in comparison with the first test. They found the formative confidence-based assessment useful in improving learning.

In a qualitative study, Florian (2010) added the confidence-based assessment (CBA) to a learning management system (LMS) used for secondary high school students. The data for this study was gathered from teachers, administrators, and content developers through open-ended questionnaires and interviews. Most of the participants referred to the usefulness of CBA-embedded LMS as a tool for improving learning.

Cash, Mitchner, and Ravyn (2011) used Online Confidence-Based Learning software as a tool for improving learning by students of medicine. Most of the participants (61.8% in one course and 98.2% in another course) achieved mastery after using this method. The authors conclude that using Confidence-Based Learning is useful in addressing knowledge and confidence gaps.

The use of formative confidence-based assessment has also been shown to improve motivation to learn. Nix and Wyllie (2011) were after finding ways to enhance learner' motivation to use the formative computer-based assessment (CBA). They showed that adding a confidence indicator tool, which allowed learners to indicate their confidence in the correctness of their answer before answering a question, was effective in enhancing learners' motivation to use CBA and improving learning as a result voluntarily.

Another study which collected its data by questionnaires was Schoendorfer and Emmett (2012). They provided free online formative assessment tools for the second-year medical students at the University of Queensland. Out of the 400 students, about 10% opted for certainty-based marking, while about 43% preferred the usual multiple-choice tests. However, most of the students either agreed or strongly agreed that certainty-based marking was useful (57%) and easy to understand (52%), but took more time (67%). Based on these findings, the authors concluded that using these practices in educational settings is conducive to deeper levels of learning.

The most recent study conducted on the formative confidence-based assessment is the one conducted by Kampmeyer, Matthes, and Herzig (2015). They compared the usefulness of formative confidence-based assessment for 3rd- and 5th-year medical students and showed that the confidence-based assessment was more useful for advanced students than for beginners. The studies conducted on confidence-based assessment in recent years show that the interest in confidence-based assessment is still alive. However, the number of studies dealing with the formative use of confidence-based assessment is less than those dealing with its summative use. Furthermore, not many of these studies are related to language learning. This shortage of studies was the motive for conducting the present study.

2. Research Question

This study tried to address the following research question:

- Is there a significant difference between the effect of the formative use of confidence-based scoring and conventional scoring on Iranian EFL learners' grammar improvement?

In other words, the aim of this study was to see whether formative scoring tests based on the students' confidence levels are more efficient than conventional scoring in improving learning and leading to better final exam scores.

3. Method

3.1 Participants

The participants of this study were English translation freshman students. The participants were both male and female, and their ages ranged from 19 to 35. The study was conducted in three classes of English Grammar 2 at the Islamic Azad University, Shahr-e-Qods Branch, Iran, and all the three classes had the same instructor. The instructor was utterly familiar with the confidence-based assessment and the scoring method used. The reason for choosing English Grammar 2 rather than English Grammar 1 was that in English Grammar 1 classes, the students were less homogeneous. However, the students in English Grammar 2 classes were more homogeneous because some of the less proficient students failed Grammar 1 and were not allowed to enroll in Grammar 2.

To examine the usefulness of confidence-based assessment as a formative assessment tool, the three classes of students were divided into two groups. The first group, which received conventional formative tests, consisted of one class of 29 students. The second group, which received confidence-based formative tests, consisted of the other two classes of 49 students on the whole.

The rationale for dedicating one class to Group 1 and dedicating two classes to Group 2, rather than dedicating one class to each group, was that the two groups were going to be homogenized based on the students' score on a TOEFL test which would be given to them at the beginning of the semester, and the researchers had to exclude some students from the study in order to have two homogeneous groups. If one class had been dedicated to each group and some students had been excluded from each class so that the two groups would be homogeneous, too few students would have remained in each group to conduct this study. Therefore, it was decided to dedicate one class to Group 1 and two classes to Group 2, keeping Group 1 unchanged, and excluding some students of Group 2 from data analysis.

