

RESEARCH

Open Access



Integrating diagnostic assessment into curriculum: a theoretical framework and teaching practices

Tingting Fan^{1*} , Jieqing Song² and Zheshu Guan³

* Correspondence: tingtingfan121@hotmail.com

¹College of Foreign Languages, Nanjing University of Aeronautics and Astronautics, 29 Jiangjun Ave., Nanjing 211106, People's Republic of China

Full list of author information is available at the end of the article

Abstract

Currently, much research on cognitive diagnostic assessment (CDA) focuses on the development of statistical models estimating individual students' attribute profiles. However, little is known about how to communicate model-generated statistical results to stakeholders, and how to translate formatively diagnostic information into teaching practices. This study proposed an integrative framework of diagnosis connecting CDA to feedback and remediation and, meanwhile, demonstrated empirically the application of the framework in an English as a Foreign Language (EFL) context. Particularly, the empirical study presented procedures of integrating diagnostic assessment to EFL reading curriculum through four phases of planning, framing, implementing, and reflecting. The results show that these procedures, influenced by the teacher's orientation to diagnostic assessment and approach to EFL teaching, affected students' perceptions of diagnostic assessment, their attitudes toward remedial instructions, as well as their learning outcomes on EFL reading. The results also provide evidence to the effectiveness of the integrative framework proposed in this study, showing that the framework could serve as practical guidance to the implementation and use of diagnostic assessment in the classroom. Overall, this study indicates that the diagnostic approach is a more effective way to provide instructionally useful information than other test and assessment approaches that do not differentiate strengths and weaknesses among students with the same total score.

Keywords: Cognitive diagnostic assessment (CDA), Diagnostic score report, Diagnosis-based remediation, English as a Foreign Language (EFL) reading curriculum

Introduction

Cognitive diagnostic assessment (CDA) refers to a set of cognitively grounded, diagnostic procedures attempting to pinpoint students' strengths and weaknesses in relation to their knowledge structures and processing skills (referred to as attributes) in the target domain (Lee & Sawaki, 2009).

In contrast to standardized testing that measures the educational level of students on a broad scale and reports the measurement results in a summative manner, CDA, rooted in the cognitive psychology of problem solving (Snow & Lohman, 1989), makes explicit the test developer's assumptions about the fine-grained attributes a test taker would use in a content domain, how the attributes develop, and how test takers of higher proficiency differ from those of lower proficiency (Mislevy et al., 2003). Over the past decade, the profusion of CDA research in the field of language testing has led to a more comprehensive conceptualization of the construct, and more advanced techniques of operationalization for language testers and language teachers. Although the importance of CDA in language studies has been widely recognized (Alderson et al., 2015; von Davier & Lee, 2019), how CDA could be used to inform language teaching practices in the classroom remains unexplored. Previous research suggests that the fine-grained information generated from CDA could help promote students' learning and guide further instruction (e.g., Lee, 2015). However, only little is known about how to communicate the model-generated statistical results to stakeholders, and how to translate formatively diagnostic information into teaching practices both theoretically and empirically (Mislevy et al., 2003; Stout, Henson, DiBello, & Shear, 2019).

To respond to the paucity of research in this respect, we proposed a framework of cognitive diagnosis integrating feedback, remedial teaching, and learning to diagnostic modeling procedures in the CDA system and reported an empirical study demonstrating the application of the proposed framework, particularly the diagnosis-based remediation specified in the framework, in an English as a Foreign Language (EFL) context.

Cognitive diagnostic assessment and its application on EFL teaching: a brief overview

The notion of diagnosis can be traced back to the 1950s (e.g., Cronbach & Meehl, 1955). On the basis of it, CDA emerged and discussed extensively in the areas of education, psychometrics, and applied linguistics, including books and special journal issues (e.g., Alderson et al., 2015; Lee, 2015; Lee & Sawaki, 2009; Leighton & Gierl, 2007; von Davier & Lee, 2019). As noted by previous studies, CDA differs from traditional test and assessment approaches in at least three aspects. First, it is developed from cognitive models of learning backed by empirical evidence of human information processing, which is rarely employed by either large-scale tests or classroom-based assessments, irrespective of the well-established assessment and psychometric practices that followed (Mislevy et al., 2003; Snow & Lohman, 1989; von Davier & Lee, 2019; Wang & Gierl, 2011). Second, different from traditional testing procedures ranking examinees along a proficiency continuum, CDA estimates students' abilities as a composite of fine-grained attributes, thereby facilitating unequivocal and multi-faceted measurement of students' learning (Chen & de la Torre, 2014; Leighton & Gierl, 2007). Last, providing

stakeholders with detailed diagnostic information, CDA can bring about a positive washback effect. Washback describes the influence of testing on teaching and learning (Tsagari & Cheng, 2017). Although high-stakes tests of nondiagnostic purposes can have enormous washback, this effect might be rather “general, systemic, complex, and difficult to trace” (Lee, 2015, p.8). In contrast, the impact produced by CDA is more specific and individualized. The resulting information can be used directly in the classroom to inform the design of instruction, improve students’ learning, and guide the development and reform of the curriculum.

Much of the research on CDA has hitherto been centered on the first two aspects, concerning selection and evaluation of cognitive models to define attributes and their relationships (e.g., Wang, Song, Chen, Meng, & Ding, 2015), development and extension of diagnostic classification models (DCMs) to estimate individual students’ attribute profiles from the response data of the assessment (e.g., de la Torre, 2011; Zhan, Ma, Jiao, & Ding, 2020), and generation and comparison of statistical indices to examine model robustness in tracking the current state of student knowledge and the process of knowledge progression in real-world applications (e.g., Johnson & Sinharay, 2018; Ma, Iaconangelo, & de la Torre, 2016). In the area of language testing, theoretical elaborations of CDA focus primarily on principles of language diagnosis (e.g., Harding, Alderson, & Brunfaut, 2015), and procedures of applying and verifying DCMs in language tests (e.g., Lee & Sawaki, 2009). Although a number of theoretical elaborations include diagnostic feedback and diagnosis-based remediation in their framework, these components are viewed as complements to diagnosis, not fully functioning in the CDA system. The cognitive diagnosis model of Lee and Sawaki (2009) is such an example: It divides CDA into definition of attributes, Q-matrix construction, data analysis, and score reporting. The first three components specifying the cognitive model and statistical procedures are considered as the core of diagnosis. By taking a “minimalist approach” (Lee, 2015, p.14) to remediation, the model does not go further than simply providing feedback about the examinees’ identified strengths and deficiencies in the attributes they have or have not mastered sufficiently, which might negate or weaken the mission of CDA to impact and promote subsequent teaching and learning positively.

In addition to theoretical elaborations, language testers have also applied empirically the CDA procedures in an array of reading and listening tests under different DCMs including the research from Jang (2009), Kim (2015), Chen and Chen (2016), Yi (2017), and Fan and Yan (2020). All of the above studies adopted the retrofitting approach fitting a DCM or several DCMs for cognitive diagnosis to the test data from assessments not originally designed for diagnostic purposes, such as the PISA, the LanguEdge, and the TOEFL. The emphasis was therefore put on examining the robustness of the selected DCM to the test data. For example, Fan and Yan (2020) adopted a model selection method to compare the four DCMs (including the G-DINA, the DINA, the LLM, and the RRUM) to validate the Q-matrix and fit test takers’ responses to the National Matriculation English Test (NMET) in China. After model-data fit comparisons with both internal and external measures, the RRUM was found to be the most adequate model among the four to fit the test data, and used to generate diagnostic results at both group and individual levels of high school EFL learners. Although the study produced fine-grained, diagnostic information at both group and individual levels, it did

not go further to use the diagnostic results in classroom settings. Questions remain as to how to translate the model-generated, statistically sophisticated information into teaching practices.

