Korean Journal of English Language and Linguistics, Vol 23, July 2023, pp. 535-553
DOI: 10.15738/kjell.23..202307.535

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

Enhancing EFL Learners' English Reading Performance through the Diagnostic Feedback of CDA*

Huihui Sun (Lyuliang University) **Yohan Hwang** (Jeonju University)

Received: January 28, 2023
Revised: July 6,2023
Accepted: July 12, 2023

Huihui Sun (1st author)
Assistant teacher, Department of Foreign Languages,
Lyuliang University
Email: 372741616@qq.com

Yohan Hwang
(corresponding author)
Professor, Department of English Language and Literature,
Jeonju University
Email: yvh5101@jj.ac.kr

* This research builds upon and extends the PhD dissertation of the first author.

ABSTRACT

Sun, Huihui and Yohan Hwang. 2023. Enhancing EFL learners' English reading performance through the diagnostic feedback of CDA. *Korean Journal of English Language and Linguistics* 23, 535-553.

Cognitive Diagnostic Assessment (CDA) enhances the value of assessments by accurately diagnosing learners' strengths and weaknesses in specific fields. It provides an invaluable opportunity for individualized teaching and targeted remediation. The primary objective of this paper is to enhance the reading section of the College English Test Band Four (CET-4) in China by incorporating the G-DINA cognitive diagnostic model and proposing a comprehensive teaching framework to effectively address students' weaknesses. To achieve this objective, diagnostic information is obtained using the G-DINA cognitive diagnostic model, specifically from the reading section of CET-4. In order to integrate the diagnostic information into the classroom setting, a novel strategy instruction framework is proposed. To assess its effectiveness, a comparative teaching experiment is conducted. The findings of the study highlight that the reading section of CET-4 serves as a powerful tool for identifying learners' weaknesses in reading. Furthermore, the proposed framework demonstrates significant improvements in students' reading performance. These compelling results indicate that CDA outperforms other traditional assessment methods, showcasing its superior instructive nature. Based on these insightful findings, this study suggests the adoption of CDA as a more comprehensive and impactful approach to assessment, as it not only diagnoses learners' strengths and weaknesses but also provides valuable insights for personalized instruction and targeted remediation.

KEYWORDS

retrofitting CDA, English reading comprehension, G-DINA, teaching intervention method, CET-4

1. Introduction

Cognitive Diagnostic Assessment (hereafter CDA) is a type of assessment designed to yield comprehensive insights into an individual's cognitive strengths and weaknesses related to a specific task or domain. In the context of reading comprehension, CDA can be used to identify examinees' mastery extent of specific cognitive processes or skills necessary for successful reading comprehension, such as decoding ability and inferencing skill (Fan et al., 2021). Alderson (2005) asserts that the diagnostic testing is the interface between learning and assessment. Because the assessment method provides specific diagnostic feedback for stakeholders, its significance in the language field has been widely recognized (Chen and Chen 2016, Du and Ma 2021, Toprak and Cakir 2021). It enables students and teachers to design better instructional programs. Based on the diagnostic information, teachers can develop efficient and precise educational interventions tailored to individual learners' needs, and learners can adopt more suitable learning methods for themselves (Sawaki et al., 2009). However, proficiency assessment remains the primary testing method in the language test field. This approach hinders teachers from developing accurate teaching interventions since proficiency assessment measures test-takers' general language level and reports the measurement results in a summative manner. Typically, examinees are provided with a total score, which may offer some indication of a test taker's overall reading proficiency, but does not provide additional insights into the specific areas of reading that require further improvement.

In order to address the problem by providing more concrete feedback, numerous areas of CDA have been investigated. A keyword co-occurrence network is pictured with CiteSpace II to study the research field and prominent thematic areas of CDA (Figure 1). The literature is extracted from the Web of Science, and the extraction time is from 2000 to 2022. An exact topic search for "cognitive* diagnos*" in titles, abstracts, or keywords results in 927 records. After manually filtering out non-representative records, such as, applications of CDA in the mathematical field, the final dataset comprises 405 records.

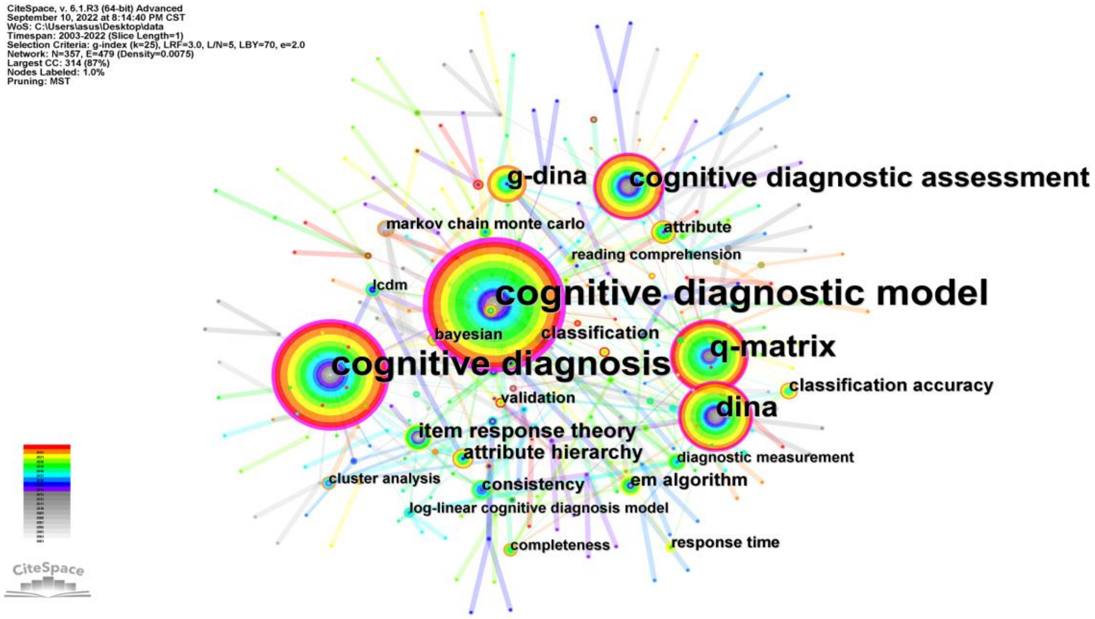
CiteSpace, v. 6.1.R3 (64-bit) Advanced
September 10, 2022 at 8:14:40 PM CST
WoS: C:\Users\asus\Desktop\data
Timespan: 2003-2022 (Slice Length=1)
Selection Criteria: g-index (k=25), LRF=3.0, L/N=5, LBY=70, e=2.0
Network: N=357, E=479 (Density=0.0075)
Largest CC: 314 (87%)
Nodes Labeled: 1.0%
Pruning: MST
g-dina
cognitive diagnostic assessment
markov chain monte carlo
attribute
reading comprehension
lcdm
cognitive diagnostic model
bayesian
classification
q-matrix
cognitive diagnosis
classification accuracy
validation
dina
item response theory
attribute hierarchy
diagnostic measurement
cluster analysis
consistency
em algorithm
log-linear cognitive diagnosis model
completeness
response time
CiteSpace