Since the participants had enrolled in any of the three classes based on their preferences and it was not possible to give them a pre-test and force them to enroll in any of the classes based on the pre-test scores, a homogenizing test was given to the participants to make the results of the two groups comparable. This test was the grammar section of a TOEFL test. After calculating the scores of the participants in this test, six of the students with the lowest scores in this test was omitted from Group 2. In this way, the means of the scores of the two groups became almost equal (Group 1=16.96, Group 2=16.95). By doing this, it was made sure that the difference between the final exam scores of the two groups was due to different kinds of formative tests given to them, not because of their previous knowledge of grammar. Therefore, the comparison between the final exam scores was done with 29 students in Group 1 and 43 students in Group 2. As mentioned earlier, the reason for choosing three classes for this study was that if two classes had been used, few students would have remained after the homogenization process. Moreover, there weren't four classes of Grammar 2 in the university where the study was conducted. Therefore, three classes were chosen: one class as Group 1, and two classes as Group 2. That's the reason for the unequal number of participants in the two groups.

3.2 Instruments

The first instrument used in this study was the Structure and Written Expression section of the paper-based TOEFL test belonging to January, 2004. It goes without saying that this part of the test included 40 multiple-choice questions. This test was used to homogenize the two groups of the students and to make sure that the participants' previous knowledge of English grammar would not affect the final results.

Other instruments used in this study included eight formative tests and a final exam. Both the formative test and the final exam were prepared by the researchers based on the material taught during the course. In Grammar 2 classes of this university, eight chapters (chapters 4 to 10 and chapter 13) of the book Communicate What You Mean (Pollock, 1997) are taught. Each of the formative tests was based on one of these chapters and included ten multiple-choice questions, and the final exam included 60 multiple-choice questions about the content of all the eight chapters.

In order to make sure of the content validity of the researcher-made tests (i.e., the eight formative tests and the final exam), three university instructors who had Ph.D. in TEFL were asked to read the tests and compare them with the content of the book. They did so and confirmed the content validity of the tests. The reliabilities of the tests were also calculated using Cronbach's alpha formula. The reliability of the formative tests, on the whole, was 0.86 and the reliability of the final exam was 0.88.

3.3 Procedures and data collection

This study was conducted in three English Grammar 2 classes at the Islamic Azad University, Shahr-e-Qods Branch. At the beginning of the semester, the three classes were divided into two groups. The next step was to give the participants a test to make sure of the homogeneity of the two groups considering their knowledge of grammar. The Structure and Written Expression Section of a TOEFL test were used for this purpose.

The next step was to give the formative tests to the participants. As mentioned before, eight chapters of the book Communicate What You Mean (Pollock, 1997) (Chapters 4 to 10 and Chapter 13) are usually taught in English Grammar 2 classes of this university. After teaching each chapter, the instructor told the students that a multiple-choice test of that chapter would be given to them in the next session. Therefore, the participants received eight formative tests during the semester.

In Group 1, the test papers were scored conventionally, and nothing was mentioned about the students' confidence in their answers. In Group 2, the students were asked about their level of confidence in their answer after each question. There were three choices for confidence level: low, mid, and high. To score these papers, the following method was used: correct answers with high confidence received 3 points, correct answers with mid confidence gained 2 points, correct answers with low confidence received 1 point, wrong answers with low confidence received 0 points, wrong answers with mid confidence received -1 point, and wrong answers with high confidence received -4 points. Before giving the first formative test to the students, this scoring method was explained to them.

In both groups, after administering each test, the papers were photocopied. One copy was kept with the researchers for data analysis, and one copy was scored based on the scoring methods mentioned above and was given back to the students so that they could work on their areas of strength and weakness. In Group 2, the copy which was given to the student was scored only based on confidence levels, but the copy which was kept with the researchers was scored both conventionally and in a confidence-based manner for further data analysis. To make sure that all students of both groups took all the tests, those students who were absent in one of the testing sessions were given the test in the next session. In this way, all students in both groups sat for all of the eight tests, and the only difference between the two groups was the method of scoring the test papers.

At the end of the semester, the students took the final exam. In this test, the students were not asked about their confidence level, and the papers were scored conventionally. At this point, the data collection stage of this study was finished, and the researchers began to analyze the data.