To address these gaps, we propose, in the following session, a framework of cognitive diagnosis aiming at integrating feedback, remedial teaching and learning to diagnostic modeling procedures in the CDA system. The framework builds on existing approaches which aim at modeling language diagnosis (e.g., Fan & Zeng, 2016; Harding et al., 2015; Lee, 2015; Lee & Sawaki, 2009), feedback (e.g., Zenisky & Hambleton, 2012), and classroom-based assessment (e.g., Hill & McNamara, 2011; Wang & Li, 2019) theoretically, as well as empirical studies that have applied CDA in the diagnosis of students' language abilities.

The integrative framework of diagnosis and specification of its components

The integrative framework of diagnosis consists of four major components, including (1) designing and conducting diagnostic assessments, (2) providing diagnostic feedback to stakeholders, (3) implementing remedial instruction and learning, and (4) validating the integrative system of diagnosis. The components that constitute the framework are visualized in Fig. 1 and explained in detail below.

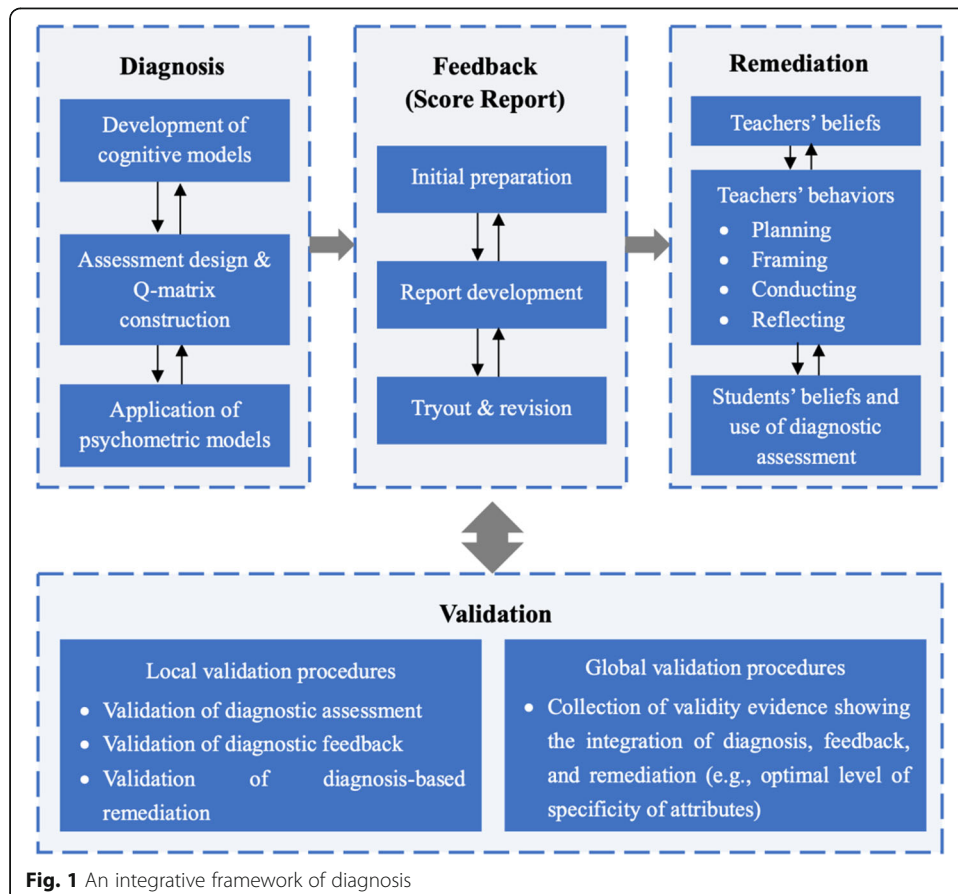


Fig. 1 An integrative framework of diagnosis

Diagnosis

Diagnosis is the first component of the framework. The purpose of diagnosis is to pinpoint individual students' strengths and, more importantly, weaknesses on the attributes that the student has not yet fully mastered. Diagnosis can be realized through developing diagnostic assessments from the ground up by following three steps, including constructing the cognitive model, designing assessment instruments and the associated Q-matrix, and applying psychometric models for data analysis (de la Torre & Minchen, 2014). Serving as the basis for principled assessment design, the cognitive model specifies the definition of attributes to be measured as well as the inter-relationship among these attributes. Oftentimes, the attributes are extracted from relevant cognitive theories or proficiency scales conceptualizing students' abilities at distinct reference levels in the target domain and validated using empirical data such as think-aloud protocols presenting students' test-taking processes (Jang & Wagner, 2013; Leighton & Gierl, 2007). The validation procedure is pivotal if the cognitive model involves the hierarchical relationship in which psychological ordering is assumed among the attributes. The results of validation and, when necessary, subsequent revisions of the hierarchy would result in a more accurate and valid cognitive model for assessment development and analysis (von Davier & Haberman, 2014; Wang & Gierl, 2011).

With the cognitive model defining the construct, diagnostic assessment instruments, for example, assessment tasks and questionnaires, can be developed, along with the Q-matrix specified to map the attributes to individual assessment items. Among a variety of item formats, discrete-point items assessing a single construct in each item are preferred by educational researchers for they allow efficient identification of weaknesses at fine-grained levels, exploration of root causes of the weaknesses, and straightforward interpretation of diagnostic results (Alderson et al., 2015; Fan & Zeng, 2016; Xie, 2019). The construction of assessment items is followed by procedures for collecting and scoring students' response data, which are subject to statistical analysis in a psychometric model. The psychometric model captures a set of statistical procedures applied on the response data, including selection of appropriate model for data analysis, generation of expected attribute response patterns, evaluation of the model-data fit, and calculation of probabilities of mastery on each attribute for individual students (Wang & Gierl, 2011). Ideally, the psychometric model would be applied iteratively to achieve an adequate fit of the selected DCM to data.

Feedback/score report

The second component is diagnostic feedback designed to describe and summarize the results of diagnosis to a variety of stakeholders, including teachers, students, policy-makers, and parents (Lee, 2015). It usually takes the form of a score report that turns statistically sophisticated assessment data into actionable information through both qualitative illustrations (e.g., verbal descriptions of attribute mastery levels) and quantitative presentations (e.g., tables and graphs; Rankin, 2016). Adapting the score report development model of Zenisky and Hambleton (2012), the feedback component in the integrative framework of diagnosis sequences through three stages of initial preparation, report development, tryout and revision. Preparation requires defining the purpose of the score report, identifying the target audience, and carrying out needs

assessments to the audience. After careful preparation, the prototype report is produced and reviewed internally within the expert panel responsible for the development and revision of the report. At this stage, questions concerning the content and design of the diagnostic score report should be addressed, including but not limited to: How does one best transform psychologically complicated information unique to CDA to either qualitative or quantitative presentations in the report so as to maximize accurate interpretation of the target audience? Is the diagnostic information conveyed through a static report (online or paper), a series of reports, or a dynamic online reporting system? How is language structured to express ideas and values that conform to the socio-cultural context from which they are produced (Hambleton & Zenisky, 2013; Roberts, 2012; Roberts, Gotch, & Lester, 2018)?