Figure 1. The Keyword Co-Occurrence Network of CDA

As shown in Figure 1, the nodes of the *cognitive diagnostic model*, *Q-matrix*, *attribute*, *attribute hierarchy*, *Markov chain monte carlo*, *EM algorithm*, *classification accuracy*, and *consistency* indicate that existing studies have mainly focused on the theoretical construction of CDA. Specifically, cognitive diagnostic models have been developed, assessed, and compared; the nature of attributes has been explored to construct a Q-matrix; and new indices and algorithms have been generated to enhance the accuracy of CDA. While expanding the theoretical foundations of CDA, researchers have also applied the theoretical findings to practical applications of English *reading comprehension*. They have developed diagnostic tests or retrieved diagnostic information from existing assessments. Nevertheless, none of the nodes are about teaching intervention. It represents that the studies have mainly been limited to the testing phase with the methods and effectiveness of using assessment results in teaching not yet fully explored. Among 405 records, only one paper is detected (Fan et al., 2021). It shows that it is time to propose a teaching method that integrates diagnostic information into the curriculum. This study attempts to survey the feasibility of incorporating diagnostic feedback of CDA into classroom teaching using the expert scaffolding framework, while the study conducted by Fan et al. (2021) focuses on enhancing the learners' weaknesses through their own repetitive drills. To this end, a new instruction framework is proposed and an empirical teaching experiment is conducted to measure the effectiveness of the proposed framework.

The diagnostic information is collected from the reading section of the College English Test Band 4 (hereafter CET-4). The selection of the CET-4 as the testing instrument is motivated by its underrepresentation in the field of CDA, despite being a national assessment in China. Most existing CDAs retrieve diagnostic information from the reading part of international exams, such as IELTS and TOEFL. Therefore, the present study seeks to achieve two objectives: firstly, examining the testing constructs of the reading section of the CET-4; and secondly, exploring the effectiveness and methods of improving learners' weaknesses in reading based on the diagnostic feedback. Specifically, the following research questions guide the current study.

1) What are the abilities required for completing the CET-4 reading tasks?

2) To what extent is the proposed instruction framework, in which diagnostic information is integrated into the curriculum, effective in remedying learners' weaknesses?

2. Theoretical background

2.1 The College English Test for Band 4

The CET-4 is a national assessment in China organized by the Ministry of Education to measure the English ability of college students. The test includes five subtests: listening, reading, writing, translation, and speaking with the latter being optional. This study focuses on the reading subtest, which is titled *Reading comprehension*. It consists of 30 multiple-choice questions distributed across three sections. Section A consists of one passage with ten blanks, and examinees are required to select one word for each blank from a list of choices given in a word bank following the passage. Section B consists of one passage with ten statements attached to it, and examinees must identify the paragraph from which the information is derived. Section C comprises two passages, each followed by five four-choice multiple-choice items, and examinees need to choose the best choice. The test aims to measure ten specific skills, as described in the test syllabus, including understanding gist, factual information, opinions, and attitudes expressed explicitly by the first author, generalizing main ideas, inferring implicit information, understanding views and attitudes expressed implicitly by the first author, guessing word meaning based on context, understanding relations across sentences, acquiring rhetorical organization through cohesions, and applying reading strategies. The time allocation for the reading part is 40 minutes.

Like other international exams, such as IELTS, an examinee's language performance in the CET-4 is generally reported through an overall score and separate scores for reading, writing, listening, and translation. However, considering that the CET-4 is intended to serve as a guide for both teachers and students, it is reasonable to utilize it for more specific diagnostic purposes. Hence, this study aims to evaluate whether retrieving diagnostic information from the reading section of the CET-4 is feasible by examining two assumptions: (1) the constructs of the reading section of the CET-4 are divisible; and (2) diagnostic information generated from the CET-4 is accurate and reliable.

2.2 Attributes of reading comprehension

A prerequisite for applying CDA to reading comprehension is that reading ability can be decomposed into a set of specific and identifiable cognitive skills that describe the processes which a learner uses to perform tasks. However, the extent to which reading ability can be identified independently remains a subject of controversy. Lunzer et al. (1979) have argued that reading ability is indivisible, representing a single, global, and integrated aptitude. Nevertheless, this notion has been challenged by numerous scholars and researchers who claim that reading ability is divisible. While believing that reading involves several different cognitive processes and skills, researchers hold different views on the number, scope, and nature of subcomponents (as cited by Jang, 2009). Alderson (2000) suggested that decoding and general reading ability are two components of reading, whereas Coady (1979) proposed that reading consists of three interactive elements: conceptual abilities, background knowledge, and process strategies. Bachman and Palmer (1996) expanded on the reading model of Coady (1979) to include word recognition, phonemic/graphemic features, syntactic feature recognition, prior knowledge, metacognition, and intratextual perceptions. A review of L2 reading research by Grabe (1991) classified the sub-components of reading into six categories, including vocabulary and structural skills, automatic recognition skills, formal discourse structure knowledge, synthesis and evaluation skills, content and background knowledge, and metacognitive knowledge and skill monitoring.

When researching L2 reading ability, reading testers also show great interest in identifying reading skills using the method of CDA. For example, Sawaki et al. (2009) identified four reading skills from the reading section of TOEFL® iBT, namely understanding word meaning, understanding specific information, connecting information, and synthesizing and organizing information. Mirzaei et al. (2020) decomposed the reading ability required for reading tasks in IELTS into six specific cognitive processes: lexico-grammatical knowledge, making inferences, scanning for specific information, skimming for general information, connecting and synthesizing information, and summarizing. Overall, the majority of theoretical and practical studies support the notion that reading ability is multifaceted, providing a prerequisite for CDA.

2.3 Cognitive diagnostic assessment

CDA consists of five main procedures: defining measured attributes or skills, designing test items, constructing a Q-matrix, analyzing examinees' response data, and reporting scores. The measured skills are typically defined by experts in a specific field (Lee and Sawaki, 2009b). Based on the identified skills, test items are designed, and a Q-matrix is built. The Q-matrix is a binary $J * K$ matrix. It relates test items to the measured skills with rows representing test items and columns representing skills. Q_{jk} equals 1 if the Kth attribute is required by the item J, and 0 otherwise. The response matrix is an $I * J$ matrix X, representing examinees' responses to the items. Each element *Xij* indicates whether examinee *I* responds to item *J* correctly (*Xij* = 1) or not *(Xij* = 0). The Q-matrix and response matrix are two critical inputs for CDA. Examinees' diagnostic information is acquired through analyzing the Q-matrix and response matrix with the Cognitive Diagnostic Model (hereafter CDM).

CDA can be divided into two categories based on the origin of test items: true CDA and retrofitting CDA. In true CDA, a set of test items is designed specifically for diagnostic purposes. Retrofitting CDA, on the other hand, aims to extract diagnostic information from existing proficiency tests. While only a few true CDAs of reading comprehension have been developed by researchers (e.g., Du and Ma 2021, Toprak and Cakir 2021), most researchers have adopted the method of retrofitting CDA (e.g., Jang 2009, Li et al., 2016, Mirzaei et al., 2020). The accurate diagnostic results obtained from retrofitting CDA prove that it is a reliable method. Furthermore, an analysis of the language skills involved in successfully performing existing proficiency tests can guide and inform the design of diagnostic tests. Therefore, retrofitting CDA is an effective method of acquiring diagnostic information and an important step in advancing CDA in the language field.