3.4 Data analysis

Before elaborating on different data analysis methods of this study, one point should be mentioned here. All the tests used in this study (i.e., the homogenizing test, the formative tests, and the final exam), were scored twice by one of the authors, and wherever a mistake was found, it was corrected.

The first step in the data analysis of this study was scoring the homogenizing test. For each correct answer in this test, one point was considered, so the range of scores was between 0 and 40. Then the mean of each group was calculated. Since the mean of the first group (16.96, N = 29) was more than that of the second group (15.98, N = 49), six of the lowest scores of the second group were omitted, and mean of the second group (16.95, N = 43) became almost the same as the first group.

At the end of the semester, after the participants took their final exam, the mean scores of the two groups in the final exam were calculated (with 29 students in group one and 43 students in group two). To test the significance of the difference between the two means, independent sample t-test was run.

4. Results

As mentioned before, to see whether the confidence-based assessment is a good formative assessment tool, eight identical multiple-choice tests were given to two groups of students. In one group, the tests were scored conventionally, and in the other group, the tests were scored based on the test takers' confidence level. Then, at the end of the semester, the final exam scores of the two groups of students were compared. However, before presenting the results of the final exam scores, some points about the scores obtained in the formative tests themselves are provided below.

Table 1 summarizes the results of the formative tests themselves in Group 1, whose formative tests were scored conventionally only, and Group 2, whose formative tests were scored both conventionally and in a confidence-based manner.

Table 1: Summary of the results obtained from the formative tests

	Mean of test 1	Mean of test 2	Mean of test 3	Mean of test 4	Mean of test 5	Mean of test 6	Mean of test 7	Mean of test 8	Mean of all tests
Conventional scores of Group One (N=29)	7.24	7.10	6.48	7.27	5.03	5.41	5.97	8.24	6.59
Conventional scores of Group Two (N=43)	7.40	6.40	5.77	6.02	4.27	5.44	6.65	7.41	6.17
Confidence-based scores of Group Two	5.05	3.45	2.52	2.76	0.92	2.48	3.86	5.17	3.28

As it can be seen in Table 1, the confidence-based scores of Group 2 are less than their conventional scores. This stands to reason entirely because, in the confidence-based marking, there are negative scores while there is no negative score in the conventional marking. Therefore, a person's confidence-based score on a test is less than or, at best, equal to his or her conventional score in the test. A more important point which can be seen in this table is that even when just the conventional scores of Group 2 are considered, the scores of this group are slightly less than those of group one. This point will be elaborated on later in the next section.

One might expect the scores of the formative tests to improve as more and more tests are given to students because students learn about their areas of strength and weakness as they take more tests. Considering this study in particular, one might expect the scores of Group 2 to increase more than those of Group 1 because, as the advocates of confidence-based assessment claim, the confidence-based assessment is better than conventional assessment in helping students learn about their areas of strength and weakness. These expectations are quite logical when each formative test covers the materials included in the previous tests. However, in this study, the materials covered in each formative test were different. In the present study, the formative tests helped student work on their areas of strength and weakness for the final exam not for the next formative test.

As mentioned earlier, after taking the formative tests during the semester, the participants of this study received a final exam at the end of the semester. The summary of the results obtained from the final exam scores is provided in Table 2.

Table 2: Final exam results of the two groups of the study

	N	Mean	Std. deviation	Std. The error of the mean
Group 1	29	30.45	9.91	1.49
Group 2	43	36.35	9.77	1.84

As it can be seen in Table 2, the mean of Group 1 (N=29) was 30.45, and the mean of Group 2 (N=43) was 36.35. To see whether the difference between the two means was significant or not, an independent sample t-test was run (see Table 3).

Table 3: Statistics of the independent sample t-test

	t	df	Sig. (2-tailed)	Mean difference	Std. the error of the difference
Final exam scores	2.50	70	.015	5.9	2.36

The result of independent-samples t-test showed that the two-tailed significance of the difference between these two means was 0.015. Therefore, considering the final exam scores, Group 2, which received the confidence-based formative assessment, significantly ($p < 0.05$) outperformed Group 1, which received the conventional formative assessment.

