The Journal of Teaching Language Skills (JTLS)
6 (4), Winter 2015, Ser. 77/4. ISSN: 2008-8191
pp. 123-152

Comparing Confidence-based and Conventional Scoring Methods: The Case of an English Grammar Class

Masoomeh Salehi*
Ph.D student in TEFL
Islamic Azad University of Shiraz
email: innocentsalehi@yahoo.com

Firooz Sadighi
Professor, TEFL
Shiraz University
email: firoozsadighi@yahoo.com

Mohammad Sadegh Bagheri
Assistant Professor, TEFL
Islamic Azad University of Shiraz
email: bagheries@gmail.com

Abstract

This study aimed at investigating the reliability, predictive validity, and self-esteem and gender bias of confidence-based scoring. This is a method of scoring in which the test takers receive a positive or negative point based on their rating of their confidence in an answer. The participants, who were 49 English-major students taking their grammar course, were given 8 multiple-choice tests during the semester. These tests were scored both conventionally and in a confidence-based manner, and the reliabilities of these two score sets were compared. Each score set was correlated with the final exam scores to compare their predictive validity. Gender and self-esteem bias of the confidence-based scores of the eight tests were also calculated. The results showed that there was no difference between the reliabilities of the two sets of scores. Confidence-based scores had better predictive validity than conventional scores, but this difference was not significant. Confidence-based scores were not biased against a specific gender and specific levels of self-esteem. The conclusion is that confidence-based scoring is as good as conventional scoring and the choice between these two scoring methods depends on the teacher's discretion and the teaching context.

Keywords: confidence-based assessment, summative assessment, scoring method

Received: 11/25/2014 Accepted: 07/12/2015
*Corresponding author

1. Introduction

Before the mid-nineteenth century, the primary means of educational testing was oral examinations, which were replaced by essay-type written tests later. In the early part of the twentieth century, studies showed that essay-type tests tended to be highly subjective and unreliable in measuring students' performance. As a result, educators were motivated to develop more objective educational measurement (Lau, Lau, Hong, & Usop, 2011). Multiple-choice tests were first used in 1917 for the selection and classification of military personnel for the United States Army (Ebel, 1979). In the 1990s, multiple-choice tests were the most widely used type of objective tests for measuring knowledge, ability, or performance (Ben-Simon, Budescu, & Nevo, 1997), and they continue to be widely used today (Frederiksen, Glaser, Lesgold, & Shafto, 2013). In spite of their widespread use, multiple-choice tests are not without their shortcomings.

Despite the advances in the area of educational measurement, some problems are still seen in this regard. One of the greatest problems that we face in the area of education is that even after students pass their exams, they lack the necessary skills to perform well in the workplace (Adams & Ewen, 2009). The challenge is to find ways to improve educational outcomes so that students can succeed in school, work, and life. Research about ways to improve assessment, and hence learning, has taken two directions. One group of scholars proposed alternative methods of assessment, arguing that traditional assessment methods place too much emphasis on assessing content and do not give enough attention to assessing creative skills and knowledge (Anderson, 1998). On the other hand, standardized achievement tests have several advantages, such as quick and easy administration, being inexpensive, easy reporting of results, high score reliability, usefulness in testing varied content, etc.(e.g. Kurz, 1999), which is why they are still widely used in spite of arguments against them. As a result of this widespread use, a second group of scholars have tried to find ways of improving objective tests (Kurz, 1999).

Confidence-based assessment, in which a student is asked not only about the correct answer of a question but also about how confident he or she feels about his or her answer, is one of the methods which improve

scoring of different types of objective tests (Gardner-Medwin, 2006). It is also claimed to be a method for improving learning (e.g. Adams & Ewen, 2009; Gardner-Medwin & Gahan, 2003). In other words, confidence-based assessment can be used both in formative and in summative assessment. In the former case, it contributes to improving learning. In the latter case, it is just a scoring method.

The present study was an attempt to investigate the usefulness of confidence-based assessment as a summative assessment tool. The purpose of this study was to see whether confidence-based scoring is really an improvement in scoring objective tests, especially multiple-choice tests. For this purpose, we compared the reliability and predictive validity of confidence-based scores and conventional scores. We also examined the bias of confidence-based scoring against the gender and self-esteem level of the test takers. As it can be seen in the next section, the previous studies are not conclusive about these issues. Moreover, few previous studies have dealt with confidence-based scoring with regard to learning language, while this is done in the present study.

2. Review of the Related Literature

For years, testing specialists have tried to minimize guesswork in assessment practices, but it seems that they have not been very successful in this regard (Adams & Ewen, 2009). In different types of the so-called objective tests, such as multiple-choice and true/false tests, guesswork comes more into play because in these types of tests, students are forced to make a decision about what they consider to be the correct answer. This forced-choice approach leads students to guess an answer even if they do not really know the right one. As a result, in current assessment practices, an individual who has correctly answered a question and really knows that it is correct is indistinguishable from a person who has guessed correctly and arrived at the same answer by sheer luck. Although the scores of these two persons are identical, it can logically be predicted that the first person will perform better than the second one in the future with regard to the topic in question (Adams & Ewen, 2009).

On the other hand, in current right/wrong assessment, a wrong answer simply means that the student is uninformed about the material and does not possess the correct information. However, there is another outcome which can be even more damaging to the student; a student may be wrong about an answer while he or she strongly believes that the wrong answer which he or she selected was correct. This high level of confidence in incorrect information leads to poor decisions and mistakes in the application of learning (Adamas & Ewen, 2009).

In sum, uncertain but correct answers, or lucky guesses, are not the same thing as knowledge, and confident wrong answers deserve special attention (Gardner-Medwin & Gahan, 2003). These considerations are ignored in current right/wrong testing procedures.

To eliminate guesswork from multiple-choice tests, or at least to minimize it, formula scoring has been proposed. In formula scoring, a wrong answer will receive a negative mark based on a specific formula which is a function of the number of choices. For example, in a four-choice question, in which a correct answer receives 1 point, a wrong answer will receive -0.33 point. However, this method has several drawbacks. This method may make the students too conservative in a way that even after recognizing one or two choices as distractors, they leave the question unanswered due to the fear of receiving a negative mark. Other examinees may exhibit the same behavior because of personality factors, such as timidity or reticence (Frary, 1988).

