THE JOURNAL OF ASIA TEFL
Vol. 17, No. 2, Summer 2020, 363-378
http://dx.doi.org/10.18823/asiatefl.2020.17.2.4.363

The Journal of Asia TEFL

http://journal.asiatefl.org/

e-ISSN 2466-1511

Doable and Practical: A Validation Study of Classroom Diagnostic Tests

Boh Young Lee
Ewha Womans University, Korea

Sang-Keun Shin
Ewha Womans University, Korea

Although many benefits have been claimed for diagnostic testing, it is utilized far less than assessments for other purposes. This study produced classroom diagnostic tests in the areas of vocabulary and grammar, in both paper-and-pencil and computer-based formats, and utilized them in high school English-as-a-foreign-language instruction. Evidence was collected to investigate whether three hypothesized claims about the advantages of diagnostic testing asserted in the literature are valid. The results showed that students' academic achievement increased when the results of instructional unit-based diagnostic tests were used to offer remediation. Although no differences in student achievement were found to result from the mode of diagnostic testing, the efficiency of diagnostic testing was greatly increased by automatic scoring and data analysis functions of the computer-based mode. A survey on the usefulness of diagnostic testing showed that when it was used, learners had higher levels of satisfaction with English instruction. However, the participants reported that when diagnostic testing pointed toward significant shortfalls in lexical knowledge, their confidence declined. Lastly, the students indicated that computer-based diagnostic testing has the advantage of being able to provide immediate feedback.

Keywords: diagnostic testing, computer-based tests, automatic scoring, feedback

Introduction

Among the various types of assessment used in schools, diagnostic testing is performed to identify the strengths and weaknesses of learners in order to make adjustments so that instruction better meets learners' needs. Learners are also provided with quick and accurate feedback on which areas they are strong in and which areas to focus on for improvement (Alderson, 2005; Bachman, 1990; Kunnan & Jang, 2009). Despite the fact that diagnostic tests can provide such useful information to teachers and students, they are sadly underused in classrooms, and there is scarce research being done on diagnostic testing (Alderson, 2005; Alderson & Huhta, 2011; Kim, 2019; Richards, 2008). Of course, even in the absence of formal diagnostic tests, a variety of methods can be used to identify learners' strengths and weaknesses (Edelenbos & Kubanek-German, 2004), but there is no guarantee that judgments based on such methods are always accurate.

The reason that diagnostic testing is not utilized much in the classroom may be that teachers have limited awareness of its usefulness because it has not been a part of their own education. A more fundamental reason is, however, that there are not many diagnostic tests available for utilization. The

primary reason for the apparent lack of diagnostic tests can be largely explained by difficulties in both its theoretical and practical aspects. First, in terms of its theoretical aspects, there is the hurdle that an understanding of the stages involved in acquiring a second language is a prerequisite for producing diagnostic tests. This is because the developmental stages must first be identified in order to discern the stages that learners have or have not reached. In terms of practical aspects, such tests are difficult to produce and to implement in the classroom because of the large number of test items required to produce a comprehensive diagnostic test.

Like other forms of language assessment, the construct of diagnostic tests should be defined by either language ability theory or class syllabi. The majority of existing studies involve the former and, as a result, few attempts have been made to develop and utilize classroom diagnostic testing based on syllabi. One way to realize the benefits of diagnostic testing in the actual classroom involves enabling teachers to administer diagnostic tests on the material to be taught and then reflect on the results in their instruction (Shohamy, 1992). Compared with large-scale diagnostic testing, classroom diagnostic testing based on a syllabus is less susceptible to the challenges present in the production and implementation of the diagnostic testing described above. In classroom language testing, the target to be evaluated is comparatively easy to define, because the construct is typically defined by the syllabus (Bachman & Palmer, 1996). If assessment is done according to the units being covered in instruction, then the practical difficulties can also be reduced, because there is limited content to evaluate.

Another measure for increasing the usefulness of diagnostic tests is to produce computer-based diagnostic tests. In addition to the automatic scoring function, test forms can be created with ease, and quick editing and revisions are also possible (Brown, 1997; Chapelle & Douglas, 2006; Dunkel, 1991). In addition, computer-based diagnostic tests can provide immediate feedback to test-takers, which is one of the most important elements in diagnostic testing (Alderson, 2005; McMillan, 2006). Due to technical constraints, only limited response items can be automatically scored. Nevertheless, because diagnostic tests characteristically measure individual items and tend to be discrete-point rather than integrative tests (Alderson, 2005; Larson & Hendricks, 2009), automatic scoring offers sufficient latitude for the required purpose. Of course, there are limitations, in that technical constraints permit automatic scoring only for limited response items. However, since diagnostic tests by nature measure individual items and tend to be produced using a discrete-point testing method, automatic scoring offers sufficient latitude for the required purpose.

Although it has often been claimed, as noted above, that diagnostic tests can enhance the efficiency of classroom instruction and assessment, there are insufficient cases in the language education field that systematically validate such claims. This study was designed to address these issues by investigating the impact of diagnostic testing on teachers' instruction and learners' academic achievements, by producing and implementing paper-and-pencil assessments as well as computer-based diagnostic tests in the context of high school English classes and then making use of the results.

Theoretical Background

Diagnostic tests are used for the purpose of discovering learners' specific strengths or weaknesses, thus obtaining necessary data for remediation. While achievement testing measures how well students have learned the content they have been taught, the focus of diagnostic testing is on identifying the areas in which students still need to improve their understanding. Detailed analysis and feedback should be immediately provided to teachers and learners, so they can act upon the feedback in their classroom teaching and learning activities (Alderson, 2005).