Thereafter, the prototype report should be trialed in field testing where audience's opinions and understanding of the report contents are investigated via controlled studies. The data are then analyzed to inform the revision of the report, which might be subject to a new round of field testing. The tryout and revision session plays a fundamental role in the development of score report because a number of existent studies have shown that many teachers, students, and parents have trouble interpreting and making appropriate use of score reports (Brown, O'Leary, & Hattie, 2019; Hambleton & Zenisky, 2013; Tannenbaum, 2019; Zapata-Rivera & Katz, 2014). The interpretability of score reports, viewed as an integral part of validity (O'Leary, Hattie, & Griffin, 2017), is more critical in the diagnostic context because CDA has a strong mission of facilitating future instruction and learning, which must be realized through careful design of the score report communicating strengths, weaknesses, and root causes of weaknesses to teachers and students effectively (Lee, 2015). Otherwise, subsequent activities cannot be acted upon to remediate the identified weaknesses and promote the overall learning potential. In addition to the score report development and revision, ongoing monitoring and maintenance are also in need to ascertain that the report remains useful and functions as intended (Zenisky & Hambleton, 2012).

Remediation

The third component is concerned with remediation (or treatment/intervention), which refers to a set of teaching and learning activities aiming at strengthening the identified weaknesses of students on the attributes in the target domain (Lee, 2015). Within this component, teachers are expected to integrate diagnostic assessment results into the curriculum to create meaningful content and concrete pedagogy. Adapting the classroom-based assessment framework of Hill and McNamara (2011) and the procedures for integrating teaching-learning-assessment of Wang and Li (2019), the remediation is centered on teachers' behaviors, comprising four processes of planning, framing, conducting, and reflecting. Planning emphasizes the decision-making process of teachers to compare and select materials, activities, and instructional approaches aligned with learning goals and diagnostic results. This category specifies information about the type and nature of planned instructional tasks for remediation and the relationship of diagnostic results to instruction. Framing investigates the extent to which the remediation is made explicit to students, along with the criteria to evaluate their progress. The purpose of this category is to share instructional intentions to students

so that they can have an appropriate orientation to the instructional tasks. Conducting specifies the process to implement the instructional plans in the classroom. At this stage, a majority of instructional activities are explicit, formal, and well-designed before the class. However, unplanned, instruction-embedded activities may also occur periodically in the classroom (Wang & Li, 2019). Reflecting captures teachers' reflections on the instructional practices they have implemented in the classroom, which is reported to have positive effects on both the academic growth of teachers and subsequent teaching adjustments (Wang & Li, 2019). Active engagement in self-reflection might help teachers exhibit an increased interest in the development of innovative teaching strategies and collaborative practices to facilitate the use of diagnostic information for improving students' learning.

Besides the behaviors of teachers, two additional elements are listed in the framework for their close relationships with remediation: teachers' beliefs, and students' beliefs and uses of the diagnostic assessment. Teachers' beliefs refer to the theory of classroom teachers about the subject, the curriculum, pedagogic principles, and assessment practices underlying their teaching practices (Hill & McNamara, 2011). In the diagnostic context, teachers' perceptions about CDA and the associated feedback are reported to have implications on each process of planning, framing, conducting, and reflecting (Doe, 2015). Therefore, teachers are worth to be investigated on their beliefs about the diagnostic assessment and provided, when necessary, with education that would equip them with the knowledge and skills to integrate the diagnostic assessment in the classroom effectively.

Students' beliefs and use of diagnostic assessment, closely related to the attitudes of learners about how the assessment is conducted, interpreted, and used in the classroom, can be investigated in teacher-to-student interactions, student-to-teacher interactions, and student-to-student interactions (Rea-Dickins, 2006). These interactions highlight discrepancies between teacher intentions and student understandings, and provide access to how students develop self-help remedial activities independent of the remediation offered by their teachers.

Validation

As the last component, validation is concerned with the process of collecting relevant and appropriate evidence to support the intended interpretation and use of diagnostic assessment scores (O'Leary et al., 2017). Generally, the validity evidence comprises two types, local and global. Local evidence is gathered within each of the three aforementioned components. In the diagnostic phase, technical evidence in support of the quality of the CDA and associated DCMs include, but not limited to the following aspects: an analysis of relationships between the content of the diagnostic assessment and the construct (i.e., the cognitive model) it is intended to measure; a comparison between examinees' actual processes of responses and the response pattern intended by the test developers; and an examination of the fit of the selected DCM to the assessment response data (Fan, 2020). In the feedback phase, validity evidence can be collected, via surveys to assessment and score report users, to justify the interpretability of the diagnostic score report; that is, the actual interpretations made and uses enacted by users are aligned with the interpretations and use of the score report intended by developers.

In the same vein, the effectiveness of remedial teaching and learning could be evaluated through such methods as interviews, classroom observations, and experimental studies in real classrooms, although only a few existent research mentioned this part.

On the other hand, global evidence emphasizes the link and integration of the three components of diagnosis, feedback, and remediation. For example, in defining the cognitive model of diagnosis, one of the core problems is to determine the level of specificity of attributes in the content domain. However, as specificity is a property that is assessed along a continuum instead of a dichotomy, it is not easy to decide on the optimal level that is sufficient for diagnosis and feedback (Lee, 2015). Therefore, evidence concerning the level of specificity should be collected within the diagnostic component as well as from the other two components of the framework, so as to answer the following questions: What is the manageable level of specificity that can be treated by DCMs appropriately in CDA? What specificity levels of diagnostic information are most effective and useful to be acted upon by teachers and learners in the classroom? In the same vein, as the bridge between diagnostic assessment and diagnosis-based remediation, diagnostic score report should communicate statistically sophisticated information from the CDA to end-users and meanwhile be effective in facilitating and promoting subsequent remedial teaching and learning. Therefore, besides the interpretability of the score report, additional evidence is required to support the use of the diagnostic report in the classroom.

Implementing the interpretive framework of diagnosis in the English as Foreign Language (EFL) course on reading comprehension

To demonstrate how the integrative framework of diagnosis can be implemented under classroom settings, in the following, we present an experimental study demonstrating the application of UDig diagnostic assessment issued by the Foreign Language Teaching and Research Press in a 12-week EFL reading course for first-year graduate students at a Chinese university. Particularly, the remediation processes were explored through self-narratives, focus groups, and individual interviews, and the remediation effects were examined by a pre-test post-test quasi-experiment. This study is part of a larger project investigating the usefulness of UDig diagnostic system in undergraduate- and graduate-level EFL courses in China. In the current study, we formulated three research questions:

RQ1: What did the teacher do to integrate UDig diagnostic assessments into the curriculum? What beliefs were underpinning the teacher's behaviors?

RQ2: What were the students' beliefs of diagnostic assessment and their perceptions about diagnosis-based remediation?

RQ3: Was the integration effective in improving students' EFL reading abilities?

Method

Participants

There are two groups of participants involved in the present study. The first group includes Flora (i.e., pseudonym of the first author), a female teacher responsible for an entry-level EFL reading course designed for first-year graduate students. Flora who

completed her doctoral degree in applied linguistics had, at the time of this study, 6 years of experience in diagnostic assessment research and 3 years of experience in the EFL program for entrance-level graduates.

The second group of participants includes 83 graduate students from four intact classes. Averaged at 23 years old, they came from non-English majors, including science, engineering, and social sciences. All of them were registered in the EFL reading course with relatively low reading proficiency as demonstrated by their performance on the College English Test Band 6 (CET6). The average score of these students on CET6 was 420 out of 710, 5 points below the cut-off score of the test. Among these participants, twenty students participated voluntarily in focus-group and individual interviews. Table 1 shows the profile of these participants; codes are used for anonymity's sake.

Materials and instruments

Diagnostic assessments on EFL reading comprehension The EFL reading proficiency of participants was assessed by the UDig, a computer-based, diagnostic assessment system published by the Foreign Language Teaching and Research Press in China. Lasting about 75 min, the UDig system for English reading consists of two parts, a placement assessment allocating the test taker into one of four levels of EFL reading proficiency and a diagnostic assessment exploring strengths and weaknesses on reading at the particular level that the test taker has been placed.