According to Kurz (1999), shortcomings of number-right scoring gave rise to the development of formula scoring. However, since formula scoring could not take partial knowledge into account, testing specialists tried to develop other scoring algorithms in which partial knowledge was taken into account. Kurz mentions five such algorithms: (1) confidence weighting, (2) answer-until-correct scoring, (3) option weighting, (4) elimination and inclusion scoring, and (5) multiple answer scoring. Kurz reviews the advantages and disadvantages of each of these scoring methods. Finally, he refers to the fact that empirical studies on these scoring methods show *slight* improvements over conventional scoring and concludes that this slight improvement cannot justify the replacement of conventional scoring with other scoring methods, considering the disadvantages of these methods.

These disadvantages include the complexity of administering and scoring the tests, as well as increased cost and time to administer the tests.

Despite Kurz's (1999) conclusion, the renewed interest in studying methods of measuring partial knowledge (e.g. Fahim & Dehghanker, 2014; Lau et al., 2011) shows that some scholars still believe in the merits of these methods. Confidence-based assessment is no exception, and there are a number of studies conducted on this topic in recent years (e.g. Barr & Burke, 2013; Davies, 2002; Gardner-Medwin, 2006; Gardner-Medwin & Gahan, 2003). This renewed interest in confidence-based assessment was a trigger for conducting the present study. Since this study is about reliability, predictive validity, and gender and self-esteem bias of confidence-based scores, the literature in this regard is reviewed in the following sections after elaborating on different methods of scoring confidence-based tests.

2.1 Different marking schemes in confidence-based assessment

There are different types of marking schemes in confidence-based assessment. The main sources of difference among marking schemes are the number of certainty levels that students are to choose among and the way different certainty levels are marked for correct and wrong answers. In most marking schemes, there are three certainty levels: high, mid, and low (C=1, C=2, and C=3). In fact, after answering each question, a student should choose one of these three certainty levels. In a marking scheme developed by Gardner-Medwin (1995), which was initially proposed for true/false questions, a correct answer with high, mid, and low certainty levels will receive 3, 2, and 1 point(s) respectively. A wrong answer, on the other hand, with high, mid, and low certainty levels will receive -6, -2, and 0 point(s) respectively.

The reason for choosing such marking is that it motivates students to report their real level of confidence. The negative scores for wrong answers with mid and high certainty levels guarantees that students will not report high levels of confidence when they are not that confident about their answer (Gardner-Medwin, 2006; Gardner-Medwin & Gahan, 2003). As Gardner-Medwin (2006) mentions "this is the motivating characteristic of

the mark scheme, rewarding the student's ability to judge the reliability of an answer, not their self-confidence or diffidence (p. 144)."

One might think that the negative points are too high and that it would be better to choose lower negative points. As Gardner-Medwin (2006) argues, in true/false questions, the probability of answering a question by pure chance and getting the answer right is 50 percent. Therefore, even when we talk about low level of certainty, it is above 50 percent. As a result, the penalties should be great. According to Gardner-Medwin, if a student is less than 67 percent sure of his or her answer, the best confidence level to choose is low (C=1). If confidence is between 67 and 80 percent, the best choice is mid confidence (C=2), and if the confidence is above 80 percent, the best choice is high confidence (C=3). This is shown in Figure 1. On this scoring scheme, it is never best to give no reply because an answer at C=1 has the possibility of gaining a mark with no risk of losing anything (Gardner-Medwin & Gahan, 2003).

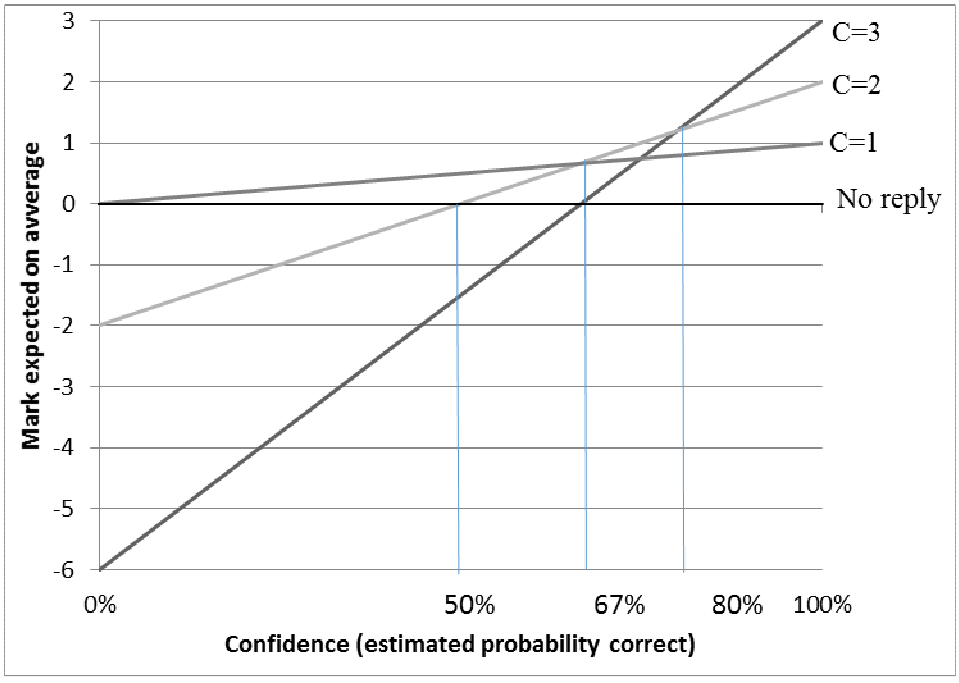

Figure 1. Gardner-Medwin's (2006) marking scheme for true/false questions and its estimated probabilities

Gardner-Medwin (2006) proposes a slightly different marking scheme for other types of questions, especially multiple-choice questions, in which the probability of answering the question by chance and getting the answer right is less than 50 percent. On this scheme, a correct answer at C=1, C=2, and C=3 receives 1, 2, and 3 point(s) respectively. Up to here, it is similar to the previously-mentioned marking scheme. What makes this scheme different from the previous one is the penalties. Here, a wrong answer at C=1, C=2, and C=3 receives 0, -1, and -4 point(s) respectively. In such questions, when the student is less than 50 percent confident of his or her answer, he or she should choose C=1. When confidence is between 50 and 75 percent, the best confidence level to choose is C=2, and when the confidence is more than 75 percent, C=3 should be chosen. This is shown in Figure 2.