In contrast to recent language assessment trends, diagnostic tests are often produced by adopting a discrete-point testing format to identify learners' strength and weaknesses (Alderson, 2005). For ease of scoring, the majority of items in such tests are limited-response items such as multiple-choice, and this poses a risk of measuring language knowledge rather than language use ability. Nevertheless, a discrete-

point approach may be inevitable for accomplishing the purpose of diagnostic testing (Alderson, 2005). Larson and Hendricks (2009) found that when diagnostic tests employ a contextualized method of testing rather than discrete-point testing, the exam takes almost four times as long and leads to very high frustration levels in test-takers, demonstrating the necessity of discrete-point testing.

In recent years, there have been many attempts to utilize the cognitive diagnostic approach (CDA) for providing diagnostic feedback on language use skills (Buck & Tatsuoka, 1998; Jang, 2009; Lee & Sawaki, 2009). By conducting content analysis to identify the skills being assessed by each item, CDA can be used to construct student skill profiles as well as to provide fine-grained diagnostic feedback on test-takers' mastery of subskills that underlie the test. Because CDA can provide diagnostic information at the level of the four language skills, it is regarded as an important breakthrough for overcoming the limitations of existing diagnostic tests that have tended to be discrete-point tests primarily in the areas of vocabulary and grammar. However, many limitations still stand in the way of its widespread adoption in the classroom. First and foremost, the process of content analysis is very complex, and among the published studies, there are discrepancies in the number of subskills that are identified even within the same test (Alderson, 2010). Because the statistical operations required for data analysis are quite complex, this process is also likely to be absolutely opaque to classroom teachers (Davidson, 2010).

As noted above, the practicality of diagnostic testing is an issue. For an assessment to have practicality, the time taken to go from test production to results analysis should not be too long, and the process should not be costly. However, producing a diagnostic test in a given area requires creating an extremely high number of test items. In order to control for instances where test-takers choose the correct answer by chance, more than one test item must be written for each target item. Even if only two or three items are written for each learning target, an enormous number of test items are entailed when dealing with a very large number of individual lexical or grammar items. As such, it would be impractical to routinely administer comprehensive diagnostic tests (Hughes, 2003).

Computer-based classroom diagnostic tests can be considered as an alternative for resolving the theoretical and practical issues described above. First, from the perspective of construct, diagnostic tests are easier to produce in situations where construct is defined by the syllabus as opposed to large-scale assessment situations where the construct is defined by theories. In addition, in cases where the diagnostic tests are produced according to sections covered in class, like textbook units, it is easier to construct diagnostic tests because the assessment content is limited, and the burden of administration is also reduced. When diagnostic tests are computer-based, their usefulness can be increased, since the automatic scoring function enables timely feedback in a short period of time. Since the automatic scoring function of computer-based testing (CBT) allows scoring to be completed within a relatively short period of time compared to human scoring, it enables the test-takers to receive immediate feedback. This allows learners and teachers to immediately begin using the information from the test results for studying and learning activities (Douglas, 2010).

The literature stresses the importance of immediate feedback in order to take advantage of the benefits of diagnostic testing (Alderson, 2005). Rather extensive prior research in psychology on feedback timing has produced mixed results. In the case of applied studies employing classroom quizzes or learning materials in the classroom, it has been revealed that immediate feedback is effective, whereas studies involving the acquisition of test content conducted in experimental environments have shown that delayed feedback is effective (Kulik & Kulik, 1988; Shute, 2008). Feedback must be provided promptly if teachers are to be able to adjust lesson plans and buy time to prepare teaching materials tailored to the test results. For learners, too, only when feedback is provided in the shortest amount of time possible would they be able to identify and correct the problem areas in their performance. Test-takers' interest in their performance is likely to dwindle over time (Norrish, 1990), and moreover, they are also apt to forget the problem-solving process. Therefore, it appears likely that immediate feedback is more helpful than delayed feedback; however, research is needed in order to confirm this prediction.

In the field of language education, studies on feedback have primarily focused on how it is provided in the areas of speaking and writing, or on the quality and effects of feedback. However, very little research

has been done on the timing of feedback. Researchers who favor immediate feedback argue that it prevents errors from being encoded in one's memory, while researchers who advocate delayed feedback suggest that immediate feedback can lead to interference-perseveration, in which the memory of an incorrect answer interferes with remembering the correct information (Kulhavy & Anderson, 1972). Such disparate findings may be accounted for by important differences in circumstances like the research settings and research instruments employed in each study. Research is needed in order to determine whether these findings apply to diagnostic language testing in the classroom, and in what ways.

Adopting Bachman and Palmer's (2010) Assessment Use Argument framework, this study sought to analyze the usefulness of diagnostic tests. In order to verify whether the interpretive argument that performing a diagnostic test enhances the effectiveness of teaching and learning activities is true, this study collected and analyzed a number of backings that warrant these claims and rebuttals based on alternative rationale, which might refute these claims. Specifically, this study started with the following three claims about the usefulness of classroom diagnostic tests and then collected relevant backings and rebuttals.

Claim 1: By utilizing diagnostic tests, teachers are able to identify learners' strengths and weaknesses and thus provide more efficient instruction.
Claim 2: When diagnostic testing is utilized, academic achievement increases.
Claim 3: When immediate feedback is provided, the effectiveness of diagnostic testing will be even greater.