The diagnostic assessment was developed according to the cognitive model extracted from the reading scales of the China's Standards of English Language Ability (CSE), a Chinese version of the Common European Framework of Reference for Languages (Zeng & Fan, 2017). In the CSE scales, EFL reading abilities are defined and specified according to the revised taxonomy of educational objectives (Anderson et al., 2001), based upon which English reading abilities were divided, in the UDig system, into multiple sub-constituents, for example, summarizing the main idea, differentiating facts and opinions, and making inferences about the author's feelings and attitudes. Thereafter, test specifications were constructed, and sample items were written and reviewed internally by a panel of experts on language testing and EFL education. To meet the fine-grained requirement of diagnosis, each item was designed to evaluate a single ability of English reading (see Fig. 2 below for an example question examining students'

Table 1 Profile of the interview participants

No.	Code	Gender (number)	Average age	Major	Type of interview
1	G1	Male (4); female (1)	22	Science and engineering	Focus group
2	G2	Male (3); female (2)	23		
3	G3	Male (5)	23		
4	S1	Female	22	Science and engineering	Individual interview
5	S2	Female	23		
7	S3	Male	24		
6	S4	Female	23	Humanities and social sciences	
8	S5	Male	22		

Directions: In this section, you will read three passages. At the end of each passage, four questions will be asked. Please choose the best answer to each question.

22. An introductory sentence for a brief summary of the passage is provided below. Complete the summary by selecting the THREE answer choices that express the most important ideas in the passage. Some answer choices do not belong in the summary because they express ideas that are false or not presented in the passage.

This passage discusses three major benefits of running marathon.

- _____
- _____
- _____

- A. Challenging oneself.
- B. Exploring big cities.
- C. Boosting local tourism.
- D. Promoting sportsmanship.
- E. Getting to know top runners.
- F. Enlarging the circle of friends.

Fig. 2 An example question from the UDig reading assessments

abilities on summarizing the main idea of the text). In addition, as each item assessing a single construct, Item Response Theory (IRT) was employed to be the primary method for data analysis. However, the G-DINA, which is a commonly used DCM in CDA studies and has proven to be robust in diagnosing EFL reading abilities of Chinese test takers (Chen & Chen, 2016), was also applied iteratively in one of the parallel assessment papers. The results provided evidence to both the validity and the feasibility of the assessment to diagnose test takers' reading abilities (Sun, 2020). Validity of the diagnostic assessment was also supported by other types of evidence, including the fit between the construct of assessment items and the response processes actually engaged by the test taker (Sun, 2019). In this study, the assessment was administered by Flora and manager of the UDig system (i.e., the second author) collaboratively. Score reports of student and teacher versions were generated automatically from the system.

Diagnostic score reports The results of the diagnosis were communicated, within the UDig system, through two types of online score reports designed for students and teachers respectively. The student's report, retrieved immediately after the student has completed the assessment, covers six sections, including the English ability level (corresponding to the CSE level), verbal descriptions of the students' overall performance, a bar diagram showing the student's mastery of each attribute involved in the assessment, the definition of the attributes and examples showing the relationship of the attributes and assessment items, a line chart tracking the student's performance on, if any, multiple reading assessments, and suggestions for future studies. The teacher's report provides a summary of the performance of students in the class. These reports were designed with reference to guidelines of score report development (e.g., Hattie & Timperley, 2007; Roberts, 2012; Zenisky & Hambleton, 2012), as well as examples of diagnostic score reports in the existent literature (e.g., Jang, 2009; Roberts & Gierl, 2010). Table 2 shows an analysis of the report contents on the basis of a model of feedback to enhance learning (Hattie & Timperley, 2007). The interpretation of the student's report

Table 2 Dimensions of UDig diagnostic score reports on reading comprehension (adapted from Fan, [accepted](#))

A model of feedback to enhance learning (Hattie & Timperley, 2007)		UDig Report
Purposes	To reduce discrepancies between current understandings/performance and a desired goal	√
Objects	Student	√
	Teacher	√
Contents	Feed up: Where am I going? (the goals)	√
	Feed back: How am I going?	√
	Feed forward: What to next?	√
Levels	Task level: How well tasks are understood.	√
	Process level: The main process needed to understand/perform tasks.	√
	Self-regulation level: Self-monitoring, directing, and regulating of actions.	×
	Self-level: Personal evaluations and affect (usually positive) about the learner.	×

was ascertained from both qualitative and quantitative data. However, empirical investigations indicated that intended interpretations and use could be better achieved with the assistance of teachers, particularly for students with a low level of learner autonomy (Fan, [accepted](#)).

Interview questions Focus groups and individual interviews were designed as part of the larger project investigating the usefulness of UDig diagnostic system in China. Some results on students' understandings of UDig diagnostic score report were reported by Fan ([accepted](#)). In the current study, both methods were used to qualitatively investigate students' beliefs of diagnostic assessment, and their perceptions of diagnosis-based remediation either through interactive and directed discussions, or by individual interviews. Interview questions were adopted from Doe (2015), Jang (2009), and Yin, Sims, and Cothran (2012). Below is a list of example questions.

1. Would you say that the diagnostic assessment has had an effect on your learning of English? If yes, in what ways?
2. Do you think you have improved your English skills since the beginning of the term? How have you improved?
3. Do you think that the remediation activities (i.e., group discussions, individual tasks) in the classroom useful for helping you connect the feedback to your English learning, and for improving your English skills? If yes, which activity do you think is the most useful?
4. Do you think that the diagnostic assessment has been well integrated into the reading curriculum? If yes, in which ways? If not, why not?

Research procedures and data analyses

Research procedures are illustrated in Table 3. Student participants were divided into two groups: the experimental group and the control group. Thirty-nine students from two classes were allocated into the experimental group ($N_{\text{exp}} = 39$), whereas forty-four students from the other two classes were assigned to the control group ($N_{\text{con}} = 44$).

Table 3 Research procedures

Week	Stage	Group		Duration of time	Data collected
		Control group	Experimental group		
Week 1	Pre-test	Receiving the pre-test	Receiving the pre-test	75 min	Pre-test results
Week 2	Treatment session 1	(1) Explaining and discussing the diagnostic score report (2) Dividing students into small groups based on the pre-test results	(1) Dividing students into small groups randomly	15–30 min	Classroom observations (e.g., teacher-student interactions)
Week 3-11	Treatment session 2	(3) Assigning and evaluating remedial learning tasks of each group (4) Assigning and evaluating remedial learning tasks of individual students	(2) Assigning and evaluating remedial learning tasks of each group (3) Assigning and evaluating remedial learning tasks of individual students	50 min per week 30 min per week	Classroom observations (e.g., student-student interactions); teaching materials; teacher's self-reflections
Week 12	Post-test	Receiving the post-test	Receiving the post-test	60 min	Post-test results; teacher's self-reflections; students' interview protocols

Note. The time durations of the pre- and post-tests are different, because the pre-test consists of a placement assessment and a diagnostic assessment, whereas the post-test includes only the diagnostic assessment

Before the study, informed consent was obtained from all participants, who were voluntarily joined in the study, and could withdraw at any stage of the research.

Using a pre-test post-test quasi-experimental design, students completed two parallel diagnostic reading tests in the first and last week respectively. From weeks 2 to 11, the experimental group was required to take group-level intervention based on their pre-test results, whereas the control group did not receive planned remediation. Instead, they were allocated to small groups randomly, and required to complete, both during the class and after the class, learning tasks not specifically tailored to their weaknesses on English reading.

Data of teaching behaviors and the teacher's beliefs were collected during the 12-week assessment and remediation sessions. Data comprised teaching materials of each week for both groups (including course syllabus, lesson plans, the textbook and reference books, and power point slices used in the classroom), teacher's self-reflections on the design of the remediation as well as her understandings of diagnosis in language education, classroom observations displaying the teaching processes, students' reactions, teacher-student interactions, and student-student interactions during the small-group discussion. Data of students' beliefs and perceptions were collected after the post-test in the form of interview protocols.

A three-cycle coding method was used to analyze the qualitative data (Saldaña, 2009). First, open coding of the qualitative data was conducted. Second, all data in relation to the research questions and the proposed framework were categorized and analyzed. Finally, as a post-coding session, the general categories produced at the second step were transformed to themes corresponded to the remediation component and sub-components listed in the proposed framework. The validity of the coding was tested by check of transcripts by participating students, comparison of teaching materials,

observation data, teacher's self-reflections, and students' interview protocols, as well as comparison and discussion of the coding results among the authors of this article.

To investigate the effects of diagnosis-based remediation, both descriptive and inferential statistics were analyzed to explore whether the English reading ability changed significantly in the pre- and post-tests for each group. One-way ANCOVAs were performed to examine the statistically significant difference between the remediation of the experimental group and the control group on post-test means controlling for pre-test scores.

Results

RQ1: Diagnosis-based remediation practices and the teacher's beliefs

Teaching practices integrating diagnostic assessment into EFL reading curriculum

The remediation of the experimental group was conducted through four phases of planning, framing, implementing, and reflecting. In the planning phase, Flora made a detailed analysis of the current curriculum, teaching objectives, and students' needs. The purpose of the course was to prepare first-year, non-English major graduate students with relatively low reading proficiency for the study of academic English in the following semesters. Aligned with the CSE scales, the course was organized to improve the reading skills required by the Level 6 of CSE targeting at senior undergraduate students and entrance-level graduate students. The teaching materials were texts from *An English Reader for Postgraduates* (Zhu, 2011), covering a variety of text types listed in the CSE, including narration, description, exposition, and argumentation. Small-group discussions were used as a primary teaching method, as discussed and determined by all teachers in the department who have experience in teaching the course. They believed that small-group discussion could stimulate students' text understanding, problem solving, and critical thinking. During the discussion students presented multiple points of view, responded to the ideas of others, and reflected on their own ideas in an effort to build their knowledge, understanding, or interpretation of the text at hand.

Flora also investigated the UDig diagnostic assessment on EFL reading ability, its match with the current curriculum of graduate course, and students' trials and responses in the previous semester. On the basis of multiple analyses, plans of integrating the UDig diagnostic assessment into the curriculum was formed for the experimental group, and is described in the implementing section below.

As part of the framing process, Flora introduced, via emails, the UDig system to the students of the experimental group at the beginning of the semester. The instructional plans were also communicated with students after they had completed the diagnostic assessment.

In the implementing phase, students of the experimental group took the UDig assessment online before the course began (i.e., the pre-test), and obtained individual score reports presenting their strengths and weaknesses on English reading. During the class, discussions were organized to facilitate students' understandings of the diagnostic score report. Questions and misunderstandings about the score report were solved through discussions and teacher's explanations. Below present two examples of the teacher-student interaction:

(1) Student A: According to the report, I’m at the Level 5 of English reading proficiency. What does Level 5 mean?

Teacher: The UDig diagnostic assessments divide test takers’ English reading proficiency into four levels, corresponding to levels 4, 5, 6 and 7 of the CSE. [Flora had introduced the CSE and its relations to the curriculum at the beginning of the semester.] Moreover, in the UDig diagnostic report, you can find detailed descriptions of the level you’re placed at, including the reading skills you’ve mastered, and types and language complexity of the reading texts involved in the assessment. You can also find descriptions of a higher level, which might serve as your learning goals in your future studies.

(2) Student B: When I received this report, I was looking for the item review section that would help me identify my incorrect responses. But I cannot find this information, and I’m a bit confused about it.

Teacher: You might be expecting the item reviews as in a traditional test report, but the main purposes of diagnostic assessment and diagnostic feedback are to examine your English reading abilities in relation to the skills you have or have not mastered, instead of your accuracy in answering each test item. For example, you can learn your mastery of reading skills from either the “General Description of your Performance” section on page one, or the bar chart on page 3. Based on your skill mastery, suggestions for your future studies are also presented in the last section of the report.

Thereafter, informed with their mastery of attributes, students were divided into small groups. Table 4 summarizes the division of groups in the two experimental classes. Each group, consisting of five to seven students and sharing two attributes scored the lowest in the pre-test, was required to complete remedial activities collaboratively.

Table 4 Group divisions of the two classes in the experimental group

No.	Weaknesses		Number of groups (number of students)		
	Weakness I	Weakness II	Class I	Class II	Total
1	Analyzing logical relationships between ideas	Summarizing the main idea of the text	1 (4)	2 (9)	3 (13)
2	Making inferences about the author’s feelings and attitudes	Locating the target information by skimming, scanning, or browsing	0 (0)	1 (7)	1 (7)
3	Making inferences about the author’s feelings and attitudes	Summarizing the main idea of the text	1 (4)	0	1 (4)
4	Locating the target information by skimming, scanning, or browsing	Making inferences about the author’s feelings and attitudes	1 (4)	0	1 (4)
5	Analyzing logical relationships between ideas	Differentiating facts and opinions/ primary and secondary information from the text	0 (0)	1 (4)	1 (4)
6	Differentiating facts and opinions/ primary and secondary information from the text	Extracting detailed information from the text	1 (3)	0	1 (3)
7	Summarizing the main idea of the text	Analyzing logical relationships between ideas	1 (4)	0	1 (4)
Total			5 (19)	4 (20)	9 (39)

The group discussion was organized through four steps as follows: (1) Each group was assigned two tasks for discussion which were aligned with the target attributes (i.e., weaknesses) of the group. Please see Table 5 for examples of discussion tasks categorized by reading attributes. (2) After the group discussion, representatives of each group were required to summarize the discussion process and provide answers to the questions. (3) The representatives answered questions raised by students in other groups. (4) The group discussion and presentation were reviewed and evaluated by the teacher, who would further explain the questions remained unsolved during the discussion session. The small-group discussions, lasting about 50 min, were conducted once a week from weeks 3 to 11.

After the class, students were assigned exercises targeting on their individual weaknesses of English reading each week. Sources of tasks include exercises from the textbook, and items from reading tests that have been used for diagnosis in previous research (e.g., the reading section of TOEFL test). The quality of task performance was evaluated by a teaching assistant (i.e., the third author). At the end of the semester, students took an equivalent form of UDig assessment online (i.e., the post-test) and obtained the diagnostic feedback immediately after they finished the assessment.

In the course of the remediation, Flora, the test administrator, and the teaching assistant held meetings regularly to communicate results of two diagnostic assessments, plans of remedial instructions, and students' performance during group discussions and follow-up exercises, based upon which teaching practices were implemented and adjusted.

Table 5 Examples of reading tasks for small-group discussion in the classroom

No.	Attribute	Example task for small-group discussion
1	Locating the target information by skimming, scanning, or browsing	(1) The title of the text is "The Three New Yorks". Please scan the text and find out what "The Three New Yorks" refer to. What are the major characteristics of the three New Yorks?
2	Extracting detailed information from the text	(2) Please identify the word "commuter" in the text and find out the author's detailed description of commuters. (3) The author uses examples to describe the current state of settlers in New York. Please find out these examples and explain the function of them.
3	Summarizing the main idea of the text	(4) Try to find out the topic sentence of each paragraph, or to summarize the main idea of each paragraph in one or two sentences. (5) Try to summarize the main idea of the text.
4	Analyzing logical relationships between ideas	(6) Please find out, in the second paragraph, the connective words that help develop the logical relationship of the paragraph. (7) Draw a mind map to show how the author develops the topic and organizes the ideas in the text.
5	Making inferences about the author's feelings and attitudes	(8) Try to make inferences about the author's attitudes toward the three groups of New Yorks discussed in this text.
6	Differentiating facts and opinions/ comparing the opinions and attitudes of different authors	(9) This unit comprises two texts about urban life (i.e., "The Three New Yorks" and "Loving and Hating of New York"). Please read the two texts and compare the opinions of the two authors about New York city and life in the city.