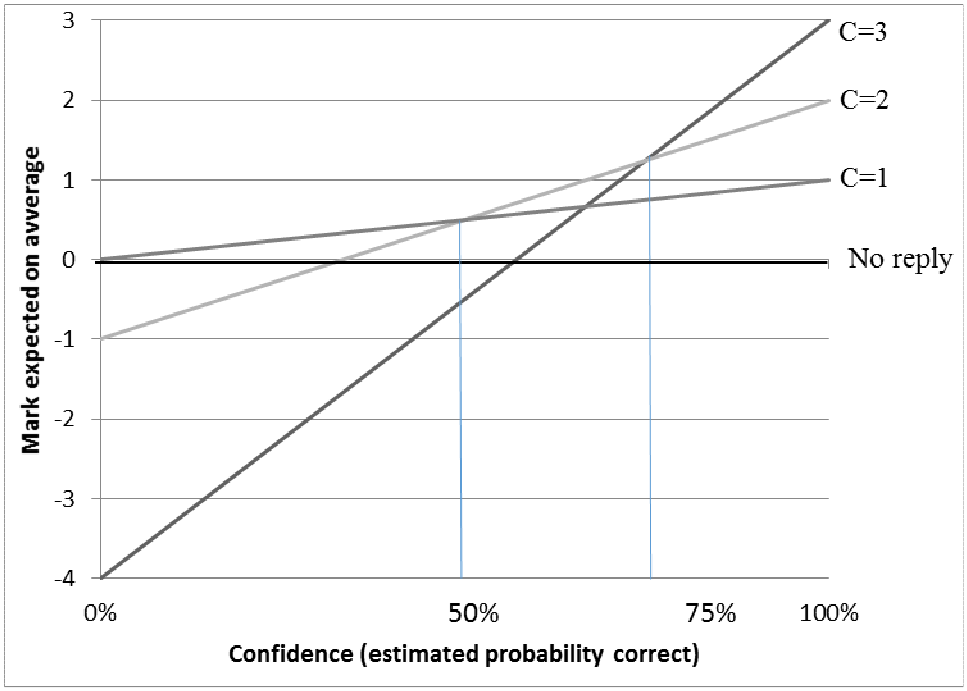

Figure 2. Gardner-Medwin's (2006) marking scheme for multiple-choice questions and its estimated probabilities

In another marking scheme used by Davies (2002) for multiple-choice questions, the penalties are equal to positive scores. In other words, a correct answer at C=1, C=2, and C=3 levels of confidence will receive 1, 2, and 3 point(s) respectively, and a wrong answer at C=1, C=2, and C=3 certainty levels will receive -1, -2, and -3 point(s) respectively. According to Gardner-Medwin and Gahan (2003) and Gardner-Medwin (2006) this marking scheme is not motivating because in this scheme, the best mark will be achieved if a student always uses high confidence (above 50%) or does not answer the question at all. Neither of the lower confidence levels is useful, and students who choose C=1 and C=2 confidence levels based on their teacher's advice are disadvantaged. In other words, this marking scheme rewards high confidence (whether reported honestly or dishonestly) not true and honest reporting of confidence. This marking scheme is illustrated in Figure 3.

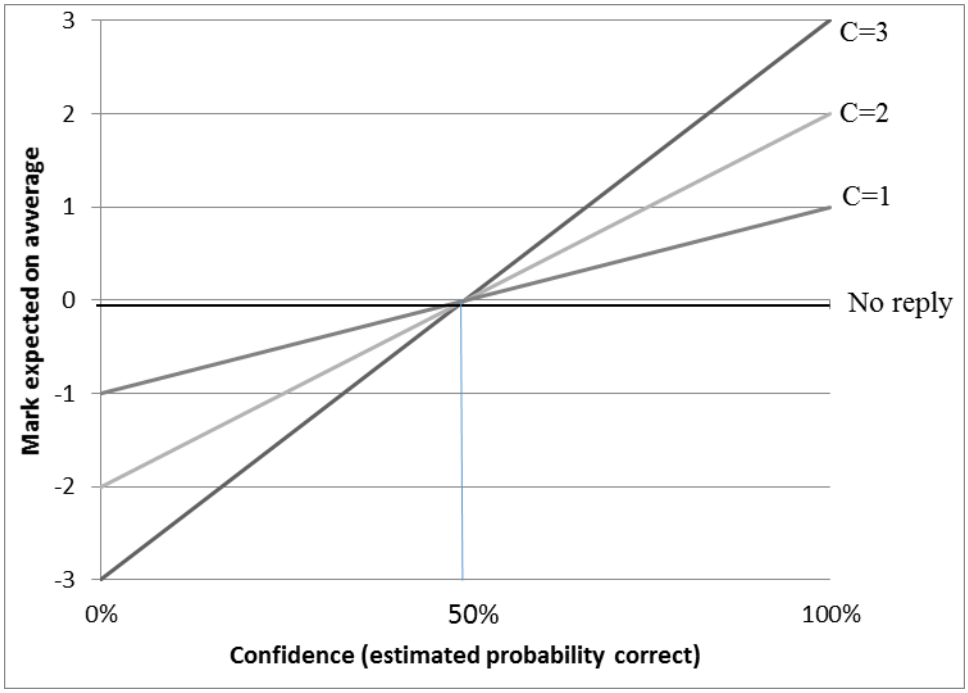

Figure 3. Davies' (2002) marking scheme for multiple-choice questions and its estimated probabilities

Another marking scheme proposed by Hassmen and Hunt (1994) includes five certainty levels, instead of three. In this scheme, which is proposed for multiple-choice questions, marks for correct answers are 20, 54, 74, 94, and 100, and marks for wrong answers are 10, -8, -32, -64, and -120. Gardner-Medwin and Gahan (2003) believe that this scheme is motivating in principle, but it is complex and inflexible. Gardner-Medwin (2006) also mentions that this scheme is hard for the students to remember and understand. For the same reason, it has been sometimes used without the students being aware of the marks associated with different confidence levels. According to Gardner-Medwin, if we want to obtain full engagement of students to improve their study habits and assessment, we need a simple and transparent marking scheme, but this scheme lacks these features.

Based on the above discussion, it seems that the best marking scheme proposed up to now for multiple-choice questions is the one developed by Gardner-Medwin (2006) (the one with the following scores: 3,2,1,0,-1,-4). For the same reason, this is the marking scheme which will be used in the present study.

2.2 Reliability and validity of confidence-based assessment

Advocates of confidence-based assessment have tried to show that confidence-based marking of a test produces more reliable results than number-right scoring of it. Ahlgren (1969) has summarized the results of different studies in this regard. The reliability changes of these studies are shown in Figure 4.