As noted above, diagnostic tests are not being widely implemented in real-world classrooms, and there are almost no studies implementing them in actual classroom instruction to examine their usefulness. In addition, there has been almost no empirical research to verify the benefits that CBT offers. In short, while claims have been put forth that the three claims presented above are true, empirical evidence has not been presented. In order to encourage wider use of diagnostic testing in classrooms, empirical evidence demonstrating its educational benefits would seemingly need to be presented. To meet this need, this study intends to confirm whether the above three claims are true by implementing diagnostic testing in high school English instruction.

Method

Participants

The participants of the study included high school freshmen from three classes at a high school in Korea's Gyeonggi Province. Each of the three classes was assigned a different set of conditions: a control group that was not administered any diagnostic tests; a group assessed using paper-and-pencil tests; and a group that received computer-based diagnostic tests. A total of 83 students participated in the experiment, including 27 in the control group, 26 in the paper-and-pencil group, and 30 in the CBT group. Each of the three groups received 18 instructional sessions. Since the experiment was conducted as part of actual instruction in intact classes, its ecological validity (Brewer, 2000) can be considered high.

To verify that the students in each group were homogeneous, the mean score of each group on the English proficiency test administered at the provincial level were analyzed. The results of one-way ANOVA indicated no statistically significant difference among the test scores of the three groups, with 97.93 for the control group, 97.54 for the paper-and-pencil group, and finally 98.93 for the CBT group. Thus, the test confirmed that the students in the three classes were at the relatively same English proficiency level.

Instruments

The following instruments were employed to test the validity of the three claims.

Diagnostic tests

During the experimental period, one diagnostic test was developed for each of the four units to be covered by the students, making a total of four diagnostic tests. The target vocabulary and grammar dealt with in the diagnostic tests were selected according to the following criteria: (1) new vocabulary items and target grammar points in the pertinent units, and (2) vocabulary and grammar in the previous units that students often find confusing. Before the diagnostic tests were produced, the teacher used these three criteria to select approximately sixty vocabulary items and about ten grammar points per unit. She then consulted with her colleagues to narrow these down to 40 target vocabulary items and 5 to 10 grammar points from each unit.

In order to clarify which items test-takers found difficult, discrete-point tests were produced so that each learning target could be measured individually. More specifically, the tests consisted of multiple-choice, matching, choosing, short answer, error correction, and fill-in-the-blank items. Each grammar and vocabulary item was assessed at least twice, to increase the dependability of the determination of the degree of learning. As for short-answer items, the teacher determined in advance what would constitute correct answers, acceptable answers, and answers warranting partial credit, and the marking schemes were uploaded to the system. Later, after the scoring was finished, the incorrect answers for these items were collected and checked to determine whether any of them should receive full or partial credit. The internal consistency of the diagnostic tests was found to be high, with Cronbach's alpha coefficients ranging from .85 to .95. An example of a diagnostic test appears below in Figure 1.

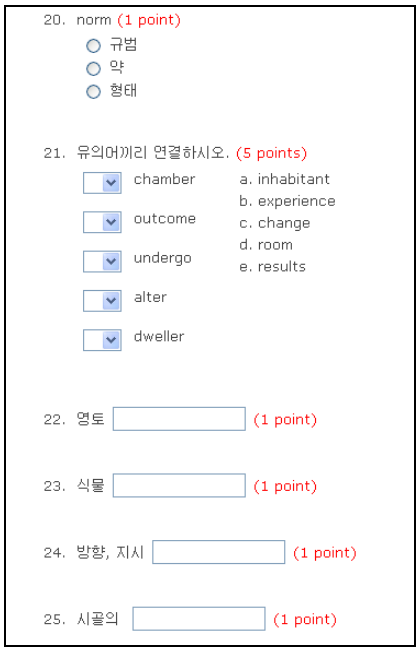

Figure 1. Screenshot of diagnostic test 4.

The diagnostic tests for the two experimental groups were identical in terms of content and test item formats. The computer-based tests were administered in a computer lab. The automatic scoring function was used to provide immediate feedback to the CBT group following the test, while the paper-and-pencil group received their results two to three days later.

When the test was over, students in the experimental group were able to click a review button to see the questions, the correct answers, their own answers, and the scoring results. To provide metalinguistic feedback for all items (Lyster & Ranta, 1997; Sheen, 2007), the correct forms and metalinguistic information were provided. An example of feedback on a short-answer item is shown below in Figure 2.

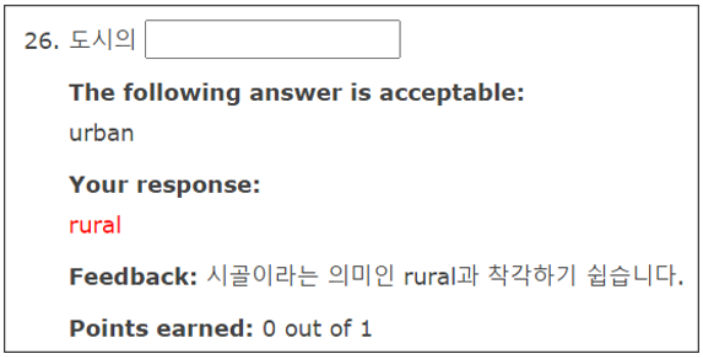

Figure 2. Screenshot of diagnostic feedback on short answer items.