2.4 Selecting CDM and building Q-matrix for the reading test

A good fit between CDM, Q-matrix, and participants' response data is the basis for acquiring accurate diagnostic feedback. CDMs are psychometric models that have been developed specifically for CDA (Chen and Chen 2016). According to assumptions and performances of models, CDMs are divided into three types, namely, compensatory, non-compensatory, and saturated models (Yamaguchi and Okada 2018). The compensatory models allow for the possibility of mastered attributes compensating for others that have not been mastered, whereas non-compensatory models do not allow for such compensation, requiring all required skills to be mastered before a test-taker can answer an item correctly. The saturated models allow for both compensatory and non-compensatory relationships between attributes within the same test (Mirzaei et al., 2020, Yun 2017).

The point that researchers have argued is which type of CDM is most appropriate for reading tests. Their debates relate to their different perceptions of the nature of the reading process. Researchers, such as, Bernhardt (2005) and McNeil (2012), claim that reading skills are not independent, but rather compensatory. Having sufficient knowledge of one reading skill can assist in other skills that are inadequate. However, studies solely employing compensatory models in CDA are infrequent. On the other hand, directly employing non-compensatory models in assessments is common. This is because non-compensatory relationships between attributes are more compatible with the cognitive assumption of CDA. These assumptions require examinees to perform all of the attributes required by an item to obtain a correct answer (Mirzaei et al., 2020). While most studies relied on a single predetermined model, several scholars have compared the functioning of different models to determine the true model. Li et al. (2016), for example, compared the performances of five CDMs on the reading test of the Michigan English Language Assessment Battery (MELAB), including a saturated model, two compensatory models, and two non-compensatory models. Lee and Sawaki (2009a) investigated the performance of three models on reading and listening assessments of the TOEFL iBT. Among the three models, the general diagnostic model is a compensatory model, the fusion model and latent class analysis are two non-compensatory models. However, their comparison results are inconsistent with Li et al. (2016) reporting that the saturated model and compensatory model had better fitness than the others, while the other study reported that all measured models produced similar results (Lee and Sawaki, 2009a). These inconsistencies suggest that the appropriate CDM should be chosen based on the specific testing context through relative and absolute fitness tests. The former test indicates the most appropriate model when competing models are available, whereas the latter evaluates whether the model itself fits the data.

As for building the Q-matrix, there are various available sources (e.g., reading ability models, previous CDA research, and think-aloud protocol). Scholars have taken different approaches to build the Q-matrix. Buck et al. (1997) identify a set of important item characteristics involved in the necessary information that the reader must understand to be certain of the correct answer. They then code the test items for item characteristics and analyze the abilities needed to perform the characteristics, ultimately building an initial Q-matrix that is refined by examining the statistics of each attribute. In contrast to establishing the Q-matrix depending on experts' analyses for test task characteristics, Jang (2005) develops her initial Q-matrix with verbal reports from 12 participants. Mirzaei et al. (2020) adopt the same method. They determine the final Q-matrix with expert judgments and empirical validation with examinees' response data. When disagreements are observed between experts' judgments and students' introspections, experts' disagreements are resolved in favor of the students' introspections. These examples highlight the importance of considering multiple sources when building a Q-matrix, especially students' actual performance on a given task.

3. Methods

This study consists of two phases. Phase 1 aims to retrieve diagnostic information from the reading section of the CET-4. The research procedure of this phase is iterative due to the complexity of building the Q-matrix. The initial Q-matrix is constructed based on the judgments of three judges regarding the test task characteristics of the items. Subsequently, the Q-matrix is refined using data from think-aloud protocols. To ensure objectivity in building the Q-matrix, the revised Q-matrix is evaluated using the Q-matrix validation function of the G-DINA package in R. After obtaining a Q-matrix of satisfactory quality, a CDM is chosen, and diagnostic information is acquired.

In Phase 2, a strategy instruction framework is proposed, and a case experiment is conducted to measure the effectiveness of the framework in improving learners' reading weaknesses. The six-week experiment includes two experimental groups and one control group, with each class lasting for two hours and held twice a week. The experimental groups receive their diagnostic information, while the control group does not. In addition, experimental group one is not taught with the proposed framework, while experimental group two is taught under this notion. The teaching procedures for all groups are the same. During the first hour, students engage in text discussion within groups while their performance in solving the tasks is evaluated by themselves, peers, and the teacher in the second hour. Participants' weaknesses are monitored through pre- and post-tests administered in the first and sixth weeks, respectively. The test results are analyzed with the paired-samples t-test and independent sample t-test to compare the results.

3.1 Participants

This study has a diverse group of participants. Specifically, the test-takers whose responses are used as one of the inputs comprise a sample of 1,875 examinees who are randomly selected from a local university in China. These examinees represent a variety of academic majors, for example, chemical engineering, mining engineering, computer science and technology, education, and Chinese language and literature. Their performance on the College Entrance English Test is varied, ranging from less than 70 to more than 120 points out of a total score of

150 points. Three raters, who are doctoral students with more than eight years of college English teaching experience and familiar with reading skills required by the CET-4 reading section, are recruited to judge the items' test task characteristics. 16 pre-CET-4 participants with different English proficiency are selected to take part in think-aloud reports. To ensure that the participants could represent the population, they are chosen with the stratified random sampling method, which involves dividing the population into three subgroups (i.e., strata) based on their total scores in the reading section of the CET-4. Then, a random sample is selected from each stratum with the proportion of individuals sampled from each stratum determined by the proportion of that stratum in the overall population. In this study, individuals ranking in the top 25% and bottom 25% of the total score are classified as high- and low-level groups, respectively, and the rest are classified as the middle group. Thus, out of 16 participants, four are from the low-level group, eight are from the medium-level group, and four are from the high-level group. The advantages of stratified random sampling are significant, including improving the representativeness of the sample and reducing the sample size requirements. Moreover, it can help ensure the comparability of different groups in the analysis by ensuring each stratum is represented in the sample.

In Phase 2, 36 participants are recruited for a case study to measure the effectiveness of the proposed framework. Similar to Phase 1, participants are chosen with the stratified random sampling method with 12 participants selected from each strata. All teaching activities are managed by a female teacher who is familiar with the proposed instruction framework and has rich experience in teaching English reading. To ensure confidentiality, the study only contains completely anonymous data with prior agreement and full approval from all the participating students.