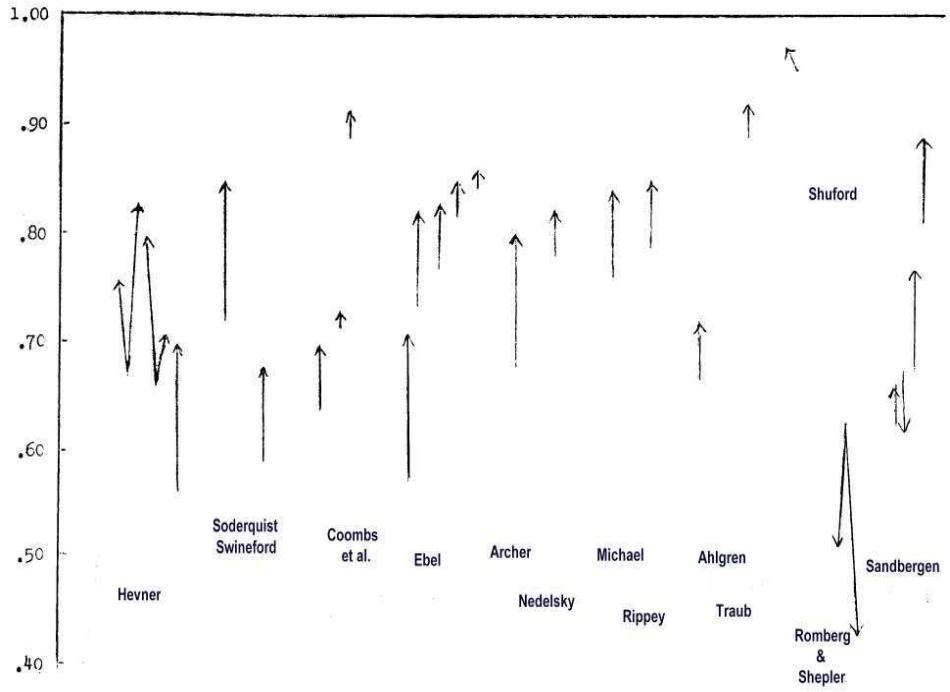

Figure 4. Reliability changes from confidence weighting (taken from Ahlgren, 1969)

In most of the summarized studies, the reliability of test scores increased when the test was marked based on the confidence level in comparison with the time when the test was scored conventionally. Only in 2 of the studies, the reliability decreased. Ahlgren attributes this decrease to the specific method of confidence-based marking. Reliability in this context refers to the internal consistency, and the studies mentioned by Ahlgren estimated reliability by split-half correlations or Kuder-Richardson formula.

Ahlgren (1969) states that validity reports in the literature are less frequent than reliability reports. Among the studies summarized by Ahlgren, only 6 studies estimated validity as well as reliability. The validities in these studies were Pearson correlations with some criteria. In four of the six studies, validity increased by confidence weighting of scores. In one study, the validity did not change significantly. And in one of the studies, validity increased in one subtest and decreased in the other one.

Ahlgren (1969) believes that in researching the benefits of confidence-based assessment, it is not enough to test only reliability. He states that most instruction is intended to have long-term effects. However, most achievement tests are given immediately after the end of a course. Ahlgren believes that a substantial part of knowledge measured by achievement tests is short-lived knowledge, stored temporarily for the purpose of taking the test. If confidence-based testing let us weight heavily on long-term knowledge and weight lightly on transient knowledge, then the weighted score might predict much better the state of knowledge at a later time. With these arguments, Ahlgren tries to say that to advocate confidence-based assessment, we should show that confidence-based marked scores have better predictive validity than conventional scores.

In his own study, Ahlgren (1969) gave a mid-term exam to 160 high school physics students. The test was both confidence-based marked and conventionally marked. At the end of the semester, the final exam of the students was just conventionally marked. The confidence-based scores of the mid-term exam had higher reliabilities and correlated significantly higher with the final exam grades. In another part of this study, 320 high school physics students were given a test, which was confidence-based marked and conventionally marked. Four months later, the students were given a parallel-form retest, which was just conventionally scored. This time, the reliability of weighted scores did not go up, but the predictive validity did increase.

Hopkins, Hakstian, and Hopkins (1973) gave a 65-item multiple-choice test to 63 graduate students taking elementary statistics course in education and having previous experience with confidence weighting. The test was both confidence-weighted scored and conventionally scored. They also gave a short-answer examination covering the same material. The second test served as the criterion because the authors believed that its response style and chance had little effect on performance. Their results showed that the reliability of confidence-weighted scores was slightly higher than conventional scores (0.91 vs. 0.88). However, the validity coefficient for the confidence-weighted scores was slightly lower than for conventional scores (0.67 vs. 0.70). The authors conclude that the added reliable variance often

observed in confidence-weighting studies may be irrelevant style variance and does not increase validity. In fact, it may actually decrease the validity.

In another study, Pugh and Brunza (1975) compared the reliability of conventional scores and confidence-weighted scores of a vocabulary test. In this study, the reliability of conventional scores was 0.57 while the reliability of confidence-weighted scores was 0.85. Moreover, they found no significant interaction between the difficulty of test items and the type of scoring system. The authors also concluded that validity must also have improved because the weighted scores showed no personality bias. However, they gave no direct evidence in support of the claim that validity can be improved using confidence-weighting.

Bokhorst (1986) gave a multiple-choice test to students of psychology taking an introductory psychology course. The test was once conventionally scored with a penalty for guessing and once scored based on confidence levels. The results showed that the reliability of confidence-weighted scores was more than that of conventional scores. Considering validity, the author examined the criterion validity of the confidence-weighted scores. The criterion was the overall achievement for the academic year in the psychology. Despite reliability, no improvement in validity was found.

More recently, Gardner-Medwin and Gahan (2003) tested the difference between reliability of confidence-based scores and conventional scores of 6 medical exams, each with over 300 students, and 25-300 questions. For testing reliability, they ran correlations between odd- and even-numbered questions in each test. The results showed that confidence-based scores were significantly more reliable than conventional scores. To demonstrate the validity of confidence-based marked (CBM) scores, Gardner-Medwin and Gahan divided the test in two halves (odd- and even-numbered). One half was conventionally scored once and CBM scored the other time. The other half was just conventionally scored. Running correlations, they showed that CBM scores of the first half were better predictors of the second half, which was conventionally scored. The authors concluded that CBM scores were more valid than conventional ones.

Hunt (2003) talks about the relationship between confidence and retention of materials, which is very similar to what Ahlgren (1969) calls

"predictive validity" of confidence-based scores. Hunt cites Cabigon (1993) as showing that when people are "not sure at all" of their correct answer, then a week later they can only remember 25% of the material. If learners are "extremely sure" of their correct answer, they retain 91% of the information they have learned.

Omirin (2007) gave three mathematics tests to 450 secondary students and compared the test-retest reliability and content validity, determined by mathematics experts, of conventional scores and confidence-based scores in these tests. The results showed that in all of the three tests, the reliability of confidence-based scores was higher than that of conventional scores. Considering validity, except one of the tests, the validity of confidence-based scores was higher than that of conventional scores.