Note. The example tasks for small-group discussion are adapted from "Unit Three – The Three New Yorks" in the textbook *An English Reader for Postgraduates* (Zhu, 2011)

The teacher's beliefs underpinning teaching practices Flora's beliefs informing her use of UDig reading assessment in her classroom instruction centered on her orientation to diagnostic assessment, her approach to EFL teaching, and her views on English teaching and diagnosis in the Chinese context. Informed by her previous experience of diagnostic assessment research, Flora believed that, instead of providing a summative score for test takers, CDA can explore students' knowledge structures and cognitive processes underlying test-taking behaviors and can be "directly used to derive remedial instruction and learning that targets students' deficiencies". To guide follow-up remediation, Flora preferred to use, in the classroom, diagnostic assessments suitable to the specific language learning context and curriculum. The UDig reading assessment designed on the basis of the reading scales of CSE and trialed to diagnose English reading deficiencies for Chinese students was aligned with the goals of the entry-level EFL reading course, and was therefore selected as the diagnostic tool in the current study.

Although Flora had some research experience on diagnostic assessment, she had not used it in the classroom to help students. In the course of this study, Flora was committed to integrate the diagnostic assessment into reading curriculum, including selecting the diagnostic assessment suitable to students English proficiency level and the teaching objectives, sharing the diagnostic information and remedial plans with students, implementing the diagnostic assessment in the classroom, and reflecting the remediation process during and after the semester.

Related to the instruction of EFL reading, Flora believed that small-group discussion was an effective way of remediation because of its familiarity to both the teacher and students, and characteristics of the Chinese context. Small-group discussion has long been adopted as the major teaching method in the graduate EFL course at the university. It could serve intellectual, emotional, and social purposes required by the course objectives. Moreover, Flora had positive attitude toward group-level diagnosis and remediation. She agreed that students needed to be treated as individuals based on their mastery profiles for more targeted help. However, in the Chinese context where a teacher must take care of typically more than eighty students at the same time, the remedial instruction conducted for individual students would be castle in the air. Group-level remedial instructions could be a possible substitute. In this way, the quality of learning outcome could be enhanced effectively, and resources could be allocated by curriculum developers and school administrators strategically. Informed by these views, during the semester, the curriculum and teaching activities were tailored to the needs of students who shared the similar weaknesses on English reading.

RQ2: Students' beliefs and perceptions

Twenty students participated in focus-group and individual interviews investigating their thoughts and attitudes toward the diagnostic assessment, as well as their perceptions of diagnosis-based remediation. Generally, the majority of students' attitudes toward diagnostic assessment were positive. However, a few students viewed it as unrelated to their future study of English for lack of motivation. For example, S1 commented that "my major is computer science, and I'd like to know my English skills on reading academic papers about the computer science, or at least on reading general

articles like science news and information. The diagnostic assessment should find out my weaknesses in that area.”

Students’ perceptions of the diagnosis-based remediation were mixed. The majority of students reported that the remedial instruction was quite useful, because “it can help me understand specific skills in English reading, and pinpoint my areas of improvement in terms of these skills” (S2). Among a variety of instructional activities, they expressed that the explanation of the purpose of diagnostic assessment and the question-answer session about the diagnostic score report in the classroom were the most useful, as these activities could “help us understand that diagnostic assessment is different from the tests we’ve taken before” (G2). Otherwise, some students might take the assessment as the summative language test that they were familiar with.

In terms of group-level activities conducted in the classroom, most students commented that they agreed with the teacher that the small-group discussion was “a more efficient method than learning activities at the individual level” (G1), and can “improve my problem-solving skills through communication with my peers” (S5). However, a few students expressed that the individual instruction should be the major method for remediation in the classroom, complemented by small group discussions, so that the remediation would be “more targeted and effective” (G3). In addition, students suggested that, besides the pre-test and post-test, it would be better to implement more diagnostic assessments during the course, as “they could remind us of our mastery patterns of reading skills” (G1), and “I can also adjust our learning according to the test results if the test could be administered once a month” (S4).

RQ3: Learning outcomes of the experimental and control groups

Descriptive statistics and results of ANCONVA analyses are presented in Table 6. The number of participants, means, and standard deviations for total scores and attribute

Table 6 Pre-test, post-test, and ANCOVA results for the experimental and control groups

Attribute	Group	No.	Pre-test		Post-test		ANCOVA		
			Mean	SD	Mean	SD	F	p	Cohen’s d
A1	EG	34	10.59	10.13	14.65	10.77	4.80**	0.033	0.63
	CG	20	12.00	9.73	9.6	8.35			
A2	EG	34	17.79	4.17	18.06	4.74	4.58**	0.037	0.59
	CG	20	16.65	6.90	14.80	5.53			
A3	EG	10	22.56	7.50	26.11	6.01	4.32**	0.049	0.87
	CG	16	17.59	7.53	14.23	8.66			
A4	EG	34	22.58	7.68	23.00	7.00	0.78	0.383	0.29
	CG	20	19.45	9.89	19.80	10.85			
A5	EG	30	20.36	5.65	22.83	6.59	0.88	0.352	0.29
	CG	17	17.00	8.25	20.18	7.66			
Total score	EG	34	220.21	29.67	233.26	39.67	1.66	0.203	0.35
	CG	20	200.30	42.67	207.05	47.98			

A1, analyzing logical relationships between ideas (e.g., causation, transition, progression); A2, summarizing the main idea of the text; A3, making inferences about the author’s feelings and attitudes; A4, differentiating facts and opinions/ comparing the opinions and attitudes of different authors; A5, locating the target information by skimming, scanning, or browsing; EG, experimental group; CG, control group
 **p < .05

scores under diagnosis-based remediation and non-remediation conditions are reported. Specifically, although 83 students took the pre-test, only 54 students completed the post-test, among which 34 students were from the experimental group ($N_{\text{exp1}} = 34$) and 20 from the control group ($N_{\text{con1}} = 20$). In addition, placed at four different levels of reading proficiency, the participants were measured by one of the four diagnostic assessments targeting at their levels. The attributes varied as measured by different levels.

One-way ANCOVAs were conducted to determine the statistically significant difference between the teaching practices of the experimental group and the control group on post-test means controlling for pre-test scores. Outliers and assumptions for performing ANCOVA were checked using the SPSS (version 26.0). Results show that there were significant effects of the diagnosis-based remediation on three attributes as follows: analyzing logical relationships between ideas (e.g., causation, transition, progression), $F(1, 52) = 4.80$, $p = 0.033$, *cohen's d* = 0.63; summarizing the main idea of the text, $F(1, 52) = 4.58$, $p = 0.037$, *cohen's d* = 0.59; and making inferences about the author's feelings and attitudes, $F(1, 24) = 4.32$, $p = 0.049$, *cohen's d* = 0.87. However, no significant differences were found between the experimental and control group on the other attributes and the total score after controlling for the pre-test means.

Discussion and conclusions

The integrative framework of diagnosis

The present study proposed the integrative framework connecting diagnostic assessment to feedback and remediation. The framework consists of four major components, diagnosis, feedback, remediation, and validation, each playing a unique role in the framework. Diagnosis serves as the starting point, on the basis of which the other components can be built on. Feedback serves as the bridge between the CDA and a diverse audience of assessment users, translating the assessment data to understandable and actionable information for subsequent instruction and learning. Remediation is an essential component of the framework, functioning as the key to achieving the desired goals of diagnosis. Validation is also necessary as evidence must be collected and analyzed to ensure that the intended impact of the CDA is realized through close alignment among the diagnosis, feedback, and remediation.