3.2 Instruments

The test instrument for retrieving diagnostic information in Phase 1 is the reading section of the CET-4 held in June 2021. The section comprises 30 multiple-choice items based on four passages ranging in length from 210 to 1077 words. The passages cover topics in humanities, social and life sciences, and necessary background knowledge is either known to examinees or provided within the passages, as described in the test syllabus. A rating instrument is developed to analyze the content characteristics of key information for answering questions. The instrument draws on research led by Bachman et al. (1995, 1996), Carr (2006), and Freedle and Kostin (1993, 1999), as well as syntactic descriptions from Celce-Murcia and Larsen-Freeman (1999). Seven variables are considered: length and location, propositional content, focus constructions, cohesion, negations, rhetorical organization, and item type. An example item in the rating instrument is presented, which is taken from an instrument used by Carr (2006). The item is rated for the degree to which information is ambiguous: 1 = no ambiguity; very clear; 2 = slightly ambiguous; 3 = moderately ambiguous; 4 = ambiguous; 5 = highly ambiguous.

To refine the Q-matrix developed from experts' suggestions, a think-aloud protocol is employed, which asks participants to verbalize their thoughts while performing a task or solving a problem. Several factors are considered to obtain accurate verbal reports. One of the key factors is the participants' English speaking ability. Since many of them have difficulty expressing themselves fluently in English, they are allowed to use Chinese to report their thoughts and opinions to ensure that their language abilities do not interfere with the accuracy and quality of the data. Another crucial factor is the reporting time. There are two main types of think-aloud protocols: concurrent and retrospective think-aloud. Concurrent think-aloud requires participants to speak their thoughts in real-time as they perform the task, while the retrospective type involves participants recalling their thoughts immediately after completing the task. The participants are given the freedom to select their preferred method. Out of the group, only one opts for concurrent verbal reporting, while the rest prefer immediate retrospective reporting. The interviewer's role is an important factor to consider when conducting think-aloud protocol. It is essential that the interviewer minimizes interruptions to allow participants to express their thought processes fully. After considering the above factors, the tester and each participant complete their report individually. Prior to the reporting, the research purpose and think-aloud process are briefly explained to the participants, who are then asked to practice the process by solving math problems and given five minutes to review the reading material and test questions before the formal interviews begin. All recordings are transcribed into text at the item level, and two raters encode them by referring to a set of attributes acquired from the literature. The percentage agreement is computed (r = .75). The identified skills are used to refine the Q-matrix.

The G-DINA package in R is used to measure the Q-matrix and obtain diagnostic information about the examinees. This package offers the advantage of estimating diagnostic information and comparing different models, which is not available in many other computer programs that handle only one type of CDM, such as Mplus for log-linear CDM and Arpeggio Suite for noncompensatory-RUM. Furthermore, the package examines the Q-matrix using examinees' response data, which helps to minimize subjectivity in building the Q-matrix.

In Phase 2, diagnostic information is shared with the examinees and their teachers to help them implement remedial instruction. The diagnostic information for an examinee includes the list of the CET-4 reading skills with their related descriptions, his or her band score, ranking among all participants, and detailed diagnostic feedback about his or her weak areas. A total of 12 passages for training are chosen from textbooks and educational websites, and evaluated with TextEvaluator to ensure that the texts' complexity level is equivalent to that of college students' reading ability level. The average length is 1,126 words. All of the passages are expository or narratives, and they represent a range of topics, such as, the impact of climate change on environment and society, the role of education in personal and societal development, and the importance of mental health awareness and treatment. TextEvaluator is freeware designed by Educational Testing Service (ETS). It assesses the texts based on the difficulty of words, sentences, connections across ideas, and text organization manners. Furthermore, two additional assessment instruments are developed with the items sourced from the CET-4 item bank to monitor improvement in participants' weaknesses. Experts meticulously modify some questions to ensure a one-to-one correspondence between each question and the skill being measured, while ensuring that each skill is tested an equal number of times (n = 10).

3.3 Statistical Analysis

The process of content analysis starts with identifying key information and ends with determining the item types. Inter-rater agreement for two nominal variables (i.e., rhetorical organization and item type), is calculated using percentage agreement, while the intraclass correlation coefficient (ICC) is used to measure the judges' agreement on the other scale variables. When a variable is measured by more than one item, its agreement coefficient is taken as the mean of the items' coefficients. The inter-rater reliabilities *for the line number on which the key sentence(s) start(s) and the number of key sentence(s)* are .71 and .73, respectively. To increase consistency in the location and number of key sentence(s), group discussions are held. After unifying the vital information, the three judges independently evaluate the other indicators of the key information and use them as the basis for judging the item types. Thus, the coded item characteristics are interpreted in terms of cognitive abilities.

To illustrate the correspondence between them, an example is provided. The item stem of Question 21 is *what does the author say about educator*. The key sentences of this question are *educators and business leaders have more in common than it may seem. Teachers want to prepare students for a successful future.* The correct answer is *they help students acquire the skills needed for their future success*. Three judges agree that the item belong to the detail explicit item type which requires a reader to recall or identify specific details that are directly stated in a passage because the unfamiliar information is explained fairly well, and the perceived complexity of the grammar is simple. The rating data are organized with variables as the measurement objects for statistical convenience, and high inter-rater reliabilities are obtained for propositional content (.67), focus constructions (.81), cohesion (.83), negations (.92), rhetorical organization (.81), and item types (.79). Inconsistencies in the variable of item type are discussed and resolved. The initial Q-matrix is built with the raters' judgments on the item types, and then refined by analyzing data derived from the students' think-aloud protocols.

The verbal data analysis shows that although the majority of test takers' responses align with the expected ones, there are notable variations. For example, experts and students utilize different reading skills to solve reading tasks in the gap-filling section. While experts use semantic clues to complete the tasks, students report relying on both semantic cues and grammatical knowledge. They first identify the part of speech of the blank and then select the correct answer from corresponding words with the same part of speech. For example, in Question 8, participants determine that an adjective is needed because the blank is situated between *a* and *knife*. As all of the available options for this blank are adjectives, they ultimately select *tiny* because the subject of the text is ants and the knife is used to remove the hair of the ants. The knife is likely small based on prior knowledge. Hence, the students' verbal data demonstrate that syntactic knowledge is a crucial reading skill required by the reading section of CET-4. Another example of disagreement between the experts and students is in Question 12, which asks for matching the statement *according to one study, students' academic performance is not the only decisive factor of their stress responses* with one of the available paragraphs. Participants report that they do not conclude the comparison study before the key sentence *this was an exciting result because it showed that the body's stress response are not determined solely by one's grades*, instead, they locate the key sentence through recognizing the overlapping phrases between two sentences, i.e., *stress responses*, and matching the meaning of two sentences. In contrast, the experts assign both summarizing the results of the study and locating specific information as the primary required skills. Besides, participants indicate that they need to infer the meaning of *academic*. Thus, three skills are assigned to Question 12, namely, summarizing the main idea, locating specific information, and inferring word meaning with contextual information. Detailed explanations of the other changes are not provided because of space limitations. The principle of the changes is that when there is a conflict between the experts' and students' opinions, both are retained and submitted to the GDINA package for evaluation. Q1 is revised as Q2 in this way (Table 1). New additions are labeled with "#".