Yen, Ho, Chen, Chua, and Chen (2010) compared ordinary (conventionally-scored) Computerized Adaptive Testing (CAT) with Confidence-Weighted Computerized Adaptive Testing (CWCAT). Both of these systems were used to test the English vocabulary knowledge of senior high school students. To compare the predictive validity of the two testing systems, the authors used English term scores as the external criteria. The results showed that the predictive validity of CWCAT scores was more than CAT scores. Furthermore, the authors concluded that CWCAT was more precise and efficient than CAT because the mean ability estimated for CWCAT was slightly higher than for CAT, standard error of estimation for CWCAT was lower than for CAT, and the test length of CWCAT was significantly less than that of CAT.

2.3 Gender and personality bias in confidence-based assessment

Besides investigating the reliability and validity of confidence-based scores, a number of studies tried to investigate the relationship between confidence-based assessment and specific personality traits. The reason for conducting such studies was that confidence-based assessment was criticized (e.g. Jacobs, 1971) for introducing bias into assessment procedures by favoring one gender or specific personality types.

Ahlgren (1969) believes that, undoubtedly, personality has an effect on confidence-marking. However, what we should be after is whether

personality differences result in weighted scores having an unfair bias. Ahlgren believes that when we use confidence-based assessment for instructional purposes (formative assessment), then personality bias is not a disadvantage. On the contrary, it can have worthwhile educational effects. With regard to confidence-marking uses in summative assessment, Ahlgren investigated the effects of personality on test scores. As mentioned before, Ahlgren gave a test and a parallel-form retest to high school students with a four-month interval. The first test was scored both with CBM and conventional methods, while the second test was scored only conventionally. He classified his subjects to subgroups based on gender, initial test score, average confidence, appropriateness of confidence, general test anxiety, and general defensiveness. He also classified some subgroups based on pairs of these variables. The result was that in many subgroups there was prediction bias when using the weighted scores. But the important point is that for these same groups, there was also bias when using the conventional scores. Moreover, the bias for weighted scores was never more, and was often less, than the bias for the conventional scores.

Echternacht, Boldt, and Sellman (1972) looked at personality bias of confidence testing in a different light. They stated that advocates of confidence testing have claimed that the effects of personality variables on confidence testing scores can be reduced by practice. Echternacht et al. tried to examine this claim in their study and found that it is in fact the case. They found significant correlations between personality variables and confidence testing scores, but these correlations disappeared with replication. In other words, as more confidence tests were given to the participants and they got acquainted with confidence testing, the relationship between personality variables and confidence scores was not significant any more.

Gardner-Medwin and Gahan (2003) also refer to the commonly-held view that confidence-based assessment introduces a bias into assessment which favors one or the other gender, or certain personality types. Gardner-Medwin (2006) states that the view about gender bias in confidence-based assessment is the result of the personality type bias. According to him, some people believe that confidence-based assessment might disadvantage diffident or risk-averse personalities, which is supposedly more common

among females. To examine this view, Gardner-Medwin and Gahan conducted a study on 331 first-year medical students at University College London (UCL). They separated the questions which were answered at different confidence levels (low, mid, and high). At none of the confidence levels, the number of correct answers was different between male and female participants.

The idea that confidence-based assessment may favor male test takers because boys and men usually have more self-confidence or are better risk takers can also be seen in a different light. If we can show that self-confidence or risk taking tendencies of girls and women is no less than boys or men, then the idea of gender bias in confidence-based assessment will be challenged. One such study was conducted by Lenny (1977). According to her, although low self-confidence is a frequent problem among women, their self-confidence is not lower than men in all achievement situations. Lenny argues that the nature of this gender difference depends on such situation variables as the specific ability area, the availability of performance feedback, and the emphasis placed on social comparison and evaluation.

Even if a study shows that confidence-based assessment is biased against one gender, we cannot conclude that conventional assessment is better than confidence-based assessment with regard to gender. The reason is that conventional assessment itself may be biased against one gender. Ben-Shakhar and Sinai's (1991) study supports this view. The authors of this study investigated the answers of male and female participants in two conventionally-scored multiple-choice tests which tested a wide range of different subject areas. The instructions of one of the tests encouraged the participants to guess, and the other test had no specific instruction as to guess or not. However, many test takers omitted some of the questions. The number of omitted questions in female participants was significantly more than those of male participants, which shows that male test takers were more inclined to guess than females. The results of this study suggest that even conventionally-scored multiple-choice tests may introduce a kind of gender bias into the assessment procedure.

Hunt (1993, 2003) also believes that conventional multiple-choice tests are biased against female test takers while self-assessment responding

(confidence-based assessment) is not so. Evidence for this belief comes from Hassmen and Hunt's (1994) study, in which female test takers scored lowered than males, on the average, when a conventional multiple-choice test was used. However, when the test was scored based on test takers' confidence levels, the difference between the scores of male and female participants reduced.

3. Research Questions

There are four research questions in the present study:

1. Is there a significant difference between the reliability of number-right scoring system and confidence-based scoring system?

2. Is there a significant difference between the predictive validity of number-right scoring system and confidence-based scoring system?

3. Is there a significant difference between the means of the lost scores (the number-right score minus the confidence-based score) in male and female students?

4. Is there a significant correlation between the lost scores and the level of self-esteem of the participants?

4. Method

4.1 Participants

The participants of this study were freshman students majoring in English translation at the Islamic Azad University, Shahr-e-Qods Branch, Tehran Province, Iran. The study was conducted in two classes of English Grammar 2, both of which were taught by the corresponding author. The students themselves had enrolled in these classes based on their preferences, and there was no specific sampling procedure because this study was of a correlational design, not an experimental one. Therefore, there was no need to choose participants randomly. The total number of students in these two classes was 49 (30 female and 19 male students). In answering the first three research questions, all the 49 students were included in the study. However, in answering the last research question, 6 students were omitted from the data analysis because 3 of them did not complete the self-esteem

questionnaire and 3 others did not seem to be honest in their answers to the questionnaire.

4.2 Instruments

The main instruments used in this study were eight tests given to students during the semester and a final exam. These tests were prepared by the researchers based on the content of the course. Each of the eight tests corresponded to one of the chapters of the book Communicate What You Mean (Pollock, 1997) (chapters 4 to 10 and chapter 13), which formed the syllabus of English Grammar 2. The content of the final exam was based on the whole eight chapters. Each of the eight tests included 10 multiple-choice questions, and the final exam was a 60-item multiple-choice test.

Another instrument used in this study was the Self-Esteem Inventory (SEI) developed by Coopersmith (1967). This is a 58-item questionnaire to measure the subjects' global self-esteem. However, eight of the items are called lie scale items (items 1, 6, 13, 20, 27, 34, 41, and 48) and the answers to these questions are not considered in calculating the total score. In fact, the purpose of including these items in this questionnaire was to find out whether a participant is honest in his or her responses or not. If a participant agrees with 3 or more of these items, it suggests that he or she is trying too hard to present him or herself in a positive light. The other 50 items contain four subscales: academic tasks, social relationship, family, and self.

Considering the reliability of SEI, Coopersmith (1967) reports a test-retest correlation of 0.8. Moreover, he states a desirable value for both validity and reliability of the test (cited in Gurney, 1988). Ebrahimi (1990, cited in Sazvar, 2003) is one of the early studies which used this questionnaire with Iranian subjects. In this study, which was conducted on 200 junior students of psychology, the reliability of this questionnaire was calculated by Cronbach's alpha method. In this study, the internal consistency of the whole questionnaire was 0.85, and the internal consistencies of the four subscales were reported as 0.5, 0.8, 0.6, and 0.7, respectively.

The original questionnaire is a dichotomous one with "Like me" and "Not like me" answers. However, since it is not easy to express oneself

categorically, the response format was changed into a five-point Likert scale in order to obtain more accurate results. This is what has previously been done by Fani (2009). Moreover, Sazvar (2003) conducted a pilot study on this questionnaire. She gave the questionnaire to 45 senior students from four different universities and asked them to answer the items and comment on them if there seemed to be any ambiguity in the content of the items or if the dichotomous scale did not suffice for answering. Eighty-two percent of her subjects in the pilot study mentioned that they could not express themselves in a dichotomous format and needed a more extensive scale to answer the questions. Therefore, Sazvar also made changes in the original questionnaire for her main study, but she changed it into a four-point Likert scale.

Another modification made on this instrument was that the Persian translation of this questionnaire was used in this study. Although the participants of this study were students majoring in English translation, most of them were not of high proficiency levels. Therefore, in order to make sure of all participants' full understanding of the questionnaire, the Persian translation of SEI was given to them.

The Persian translation was taken from Sazvar (2003). In her pilot study, Sazvar used a translated version of the SEI which, according to her, is administered in all formal psychological institutions in Iran. Seventy-one percent of the participants in the pilot study stated that 15 of the items were not clear enough to them. Besides the 15 items mentioned by the participants, Sazvar herself believed that 11 more items had been translated incorrectly or vaguely. To remove this ambiguity, Sazvar gave the English version of the questionnaire to seven university instructors who were experts in translation and asked them to translate the test items into Persian. Their proposed translations subsequently replaced the items that the author and the participants had judged to be problematic. This translation was given to the same 45 subjects again. As no objection to the translated sentences was reported; this translation was used in the main study.

4.3 Procedures and data collection

This study was conducted in two classes of English Grammar 2 at Islamic Azad University, Shahr-e-Qods Branch, in the second semester of 2013-2014 academic year. The syllabus of the course English Grammar 2 at this university includes 8 chapters of the book Communicate What You Mean (Pollock, 1997) (chapters 4 to 10 and chapter 13). When teaching each chapter was finished, the students were told that they would be given a multiple-choice test of that chapter the next session. Therefore, eight tests were given to the participants during the semester. In these tests, after each question, the students were asked to say whether their confidence level was low, mid, or high. Then their scores were calculated based on one of the confidence-based marking schemes: correct and high confidence = 3, correct and mid confidence = 2, correct and low confidence =1, wrong and low confidence = 0, wrong and mid confidence = -1, and wrong and high confidence = -4. This method of scoring was explained to the participants before administering the first test. The test papers were scored both conventionally and based on confidence levels. If a student was absent in one of the sessions in which a test was administered, the test paper was given to him or her the next session. In this way, all students sat for all tests.

The participants of the study were also given the self-esteem questionnaire in the middle of the semester. It was explained to the students that their answers to the questionnaire would be used in a research study. They were also told that completing the questionnaire was not compulsory, but if they were willing to complete it, it was necessary for them to write their names on it. Fortunately, most of the participants filled out the questionnaire, and only three of them refused to do so.

At the end of the semester, the students took the final exam, which was a 60-item multiple-choice test based on the content of the whole eight chapters. In this test, the students were not asked about their confidence level, and the papers were scored conventionally. After all the data was collected, data analysis began at the end of the semester. Methods of data analysis are described in the following section.

4.4 Data analysis

To answer the first research question of this study, which relates to comparing the reliability of number-right scores and confidence-based scores, each item of all eight tests was scored twice: once in a number-right fashion (correct = 1, and wrong = 0), and once in a confidence-based manner (correct and high confidence = 3, correct and mid confidence = 2, correct and low confidence =1, wrong and low confidence = 0, wrong and mid confidence = -1, and wrong and high confidence = -4). Since each test had 10 items, the analysis was run on 80 items with 49 participants. The reliabilities of number-right scores and confidence-based scores were calculated with Cronbach's alpha formula. Since reliability is a kind of correlation, to test the significance of the difference between the two reliabilities, we can use a method which is usually used to test the significance of the difference between two correlations. This method is Fisher's R to Z transformation formula.

The second research question was answered in the following method. In each of the eight tests, each student received two different scores: a number-right score and a confidence-based scores. Since the number-right scores and confidence-based are not comparable (the maximum score in number-right scoring in each test was 10 while the maximum confidence-based score is 30), the confidence-based scores were divided by three so that the two types of scores would be comparable. After that, the means of each student's number-right scores and confidence-based scores were calculated. In other words, for each student, two means were obtained: the mean of his or her number-right scores in the eight tests, and the mean of his or her confidence-based scores in the same eight tests. Finally, two correlations were run: one between the means of number-right scores and the final exam scores, and one between the means of confidence-based scores and the final exam scores. The first correlation shows the predictive validity of the number-right scores, and the second correlation indicates the predictive validity of the confidence-based scores. To test the significance of the difference between these two correlations, Fisher's R to Z transformation formula was used.

To answer the third research question, which relates to gender bias of the confidence-based scores, the lost score of each student was calculated. The lost score refers to the mean of number-right scores minus the mean of confidence-based scores. Then, the mean of the lost scores was calculated for male and female students, and the significance of the difference was checked with independent sample t-test.