The effect size was also calculated with Cohen's d formula, and the resulting number was 0.29. Although this effect size is relatively small, it still indicates the usefulness of confidence-based formative assessment (see Section 5 for more elaboration). The results of the present study demonstrate that the confidence-based formative assessment was more conducive to students' better learning and better final exam performance as a result.

5. Discussion

The results of this study showed that the confidence-based assessment was a better tool for formative assessment than conventional assessment despite the fact that the effect size was rather small. This small effect size can be attributed to the unfamiliarity of participants with the scoring method in the confidence-based assessment. Garder-Medwin (1995) mentions that the participants of his study had difficulty in understanding the scoring system, but this problem was removed as they gained more experience in this regard.

The reason why Group 2 outperformed Group 1 is that the students in Group 2 were better able to work on their areas of strength and weakness and improve their final exam scores than Group 1. The fact that the scores of Group 2 in the formative tests themselves were not as good as those of Group 1 is a further proof of this claim. Although in Group 2, the formative test scores were lower, the students in this group obtained better scores in the final exam, which shows that they could work on their areas of weakness and improve their learning better than Group 1. This is similar to what the participants of Issroff and Gardner-Medwin's (1998) study have reported.

The results of this study are consistent with previous studies conducted on the formative use of confidence-based assessment. One of these studies whose design is the most similar to the one used in the present study is that of Davies (2002). In that study, two groups of students received formative tests, one of which was scored conventionally and the other one was scored based on confidence levels. The group which received the confidence-based formative assessment gained better scores in the final exam than the other group. As it is evident, the results of the present study are very similar to those obtained by Davies.

The results of this study are also similar to those of Gvozdenko and Chambers' (2007). Running a correction between confidence-based scores of formative tests and the accuracy of answers to another test, they showed the usefulness of confidence-based formative assessment. Although the design of the present study is different from that of Gvozdenko and Chambers, the findings of this study are consistent with theirs because both of them show the usefulness of confidence-based formative assessment.

Some other studies cited in Section 1 also reported the usefulness of formative confidence-based assessment in different contexts (Cash et al., 2011; Hunt, 1982; Khan et al., 2001, Serradell-Lopez et al., 2010). Therefore, the results obtained here agree with those of the cited studies.

Some other studies cited in Section 1 used questionnaires or interviews to elicit the participants' opinions about formative confidence-based assessment (Florian, 2010; Issroff & Gardner-Medwin, 1998; Schoendorfer & Emmett, 2012). In these studies, most of the participants agreed that the formative confidence-based assessment was conducive to better learning. Although these studies had a method different from that of the present study, the results of all of them show that the formative confidence-based assessment is effective in improving learning.

The results obtained in this study can be explained in this way: in the conventional formative assessment, when an item is answered correctly, no further practice is done by the students and the teacher because they take it for granted that the material in question is learned. However, in the confidence-based formative assessment, a correct answer with mid or low certainty level is considered as partial knowledge, so both the students and the teacher try to change it into full knowledge through more practice. The students will go back to the point in question and work on it until they change their partial knowledge into full knowledge. Moreover, the teacher will probably reteach that point and provide the students with more exercises (Garder-Medwin, 1995; Issroff & Gardner-Medwin, 1998).

On the other hand, in the conventional formative assessment, a wrong answer is considered as lack of knowledge, while it may mean wrong information. In the confidence-based formative assessment, a wrong answer with mid or high certainty level means that the student has a misconception or delusion (Gardner-Medwin & Gahan, 2003). Therefore, the students try to find the reason for their misconception or fallacy, and the teacher realizes that most probably there had been a mistake in the teaching method with regard to the material in question and tries to remove those misconceptions and delusions (Gardner-Medwin & Curtin, 1996, cited in Issroff & Gardner-Medwin, 1998).