Table 1. Q-Matrix Revised Based on Verbal Reports (Q2)

Item	Attribute1	A2	A3#	A4	A5	A6	Skill
1	0	1	1#	0	0	0	2
2	1	0	1#	0	0	0	2
3	0	1	1#	0	0	0	2
4	1	0	1#	0	0	0	2
5	0	1	1#	0	0	0	2
6	0	1	1#	0	0	0	2
7	1	0	1#	0	0	0	2
8	0	1	1#	0	0	0	2
9	1	0	1#	0	0	0	2
10	1	0	1#	0	0	0	2
11	0	0	0	1	0	0	1
12	1#	0	0	1#	0	1	3
13	0	0	0	1#	0	1	2

14	0	0	0	1	0	0	1
15	1#	0	0	1	0	1	3
16	1#	0	0	1	0	1#	3
17	0	0	0	1	0	1#	2
18	0	0	0	1	0	0	1
19	0	0	1#	1	0	1#	3
20	0	0	0	1	0	0	1
21	0	0	0	1	0	0	1
22	0	0	0	1#	1	0	2
23	0	0	1#	0	1	1#	3
24	0	0	0	1#	1	0	2
25	0	0	0	1	1#	0	2
26	0	0	0	0	1	0	1
27	0	0	1	1	0	0	2
28	0	0	0	0	1	1	2
29	0	0	0	0	1	0	1
30	0	0	0	1	0	0	1

Note. A1 = inferring word meanings from contextual clues, A2 = deducing word meanings with background knowledge, A3 = recognizing syntactic elements and discourse markers, A4 = locating and understanding specific information, A5 = paraphrasing implicit information, A6 = summarizing the main ideas.

After revising the Q-matrix based on verbal reports, it is refined with the GDINA package. The results show that the matrix is well-suited to the data, thereby supporting the use of Q2 as the final Q-matrix. The assessment instrument consists of 30 items, with 8 items testing only one skill, 17 items testing two skills, and 5 items testing three skills. Once the Q-matrix is decided, the next step is to select the most appropriate model from six rival CDMs, namely, G-DINA, DINA, DINO, R-RUM, A-CDM, and LLM, through both relative and absolute fitness measures. The relative fit indicators for the models are presented in Table 2. There is no cut-off value for these indicators. It is better to use the model that produces lower values.

Table 2. Statistical Results of Relative Test Fit

	DINA	DINO	R-RUM	LLM	A-CDM	G-DINA
Deviance	50182	52174	51026	53217	52631	49721
AIC	50722	52176	51724	52273	53628	49411
BIC	50260	51062	52018	52126	52429	49677

As shown in Table 3, G-DINA has the lowest values of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Deviance. Therefore, its fitness is better than the other rival models. To further evaluate the fit of G-DINA itself to the data, its absolute fitness is computed using the transformed correlation (r), proportion correct (p), and log-odds ratio (l) statistics provided by the G-DINA package. To evaluate the model-data fit, corresponding *p-values* and adjusted p-values of the maximum Z-score of each statistic are assessed. The null hypothesis of the test is that the model fits the data. A larger *p-value* and *adj. p-value* indicate stronger evidence to retain the null hypothesis.

Table 3. Statistical Results of Absolute Test Fit

	mean	max	max(z)	p-value	adj.p-value
r	.0008	.0034	.2061	.8412	1
p	.0012	.0064	.2813	.7618	.9
l	.0009	.0036	.2104	.8325	1

Note. *p*-value and *adj.p*-value are associated with max*(z)*; *adj.p*-values are based on the Bonferroni method.

As shown in Table 5, the *adj.p*-values of maximum *z*-scores of *r*, *p*, and *l* are larger than the significance level 0.05. These findings retain the null hypothesis, indicating that the G-DINA itself fits the data. Therefore, based on both relative and absolute measures, it can be concluded that G-DINA is the most appropriate model for acquiring the diagnostic information in the subsequent analysis.

4. Results

4.1 Diagnostic information

The RQ 1 aims to identify the specific abilities required for successful completion of the CET-4 (College English Test - Level 4) reading tasks. By understanding the underlying abilities necessary for achieving proficiency in CET-4 reading, educators and curriculum developers can tailor instructional strategies to enhance students' reading skills effectively (Jang et al., 2013; Kim, 2015). Specifically, in this study, the six reading skills are identified: inferring word meaning from context, deducing word meaning out of context, syntactic knowledge, locating and understanding specific information, making inferences, and summarizing major ideas. Descriptions for them are listed in Table 4. Some of the descriptions adopt the explanations of Jang (2009) for the primary reading skills involved in the TOEFL reading section.

Table 4. Reading Skills Involved in the Reading Section of CET-4

Skill	Description
Inferring word meaning from context	Deducing word meanings by using collocation boxes and contextual clues
Deducing word meaning out of context	Determining word meaning with background knowledge
Recognizing syntactic elements and discourse markers	Using syntactic knowledge, such as parts of speech, sentence components, and discourse structure markers, to understand words or complex sentences
Locating and understanding specific information	Quickly locating and understanding specific information through skimming and scanning
Making inferences	Paraphrasing implicit information from explicit information
Summarizing major ideas	Analyzing the relative importance of the text's information by separating its main ideas from its supporting information and identifying its significant contrast.

As shown in Table 4, using context clues or background knowledge to infer word meanings is a vocabulary skill involved in the reading task of CET-4. Deliberately using syntactic knowledge, such as, parts of speech, sentence structures, and discourse markers, is also an effective strategy for comprehending complex sentences or tasks. Locating and understanding specific information involves rapid reading processes, where skimming enables readers to develop a basic understanding of the text, while scanning is useful for identifying specific graphic forms. The combination of these skills allows readers to efficiently search for specific information. Making inferences is an advanced skill that requires readers to infer the deep meanings from literal information or connect prior experiences to comprehend the text information. Summarizing main ideas requires readers to synthesize information from various parts of a text. Readers must distinguish main ideas from supporting information, and then create their own coherent organizational frame. This skill represents a more complicated and complex task for readers than other skills.

After reading skills are retrieved, diagnostic information is interpreted at both the population and individual level. Table 5 and Table 6 present the respective results. In Table 5, overall examinees' reading performance in the CET-4 reading section is evaluated by analyzing their mastery probabilities of the six reading attributes.

Table 5. Overall Examinees' Attribute Mastery Probability

CET-4 Reading Attributes	Attribute Mastery Probability
Deducing word meaning from context	.581
Deducing word meaning out of context	.512
Recognizing syntactic elements and discourse markers	.547
Locating and understanding specific information	.739
Making inferences	.628
Summarizing major ideas	.462

As shown in Table 5, the population's attribute mastery probabilities range from .462 to .739. That is, 58.1% of the examinees have mastered deducing word meaning from context; 51.2 % have mastered deducing word meaning out of context; 54.7% could recognize syntactic elements and discourse markers; 73.9% could process explicit information; 62.8% could make inferences; and 46.2% have mastered summarizing major ideas. Processing explicit information is the easiest CET-4 reading skill in that 73.9% of the examinees have mastered it. Summarizing major ideas is the most difficult CET-4 reading attribute because it has the lowest mastery probability. In order to understand individuals' skill mastery probabilities, attribute mastery patterns are assigned to each test-taker (see Table 8), which describes an individual's mastery state of all skills. The mastered skills are marked as 1, while skills that are not mastered are marked as 0. Attribute mastery patterns of three participants with the same score (21 points) are shown in Table 6.