Questionnaire

In order to elicit test-takers' perceptions of the usefulness of diagnostic testing in English instruction, the students were asked to respond to the post-experiment questionnaire. The questionnaire consisted of five closed-ended and two open-ended items. All three groups responded to the closed-ended items, which inquired about their interest in the English class, their overall satisfaction with the English class, the teacher's ability to identify their strengths and weaknesses, how well their weak points were addressed in class, and how successfully they learned. All items were answered on a five-point Likert scale ranging from "not at all" to "very well." Both experimental groups answered an open-ended item about the pros and cons of utilizing diagnostic tests, and the computer-based diagnostic testing group responded to another open-ended item about the advantages and disadvantages of computer-based diagnostic tests.

Audio-recording of lessons

After audio-recording the entire lessons, the researcher measured the time the teacher spent explaining language items covered in the diagnostic tests. Specifically, 12 hours and 8 minutes of instruction were recorded for the control group, 11 hours and 25 minutes for the paper-and-pencil assessment group, and 11 hours and 8 minutes for the CBT group. Since class was not recorded during diagnostic testing or while post-testing or surveys were being conducted, the total amount of instruction recorded for the two experimental groups was less than for the control group.

Post-test

For the purposes of post-testing, the 60 vocabulary and 20 grammar items showing the lowest correct response rates were selected from the total of 160 vocabulary and 22 grammar items. The post-test contained 48 multiple-choice items and 32 short-answer items, with one point allotted to each item for a total possible score of 80 points. The Cronbach's alpha reliability for each group's post-test was high, ranging from 89 to 92.

Procedures

As mentioned above, this study was carried out as part of high school instruction. Prior to the start of each new unit, the teacher did lesson planning for the upcoming unit, and instruction for the control group was implemented according to these lesson plans. The two experimental groups were given a unit-specific diagnostic test each time they began a new unit. The lesson plans were revised based on the results of the diagnostic tests of the two groups.

Whenever a unit was completed, teachers would create lesson plans for the following unit, after which diagnostic testing would be administered to the two experimental groups. The lesson plans would subsequently be revised to reflect the test results. The 13-minute diagnostic tests for the following unit were administered late in the final session of each unit. The CBT group would hold class in the computer lab for that session. To offset the possibility of the experiment's results being affected by the experimental groups' early exposure to target vocabulary and grammar during diagnostic testing, the control group was also exposed to the same target vocabulary and grammar that appeared in the diagnostic tests, through activity sheets given to them for practice.

The following data were collected to evaluate the validity of the three claims. A post-test was administered to analyze whether diagnostic tests lead to achievement differences. In addition, post-experiment questionnaires were distributed right after the students completed the post-test to elicit test-takers' perceptions of diagnostic testing in English instruction. In order to identify the impact of diagnostic testing on teaching activities, audiorecordings were made of the lessons presented to the three groups. Since class was not recorded during diagnostic testing or while post-testing or surveys were being conducted, the total amount of instruction recorded for the two experimental groups was less than for the control group.

Results

Enhanced Instructional Quality

Claim 1 postulates that teachers are able to identify learners' strengths and weaknesses and thus provide more efficient instruction by utilizing diagnostic tests. If this first claim is true, the teacher should have been able to perform remediation based on the diagnostic test results, allocating more time to areas of weakness. In order to determine whether Claim 1 was true, four backings were collected and examined

Relationship between item difficulty and instructional time allocation

In order to determine the impact of the diagnostic test results on the teacher's instruction, product-moment correlation coefficients were estimated between each group's correct answer rate and the amount of time teachers subsequently spent explaining the associated items. Since a high item difficulty value signifies a higher level of knowledge in students, a teacher utilizing the diagnostic test results would be expected to reduce the amount of expository instruction on that point. Conversely, the lower an item's correct answer rate is, the more time the teacher would spend explaining it.

As shown in Table 1, a statistically significant negative correlation was found between the two variables in both diagnostic testing groups, indicating that the amount of time the teacher spent explaining items with low correct answer rates was comparatively longer, while items with high correct answer rates received comparatively shorter amounts of instruction time. This finding may be viewed as evidence that she utilized the diagnostic test results and took learners' strengths and weaknesses into consideration when implementing instruction.

TABLE 1
Relationship between Correct Answer Rate and Instruction Time

Diagnostic tests	Area	Paper-and-pencil group	CBT group
1st unit test	Vocabulary	-.548*	-.578*
	Grammar	-.936*	-.931*
2nd unit test	Vocabulary	-.681*	-.579*
	Grammar	-988*	-.780*
3rd unit test	Vocabulary	-.325*	-.342*
	Grammar	-.848*	-.916*
4th unit test	Vocabulary	-.451*	-.251*
	Grammar	-.981*	-.901*
Average	Vocabulary	-.511*	-.432*
	Grammar	-.615*	-.622*

* $p < .05$

Comparison of instructional time allocation in three groups

While the analysis for the experimental groups found a negative correlation between correct answer rate and explanation time, the possibility could not be ruled out that the teacher might have shown similar teaching patterns in the control group. Thus, the expository instruction time the teacher allotted in the three groups were compared to identify differences in the amount of time across the groups. Specifically, the vocabulary and grammar items with low or high correct answer rates were selected, after which the amount of time spent by the teacher on explaining these items to the three groups was compared and analyzed. If the diagnostic testing had performed its intended function, the expected finding would be that, for the items that were the most problematic in the diagnostic testing, the teachers' explanations in the two experimental groups would be longer than in the control group, and the amount of time spent explaining the items that the students are already familiar with would be correspondingly shorter than that of the control group.