Compared with previous models on a similar attempt, the integrative framework of diagnosis has a number of advantages. Firstly, as a response to the increasing calls in CDA to design assessment tasks for diagnostic purposes, this framework adopts a systematic item development approach starting with the construction of the cognitive model, that is, the construct of the assessment items to be designed and used. Although the retrofitting approach using existing unidimensional tests originally not designed for diagnostic purposes is preferred by previous models of CDA (e.g., Lee & Sawaki, 2009), we argue that items developed with cognitive characteristics would extract micro-level information that is unlikely to exist on general-purpose tests, and can thus gear to the needs in the educational micro-environments to monitor the teaching and learning process at the classroom level. Compared with the retrofitting approach, the diagnosis of the item development approach is also expected to achieve finer granularity on the attributes defined in the cognitive model, better model fits to examinees' response data,

and more straightforward interpretations of diagnostic results (de la Torre & Minchen, 2014; Deonovic, Chopade, Yudelson, de La Torre, & von Davier, 2019). This approach has been proved feasible by the UDig assessment used in the present study as well as other empirical studies of cognitive diagnosis (e.g., Xie, 2019).

Secondly, in contrast to previous models that viewed feedback and remediation as subsidiary components to CDA, this framework recognizes feedback and remediation as having equal importance with diagnosis, because the positive impact of CDA cannot be fully realized without the accurate presentation of the diagnostic results and the effective implementation of remedial activities (Lee, 2015). Given the paucity of research on diagnostic feedback and diagnosis-based remediation, it is worth drawing knowledge and experience from studies in other fields of education, as the practice adopted in the present study.

Finally, in addition to the components specifying the diagnostic assessment, feedback, and remediation widely discussed, validation is included as one of the major constituents in the framework, highlighting the significance of making assumptions and collecting evidence in support of the proposed interpretation and use of the diagnostic assessment. The validation procedure is both local and global. It is local in the sense that validity evidence should be collected from within each component to ensure the accuracy of diagnostic assessment, the interpretability of diagnostic feedback, and the effectiveness of remedial activities employed in the classroom. It is global since evidence should be collected to show the close alignment among the three components of diagnosis, feedback, and remediation. For example, the specificity of attributes is manageable and treatable in the CDA while can be acted upon by teachers and learners in the classroom.

The experimental study

To demonstrate how the integrative framework of diagnosis can be implemented under classroom settings, an experimental study was reported to apply the UDig diagnostic assessment in the entry-level EFL reading course for first-year graduate students at a Chinese university. Participating students were divided into experimental and control groups, and the remediation processes of the experimental group were investigated qualitatively. The results demonstrate procedures of integrating the UDig diagnostic assessment to the EFL reading curriculum for entry-level graduate students, and implementing remedial instructions through four phases of planning, framing, implementing, and reflecting.

It is worth noting that the UDig diagnostic assessment serves as different purposes at different stages in the course of the integration. The pre-test implemented at the beginning of the English reading course functions as the assessment for learning (AfL), and as learning (AaL) by providing the teacher and students with the information needed to modify instruction and learning in classrooms, whereas the post-test utilized at the end of the course functions primarily as the assessment of learning (AoL) tool for summative evaluation of how much of the goals being achieved. The three assessment methods of AfL, AaL, and AoL were integrated into a learning-oriented approach to assessment by the deployment of the diagnostic assessment in the classroom (Jones, Saville, & Salamoura, 2016). From the diagnostic assessment, students could find out

what they had achieved, learn the areas for future improvement, and develop metacognition by being involved in the assessment (Jang & Wagner, 2013).

The study also investigated beliefs of the teacher and students on language diagnosis and diagnosis-based remediation. The results show that plans of integration, use of diagnostic feedback, and procedures of remediation were influenced by the teacher's orientation to diagnostic assessment and approach to EFL teaching. Flora, in particular, took a collective view on the use of diagnostic assessment, holding that remedial teaching at the group level would be more efficient and effective compared with individualized instruction in the Chinese context where teachers are burdened with a large amount of teaching tasks. However, she acknowledged that the group-level remediation should be complemented, to some extent, by teaching and learning exercises tailored to the needs of individual students. Generally, students had similar views on diagnostic assessment and follow-up remediation with their teacher, especially after the teacher reviewed and explained the diagnostic score report in the classroom, indicating that diagnostic assessment and remediation could play a positive role in students' English learning if assisted by teachers.

To examine the effect of remediation, a quasi-experiment was implemented and reported. The study found that compared with the control group, the experimental group was improved significantly on three attributes after 12-week diagnosis-based remediation. The attributes include analyzing logical relationships between ideas, summarizing the main idea of the text, and making inferences about the author's feelings and attitudes. However, no significant improvement was observed on the total score and other attributes covered by the assessment. The results indicate that the diagnosis-based approach is a more effective way to provide instructionally useful information that can be acted upon in the classroom than other test and assessment approaches that do not differentiate strengths and weaknesses among students with the same total score. It should be noted that 29 students quit the research before the post-test, 5 from the experimental group and 24 from the control group, indicating that the implementation of remedial activities after the pre-test might exert a positive impact on the attitudes of students toward diagnostic assessment. However, the high dropout rate of the control group brings about the problem of small and unequal sample sizes in the two groups, mitigating the statistical power of the ANCOVA.

Limitations and future directions

Despite the above insights, the present study has three remaining issues that should be addressed in future research endeavors.

First, the integrative framework of diagnosis was proposed and presented following operational procedures drawn from research in the areas of CDA, score reporting, instruction, and validation. Compared with a large body of studies on CDA, research on the other three components remain scarce. As a result, the procedures listed in the present framework are subject to refinement and revision through, to the best of our knowledge, two ways: a number of empirical studies implementing the framework in a diversified context, and updated reference to increasing research in relevant areas of education.

Second, the UDig system was utilized as diagnostic tools in the classroom, showing the feasibility of using the computer-based diagnostic assessment in the classroom. In

the future, the diagnostic assessment system can be expanded to accommodate personalized learning online. In so doing, the diagnostic assessment would be connected to remedial learning more effectively, and learning data could be collected more conveniently.

Last, the present study explored and reported behaviors of the teacher in planning, framing, conducting, and reflecting the remedial activities tailored to students' needs on English reading, as well as the teacher's beliefs and students' beliefs of both diagnostic assessment and diagnosis-based remediation. These qualitative data serve as valuable sources of information. However, other types of qualitative data could be collected in future studies, for example, students' diaries and self-reflections, videos of small-group discussions, and emails between the teacher and students, so that the remediation process would be explored more thoroughly from the perspectives of both the teacher and the students.

Limitations aside, this article attempts to add to the meager literature on the integration of CDA to diagnostic feedback, and remedial teaching and learning, and provide practical guidance to the construction of diagnostic assessment, communication of diagnostic score report, and implementation of remedial activities in the classroom.

Abbreviations

AaL: Assessment as learning; AfL: Assessment for learning; AoL: Assessment of learning; CDA: Cognitive diagnostic assessment; CEFR: Common European Framework of Reference for Languages; CET6: College English Test Band 6; CSE: China's Standards of English Language Ability; DCMs: Diagnostic classification models; EFL: English as a Foreign Language; NMET: National Matriculation Entrance Test

Acknowledgements

The authors would like to thank the editorial board and anonymous reviewers for their insightful comments on the manuscript.

Authors' contributions

The first author was responsible for the construction of the theoretical framework proposed in the study, the design of empirical research, and the writing and editing of the manuscript. The second author was responsible for the administration of the diagnostic assessment used in the study, as well as the design of remedial plans implemented in the classroom. The third author contributed to data collection and part of data analysis. All three authors read and approved the final manuscript.

Authors' information

Dr. Tingting Fan has completed her doctoral degree at Guangdong University of Foreign Studies, China, and is currently an assistant professor in the College of Foreign Languages at Nanjing University of Aeronautics and Astronautics. Her areas of interest include EFL education and assessment, mostly in Chinese contexts.

Ms. Jieqing Song is a senior editor in the Foreign Language Teaching and Research Press, responsible for the management and operation of the UDig diagnostic system issued by the Press. Her areas of interest lie in language testing and classroom assessment.

Ms. Zheshu Guan is an English-major undergraduate student in the College of Foreign Languages at Nanjing University of Aeronautics and Astronautics. She was working as a teaching assistant for Dr. Fan in the course of the study.

Funding

This work was supported by the Humanities and Social Sciences Fund of the Ministry of Education in China [17YJC740017], the Foreign Language Teaching and Research Press [302019070273], and the Undergraduate Training Program for Innovation in China [2019CX01210].

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Author details

¹College of Foreign Languages, Nanjing University of Aeronautics and Astronautics, 29 Jiangjun Ave., Nanjing 211106, People's Republic of China. ²Foreign Language Teaching and Research Press, 19 XisanhuanBeilu, Beijing 100089, People's Republic of China. ³College of Foreign Languages, Nanjing University of Aeronautics and Astronautics, 29 Jiangjun Ave., Nanjing 211106, People's Republic of China.

Received: 2 September 2020 Accepted: 15 December 2020

Published online: 15 January 2021

References

- Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. New York and London: Routledge.
- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., ... Wittrock, M. C. (2001). *A taxonomy for learning teaching, and assessing: a revision of Blooms Taxonomy of Educational Objectives*. White Plains: Longman.
- Brown, G. T. L., O'Leary, T. M., & Hattie, J. A. C. (2019). Effective reporting for formative assessment: The asTTle case example. In D. Zapata-Rivera (Ed.), *Score reporting research and applications*, (pp. 107–125). New York: Routledge.
- Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly*, 13(3), 218–230.
- Chen, J., & de La Torre, J. (2014). A Procedure for diagnostically modeling extant large-scale assessment data: The case of the programme for international student assessment in reading. *Psychology*, 5(18), 1967–1978.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 79–199.
- de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Educational Psychology*, 20(2), 89–97.
- Deonovic, B., Chopade, P., Yudelson, M., de La Torre, J., & von Davier, A. A. (2019). Application of cognitive diagnostic models to learning and assessment systems. In M. von Davier, & Y. S. Lee (Eds.), *Handbook of diagnostic classification models: models and model extensions, applications, software packages*, (pp. 437–460). Berlin: Springer.
- Doer, C. (2015). One teacher's take on using a 'test' for diagnostic purposes in the classroom. *Journal of English for Academic Purposes*, 18, 40–50.
- Fan, T. (2020). Book review: Handbook of diagnostic classification models: Models and model extensions, applications, software packages. *Language Testing*, 37(4), 623–626.
- Fan, T. (Accepted). Investigating the validity of a diagnostic score report from the perspective of language learners. *Foreign Languages and Their Teaching*.
- Fan, T., & Yan, X. (2020). Diagnosing English reading ability in Chinese senior high schools. *Studies in Educational Evaluation*, 67, 1–12.
- Fan, T., & Zeng, Y. (2016). Cognitive diagnostic assessments and their applications on reading comprehension. *Foreign Languages in China*, 13(2), 82–89.
- Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: a research-based approach to score report design. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Testing and assessment in school psychology and education*, (vol. 3, pp. 479–494). Washington, D.C.: American Psychological Association.
- Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: elaborating on diagnostic principles. *Language Testing*, 32(3), 317–336.
- Hattie, J. A. C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Hill, K., & Mcnamara, T. (2011). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing*, 29(3), 395–420.
- 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., & Wagner, M. (2013). Diagnostic feedback in the classroom. *The Companion to Language Assessment*, 2, 693–711.
- Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement*, 55(4), 635–664.
- Jones, N., Saville, N., & Salamoura, A. (2016). *Learning oriented assessment*, (vol. 45). Cambridge: Cambridge University Press.
- Kim, A. H. (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. (2015). Diagnosing diagnostic language assessment. *Language Testing*, 32(3), 299–316.
- Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: an overview. *Language Assessment Quarterly*, 6(3), 172–189.
- Leighton, J., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge University Press.
- Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, 40(3), 200–217.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspective*, 1(1), 3–62.
- O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice*, 36(2), 16–23.
- Rankin, J. G. (2016). *Standards for reporting data to educators: what educational leaders should know and demand*. New York: Routledge.
- Rea-Dickins, P. (2006). Currents and eddies in the discourse of assessment: a learning-focused interpretation. *International Journal of Applied Linguistics*, 16(2), 163–188.
- Roberts, M. P. R. (2012). *Developing and evaluating student score reports for cognitive diagnostic assessment*. Doctoral Dissertation. University of Alberta.
- Roberts, M. P. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, 29(3), 25–38.
- Roberts, M. P. R., Gotch, C. M., & Lester, J. N. (2018). Examining score report language in accountability testing. *Frontiers in Education*, 3, 1–17.
- Saldaña, J. (2009). *The coding manual for qualitative researchers*. Thousand Oaks: Sage Publications Inc.

- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 263–331). New York: American Council on education and Macmillan.
- Stout, W., Henson, R., DiBello, L., & Shear, B. (2019). The reparameterized unified model system: A diagnostic assessment modeling approach. In M. von Davier, & Y. S. Lee (Eds.), *Handbook of diagnostic classification models: models and model extensions, applications, software packages*, (pp. 47–79). Berlin: Springer.
- Sun, H. (2019). Investigating students' cognitive processes in a diagnostic EFL reading test. *Foreign Language Education in China*, 4, 25–32.
- Tannenbaum, R. J. (2019). Validity aspects of score reporting. In D. Zapata-Rivera (Ed.), *Score reporting research and applications*, (pp. 9–18). New York: Routledge.
- Tsagari, D., & Cheng, L. (2017). Washback, impact, and consequences revisited. In E. Shohamy, I. Or, S. May (Eds.), *Language Testing and Assessment* (pp. 359–372). Cham: Springer.
- Von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic' classification models – a commentary. *Psychometrika*, 79(2), 340–346.
- von Davier, M., & Lee, Y. S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. New York: Springer.
- Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, 48(2), 165–187.
- Wang, Q., & Li, L. (2019). Integrating teaching-learning-assessment in the EFL classroom in the context of developing key competencies: significance, theories and methods. *Curriculum, Teaching Material and Method*, 39(5), 114–120.
- Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement*, 52(4), 457–476.
- Xie, Q. (2019). Diagnosing linguistic problems in English academic writing of university students: An item bank approach. *Language Assessment Quarterly*, 17(2), 183–203.
- Yi, Y. S. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: an application of cognitive diagnostic models. *Language Testing*, 34, 337–355.
- 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(1), 78–104.
- Zapata-Rivera, J. D., & Katz, I. R. (2014). Keeping your audience in mind: applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy and Practice*, 21(4), 442–463.
- Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, 31(2), 21–26.
- Zhan, P., Ma, W., Jiao, H., & Ding, S. (2020). A sequential higher order latent structural model for hierarchical attributes in cognitive diagnostic assessments. *Applied Psychological Measurement*, 44(1), 65–83.
- Zhu, B. (2011). *An English reader for postgraduates*. Suzhou: Suzhou University Press.
- Zeng, Y., & Fan, T. (2017). Developing reading proficiency scales for EFL learners in China. *Language Testing in Asia*, 7(8), 1–15.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)