According to Gardner-Medwin (1995), "confident belief in answers that are wrong is far worse than the recognition that one simply does not know the answer" (p. 81). Therefore, changing wrong information into correct information is more important than gaining some information which used to be lacking, and it needs much more time and effort. The reason is that in the former case, the person should delete the wrong information first and then gain the correct information. Since in conventional assessment the difference between lack of knowledge and wrong information is not clear, the students and the teacher may not allocate enough time and effort to the point in question. However, such a distinction is made in the confidence-based assessment. That's another reason why Group 2 gained better final exam results than Group 1 in the present study. Based on the results obtained here, the following conclusions are made.

6. Conclusions and Implications

The results of this study showed that the confidence-based scoring of formative tests is a useful method for increasing the effectiveness of these tests and improving learning and teaching. Maybe it is a too broad conclusion to say that the confidence-based assessment is a good formative assessment tool in all educational contexts or even in all language classes. Since not many studies have been conducted on the formative use of confidence-based assessment, the usefulness of formative confidence-based assessment cannot be generalized to all subject matters, or even to all components and skills of language. However, the most reasonable conclusion which can be drawn from the results of this study is that using the confidence-based marking informative tests is an effective way of improving learning in English grammar classes.

Based on the above conclusion, the main pedagogical implication of this study is that teachers in general, and English grammar teachers in particular, should incorporate formative confidence-based assessment into their teaching processes. Using the confidence-based assessment during a course is helpful for both students and teachers.

This method helps students understand the difference between knowledge, uncertainty, ignorance, misconception, and delusion. In a conventional assessment, when a student gives a correct answer to a test item with low confidence, he or she usually does no more practice on that concept because he or she has received the full score for that item. However, in the confidence-based assessment, such an answer does not receive the full score, which leads the student to do more practice on that concept and change his or her uncertainty into full knowledge. In the confidence-based assessment, the student knows that when he or she answers an item correctly but with uncertainty, he or she may answer a similar question wrongly the next time a similar test is given to him or her. Therefore, such a student tries to learn that concept completely to make sure of not losing any scores in other similar tests to come. This is also the case with wrong answers. In a confidence-based assessment, students know that a wrong answer with high certainty needs much more practice than a wrong answer with mid or low certainty because the probability of giving a wrong answer to a similar item in the next tests is much more in the former case.

The confidence-based assessment also helps teachers gain information about the students' areas of strength and weakness and know which parts of the course materials need more teaching and practice in class. In a conventional assessment, when a teacher sees that most students have answered an item correctly, he or she thinks that they have learned that concept. However, in a confidence-based assessment, if most of the correct answers are given with low or mid certainty levels, the teacher realizes that the concept still needs some further teaching and practice. About wrong answers, when the teacher sees that most of the wrong answers are given with high certainty, he or she should reconsider his or her method of teaching on that concept and allocate more class time to teaching that concept again. As discussed in Section 1, the difference between ignorance and delusion is great. Ignorance or lack of knowledge does not usually lead to action, while misconception and delusion usually lead to dangerous actions. When confidence-based tests are delivered to students during a course, the teacher can distinguish ignorance from misconception and delusion. Then he or she should try to prevent the occurrence of dangerous actions by removing any misunderstanding or delusion in class.

As mentioned earlier, the conclusions of this study are only generalizable to English grammar classes. However, formative confidence-based assessment seems to be useful in classes on other language skills and components and even in other subject matters. Therefore, further research can be conducted in this area to investigate the usefulness of formative confidence-based assessment in classes of different language skills and components, integrative language classes, and even in classes of other subject matters.

References

Bruno, J. E. (1987). Admissible probability measurement in instructional management. *Journal of Computer-Based Instruction, 14*(1), 23-30.

Bruno, J. E. (1995). *Information reference testing (IRT) in corporate and technical training programs.* Los Angeles: UCLA.

Bryan, C. & Clegg, K. (2006). *Innovative assessment in higher education.* London: Routledge.

Burton, R. F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Education, 36*(9), 805-811.

Cash, B., Mitchner, N. A., & Ravyn, D. (2011). Confidence-based learning CME: Overcoming barriers in irritable bowel syndrome with constipation. *Journal of Continuing Education in the Health Professions, 31*(3), 157-164.

Davies, P. (2002). There is no confidence in multiple-choice testing. *Proceedings of the* 6th *International Computer-Aided Assessment Conference,* Loughborough, 119-130.