Table 6. Attribute Mastery Pattern for Individual Test-takers

Test-Taker	Score	Attribute Mastery Patterns
1	21	011111
2	21	101110
3	21	110110

Notes. The total score is 30 points.

As shown in Table 6, the three examinees have different attribute mastery patterns, although they receive the same total score. Particularly, participant 1 has mastered all the CET-4 reading attributes except A1. Participants 2 and 3 have mastered only four skills out of six skills. The comparison results show that although the students' scores are the same, their reading performance is different. For inferences from the diagnostic results to be valid, the accuracy of diagnostic results is examined from both test and attribute level accuracy. The statistical results obtained from the G-DINA package are shown in Table 7.

Table 7. Accuracy of the Diagnostic Results

A1	A2	A3	A4	A5	A6	Mean
0.752	0.792	0.942	0.988	0.944	0.83	0.875

Note. A1 = inferring word meanings from contextual clues, A2 = deducing word meanings with background knowledge, A3 = recognizing syntactic elements and discourse markers, A4 = understanding specific information through skimming and scanning, A5 = paraphrasing implicit information, A6 = summarizing the main ideas.

As shown in Table 7, all accuracy measurement has high values (≥.70). The mean value is .875, which means overall diagnostic results are reliable and can be used to guide classroom teaching.

4.2 Method of integrating diagnostic assessment of English reading into teaching

The RQ 2 focuses on evaluating the effectiveness of a proposed instruction framework that integrates diagnostic assessment into the language learning classroom. This framework aims to address learners' weaknesses by providing targeted remedial instruction based on individual diagnostic profiles. By understanding the extent to which this instructional approach effectively improves learners' weaknesses, educators and curriculum designers can gain valuable insights into the potential benefits and limitations of incorporating diagnostic assessment into the learning process and curriculum. This knowledge can inform the development of evidence-based instructional strategies that promote more effective language learning outcomes.

4.2.1 Proposing a new teaching framework

Assuming that fine-grained diagnostic information has a positive impact on practical consumers, a teaching framework and experiment are developed and implemented to test the hypothesis within the context of classroom teaching when English is a foreign language. The new instruction framework is proposed based on the *Strategy Instruction Framework* led by Chamot, et al. (1999), and is presented in Figure 3. Since its proposal, the framework has been investigated in numerous studies for its effectiveness in developing EFL students' reading abilities (e.g., Albashtawi et al., 2016, Cubukcu 2008, Karbalaei 2011, Prakoso et al., 2016, Ravari 2014, Takallou, 2011).

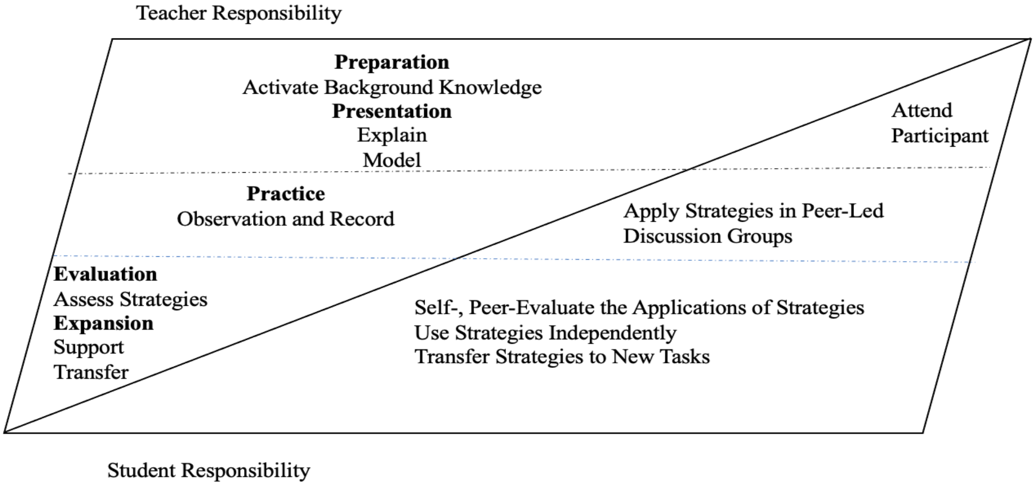
Teacher Responsibility
Preparation
Activate Background Knowledge
Presentation
Explain
Model
Attend
Participant
Practice
Observation and Record
Apply Strategies in Peer-Led Discussion Groups
Evaluation
Assess Strategies
Expansion
Support
Transfer
Self-, Peer-Evaluate the Applications of Strategies
Use Strategies Independently
Transfer Strategies to New Tasks
Student Responsibility

Figure 2. A New Strategy Instruction Framework

As shown in Figure 2, while the overall structure of this framework is consistent with the framework of Chamot et al. (1999), some minor adjustments have been made to align with the reading curriculum objectives and the actual teaching situation, which involves a large number of students. The first adjustment made to the framework pertains to the presentation section, where student strategy needs are integrated with reading course objectives, as suggested by Chamot et al. (1999). In this way, explicit strategy instruction is conducted in a meaningful context, instead of being isolated from reading instruction. To achieve this goal, this study follows Halliday's (1978) reading purpose model, which includes ideational, textual, and interpersonal meanings. The ideational meaning interprets the text in terms of what the text is about, while textual and interpersonal meanings represent how the text is structured and functioned as a communicative unit, and what the social or relational aspects of the text are, respectively. The presentation part is conducted in the form of dialogue, which facilitates meaningful interaction between the teacher and students.

The second change in the framework concerns the practice stage, where the class is divided into small groups based on complementary skills, to promote the development of student-teachers and peer-to-peer learning. Given the large number of students in the class, it is not feasible for one adult-teacher to provide individual attention to all students. If student-teachers are developed, group members can mutually teach each other, and the framework can be applied to a large classroom setting.

In the planning phase, the teacher conducts a comprehensive analysis of the students' needs and teaching objectives to identify strategies to be taught, the method of presentation, the appropriate activities for practice, and the follow-up exercises to promote the internalization of skills. In the presentation phase, the teacher thoroughly explains and models the plan-oriented strategies with the aim of fostering learners' reading and strategy awareness, involving helping learners understand how to use abstract skills to comprehend ideational, textual, and interpersonal dimensions of the text. Students actively participate in classroom activities and take notes carefully during this period. The practice stage is designed to guide learners in using the strategies they have learned to complete reading tasks. The activity is conducted in group-level form, and the members' reading skills are compensatory, as indicated by their diagnostic information acquired in Phase 1. The reading tasks for all groups are developed based on ideational, textual, and interpersonal meanings. Sample tasks are presented in Table 8. However, each member has a unique reading task tailored to remedy their individual reading weaknesses. For example, the reading task of the student who has not yet mastered the skill of summarizing the main idea is to conclude the central message or theme. The teacher's responsibility in this stage is to observe and record whether the student-teachers give correct prompts and whether students with reading weaknesses understand their cues and respond accordingly. The final stage involves helping learners reflect on their use of strategies and evaluate their effectiveness in achieving their learning goals. Extensive reading materials tailored to learners' weaknesses are recommended for further practice.