The reason for using lost scores instead of confidence-based scores here is that confidence-based scores alone are not indicators of gender bias. Imagine that we tested the difference between the means of confidence-based scores of male and female participants and showed that the means were significantly different and concluded that the confidence-based scores were biased against one gender. What if the means of conventional scores were also significantly different? Then our conclusion had been wrong because the difference between the two genders was related to better knowledge of one gender not to the bias of confidence-based scores. Therefore, to see whether confidence-based scores are biased against one gender, we should see how many points are lost by the students of different genders due to the use of confidence-based scoring rather than conventional scoring.

Finally, to answer the fourth research question, a correlation was run between the students' scores in the self-esteem questionnaire and the lost scores. The questionnaire scores were calculated in this way: in positively worded items the following scores were given: Always = 5, Usually = 4, Sometime = 3, Rarely = 2, and Never = 1. In negatively worded items, the order of the scores was reversed. Here again we performed the analysis on the lost scores rather than the confidence-based scores for the same reason mentioned in the previous paragraph.

Out of the 49 students in group two, 43 students were included in this data analysis because three students did not complete the self-esteem questionnaire and three other students did not fulfill the honesty criterion of the questionnaire. As mentioned in Section 4.2, the Self-Esteem Inventory has eight lie items, and if a respondent chooses three or more "Like me" answers in these items, it means he or she has not been honest in his or her answers. In the present study, the answers were changed to a five-point

Likert scale rather than a dichotomous one with "Like me" and "Unlike me" answers. Therefore, in the eight lie items, those students who had chosen three or more "Always" answers were omitted from the data analysis.

5. Results

5.1 Reliability of confidence-based scores

To answer the first research question, the reliabilities of the number-right scores and the confidence-based scores in the eight tests were compared. The reliability of each set of scores was calculated with Cronbach's alpha formula. To calculate the reliability of each type of scores, the eight tests were considered as one test. In other words, instead of calculating the reliabilities of eight ten-item tests and then calculating the mean of these eight reliabilities, the reliability of one 80-item test was calculated.

Surprisingly, the reliabilities of the two sets of scores were exactly the same. This is shown in Table 1. Since the reliabilities are exactly the same, it is meaningless to talk about the significance of the difference.

Table 1. Reliabilities of number-right scores and confidence-based scores

	N	N of items	Cronbach's alpha
Number-right scores	49	80	0.865
Confidence-based scores	49	80	0.865

5.2 Predictive validity of confidence-based scores

To compare the predictive validity of conventional scores and confidence-based scores, two correlations were run: one between the means of students' conventional scores in the eight tests and the final exam scores, and another one between the means of students' confidence-based scores in the eight tests and the final exam scores.

The results showed that the predictive validity of confidence-based scores was a little more than that of conventional scores. However, this difference was not significant (two-tailed significance=0.726). These results are shown in Table 2. It is worth mentioning that each of these correlations

was significant at $p<0.001$. However, the difference between these two correlations was not significant.

Table 2. Predictive validity of number-right and confidence-based scores shown in the form of correlations with the final exam scores

Correlation coefficient	Final exam scores
Means of number-right scores	0.833
Means of confidence-based scores	0.854

5.3 Gender bias of confidence-based scores

To answer the third research question, a comparison was made between the means of lost scores (conventional score minus confidence-based score) in male and female participants. As it can be seen in Table 3, the mean of lost scores in male participants was a little more than that of female students, but this difference was not statistically significant (two-tailed significance=0.499). This shows that confidence-based scoring in our study was not biased against one gender.

Table 3. Gender bias of confidence-based scores shown in the form of means of lost scores

	N	Mean of lost scores	Std. deviation
Male participants	19	2.97	0.12
Female participants	30	2.83	0.14

5.4 Self-esteem bias of confidence-based scores

To answer the last research question, the Self-Esteem Inventory scores were used. First of all, the reliability of the scores obtained from this questionnaire was calculated by Cronbach's alpha formula, which was 0.89. Then a correlation was run between the scores of the Self-Esteem Inventory and the lost scores (N=43). This correlation coefficient was -0.151, which was not statistically significant (two-tailed significance=0.332). Therefore,

our results show that confidence-based scoring is not biased against students with lower self-esteem.

6. Discussion

The results of this study showed that confidence-based scoring system is not at an advantage over conventional scoring considering reliability and predictive validity. However, the results also demonstrated that confidence-based scoring does not suffer from the shortcomings proposed by its critics, including gender and self-esteem bias.

The results related to the reliability of confidence-based scores seem not to be very consistent with most of the previous studies. However, the literature is not conclusive in this area. In most of the studies summarized by Ahlgren (1969), the reliability of confidence-based scores was more than that of conventional scores. However, in two of the summarized studies, the opposite was true. Moreover, Ahlgren does not talk about the significance of the difference between two reliabilities of the cited studies. Some of the reliability increases are very slight and do not seem to be significant. In the empirical part of his own study, Ahlgren once showed that reliability of confidence-based scores was significantly more than that of conventional scores. However, in another part of the same study, the reliability of the confidence-based scores was not more than conventional scores. The other studies mentioned in Section 2 which showed a reliability increase of confidence-based scores in comparison with number-right scores have not mentioned anything about the significance of the difference between the two reliabilities. In one of these studies (Hopkins et al., 1973), in particular, the reliability increase is very low. It seems that more research in this area is needed before we can conclude for sure that the reliability of confidence-based marking is more or less than the number-right scoring.

Considering the consistency of predictive validity results with the previous studies, we should note that only a few studies have previously dealt with the predictive validity of confidence-based scores (Ahlgren, 1969; Yen et al., 2010). These studies showed that the predictive validity of confidence-based scores was more than that of conventional scores. However, they did not mention anything about the significance of the difference between the two validities. Therefore, we can say that our results are consistent with previous studies because we did show that the predictive

validity of the confidence-based scores was more than that of conventional scores, though this increase was not significant. However, since not many studies have been done in this area, we cannot draw a definite conclusion about the predictive validity of confidence-based scores.

The results related to gender bias confirm the results obtained by Ahlgren (1969) and Gardner-Medwin and Gahan (2003), who showed that confidence-based assessment was not biased against one gender. Another important point to mention here is that in our results, the female participants lost fewer scores than male participants as a result of confidence-based marking. Although this difference was not significant, this is against the claims of opponents of confidence-based assessment (cited in Gardner-Medwin & Gahan, 2003), who believe that confidence-based assessment might disadvantage female students. Our results showed that it was not the case and that confidence-based scoring favored female students a little more than male ones. Therefore, even if somebody still believes that one gender is at a disadvantage over the other one as a result of confidence-based scoring, he or she should think of this bias against any of the genders not just against female students.

To compare our self-esteem results with previous studies, we can say that they are almost in line with Ahlgren's (1969) study, which showed that confidence-based assessment is not biased against specific personality types, including general confidence. No previous study has dealt with the relationship between self-esteem and confidence-based scores. However, since self-esteem and self-confidence are very similar concepts, we can say that our results are consistent with Ahlgren's results.

The fact that confidence-based scoring is not biased against one gender or different levels of self-esteem seems reasonable. The scores given to the right and wrong answers with different levels of confidence are designed in a way that they do not award or penalize over-confidence or diffidence. Therefore, learners with different levels of self-esteem and different genders are not at an advantage or disadvantage compared with others.

To sum up, with regard to gender and self-esteem bias, the results of this study were both logical and consistent with previous studies. Considering the reliability and predictive validity, the results did not confirm what the advocates of confidence-based assessment claim. Of course, the

literature is still inconclusive in this area. Based on the obtained results, conclusions and implications are presented in the next section.

7. Conclusions and Implications

Drawing a definite conclusion from the results of this study is a difficult task. On the one hand, our results showed that the reliability and predictive validity of confidence-based scores are not significantly higher than those of number-right scores. On the other hand, we showed that confidence-based scores are not biased against a specific gender and against different levels of self-esteem. The best conclusion we can draw from these results is that confidence-based scoring is as good as conventional scoring. Although in the present study, the reliability and predictive validity of confidence-based scores were not significantly more than those of number-right scores, they were not less either. Our results also showed that confidence-based scoring is not biased against one gender and against a specific personality type, namely self-esteem. These results indicate that unlike what some of the opponents of confidence-based assessment claim, this scoring method does not favor a specific gender and people with different levels of self-esteem, so it is not worse than conventional scoring.

Since the results of this study demonstrated that confidence-based and conventional scoring methods are almost of equal merits, it is difficult to mention definite pedagogical implications. In other words, we cannot recommend teachers to use any of the scoring methods. We should leave it to teachers themselves to choose between these two scoring methods before more research is done and the merits of one of the methods are indicated. Before that time, conventional scoring and confidence-based scoring can be used on a par depending on the teaching context and the teacher's discretion.

Maybe the decision as to use confidence-based scoring as a summative assessment tool or not depends on the subject matter and its importance in the lives of people. The majority of recent studies on confidence-based assessment are in the field of medicine (e.g. Barr & Burke, 2013; Cash, Mitchner, & Ravyn, 2011; Gardner-Medwin, 1995; Gardner-Medwin & Gahan, 2003; Issroff & Gardner-Medwin, 1998; Khan, Davies, & Gupta, 2001). The reason is probably that in medicine, it is very important that the students who are allowed to go to higher levels have full knowledge about the subject matter because in medicine, we deal with lives of people. As

Hunt (1993) mentions, confidence in misinformation will lead to taking actions which are usually negative and potentially dangerous actions. In medicine, confidence in misinformation will endanger the lives of people. Therefore, it seems logical to recommend summative confidence-based assessment to high-stakes testing situations. However, some language courses are not so high-stakes. In such language courses, it seems better to leave it to teachers themselves to choose between confidence-based scoring and conventional scoring.

References

Adams, T. M. & Ewen G. W. (2009). The importance of confidence in improving educational outcomes. *Proceedings of the 25th annual conference on distance teaching and learning,* pp. 1-5.

Ahlgren, A. (1969). Reliability, predictive validity, and personality bias of confidence-weighted scores. *Proceedings from American Educational Research Association symposium "Confidence on Achievement Tests-Theory, Applications"*. Retrieved from www.p-mmm.com/founders/AhlgrenBody.htm

Anderson, R. S. (1998). Why talk about different ways to grade? The shift from traditional assessment to alternative assessment. *New Directions for Teaching and Learning, 74,* 5-16.

Barr, D. A. & Burke, J. R. (2013). Using confidence-based marking in a laboratory setting: A tool for student self-assessment and learning. *The Journal of Chiropractic Education, 27(1),* 21-26.

Ben-Shakhar, G. & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28(1),* 23-35.

Ben-Simon, A., Budescu, D. V. & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21(1),* 65-88.

Bokhorst, F. D. (1986). Confidence weighting and the validity of achievement tests. *Psychological Reports, 59,* 383-386.

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.

Coopersmith, S. (1967). *The antecedents of self-esteem*. San Francisco: W. H. Freeman & Company.

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

Ebel, R. L. (1979). *Essentials of educational measurement (3rd ed.)*. Englewood Cliffs, NJ: Prentice Hall.

Echternacht, G. J., Boldt, R. F., & Sellman, W. S. (1972). Personality influences on confidence test scores. *Journal of Educational Measurement, 9(3),* 235-241.

Fahim, M., & Dehghankar, A. (2014). Towards fuzzy scores in language multiple-choice tests. *International Journal of Language Learning and Applied Linguistics World, 6(2),* 291-308.

Fani, T. (2009). *The role of foreign language anxiety, motivation, and self-esteem in the accuracy of self-assessment of EFL students' reading comprehension abilities*. Unpublished master's thesis, Allameh Tabataba'i University, Tehran, Iran.

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

Frederiksen, N., Glaser, R., Lesgold, A., & Shafto, M. G. (Eds.). (2013). *Diagnostic monitoring of skill and knowledge acquisition*. Routledge.

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

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.

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

Gurney, P. W. (1988). *Self-esteem in children with special educational needs*. London: Routledge.

Hassmen, P. & Hunt D. P. (1994). Human self-assessment in multiple-choice testing. *Journal of Educational Measurment, 31,* 149-160.

Hopkins, K. D., Hakstian, A. R., & Hopkins, B. R. (1973). Validity and reliability consequences of confidence weighting. *Educational and Psychological Measurement, 33*(1), 135-141.

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.

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.

Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Educational Technology and Society, 14(4),* 99-110.

Lenney, E. (1977). Women's self-confidence in achievement settings. *Psychological Bulletin, 84(1),* 1-13.

Omirin, M. S. (2007). Validity and reliability indices of three multiple-choice tests using the confidence scoring procedure. *The Social Sciences, 2(1),* 20-23.

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

Pugh, R. C., & Brunza, J. J. (1975). Effects of a confidence weighted scoring system on measures of test reliability and validity. *Educational and Psychological Measurement, 35(1),* 73-78.

Sazvar, A. (2003). *The impact of self-esteem on authentic material use: A case of Iranian non-English major students/graduates*. Unpublished doctoral dissertation. Allameh Tabataba'i University, Tehran, Iran.

Yen, Y. C., Ho, R. G., Chen, L. J., Chou, K. Y., & Chen, Y. L. (2010). Development and evaluation of a confidence-weighting computerized adaptive system. *Educational Technology and Society, 13(3),* 163-176.