Using the results from the diagnostic tests for all four units, the vocabulary and grammar items with a correct answer rate of 30% or lower in the two experimental groups were selected and then compared with the amount of time each group's teachers spent explaining each item. As shown in Table 2, the teacher spent significantly more time explaining the items most problematic to students in the two experimental groups than in the control group. The results of the one-way ANOVA indicated that there was a significant difference in the teacher's explanation times for problematic items between the groups. The results of the post hoc Tukey test on the differences among the three groups' means indicated statistically significant differences between the control group and the two experimental groups in each diagnostic test. There was no significant difference in the amount of time the teacher spent in the two experimental groups.

TABLE 2
Explanation Time for Items with Low Correct Answer Rates (unit: seconds)

Diagnostic tests	Control group		Paper-and-pencil group		CBT group		*F*	*p*
	M	*SD*	*M*	*SD*	*M*	*SD*		
1st unit test	10.00	6.93	50.60	19.72	47.86	16.69	15.100	.000
2nd unit test	14.33	3.14	52.33	22.00	61.50	13.65	16.555	.000
3rd unit test	21.00	6.63	55.40	21.90	53.80	22.35	5.525	.020
4th unit test	19.67	9.14	50.50	19.12	41.33	8.91	8.539	.003

Meanwhile, in order to determine whether any difference was present between the two experimental groups using different testing modes in terms of instruction time for vocabulary items with high correct answer rates, vocabulary and grammar items with correct answer rates of at least 80% in both experimental groups were selected and then the two teachers' explanation times for each item were compared. While the control group exhibited longer explanation times for items with high correct answer rates than the two experimental groups, as seen in Table 3, the results of the one-way ANOVA indicated that there was no statistically significant difference in the amount of time the teacher spent in the three groups.

TABLE 3
Instruction Time for Items with High Correct Answer Rates (unit: seconds)

Diagnostic tests	Control group		Paper-and-pencil group		CBT group		*F*	*p*
	M	*SD*	*M*	*SD*	*M*	*SD*		
1st unit test	7.4	4.71	6.4	5.23	6.90	5.90	.089	.915
2nd unit test	8.63	4.84	8.33	7.81	5.50	3.07	.741	.488
3rd unit test	23.10	18.71	18.70	15.41	18.00	15.57	.276	.761
4th unit test	20.86	13.70	15.14	10.38	17.57	14.73	.337	.718

These results show that the diagnostic testing provides greater help for teachers in providing remediation by identifying content that students do not know very well, rather than what they already know well.

Impacts on teaching strategies

Up to this point, the analysis has tended to focus on comparing instruction times, without examining how teachers actually utilized the diagnostic testing data in their instruction. In order to find out how the teacher actually utilized the diagnostic testing results, the recordings were analyzed to examine if there is any difference in how the teacher dealt with the associated items. The analysis showed that the teacher utilized the results of diagnostic testing chiefly in two ways. The first difference was that, in contrast to the control group, the diagnostic test results were mentioned in the two experimental groups. For example, in those two classrooms, the teacher shared such information as 'the results of your diagnostic tests showed that many of you don't know this word (or grammar point),' or 'only two students knew this word (or grammar point),' or 'the correct answer rate was only 20%.' These comments served to call attention to the material to be learned. Examples are presented below.

> This word 'sedentary' is a high-difficulty word, with a correct answer rate of only 20%, so you really should become familiar with it. (paper-and-pencil group)

> Only two students knew the meaning of 'uncharted.' A lot of you must be encountering it for the first time. (CBT group)

The second difference revealed in the teacher's explanation methods was that, unlike the control group, in the experimental groups, the teacher presented learners with the incorrect answers they had given in the diagnostic tests, and explained details of the errors.

> So, a lot of you selected the definition of 'insert' for 'insult.' The pronunciation and spelling of the two words are similar, right? But, the word 'insert' means '삽입하다.' (paper-and-pencil group)

> It turned out that some of you were confused about 'slum' and 'slam'. The world 'slam' means 'closing something like a door so hard that it goes KWANG!' I hope you won't get them confused. (CBT group)

As shown above, the diagnostic tests enabled the teacher to identify the material that students did not know very well or misapprehended, while stressing the importance of and taking corrective action toward said material by mentioning specific errors to students.

Learners' perception of the quality of instruction

In order to determine the impact of diagnostic testing on students' perceptions of English class, the students were asked to respond to a post-experiment survey. The survey questions inquired about students' interest in English class, overall satisfaction with English class, the teacher's ability to identify learners' strengths and weaknesses, and how well the students' weak points were addressed in class. The students were requested to select a response for each item from a five-point Likert-type scale, ranging from 1 (not at all) to 5 (very well). As seen in Table 4, the survey results showed that the scores of the experimental groups who took diagnostic tests were significantly higher than the control group with regard to interest in and satisfaction with English class.

TABLE 4
One-way ANOVA Comparison of Groups' Satisfaction with English Class

	Control group	Paper-and-pencil group	CBT group	*F*	*p*
Interest in English class	3.15	3.81	3.43	4.018	.022
Satisfaction with English class	3.70	4.42	4.15	11.972	.000
the teacher's ability to identify learners' weaknesses	3.33	4.12	4.03	7.473	.001
how well the teacher address students' weak points	3.59	4.00	4.10	3.259	.044
how well students became aware of the content they didn't know	3.56	4.00	4.27	6.841	.002

Notably, there were statistically significant differences in the control group's and the experimental groups' responses to the questions, which indirectly explains why the implementation of diagnostic testing has a positive impact on student learning outcomes.

Impact on Student Achievement

Claim 2 posits that when diagnostic testing is utilized, academic achievement is increased. If this hypothesis is true, it is expected that learners' academic achievements will rise when they receive instruction that meets their needs as a result of diagnostic testing. In order to determine whether diagnostic testing led to academic achievement, the post-test was administered to the three groups. The 80-point post-test was produced by selecting the 60 vocabulary items and 20 grammar items that had the lowest correct answer rates from the 160 target vocabulary and 40 target grammar items covered in diagnostic testing. The Cronbach's alpha reliability for each group's post-test appeared high, with coefficients of .89, .92, and .91.

Comparison of student achievement in three groups

The results of the one-way ANOVA indicated that a significant difference existed between the three groups' total post-test scores, as reported in Table 5. The results of the post hoc Tukey test revealed that the paper-and-pencil group and CBT group were grouped together while the control group fell into a different group and that no difference was found in the academic achievement of the two experimental

groups. This finding suggests that the implementing diagnostic testing had a positive impact on academic achievement regardless of the diagnostic test mode.

TABLE 5
Post-test Results among Groups

Assessment areas	Control group		Paper-and-pencil group		CBT group		*F*	*p*
	M	*SD*	*M*	*SD*	*M*	*SD*		
Vocabulary	37.59	12.81	52.12	9.07	51.00	8.65	16.718	.000
Grammar	12.96	3.80	13.35	3.44	16.50	2.62	10.071	.000
Total	50.56	15.59	68.46	11.47	67.50	10.41	17.316	.000

Students' perceptions of the benefits of diagnostic feedback

In order to find out learners' opinions on using diagnostic testing, the two experimental groups were first asked whether diagnostic testing had been helpful in learning. The paper-and-pencil group (M = 4.19) and CBT group (M = 4.20) gave positive responses, with both having an average of over 4 out of a possible 5 points, and the results of the independent sample t-test showed that there was no statistically significant difference between these two groups' responses (t = -.055, p = .956).

The participants were also asked to provide their perspectives on the benefits of diagnostic testing, and both groups gave similar responses. As shown in Table 6, the main advantages of diagnostic testing indicated most often by students in both groups were that it enabled them to identify their current ability and areas of weakness, and that it enabled them to remember information better, since, following the diagnostic testing, these items could be reviewed in class. Some respondents also indicated that they started paying more attention during class and experienced increased motivation to learn. They also liked the fact that previously confusing items became more understandable, and that the teacher's instruction addressed areas of less familiarity.

TABLE 6
Opinions on the Pros and Cons of Diagnostic Testing

	Opinions (total number of responses from both groups)
Positive opinions	I can identify my current ability and areas of weakness. (32) I can remember things better, because the items can be checked and reviewed in class after diagnostic testing. (30) I pay more attention in class, and have increased motivation to learn. (15) Things that used to be confusing are more understandable. I like that the teacher seems to run the class with awareness of which areas students know or do not know very well, and which points are confusing. (13) Since important vocabulary and grammar is previewed beforehand, it is helpful for reading comprehension. (11) I can find out about upcoming vocabulary items and their importance before the next English unit. (2) I don't feel any pressure about it. (1)
Negative opinions	I feel anxious and lose confidence because I don't know a lot of the vocabulary and grammar on the tests. (12) Since diagnostic testing takes place during class, our actual class time is reduced. (4) Seeing the vocabulary beforehand interferes with reading comprehension. (2)

These responses suggest that, by clearly informing learners of their areas of weakness, diagnostic testing provides them with an opportunity for review and self-reflection. Yet, because they are not high-stakes exams, diagnostic tests create little or no anxiety and can heighten learners' motivation. Such findings are consistent with the opinions of various scholars (Alderson, 2005).

Students' perception of the disadvanges of classroom diagnostic testing

The participants were also asked to comment on the problems and disadvantages of classroom diagnostic tests. As Table 6 shows, when asked about the disadvantages of diagnostic testing, nearly 20% of the students indicated that they felt some anxiety and decreased confidence when they learned that they did not know a lot of the vocabulary and grammar on the tests. Such a response can be interpreted in the same context as the loss of confidence that accompanies immediate feedback reported in Alderson (2005). In addition, although only 13 minutes of class time was spent on diagnostic testing for each unit covered, the fact that some respondents had a negative perception of this reduction in actual instruction time is an issue needing serious consideration before diagnostic testing can become a routine classroom practice.

Benefits of Immediate Feedback

Claim 3 is based on the premise that when immediate feedback is provided, the effectiveness of diagnostic testing will be even greater. One of the most significant differences between the two experimental groups was the difference in how quickly feedback was received. Feedback was provided to the paper-and-pencil group 2 to 3 days after the test, while test-takers in the CBT group were able to check their results immediately after answering the questions. This study analyzed whether immediate feedback resulted in higher achievement.

Comparsion of student achievement in two diagnostic testing groups

When the post-test scores of the two experimental groups were examined to determine whether this divergence produced a difference in learning results, no significant difference was found in the two groups' scores, as shown in Table 5. Therefore, Claim 3 was not supported. Of course, the possibility must be considered that differences in other aspects of the two groups' diagnostic tests may have affected the post-test results.

Students' perception of the benefits of computer-based diagnostic testing

The participants in the CBT group were asked for their views on computer-based diagnostic testing in an open-ended questionnaire. As reported in Table 7, some responses indicated that inputting answers on the computer was convenient and fun, and that they liked being able to receive the correct answers and scores immediately.

Students' perception of the limitations of computer-based testing

As seen in Table 7 above, the negative opinions regarding computer-based diagnostic testing involved the inconvenience of moving to the computer lab, the potential for distractions on the computer, and that while CBT is sophisticated, it is not all that different from paper-based assessment. The comment that the item format in the computer-based tests were the same as a paper-and-pencil test warrants attention, as some may feel it is not worth the trouble of using CBT when the same exam could be taken on paper.

Table 7
Opinions on the Pros and Cons of CBT Diagnostic Testing

	Opinions (total number of responses from both groups)
Backing	Inputting the answers is convenient and saves time. (12) Scoring is fast, and I can directly see the assessment results and correct answers. (9) Using the computer program seems fun and beneficial. (8) I pay more attention in class, and memorizing words is easier. (5) The scoring is accurate. (2)
Rebuttal	It is a hassle moving to the computer lab every time we have to take a diagnostic test. (11) The computers might be used for unrelated purposes. Students might be distracted by games or using the Web. (10) The procedure of booting up the computer, accessing the site, and logging in is cumbersome and wastes time. (8) It's essentially not that different from paper-based assessment. (7) It's inconvenient to have to type instead of writing by hand. (5) It's hard to concentrate. (3) Since scoring is automatic, it is less accurate for open-ended questions. (2)

Discussion and Conclusion

This study is one of the first attempts to investigate the usefulness of computer-based classroom diagnostic tests with an argument-based validation approach. The most important finding is that diagnostic testing provides teachers with data that can enhance the efficacy of instruction, as asserted in the literature. By identifying the strengths and weaknesses of students, diagnostic testing provides data that teachers can use to adjust and modify instructional content, thus having a positive influence on academic achievement. The difficulties of construct definition for assessment and the practical issues can both be mitigated if diagnostic tests are created and administered according to individual units, as in this study. Thus, there is a need to actively utilize diagnostic testing in the actual field of education. Producing and implementing diagnostic tests can be burdensome for many teachers who are already overloaded with work. Rather than creating and developing diagnostic tests on their own, teachers could explore the option of group collaboration with other teachers to lessen the individual burden. Taking this one step further, an option to consider is to have textbook writers construct diagnostic tests and make them available to teachers.

The results of this study demonstrate that diagnostic tests enable teachers to pinpoint specific gaps in student knowledge that need to be targeted. Also, the positive impact of diagnostic testing on student achievement suggests that diagnostic testing should be actively utilized in the actual classroom. Producing and implementing diagnostic tests can, however, be burdensome for many teachers who are already overloaded with work. Rather than creating and developing diagnostic tests on their own, teachers could explore the option of group collaboration with other teachers in order to lessen the individual burden. Taking this one step further, another option to consider is having textbook publishers develop and provide diagnostic testing items when they are creating instructional materials.

As reported in the study's findings, there were some students who felt it was a drawback that diagnostic testing during class hours cut into regular class time that could have been spent studying. In cases where there is insufficient time for diagnostic testing in the classroom, an idea worth considering is to implement computer-based diagnostic testing as a self-assessment tool in an online test format, such as DIALANG, capitalizing on the advantages of the 'anytime, anywhere' feature of online assessment (Alderson, 2005; Roever, 2001). Because security is not an issue for low-stakes exams such as these, online diagnostic tests could be assigned as part of students' homework, provided that the purpose of the exam is clearly explained in advance.

Although some learners expressed the opinion that immediate feedback was helpful, no significant difference was identified in academic achievement resulting from the provision of immediate feedback. Possible reasons for the lack of effect are that learners did not properly understand the feedback, or may

have lacked the ability to use the feedback to revise and make fresh plans for self-learning strategies. Alternatively, because diagnostic tests are not high-stakes exams, they might not have analyzed the feedback with great attention. The way learners interpret and utilize feedback is a crucial research area in second and foreign language education (Hyland, 1998), and a few studies have found that students pay little attention to feedback (Crooks, 1988; Hounsell, 1987). If students do not refer to feedback, it will be difficult to achieve one of the key goals of diagnostic testing. Therefore, research is essential for identifying teaching strategies to help students use the feedback provided by diagnostic tests in order to make learning plans. Since the diagnostic tests in this study were administered in the classroom setting, the teacher had time to use the results to provide explanations during class time. In situations where no explanations are offered, however, there could be differences between the two test modes. It would be helpful to investigate this issue in follow-up studies. In addition, the impact of diagnostic testing in this study was found to vary between the two language skill areas of vocabulary or grammar, and this finding merits further research to investigate whether a relationship exists between the diagnostic test mode and language learning area.

Another fruitful avenue for future studies may be to determine if there is a threshold of time after which a difference would become apparent. In this study, each of the two experimental groups received feedback within a period not exceeding 2 to 3 days. In the case of CBT, feedback could be provided immediately after each individual item or could alternately be withheld until all the test questions had been answered. Some students pointed out that they got distracted easily on the computer. Another possibility is that, as suggested by Yin, Sims, and Cothran (2012), different learners may prefer different types of feedback or prefer to receive it at different times. Future studies on these issues would help us to better understand the effects of the timing and mode of diagnostic feedback on student achievement.

Because it was conducted as part of the instruction in intact high school classes, this study is expected to have high ecological validity (Brewer, 2000). However, it has certain limitations. First, the students were not randomly assigned to the three groups. Next, given the small number of participants, it could be difficult to generalize these results across classroom settings of other education environments. The possibility of a ceiling effect also exists, considering that almost all students received extremely high scores in the pre-test. In addition, the effect of research bias cannot be entirely ruled out, since the same teacher was in charge of instruction in all three classrooms, and this teacher also participated in the task of developing the diagnostic tests, so it was not a blind design. Another limitation is that the areas tested were limited to grammar and vocabulary. In order to maximize the effect of diagnostic testing in actual classrooms, additional research is needed in order to develop diagnostic testing that can be applied to all four language skills. Finally, the large number of vocabulary and grammar areas to be assessed precluded the possibility of conducting a comprehensive pre-test during regular class hours. However, if such a pre-test had comprehensively examined vocabulary and grammar, a more precise conclusion could have been derived by identifying the impact of the experimental treatment. The internal validity of future studies could doubtless be further enhanced if both pre- and post-tests are administered targeting the grammar and vocabulary items to be covered in the diagnostic tests.

This study is significant in the sense that it explored whether purported advantages of diagnostic testing -which are accepted as a matter of course in the field of language education, even without empirical support - were in fact valid. Davies (1984) stated that “diagnostic testing has to date been highly desirable but somehow unattainable” (p. 43). The results of this study demonstrate that classroom diagnostic testing is indeed doable and beneficial. It is hoped that this study will serve as a guide in encouraging a wider use of diagnostic testing in language classes.

The Authors

Boh Young Lee (first author) is a graduate student in English education program at the Graduate School of Education in Ewha Womans University. Her main research interests include assessment for learning, ELT methods, and computer-assisted language teaching.

Department of English Education
Ewha Womans University
Email: bboh612@ewhain.net

Sang-Keun Shin (corresponding author) is a professor of Applied Linguistics at Ewha Womans University. His main research interests are language assessment, multimedia-assisted language teaching, and second language teacher education.

Department of English Education
Ewha Womans University
Email: sangshin@ewha.ac.kr

References

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

Alderson, J. C. (2010). Cognitive diagnosis and Q-matrices in language assessment: A commentary. *Language Assessment Quarterly, 7*(1), 96–103.

Alderson, J. C., & Huhta, A. (2011). Can research into the diagnostic testing of reading in a second or foreign language contribute to SLA research? *EUROSLA Yearbook, 11*(1), 30–52.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14*(3), 191–205.

Brewer, M. (2000). Research design and issues of validity. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3–16). Cambridge, England: Cambridge University Press.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 44–59.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*, 119–157.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. New York, NY: Cambridge University Press.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*(4), 438–481.

Davidson, F. (2010). Why is cognitive diagnosis necessary? A reaction. *Language Assessment Quarterly, 7*(1), 104–107.

Davies, A. (1984). Computer-assisted language testing. *CALICO Journal, 1*(5), 41–43.

Douglas, D. (2010). *Understanding language testing*. London, England: Hodder Education.

Dunkel, P. (1991). *Computer-assisted language learning and testing: Research issues and practice*. New York, NY: Newbury House.

Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence.' *Language Testing, 21*(3), 259–283.

Hounsell, D. (1987). Essay writing and the quality of feedback. In J. T. E. Richardson, M. W. Eysenck, & D. Warren-Piper (Eds.), *Student learning: Research in education and cognitive psychology* (pp. 109–119). Milton Keynes: Open University Press and Society for Research into Higher Education.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.

Hyland, F. (1998). The impact of teacher-written feedback on individual writers. *Journal of Second Language Writing, 7*(3), 255–286.

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

Kim, Y.-H. (2019). Developing and validating empirically-derived diagnostic descriptors in ESL academic writing. *Journal of Asia TEFL, 16*(3), 906–926.

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 63*(5), 505–512.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79–97.

Kunnan, A., & Jang, E. E. (2009). Diagnostic feedback in language testing. In M. Long & C. Doughty (Eds.), *The handbook of language teaching* (pp. 610–625). Oxford, UK: Blackwell Publishing.

Larson, J. W., & Hendricks, H. H. (2009). A context-based online diagnostic test of Spanish. *CALICO Journal, 26*(2), 309–323.

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

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition 19*, 37–66.

McMillan, J. H. (2006). *Classroom assessment: Principles and practice for effective standards-based instruction*. Boston, MA: Allyn & Bacon.

Norrish, N. (1990). An experiment in individualization using technical support. In H. A. de Jong & D. K. Stevenson (Eds.), *Individualizaing the assessment of language abilities* (pp. 154–165). Bristol, PA: Multilingual Matters.

Richards, B. J. (2008). Formative assessment in teacher education: The development of a diagnostic language test for trainee teachers of German. *British Journal of Educational Studies, 56*(2), 184–204.

Roever, C. (2001). Web based language testing. *Language Learning and Technology, 5*(2), 84–94.

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly, 41*(2), 255–283.

Shohamy, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal, 76*, 511–521.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Yin, M., Sims, J., & Cothran, D. (2012). Scratching where they itch: Evaluation of feedback on a diagnostic English grammar test for Taiwanese university students. *Language Assessment Quarterly, 9*(3), 78–104.