Table 8. Examples of reading tasks presented to the learners

No.	Perspective	Question
1	Writing purpose	What's the author's main purpose of writing this passage?
2	Author's key questions in mind	What question is the author trying to answer?
3	Fact, experiences, data used	What's the most important information is the author using?
4	Author's key conclusion	What's the author's important conclusions or inference?
5	Author's key/basic concept	What are the author's basic concepts?
6	Intended readers	To whom does the author write the passage?

7	Positive consequences	What consequences are likely to follow if people take the author's line of reasoning seriously?
8	Negative consequences	What consequences are likely to follow if people ignore the author's reasoning?
9	Author's main point of view	What's the main point of view presented in the passage?
10	Possible source	Where can you possibly read the passage?
11	Structure organization	What's the logical structure of the passage?

4.2.2 A case study based on the new instruction framework

A comparative teaching experiment is conducted to evaluate the effectiveness of the new teaching framework in addressing learners' reading weaknesses. To ensure that the only variable affecting the results is the teaching method applied, various other factors are controlled, such as teaching materials, teaching tasks, and distribution of group members. All groups are taught to interpret text from ideational, textual, and interpersonal perspectives, while only the experimental groups receive diagnostic information. The difference between the two experimental groups is that experimental group two (EG2) receives explicit guidance from both adult- and student-teachers on how to apply and utilize the specific strategy effectively, while the members in experimental group one (EG1) do not receive such guidance. Instead, they rely on repetitive drills to develop their own experience in applying the strategies to new tasks or situations. The sample prompts and guidance are presented in Table 9, with some questions taken from the study of Chamot et al. (1999).

Table 9. Sample questions for reader's response

No.	Question
1	Does this story remind you of anything in your own life? Did you use background knowledge while you were reading?
2	Are there any new words in this story? What strategy can help you figure out what they mean?
3	Did you use the strategy prediction while you were reading this story? Find the places in the story where you used this strategy.
4	Can you think of a different ending for the story?
5	What is your favorite part of the story? What strategy or strategies did you use to understand the story?
6	Pretend you are a character other than the protagonist in the story. Tell the story from this character's point of view. What learning strategies can help you?

As shown in Table 9, reading instruction and explicit strategy instruction are integrated through the use of specific questions. For example, the first question stimulates students to activate their prior knowledge and relate it to the story, thereby enhancing their comprehension and recall of the content. This question also helps students develop their metacognitive awareness of their own learning processes. The second question prompts students to identify unfamiliar words and use various strategies to determine their meanings, such as using context clues or word parts. Consequently, this question supports the development of vocabulary skills. The third question is designed to prompt students to recognize and reflect on their use of the prediction strategy when reading. This question encourages the development of metacognitive skills by fostering students' awareness of their own thinking processes. The fourth question is intended to cultivate students' imaginative and creative abilities, as well as promote critical thinking and comprehension skills. The fifth question prompts students to identify their personal interests and preferences, and reflect on the strategies they used to comprehend the story. Lastly, the final question is designed to enhance higher-order thinking skills and encourage perspective-taking abilities, while also encouraging students to think about how they can use reading strategies to support their understanding of different perspectives in the story. In addition to in- and after-class activities, preview tasks are broken down into six parts according to diagnostic results and instructional objectives to target the participants' weaknesses throughout the teaching process. Different preview tasks are assigned to students based on their skill mastery patterns, aiming to improve their weaknesses from the beginning.

4.2.3 Evaluating teaching effects

In Phase 2, two targeted tests are conducted to evaluate the effectiveness of the framework in improving learners' weaknesses. The tests include pre- and post-tests, with the aim of comparing the participants' performance on their identified weaknesses. The Paired-Samples T-test is used to determine whether participants' weaknesses are significantly improved before and after teaching. The null hypothesis of the Paired-Samples T-test is that there is no significant difference between the means of the pre- and post-tests. The commonly used significance level is 0.05. If the significance level is less than 0.05, the null hypothesis is rejected, indicating that participants have significantly improved their weaknesses after teaching. The statistical results of the three groups are presented in Table 10.

Table 10. Statistic results of paired-samples t-test

Group	N	Mean of pre-test	Mean of post-test	Std. deviation	t-value	Sig.(2-tailed)
EG1	12	2.08	5.17	.79	-13.47	.000
EG2	12	2.08	7.50	.99	-18.84	.000
CG	12	2.16	2.67	.90	-1.915	.082

Note. EG1 and EG2 refer to the experimental group one and experimental group two; and CG refers to the control group. The total score is 10 points.

As shown in Table 10, the mean scores of all groups improve after the teaching intervention. However, the p-values of the two experimental groups (*p1* = .000, *p2* =.000) reject the null hypothesis, while the p-value of the control group (*p* = .082) retains the null hypothesis. These results indicate that statistically significant differences are found for the experimental groups, while no significant difference is found in the control group. It can be inferred that students in the experimental groups make more progress than those in the control group. To further compare the difference between the two experimental groups, an independent sample t-test is conducted. The null hypothesis of the test is that there is no significant difference between the means of the groups. The statistical results show that when the two groups have the same pre-test level (*m* = 2.08), there is a significant difference between them when their post-test scores are compared (*t* = -5.40; *p* = .000). These findings suggest that EG2 shows more progress compared to the other group.

5. Discussion and conclusion

The current study attempts to measure the effectiveness of integrating diagnostic information into teaching to enhance learners' reading weaknesses. To achieve this goal, the researchers first identify the attributes necessary for success in the reading section of the CET-4, and conduct a diagnostic analysis to determine the examinees' strengths and weaknesses. The survey results suggest that the CET-4 reading section requires six reading skills, namely, inferring word meanings from context clues, deducing word meaning based on background knowledge, syntactic knowledge, locating and understanding specific information, making inferences, and summarizing the main idea. These skills serve as a framework for the subsequent diagnostic analysis designed to identify the strengths and weaknesses of test takers in the CET-4 reading section. The diagnostic results indicate that summarizing major ideas is the most difficult CET-4 reading attribute. This finding is inconsistent with Jang's (2009) finding, which reports that summarizing has the greatest proportion of mastery. However, this research result aligns with previous investigations into the testing constructs of reading comprehension (Jang et al., 2013, Kim 2015).

The six reading skills further demonstrate that the constructs of the reading section of the CET-4 are divisible. The divisible constructs and accurate diagnostic information generated from CDA together support the notion that the reading section of CET-4 can be an instrument for retrofitting CDA. However, the study only focuses on the reading section. Future research could investigate other sections of the test, such as writing, listening, and translation. The G-DINA model is determined to be the best model for representing the nature of English reading comprehension among the six rival models. This result aligns with previous studies (e.g., Chen and Chen 2016, Javidanmehr and Anani Sarab 2019, Mirzaei et al., 2020, Ravand and Robitzsch 2018), which suggest that the G-DINA model is appropriate for representing the cognitive processes involved in English reading comprehension. Therefore, it can be inferred that the connections between reading subskills cannot be adequately represented by cognitive diagnostic models that assume either a non-compensatory or compensatory relationship. A more flexible cognitive diagnostic model that includes both compensatory and non-compensatory relationships within the same test would be more suitable.

