Yi *Language Testing in Asia* (2021) 11:13
https://doi.org/10.1186/s40468-021-00127-4

Language Testing in Asia

RESEARCH Open Access

On the usefulness of CDA-based score reporting: implications for self-regulated learning

Yeon-Sook Yi

Correspondence: uiuc99110@hotmail.com; ysyi@smu.ac.kr
SangMyung University, 214 Education Building, 20, Hongimun 2-Gil, Jongno-Gu, Seoul 03016, South Korea

Abstract

In cognitive diagnostic modeling research, one area that has not had enough research interests is remedial learning or instruction based on the information provided by cognitive diagnostic assessment (CDA). The present study tries to address this research gap by looking into the usefulness of the fine-grained score reports based on CDA in two different ways, i.e., a post-test and a survey inquiring about the perceived effectiveness of the score report that provided the skill profile of individual students. Another significance of the current research is that it attempted to introduce cognitive diagnostic assessment into a regular school exam unlike most previous studies that retrofitted to the existing tests. College students in Korea participated in the study, who were encouraged to do self-regulated learning utilizing the detailed information in the CDA-based score report. The results of the post-test and the survey were positive overall, supporting the utility of CDA-generated performance reports. The article ends with some suggestions for future research based on the limitations of the study.

Keywords: Cognitive diagnostic assessment, EFL test, School test, Detailed score report, Self-regulated learning

Introduction

The purpose of this article is to evaluate the usefulness of score reporting based on cognitive diagnostic assessment (CDA) in actual classroom settings. It aims to provide implications for developing more detailed score reports as well as reporting strategies in order to elicit students' self-directed learning acted on the fine-grained information that CDA can produce for stakeholders of language testing.

Cognitive diagnostic assessment is an evolutionary outcome in psychometrics, born out of user demands. In the measurement field, item response theory (IRT) has overcome some limitations of classical test theory (CTT) and has been successfully applied for the unidimensional, continuous scaling of examinees in major subject or cognitive areas. Though such psychometric modeling is very useful and psychometrically reliable for summative assessment, it could not solve an essential problem in educational assessment. For more than three decades, researchers have suggested that tests be

designed for formative assessment where the results of the assessment are directly used to guide teaching and learning (Bejar 1984). To address such challenges, measurement researchers have tried to develop a test procedure that provides fine-grained diagnostic feedback about learners' mastery levels of cognitive attributes in order to help them remedy the deficiencies in the skills that they have not mastered sufficiently.

In language testing, cognitive diagnostic assessment has also drawn much interest recently, as reflected in the main theme of the annual international language testing conference (LTRC) in 2013 and special issues in major international language testing journals (Special Issue: Cognitive Diagnosis and Q-Matrices in Language Assessment in *Language Assessment Quarterly*, 2009 and Special Issue on Future of Diagnostic Language Testing in *Language Testing*, 2015).

Despite recently growing interest in CDA in educational measurement as well as in language testing, there is a paucity of research that examines the usefulness of cognitive diagnostic score reporting by looking into learner progress and/or perceptions of the score report. How to provide such in-depth feedback to students is another issue related to the effectiveness of CDA-based score reporting. In this research context, the present study purports to bridge the research gap. The significance of this study is that it is one of the first studies investigating cognitive diagnostic score reporting in the context of classroom language assessment. The study also explores subskills or attributes[1] not addressed in previous cognitive diagnostic modeling studies.

Previous studies on the empirical utility of cognitive diagnostic score reporting

The primary merit of CDA is that it can go deeper into the attributes that compose the item as opposed to the traditional item statistics which only focus on the item for the assessment and instruction. In order to apply the skills diagnosis approach to a real classroom context and put the diagnostic information to actual use, developing a comprehensive and user-friendly diagnostic score reporting card is an important task. This type of reporting provides a diagnostic profile of cognitive strengths and weaknesses. A student or parent can use this document as a starting point for discussions with a teacher or tutor regarding areas requiring further instruction or study.

In a paper that reviews the CDA literature published in peer-reviewed journals in 2009 or later, Sessoms and Henson (2018) revealed that only 8% used the CDA results to provide feedback to students or teachers. They deemed the limited use of CDA results as important shortcomings of CDA applications. Among a handful of such studies, Sun and Suzuki (2013) used a CDA approach to fraction problems in math for 144 sixth grade students in a primary school in Japan. Showing how CDA can provide detailed information about students' strengths and weaknesses, they asked the teachers whether they thought the results would be useful and discussed the applicability of CDA for providing effective feedback for teachers to improve their teaching. In the same math domain, Wu (2018) developed an online individualized tutor program for

[1]In the literature, different terms have been used to refer to dimensions of cognitive constructs, such as attributes, skills, factors, traits, and subskills. Though slightly differently defined, attributes and skills are mostly used interchangeably in the literature. Alternatively, subskills and dimensions are also used at times to convey the same meaning as attributes. (See Buck and Tatsuoka (1998); Rupp et al. (2010); and Li (2011) for detailed definitions of some of these terms.)

fourth grade students in an attempt to explore the effectiveness of the remedial program based on the cognitive diagnostic profiling. The study found that the online individualized tutor program outperformed the traditional remedial instruction program, especially for medium- and low-achieving students.

The literature in language testing also confirms this unbalance in the research. Lee (2015) describes it as "grappling with a fundamental dilemma in meeting two conflicting requirements", which were satisfying psychometric requirements of skill profiles of learners on the one hand and making CDA-based feedback effective for subsequent learning activities, on the other. Elaborating on three core components of diagnostic language assessment, Lee (2015) also mentions that remedial learning or instruction has not yet been studied much. Van der Boom and Jang (2018) report positive research findings experimenting with young learners. Motivated by the lack of research evidence in cognitive diagnostic assessment, their study found that customized diagnostic feedback based on multiple sources such as students' interests, learning preferences, and reading readiness levels gave them a much better understanding of their strengths and weaknesses and how to target these areas than just providing an achievement level. They conclude that providing students with skills and strategies through feedback allows them to increase their self-regulation and motivation to learn. Slightly from a different angle, Jang et al. (2015) looked at how holistic diagnostic feedback was interpreted by young learners with different profiles of reading skills, goal orientations, and perceived ability. What changed students' responses to diagnostic feedback were not their reading mastery levels, but psychological factors which strongly impacted the efficacy of feedback processing. Because they were young learners, their parent goal orientations, not their own, also showed significant relationships with their skill mastery levels. The study maintains that young learners' perceptions about ability and orientations to learning dynamically influence the ways they process and use diagnostic feedback and that understanding their interactions with diagnostic feedback is critical for maximizing the effect of the feedback.

Earlier than these studies, Jang's study (Jang 2005) aimed to evaluate the validity of the CDA application to a non-diagnostic L2 reading comprehension test, focusing on the dependability of the Fusion Model's skill profiling, the characteristics of L2 skill profiles, and the diagnostic capacity of the existing test. Her research also examined the validity arguments from the users' perspective on the usefulness of the diagnostic feedback. When creating score reports, she used a reporting format similar to the College Board's Score Report Plus for the PSAT/NMSQT that reports diagnostic information based on an analysis of examinee responses using a modified Rule Space Model (Roberts and Gierl 2010). Cognitive diagnostic feedback was provided in the form of language skill descriptors, discriminatory power of items, and skill mastery probabilities. The results of the study were that the CDA approach could provide more fine-grained diagnostic information about reading skills than traditional test scoring, offering useful information for future application of cognitive diagnostic assessment. This study seems to be the only study to date that inquired into the utility of CDA-generated diagnostic feedback with adult English learners.

In this research context of language testing, the uniqueness of the current study is multi-fold. First, the test was developed with the intention of applying cognitive diagnostic assessment from the onset of the project, not retrofitting to the existing non-

diagnostic test. Second, the research significance also lies in the fact that the present study brings CDA into the classroom. The study also experimented with adult EFL learners with language attributes not addressed in previous cognitive diagnostic modeling research. The current study proposes to answer the following research questions.

1. Do the participants show progress in their scores after self-regulated learning informed by the detailed score reporting?
2. Do the participants feel the information in the cognitive diagnostic score report was helpful? If so, how useful was it according to their perception?

Methodology

Participants and the instrument

The students who participated in the study were first-year college students in Seoul, Korea, who were taking a general English class. Cognitive diagnostic assessment is a measurement method that works with a relatively large number of examinees. For this reason, instructors were contacted and asked if they could participate in this research study. Consequently, 13 intact classes which were about one third of the entire first-year students took part in the experiment. Of the initial 474 participants, 52 students were international, whose English language performance (mean score = 16.038, SD = 6.745) was generally lower than their Korean counterparts. As for the Korean participants, their English proficiency can be indirectly indicated by the national College Scholastic Ability Test (CSAT) levels of the entire freshman cohort of the university that year (Table 1).

The class used a textbook from a globally well-known ESL material publisher (C publisher) which taught English for general purposes with sections for vocabulary, conversation, grammar, listening, and communicative activities. The participants took a 35-item test of English as an achievement test (mid-term test), which had items of grammar and vocabulary taught in class. The test was created by the researcher in consultation with the instructors participating in the study. Table 2 describes the key features of the test. The attributes required to answer the questions were nine and they were the following: S-V agreement, simple present vs. present continuous, zero conditional, simple past vs. past continuous, past vs. present participle, relative pronoun, used to/did not use to/Did you use to?, words, and expressions (See Q-Matrix in Appendix IV).

Multiple steps of data collection and analysis

In the initial stage of the research, 474 respondents first took a 35-item test on the 8th week of the semester. Their performance data were analyzed with a cognitive diagnostic model and the resultant detailed diagnostic score report was given to about half of the test-takers, i.e., 239 students (for the detailed score report of the test, see APPENDIX

Table 1 CSAT English levels of Korean participants

CSAT level	1 (100)	2 (95)	3 (88)	4 (76)	5 (59)	6 (39)	Unknown	Total
Percentage	18.11	44.15	23.77	6.79	2.26	0.75	4.15	100

Note. Numbers in parentheses are the percentile scores provided by the KICE (Korea Institute for Curriculum and Evaluation at www.kice.re.kr)

Table 2 Key features of the test instrument

Skills	Time	Item type1	Item type2	Scoring
Grammar, vocabulary	Max. 2 h	Multiple-choice	Grammaticality judgment, gap-fill for vocabulary	Same point (no weighted scoring)

II). The fine-grained diagnostic score report was provided around the 13th week of the semester. Then, those who received the score report were given 1 week to prepare for another test that assessed their progress with the nine attributes. The two tests were virtually identical, because only proper nouns were replaced and a few words from a couple of distractors were removed, which did not affect the meaning and answer key (for sample items, see APPENDIX I).

The second test was given on the 14th week of the same semester, with the interval between the tests being 6 weeks. However, this time lapse cannot be considered remedial intervention because participants were not instructed to relearn the attributes assessed in the first test during the period. Rather, the 6-week interval was given in order to minimize the practice effect of the test. According to the literature (Brown et al. 2008), if a test is repeated sometime between 3 and 6 weeks, the learning effect from practice or recall of initial response is not significant.

Out of the 239 EFL learners, 217 students took the second test. After the second test, they were asked to complete a survey. The survey questionnaire was designed to ask about the overall effectiveness of the CDA feedback as well as the usefulness of each component of the CDA score report. Of those who received the report and took the test, the number of students who actually read the report and responded to the questionnaire was only 205 learners (Fig. 1).

In order to capture the description of the performance of the entire group, the test-taker responses of 474 students were calibrated with methods of the classical test theory (CTT) and the 2-parameter item response theory (2PL IRT) model. Before discussing the analysis results by the CDA method, characteristics of the performance data are provided in terms of the CTT and 2PL IRT. These methods will help interpret the

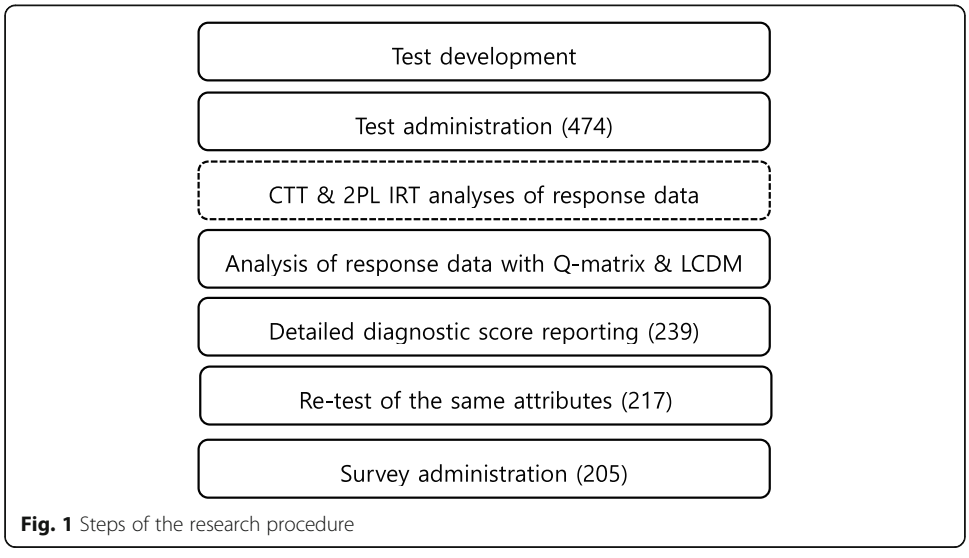

Fig. 1 Steps of the research procedure

whole picture of the data in terms of the well-known indices such as test difficulty and discrimination.

Table 3 provides the basic item statistics based on the analysis by the CTT. According to Cangelosi's (1990) criteria, the average item difficulty of 0.636 is within the appropriate item facility range (0.25–0.75) and thus deemed as acceptable. The mean discrimination of the test items (0.389) can be considered "good with possibilities for improvement" according to Ebel and Frisbie's (1986) recommendations. According to these criteria, excellent item discrimination value is 0.39 or higher.

The analysis results by the two-parameter IRT are given in Table 4. The mean difficulty of − 0.561 indicates that the test was slightly on the easy side, as an item with a difficulty (b-parameter, which is usually from − 3 to 3) of 0 is usually considered to be of average difficulty. The mean discrimination of 1.009 is considered moderate according to Baker's interpretation guide, which says discrimination values between 0.65 and 1.34 are moderate and values between 1.35 and 1.69 are high (Baker 2001). Lastly, the Cronbach alpha for the reliability of the test was 0.78, which is higher than 0.7 and thus acceptable according to the commonly accepted rule of thumb for describing the internal consistency of a test.

Cognitive diagnostic model and the software

The cognitive diagnostic model employed in the study is the Log-linear cognitive diagnosis model (LCDM) which was proposed recently by Henson et al. (2009). As a general model, the strength of the LCDM is that it deals with each item more flexibly and identifies the optimal model for each item even in the same test data. The model was also found to be the best fit to a set of ESL grammar test data in a recent study (Yi 2017). The software program used to estimate this model was Mplus (Muthén & Muthén, 2017), which is a latent variable modeling program for Windows and used for a wide variety of statistical analyses, not only for cognitive diagnostic modeling.

Results

Reliability of the parameter estimates (model fit to the data)

Before looking into the results of the analysis on test-taker ability, model fit indices for the analysis must be examined to see if the numerical findings are reliable. One quick measure of the model fit to the data is investigating the significance of the parameter estimates of the attributes that are defined for each test item. The LCDM takes a similar approach to ANOVA and the attributes in each item are statistically tested with intercepts, main effects, and interaction parameters. In order to prove that the attributes identified for each item are valid, the two types of parameters except the intercepts must be statistically significant.

Table 3 Learner response data analysis by CTT

No. of items	Mean difficulty	Mean discrimination	Measurement error	Examinees
35	0.636	0.389	2.486	474
Mean score	Standard deviation	Min. score	Max. score	Median
22.276	5.982	4	34	23

Table 4 Learner response data analysis by IRT

	Mean difficulty	Mean discrimination	Test-taker ability
Mean	−0.561	1.009	0.078
Standard deviation	1.005	0.413	0.905

A total of 43 main effects and interaction parameters were generated in this study and 37 parameters turned out to be statistically significant with the alpha level at 0.05 in both the first and second tests (35 main effects and 2 interaction parameters). Table 5 shows the insignificant parameters representing attributes and interactions between the attributes. Two parameters for Attributes 5 (S-V agreement) and 6 (relative pronoun) were insignificant in items 18–20. Despite the insignificance of these parameters, however, the initial estimation results were maintained for the following reasons: First, out of the five items coded for Attribute 5 (S-V agreement) and four items coded for Attribute 6 (relative pronoun), three parameters for Attribute 5 and two parameters for Attribute 6 were still significant. Second, in the Tests of Model Fit section of the Mplus output, both the Pearson Chi-Square Test of Model Fit and the Likelihood Ratio Chi-Square (G statistic) yielded the p-value of 1.0000. This means that there are statistically insignificant discrepancies between the actual observed distribution of respondents and what we should see under our model. In other words, the model-generated attribute profile of the test-takers would predict the observed response patterns well, which means the statistical model fits the data (Rupp et al. 2010). This translates into the fact that the analyses of examinees' mastery status of each of the nine attributes will be reliable.

Results of cognitive diagnostic analyses

The main components of the typical output of cognitive diagnostic assessment are the mastery probability of each attribute required to endorse each item and the learner profile that is based on the mastery probability of the attributes. The LCDM output produces a separate file that identifies the learner profile which can also give the mastery probability of each subskill of the test item. The cognitive diagnostic assessment literature defines the status of mastery or non-mastery with the probability, i.e., probability smaller than 0.4 as non-mastery, from 0.4 to 0.6 as unidentified, and from 0.6 to 1 as mastery of the attribute. Table 6 shows the mean mastery probabilities of all of the respondents (474) pertaining to each attribute required in the test. It also gives the proportions of test-takers who mastered or did not master the attributes.

Some notable findings are observed in the table. First, the knowledge state of the majority of test-takers are determined either as non-mastered or mastered regarding each attribute. That means only a small proportion of the test-takers are in the gray area

Table 5 Insignificant parameters of main effects and interactions

	First test	Second test
Main effect parameters	Item 18 Attribute 5 (SV agreement) Item 19 Attribute 6 (Relative pron.) Item 20 Attributes 5 and 6	Item 18 Attributes 5 and 6 Item 19 Attribute 6 Item 20 Attribute 5
Interaction parameters	Item 18 interaction Item 19 interaction	Item 18 interaction Item 19 interaction

Table 6 Mean mastery probabilities and 474 examinee percentages of attributes

	Mean	Non-mastery (0–0.4)	0.4–0.6	Mastery (0.6–1)
1. Simple present/ progressive	0.589	157 (33.1%)	64 (13.5%)	253 (53.4%)
2. Zero conditional	**0.287**	341 (71.9%)	6 (1.3%)	127 (26.8%)
3. Simple past/ progressive	0.489	220 (46.4%)	38 (8.0%)	216 (45.6%)
4. Participles	0.714	125 (26.4%)	11 (2.3%)	338 (71.3%)
5. Subject-verb agreement	0.557	177 (37.3%)	60 (12.7%)	237 (50%)
6. Relative pronoun	0.584	133 (28.1%)	50 (10.5%)	291 (61.4%)
7. Used to	**0.380**	284 (59.9%)	25 (5.3%)	165 (34.8%)
8. Vocabulary	**0.807**	78 (16.5%)	12 (2.5%)	384 (81.0%)
9. Expressions	**0.755**	96 (20.2%)	25 (5.3%)	353 (74.5%)

where their skill mastery is not clearly identified. Second, the mean mastery probability of the attributes follows the order of the attributes, 8 > 9 > 4 > 1 > 6 > 5 > 3 > 7 > 2, which is interpreted as vocabulary being the least difficult and "used to" and zero conditional the most difficult attributes to master.

One of the points of focus of this research is to examine the difference between the two tests administered before and after giving the cognitive diagnostic score reports. Out of 474 learners who took the test, 217 students took the re-test after being given the diagnostic score report. Tables 7 and 8 provide the mean probabilities of the nine attributes defined for the test. The tables also inform us of the proportion of the non-mastery and mastery learners of these sub-skills.

Important observations in these tables are the change in each mean mastery probability and the proportion of three examinee groups, i.e., non-mastery, undefined, and mastery groups. What is notable when comparing the two tables (Tables 7 and 8) is that the mean mastery probabilities of the two italicized attributes (zero conditional and simple past/progressive) somewhat decreased after the students were provided with diagnostic information on the score report. In order to see if these increases or decreases in the mastery probabilities are statistically meaningful, paired dependent t-tests were conducted and the results are as follows in Table 9. The percentages of three respondent groups (non-mastery, undefined, and mastery) indeed reflect these changes in the mean mastery probabilities. That is, the proportion of non-mastery students drops and that of mastery learners rises as the mean probabilities go up, and vice versa (Fig. 2).

Table 7 Mean mastery probabilities and 217 examinee percentages on the first test

	Mean	Non-mastery (0–0.4)	0.4–0.6	Mastery (0.6–1)
1. Simple present/ progressive	0.632	64 (29.5%)	25 (11.5%)	128 (59.0%)
2. *Zero conditional*	0.340	144 (66.4%)	3 (1.4%)	70 (32.3%)
3. *Simple past/ progressive*	0.539	87 (40.1%)	17 (7.8%)	113 (52.1%)
4. Participles	0.749	50 (23.0%)	5 (2.3%)	162 (74.7%)
5. Subject-verb agreement	0.446	112 (51.6%)	23 (10.6%)	82 (37.8%)
6. Relative pronoun	0.538	83 (38.2%)	16 (7.4%)	118 (54.4%)
7. Used to	0.317	145 (66.8%)	8 (3.7%)	64 (29.5%)
8. Vocabulary	0.860	26 (12.0%)	3 (1.4%)	188 (86.6%)
9. Expressions	0.794	36 (16.6%)	8 (3.7%)	173 (79.7%)

Table 8 Mean mastery probabilities and 217 examinee percentages on the second test

	Mean	Non-mastery (0–0.4)	0.4–0.6	Mastery (0.6–1)
1. Simple present/ progressive	0.677	60 (27.6%)	11 (5.1%)	146 (67.3%)
2. *Zero conditional*	0.323	148 (68.2%)	3 (1.4%)	66 (30.4%)
3. *Simple past/progressive*	0.472	100 (46.1%)	26 (12.0%)	91 (41.9%)
4. Participles	0.762	45 (20.7%)	7 (3.2%)	165 (76.0%)
5. Subject-verb agreement	0.588	71 (32.7%)	34 (15.7%)	112 (51.6%)
6. Relative pronoun	0.612	49 (22.6%)	24 (11.1%)	144 (66.4%)
7. Used to	0.389	129 (59.4%)	12 (5.5%)	76 (35.0%)
8. Vocabulary	0.873	21 (9.7%)	5 (2.3%)	191 (88.0%)
9. Expressions	0.831	27 (12.4%)	12 (5.5%)	178 (82.0%)

Out of the changes in the two attributes whose means declined unlike was expected, the decrease in the *zero conditional* attribute was not statistically significant ($p = 0.526$), which means the students' understanding of *zero conditional* did not show any improvement or deterioration after the cognitive diagnostic score reporting. On the contrary, their mastery of *simple past/ progressive* worsened as the mean mastery probability declined (0.539 to 0.472) and the decrease proved statistically significant. The other seven attributes also showed progress after the participants were informed of their deficiencies in the score report, except *vocabulary* whose growth in the mean did not prove statistically significant ($p = 0.229$). The reasons for these results will be addressed in the discussion section.

Results of the survey

The survey asked six questions, among which two questions were open-ended asking for a short answer (see Appendix III). The second question in the survey sought a response about the usefulness of the information provided in the cognitive diagnostic score report. It asked which information was the most helpful out of the three types: (1) the table that shows each correct (O) and incorrect (X) item and the attribute(s) of each item; (2) the mastery profile produced by the cognitive diagnostic assessment, informing their strengths and weaknesses and thus giving directions for future studies; and (3) a table that provides the mastery probability of each attribute estimated by the cognitive diagnostic assessment. Table 10 shows that out of the 205 students who

Table 9 Comparisons of mean mastery probabilities in the first and second tests

	Mean/SD (1st test)	Mean/SD (2nd test)	t-statistic (df)	p-value
1. Simple present/ progress	0.632/0.364	0.677/0.353	−2.027 (216)	0.044*
2. *Zero conditional*	0.340/0.405	0.323/0.381	0.636 (216)	0.526
3. *Simple past/progressive*	0.539/0.389	0.472/0.366	2.970 (216)	0.003*
4. Participles	0.749/0.387	0.762/0.376	−2.464 (216)	0.015*
5. Subject-verb agreement	0.446/0.356	0.588/0.359	−5.621 (216)	0.000*
6. Relative pronoun	0.538/0.338	0.612/0.317	− 2.801 (216)	0.006*
7. Used to	0.317/0.403	0.389/0.415	− 2.740 (216)	0.007*
8. Vocabulary	0.860/0.292	0.873/0.278	− 1.207 (216)	0.229
9. Expressions	0.794/0.327	0.831/0.305	−3.020 (216)	0.003*

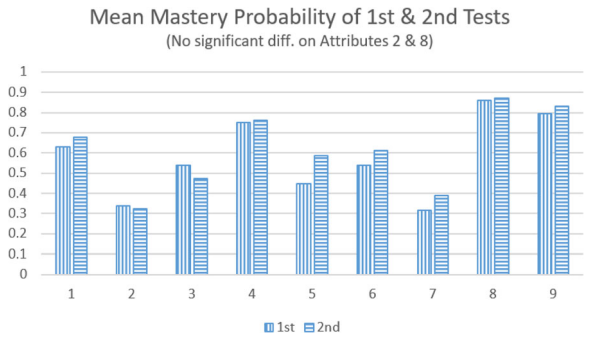

Fig. 2 Mean mastery probability of the first and second tests. *Note.* 1. simple present/ progressive; 2. zero conditional; 3. simple past/progressive; 4. participles; 5. subject-verb agreement; 6. relative pronoun; 7. used to; 8. vocabulary; 9. expressions

actually read the score report and responded to the particular question in the survey, 1–2-3 and 2–1-3 garnered the greatest number of choices, accounting for almost 60% of the responses. It means that respondents perceived the first set of information (in a table showing all the correct and incorrect items and their required attributes) and the learner profile yielded by the cognitive diagnostic analysis as the most useful and the third type of information (the estimated mastery probability of each attribute) the least helpful. However, in 31.2% of the responses, the third type of information, the mastery probability of attributes, took the second place (i.e., 2–3-1 and 1–3-2) in terms of its utility.

In response to the question asking about the language learning area in which this type of detailed score report will be the most helpful, learners answered that it will be very useful in grammar, as seen in Table 11, followed by vocabulary and writing.

In their responses to the two short answer questions (questions 3 and 4), clear patterns were discerned with the most frequent answers. Tables 12 and 13 summarize the proportion of the most frequent reactions by the participants.

Besides these repeated answers, each of the following responses was given by one student each, which took up about 3.3% of all of the responses.

1. No deficient attribute was reported so the score report did not help.
2. It helped me make an organized plan for studying
3. More intuitive information would help.
4. The report motivated me.
5. It informed me about my priorities in studying.

What is very prominent in these answer patterns is that a negative reaction (Part3 was not very helpful) was the most frequent, accounting for about 40% of the responses. Another interesting opinion is that they thought it would have helped more if they had been able to see the actual test items with the probabilities. Other answers

Table 10 Participant perception of the usefulness of the three types of information in the score report

Choices	1–2-3	2–1-3	2–3-1	1–3-2	3–2-1	3–1-2
Frequency	61 (29.8%)	60 (29.3%)	41 (20%)	23 (11.2%)	16 (7.8%)	4 (1.9%)

Table 11 Participant perception of the usefulness of CDA score report for each language area

	Q1	Q6–1 (S)	Q6–2 (W)	Q6–3 (R)	Q6–4 (L)	Q6–5 (V)	Q6–6 (G)
Mean	3.908	3.085	3.690	3.587	3.131	3.881	4.355
SD	0.905	1.095	0.814	0.896	1.608	0.789	0.751

Note. *S* speaking; *W*, writing; *R*, reading; *L*, listening; *V*, vocabulary; *G*, grammar

included "Part2 is more helpful," "non-mastery probabilities (or ability indices) rather than mastery probabilities would be more helpful," and "Part3 was like part of the CSAT score report, which was why I didn't like it."

Discussion and conclusion

The cognitive diagnostic assessment method has the potential to be a powerful tool that can bridge the gap between teaching and testing. Amid the prior research background in which only a small proportion of cognitive diagnostic assessment studies actually provided feedback to students or teachers, despite the promises that CDA can be used to inform teaching and provide detailed feedback about strengths and weaknesses, the current study attempted to fill this gap in the research.

Answering the first research question of whether the fine-grained information in the CDA score report helped with the learners' progress in learning, the t-test results between the first and second tests uncovered that students made statistically significant progress on six attributes out of nine. However, they did not show any improvement or regress on the attribute, *zero conditional*, while their mastery of *simple past* vs. *progressive* worsened as the mean mastery probability showed a statistically significant decline (0.539 to 0.472) even after the diagnostic score reporting. It is obvious that these two attributes are more challenging to master than the other attributes. Particularly, their knowledge of distinctive use of *simple past* vs. *past progressive* weakened as time passed by and was not reinforced by self-directed learning informed by the detailed CDA score reporting. For these challenging subskills, more effective reporting strategies seem clearly necessary, getting beyond a minimalist approach to feedback in order to make feedback an "inherent catalyst" for self-regulated learning activities (Butler and Winne 1995).

The remaining seven attributes also showed improvement after their score report informed their deficiencies, except *vocabulary* whose growth in the mean was not supported statistically ($p = 0.229$). We could make an assumption that the reason for this is that, unlike other language subskills in the test, learning words is not studying some kind of pattern or rule, which means mastering words from the previous test does not guarantee endorsing other vocabulary items in the subsequent test if the words are not the same in the two tests.

Table 12 Reasons for perceiving Part2 (mastery profile informing strengths and weaknesses) as useful

	1	2	3	4	5	Total
Frequency	113 (74.8%)	10 (6.6%)	9 (6.0%)	8 (5.3%)	6 (4.0%)	151

Note. 1. gives a chance to know my weaknesses and remedy them; 2. helps me study more efficiently (use time more efficiently); 3. explains my deficiencies very specifically (not just the total score); 4. provides a direction for studying; 5. do not know

Table 13 Reasons for perceiving Part3 (a table providing the mastery probability of each attribute) as useful

	1	2	3	4	5	6	Total
Frequency	45 (39.1%)	40 (34.8%)	13 (11.3%)	5 (4.3%)	4 (3.5%)	4 (3.5%)	115

* Note. 1 does not really help (including "does not inform about what to study"); 2. shows my weaknesses/ what skills to strengthen; 3. more specific, accurate, objective indices; 4. shows my global knowledge state; 5. gives confidence, motivation to increase the probabilities; 6. need to see test items

With regard to the second research question of whether learners perceived the information in the cognitive diagnostic score report as being helpful, their response overall was on the positive side with a mean of 3.91 on the 5-point Likert scale. As for the more specific question of which part of the score report was the most helpful, the majority of the participants (59.1%) thought of the first set of information (in a table showing all the correct and incorrect items and their required attributes) and the learner profile yielded by the cognitive diagnostic analysis as more useful than the third type of information (the estimated mastery probability of each attribute). However, in 31.2% of the responses, the third type of information, the mastery probability of attributes, took the second place (i.e., 2–*3*-1 and 1–*3*-2) in terms of its usefulness. One possible interpretation of these outcomes is that unlike the first and second set of learner information, learners felt mere mastery probabilities of mastering subskills did not give any interpretation and thus any guidance for future remedial studies.

Their responses to open-ended questions are in line with these results. Looking at Tables 12 and 13, about three quarters (74.8%) viewed the learner profile that defined their weaknesses and strengths as useful, while only a little over one-third (34.8%) of the respondents regarded the mastery probability of each subskill as informative. Moreover, about 40% of the students perceived that knowing their mastery probability of each language attribute did not really help or inform what or how to study to remedy their deficiencies. In relation to this reaction, an interesting opinion seems worth noting, as four participants held that it would have been more helpful if they had seen the actual test items in the score report. In school exams which do not usually reuse the same test items, this idea seems feasible. However, in some large-scale commercial tests that take test items out of item pools and, thus, might use the same items in later tests, not the actual items but alternative sample items could be used to meet the needs in this case. All in all, these survey results directly indicate what type of information that can be presented from cognitive diagnostic modeling will better inform instruction and guide self-directed learning.

In response to the question asking about the language learning area in which this type of detailed score report would be the most helpful, learners answered that it will be very useful in grammar, as seen in Table 11, followed by vocabulary and writing. One could think of two possible reasons for this opinion: first, grammar attributes can be more clearly identified than the attributes of other skills such as reading or listening. Second, because many test items employed in this research required grammatical attributes, it was very likely that they affected their judgment about the optimal skills for cognitive diagnostic modeling.

The current study has some limitations which suggest some avenues for future research. First, a contrastive research design involving a control group might reveal the effectiveness of the remedial learning more lucidly. A control group which receives the

typical kind of unidimensional score report with the total score (and with the correct/incorrect item table) might more clearly highlight the beneficial effects of the fine-grained diagnostic score report. Since there was no control group and the same test was administered twice in the study, one could wonder if part of the improvement of mastery probability on the second test may be attributable to practice effect. Because the time interval between the tests was 6 weeks, any learning effect from practice or recall of initial response probably did not affect the test scores significantly (Brown et al. 2008). However, considering the usual intermission for test-retest reliability is 8 weeks, it is also true that a minimum lapse of 8 weeks would have attenuated the possible practice effect.

Alternatively, parallel or equivalent test forms could be employed in the future research in order to exclude the chance of practice effect on the test-takers. In that case, the two test forms must be validated as truly alternate so that the scores of two tests are in agreement. Pertaining to the research design of a replicated study, giving learners more time for self-regulated learning informed by the CDA score report would lead to more obvious improvement in their knowledge state. Nevertheless, considering that the students were only given 1 week to compensate for or remedy their deficiencies in this study, the resultant progress shown in the second test was not minor, yielding statistically significant change.

Second, another interesting point of inquiry would be to examine the differing effects of CDA-based detailed score reporting on different proficiency levels. What is intuitive is that a learner must be proficient enough to be able to benefit from the fine-grained CDA score report and take the initiative to engage in self-directed learning. However, there is not enough solid evidence yet garnered from scientific research on the relationship between the level of the learner and the effectiveness of the CDA score report, especially in language testing. That is, the matter of what proficiency groups could take the most advantage of such detailed score reporting has not received sufficient research interest. Thus, it will be worth the research efforts to look into the optimal level of students for CDA and/or the type and amount of diagnostic information that cognitive diagnostic modeling can provide for different proficiency groups.

Lastly, one thing that deserves our careful attention from their responses to the two short answer questions in the survey was that a negative reaction to Part3 (not very helpful) recurred the most frequently, taking up about 40% of the responses. It may imply that the CDA score reporting should go so far as to provide remedial learning activities. For example, for grammar attributes, the score report could suggest different levels of tasks that range from consciousness raising, substitution drills, sentence-level composition using the target grammar points, to even controlled/guided writing. The report card could further guide learners to openly accessible online learning materials or mobile applications, to name a few. To this end, a string of research efforts should ensue to investigate how to design, develop, and provide remedial learning activities in the CDA score reporting.

This research shows us that the type of information which provides individual students' learner profile can be significantly more helpful than the information that traditional assessment can provide. Despite the concern that the improvement of skill mastery probability discovered in the second test may be compounded with a construct-irrelevant factor, the findings about the learner perception of the overall

CDA score reporting and each component of the score report still make this study significant. The survey results of the study clearly reveal what type of information from cognitive diagnostic modeling will inform instruction and guide self-directed learning more effectively.

Cognitive diagnostic assessment enables detailed score reporting and can provide profiling of individual students even when the test-taker group is large-scale. Nonetheless, as one Korean proverb goes, even numerous precious beads do not become a jewel before they are threaded. Whatever psychometric efforts are put into making the cognitive diagnostic modeling reliable and the resultant learner skill profiles dependable, if the end-users in the educational settings are not fully aware of the usefulness of such output and/or these student profiles are not put into actual use for remedial self-studies or classroom instruction, the overall endeavor will not bear any fruit. It is hoped that the present study motivates more research on the use of CDA-generated score reports for follow-up instruction and, particularly, remedial self-regulated learning.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s40468-021-00127-4.

Additional file 1. Appendices

Acknowledgements

The author thanks all the students and teachers who helped in collecting the data.

Author's contributions

One author conducted the research and wrote this paper. The author read and approved the final manuscript.

Author's information

The author got a Ph.D. degree in language testing from the University of Illinois at Urbana-Champaign. The author had several papers published on cognitive diagnostic assessment in journals including Language Testing and Applied Measurement in Education, which are indexed by SSCI.

Funding

The author received funding for this research from Sangmyung University.

Availability of data and materials

The datasets for the current study are available on reasonable request.

Declarations

Competing interests

The author declares that there are no competing interests.

Received: 1 March 2021 Accepted: 28 May 2021
Published online: 02 August 2021

References

Baker, F. B. 2001. The basics of item response theory, ERIC Clearinghouse on Assessment and Evaluation, http://echo.edres.org:8080/irt/baker/

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement*, *21*, 175–189.

Brown, G., Irving, E., & Keegan, P. (2008). *An introduction to educational assessment, measurement and evaluation*. Pearson Prentice Hall.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*, *15*(2), 119–157.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: a theoretical synthesis. *Review of Educational Research*, *65*, 245–281.

Cangelosi, J. S. (1990). *Designing tests for evaluating student achievement*. New York: Longman.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. In *Unpublished doctoral dissertation*. Urbana-Champaign, Urbana, IL: University of Illinois at.

Jang, E. E., Dunlop, M., Park, G., & van der Boom, E. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing*, *32*(3), 359–383.
Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing*, *32*(3), 299–316.
Li, H. (2011). *Evaluating language group differences in the subskills of reading using a cognitive diagnostic modeling and differential skill functioning approach*. University Park, PA: Unpublished doctoral dissertation, Pennsylvania State University.
Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide (Version 8) [Computer software and manual]*. Los Angeles, CA: Muthén & Muthén.
Roberts, M. R., & Gierl, M. J. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice*, *29*, 25–38.
Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: The Guildford Press.
Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary, measurement. *Interdisciplinary Research and Perspectives*, *16*(1), 1–17.
Sun, Y., & Suzuki, M. (2013). Diagnostic assessment for improving teacher practice. *International Journal of Information and Education Technology*, *3*, 607–610. https://doi.org/10.7763/IJIET.2013.V3.345.
van der Boom, E., & Jang, E. E. (2018). The effects of holistic diagnostic feedback intervention on improving struggling readers' reading skills. *The Journal of Teaching and Learning*, *12*(2), 54–69.
Wu, H.-M. (2018). Online individualized tutor for improving mathematics learning: A cognitive diagnostic model approach. *Educational Psychology*. https://doi.org/10.1080/01443410.2018.1494819.
Yi, Y.-S. (2017). In search of optimal cognitive diagnostic model(s) for ESL grammar test data. *Applied Measurement in Education*, *30*(2), 82–101.

Publisher's Note