Dressel, P. L. & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement, 13*(4), 574-595.

Echternacht, G. J. (1972). The use of confidence testing in objective tests. *Review of Educational Research, 42*(2), 217-236.

Florian, T. P. (2010). *Confidence-based assessment in Moodle: Insights from teachers, administrators, and programmers.* Ann Arbor: ProQuest LLC.

Frary, R. B. (1988). Formula scoring of multiple choice tests (correction for guessing). *Instructional Topics in Educational measurement, 7*(2), 33-38.

Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *Research in Learning Technology, 3*(1), 80-85.

Gardner-Medwin, A. R. & Gahan, M. (2003). Formative and summative confidence-based assessment. *Proceedings of the 7th International Computer-Aided Assessment Conference,* Loughborough, 147-155.

Gardner-Medwin, A. R. (2006). Confidence-based marking: Towards deeper learning and better exams. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. 141-149). London: Routledge.

Gibbs, G. (2006). Why assessment is changing. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education,* (pp. 11-22). London: Routledge.

Gvozdenko, E. & Chambers, D. (2007). Applying computerized testing & certainty based assessment to reveal more about student learning: From theory to practice. *International Journal of Learning, 13*(12), 205-216.

Hevner, K. (1932). A method of correcting for guessing in true-false tests and empirical evidence in support of it. *The Journal of Social Psychology, 3*(3), 359-362.

Hunt, D. P. (1982). Effects of human self-assessment responding on learning. *Journal of Applied Psychology, 67,* 75-82.

Hunt, D. P. (1993). Human self-assessment: Theory and application to learning and testing. In D. Leclercq & J. E. Bruno (Eds.), *Item bank: Interactive testing and self-assessment,* (pp. 177-189). Berlin: Springer Verlag.

Hunt, D. P. (2003). The concept of knowledge and how to measure it. *Journal of Intellectual Capital, 4*(1), 100-113.

Issroff, K. & Gardner-Medwin, A. R. (1998). Evaluation of confidence assessment within optional coursework. In M. Oliver (Ed.), *Innovation in the evaluation of learning technology,* (pp. 169-179). London: London Press.

Jacobs, S. S. (1971). Correlates of unwarranted confidence in responses to objective test items. *Journal of Educational Measurement, 8*(1), 15-19.

Kampmeyer, D., Matthes, J., & Herzig, S. (2015). Lucky guess or knowledge: A cross-sectional study using the Bland and Altman analysis to compare confidence-based testing of pharmacological knowledge in 3rd and 5th-year medical students. *Advances in Health Sciences Education, 20*(2), 431-440.

Khan, K. S., Davies, D. A., & Gupta, J. K. (2001). Formative self-assessment using multiple true-false questions on the Internet: feedback according to confidence about correct knowledge. *Medical Teacher, 23*(2), 158-163.

Kurz, T. B. (1999). *A review of scoring algorithms for multiple-choice tests.* Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio. Retrieved from http://files.eric.ed.gov/fulltext/ED428076.pdf

Nix, I. & Wyllie, A. (2011). Exploring design features to enhance computer?based assessment: Learners' views on using a confidence?indicator tool and computer?based feedback. *British Journal of Educational Technology, 42*(1), 101-112.

Novacek, P. (2013). Confidence-based assessments within an adult learning environment. *Proceedings of IADIS international conference on cognition and exploratory learning in the digital age (CELDA 2013),* 403-406.

Pollock, C. W. (1997). *Communicate what you mean: A concise advanced grammar (2nd ed.).* New York: Longman.

Schoendorfer, N. & Emmett, D. (2012). Use of certainty-based marking in a second-year medical student cohort: A pilot study. *Advances in Medical Education and Practice, 3,* 139-143.

Serradell-Lopez, E., Lara-Navarra, P., Castillo-Merino, D., & Gonzalez-Gonzalez, I. (2010). Confidence-based learning in investment analysis. *Technology Enhanced Learning, 73,* 28-35.

Soderquist, H. O. (1936). A new method of weighting scores in a true-false test. *Journal of Educational Research, 30,* 290-292.