The second goal is to investigate the effectiveness and method of integrating diagnostic information into teaching to improve learners' weaknesses in reading. The case study indicates that integrating diagnostic information into teaching can lead to better improvement in students' weaknesses in reading. However, if learners are consciously informed about the strategies, how to express them, and where to apply them, they make more progress compared to those who have to rely on their own repetitive drills for summarization. This finding suggests that explicit strategy instruction with expert scaffolding is more effective. It proves that although diagnostic feedback from CDA plays a bridging role between assessment and instruction, its impact on instruction is influenced by the methods of remediation.

This study has implications for reading instruction and assessment, highlighting the need to incorporate diagnostic information into teaching to effectively target learners' weaknesses. By identifying the specific areas of difficulty experienced by students, educators can develop tailored interventions that address individual needs, thereby enhancing their chances of success in reading comprehension. However, it must be admitted that this experiment also has limitations. While retrofitting CDA has been favored in previous studies, it has inherent flaws. The interference options of diagnostic assessments are purposely designed to gather additional feedback, limiting the diagnostic ability of retrofitting CDA. Additionally, the current study did not assess the long-term effectiveness of CDA for learning, nor did it consider multiple test occasions. Further research in these areas would provide valuable insights into the full potential of diagnostic assessment in language learning and instruction.

References

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

Alderson, J. C. 2005. *Diagnosing foreign language proficiency: The interface between learning and assessment.* London: Continuum.

Albashtawi, A. H., P. Jaganathan. and M. K. M. Singh. 2016. Linguistic knowledge aspects in academic reading: Challenges and deployed strategies by English-major undergraduates at a Jordanian institution of higher education. *Higher Education Studies* 6(3), 61-71.

Bachman, L. F., F. Davidson., K. Ryan. and I. C. Choi. 1995. *An investigation into the comparability of two tests of English as a foreign language*. Cambridge: Cambridge University Press.

Bachman, L. F. and A. S. Palmer. 1996. *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Bernhardt, E. 2005. Progress and procrastination in second language reading. *Annual Review of Applied Linguistics* 25, 133-150.

Buck, G., K. Tatsuoka. and I. Kostin. 1997. The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning* 47(3), 423-466.

Carr, N. T. 2006. The factor structure of test task characteristics and examinee performance. *Language Testing* 23(3), 269-289.

Celce-Murcia, M. and D. Larsen-Freeman. 1999. The grammar book: An ESL/EFL teacher's course. Boston, MA: Heinle and Heinle.

Chamot, A. U., S. Barnhardt., P. B. El-Dinary. and J. Robbins. 1999. *The learning strategies handbook*. White Plains, NY: Pearson Longman.

Chen, H. L. and J. S. Chen. 2016. Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly* 13(3), 218-230.

Coady, J. 1979. A psycholinguistic model of the ESL reader. In R. Mackay, B. Barkman and R. R. Jordan, eds., *Reading in a second language,* 5-12. Rowley, MA: Newbury House.

Cubukcu, F. 2008. Enhancing vocabulary development and reading comprehension through metacognitive strategies. *Issues in Educational Research* 18(1), 1-11.

Du, W. and X. Ma. 2021. Probing what's behind the test score: Application of multi-CDM to diagnose EFL learners' reading performance. *Reading and Writing* 34(6), 1441-1466.

Fan, T., J. Song. and Z. Guan. 2021. Integrating diagnostic assessment into curriculum: A theoretical framework and teaching practices. *Language Testing in Asia* 11(2), 1-23.

Freedle, R. and I. Kostin. 1993. The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing* 10(2), 133-170.

Freedle, R. and I. Kostin. 1999. Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing* 16(1), 2-32.

Grabe, W. 1991. Current developments in second language reading research. *TESOL Quarterly* 25(3), 375-406.

Halliday, M. A. K. 1978. *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Armold.

Jang, E. E. 2005. *A Validity Narrative: Effects of Reading Skills Diagnosis on Teaching and Learning in the Context of NG TOEFL.* Doctoral dissertation, University of Illinois at Urbana-Champaign, Champaign, IL, USA.

Jang, E. E. 2009. Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing* 26(1), 31-73.

Jang, E. E., M. Dunlop., M. Wagner., Y. H. Kim. and Z, M. Gu. 2013. Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. *Language Learning* 63(3), 400-436.

Javidanmehr, Z. and M. R. Anani Sarab. 2019. Retrofitting non-diagnostic reading comprehension assessment: Application of the G-DINA model to a high stakes reading comprehension test. *Language Assessment Quarterly* 16(3), 294-311.

Karbalaei, A. 2011. Assessing reading strategy training based on CALLA model in EFL and ESL context. *Ikala, Revista De Lenguaje Y Cultura* 16(27), 167–187.

Kim, A. Y. 2015. Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing* 32(2), 227-258.

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

Lee, Y. W. and Y. Sawaki. 2009b. Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly* 6(3), 172-189.

Li, H. L., C. V. Hunter and P. W. Lei. 2016. The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing* 33(3), 391-409.

Lunzer, E., M. Waite. And T. Dolan. 1979. Comprehension and comprehension tests. In E. Lunzer and K. Gardner, eds., *The Effective Use of Reading,* 37-71. London: Heinemann Educational Books.

McNeil, L. 2012. Extending the compensatory model of second language reading. *System* 40(1), 64-76.

Mirzaei, A., M. H. Vincheh. and M. Hashemian. 2020. Retrofitting the IELTS reading section with a general cognitive diagnostic model in an Iranian EAP context. *Studies in Educational Evaluation* 64, 1-10.

Prakoso, G. H., B. Setiyadi. and Y. Yufrizal. 2016. Modified CALLA to improve students' cognitive reading strategy and reading comprehension. *UNILA Journal of English Teaching* 5(1), 1–14.

Ravand, H. and A. Robitzsch. 2018. Cognitive diagnostic model of best choice: A study of reading comprehension. *Educational Psychology* 38(10), 1255-1277.

Ravari, M. R. 2014. The effect of reading strategies training on enhancing reading comprehension and reading motivation among Iranian EFL Learners in Kerman. *International Journal of Language Learning and Applied Linguistics World* 6(1), 409-421.

Sawaki, Y., H. J. Kim. and C. Gentile. 2009. Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly* 6(3), 190-209.

Takallou, F. 2011. The effect of metacognitive strategy instruction on EFL learners' reading comprehension performance and metacognitive awareness. *Asian EFL Journal* 13(1), 275-320.

Toprak, T. E. and A. Cakir. 2021. Examining the L2 reading comprehension ability of adult ELLs: Developing a diagnostic test within the cognitive diagnostic assessment framework. *Language Testing* 38(1), 106-131.

Yamaguchi, K. and K. Okada. 2018. Comparison among cognitive diagnostic models for TIMSS 2007 fourth grade mathematics assessment. *PloS one* 13(2).

Yun, J. 2017. *Investigating Structures of Reading Comprehension Attributes at Different Proficiency Levels: Applying Cognitive Diagnosis Models and Factor Analyses.* Doctoral dissertation, Florida State University, Tallahassee, FL, USA.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary