AERA Open
January-December 2021, Vol. 7, No. 1, pp. 1–15
DOI: 10.1177/23328584211060804
Article reuse guidelines: sagepub.com/journals-permissions

Does Diagnostic Feedback Promote Learning? Evidence From a Longitudinal Cognitive Diagnostic Assessment

Fang Tang
Peida Zhan
Zhejiang Normal University

Assessment for learning emphasizes the importance of feedback to promote learning. To explore whether cognitive diagnostic feedback (CDF) promotes learning and whether it is more effective than traditional feedback in promoting learning, this study conducted a quasi-experiment by utilizing a longitudinal cognitive diagnostic assessment to compare the effect of three feedback modes on promoting learning, including CDF, correct–incorrect response feedback (CIRF), and no feedback. The results provided some evidence for the conclusion that CDF can promote students' learning and is more effective than CIRF in promoting learning, especially in more challenging areas of knowledge.

Keywords: *cognitive diagnosis, feedback, longitudinal cognitive diagnostic assessment, assessment for learning*

In educational and psychological research, feedback has been examined for its ability to inform, identify, and correct errors (Whyte et al., 1995). Assessment for learning (Wiliam, 2011) emphasizes the importance of feedback to promote learning. Specifically, it considers that students' learning status should be promptly evaluated, and corresponding feedback is given in the teaching process to promote learning.

With the development of psychometrics, cognitive diagnostic assessments (CDAs; Leighton & Gierl, 2007) or learning diagnostic assessments (Zhan, 2020a, 2020b), which objectively quantify students' current learning status of knowledge and skills (collectively known as latent attributes) and provide the corresponding cognitive diagnostic feedback (CDF), have drawn increasing interest. In CDAs, a latent attribute can either correspond to a concrete knowledge point or refer to a more abstract latent construct. CDAs offer researchers a framework of formative assessments that are capable of providing educators with detailed information about student mastery status (e.g., mastery or nonmastery) of latent attributes in a given subject area, with examples including (a) fraction subtraction (Tatsuoka, 1983), (b) rational number operations (Tang & Zhan, 2020), (c) voltage and current (Ohm's law) (Zhan et al., 2019), (d) geometric sequences (Shute et al., 2008), (e) buoyancy (Gao et al., 2020), and (f) proportional reasoning (Tjoe & de la Torre, 2013). Researchers have applied CDAs more broadly in the psychological and behavioral sciences, such as for learning mode (H. Li et al., 2020), spatial rotation (S. Wang et al., 2018), problem-solving competence (Zhan & Qiao, 2020), situational judgment (Sorrel et al., 2016), and psychological disorders (Templin & Henson, 2006).

Currently, several cognitive diagnosis models (CDMs) or diagnostic classification models (for review, see von Davier & Lee, 2019)—such as the most popular deterministic-inputs, noisy "AND" gate (DINA) model (Junker & Sijtsma, 2001) and its generalization (de la Torre, 2011)—have been proposed to provide theoretical support for CDAs. CDMs are a family of restricted latent class models that model relationships between discrete latent attributes and observed item responses. CDMs have the potential to provide richer individual-level diagnostic information (i.e., CDF) than traditional psychometric models (e.g., item response theory models; Ma & de la Torre, 2020). A previous survey has shown that teachers are eager to obtain detailed individual-level diagnostic information and corresponding remedial strategies (Huff & Goodman, 2007).

Figure 1 shows an example of how CDF can be presented in the form of a diagnostic feedback report, which was used in this study. This report's data come from a test containing 18 dichotomous scoring items, which was developed to diagnose whether students were masters or nonmasters of six latent attributes related to knowledge in rational number operations (Tang & Zhan, 2020). This report provides feedback on the correct or incorrect item responses as well as feedback on the mastery status of each attribute. Specifically, this student correctly answered nine out of 18 items, and he or she has a high probability/certainty of mastering the latent attribute associated with A1 (97%), A2 (94%), and A4 (76%). This student is almost certainly not a master of A3

Diagnostic Feedback Report

Student ID:

Part I. Your Answers

Multiple-Choice Items

Item	1	2	3	4	5	6	7	8	9	10	11	12
Answer Key	D	D	D	C	D	D	C	B	B	A	B	B
Your Answer	D	D	D	B	C	B	C	A	B	A	B	A
Accuracy	√	√	√	×	×	×	√	×	×	√	√	×

Calculation Items

Item	13	14	15	16	17	18
Correct Answer	21	–1	160	8/75	16	–13
Your Answer	21	–1	160	16/75	9	
Accuracy	√	√	√	×	×	×

Score: 9

Part II. Your Mastery Status of Knowledge Points

Knowledge Point	Mastery or Non-mastery	Probability*
A1: Rational Numbers	Yes	97%
A2: Related Concepts of Rational Numbers	Yes	94%
A3: Axis	No	11%
A4: Addition and Subtraction of Rational Numbers	Yes	76%
A5: Multiplication and Division of Rational Numbers	Yes	57%
A6: Mixed Operation of Rational Numbers	No	33%

Note, * means the degree of certainty of classification.

Attribute Mastery Status: (A1, A2, A3, A4, A5, A6) = (110110)

FIGURE 1. *Example of a cognitive diagnostic feedback card.*

and A6 because his or her mastery probability for those latent attributes is 11% and 33%, respectively. Finally, for A5, he or she has a probability of 57% of being a master, making his or her diagnosis on that latent attribute uncertain. It can be found that, for each student, in addition to the traditional correct–incorrect response feedback (CIRF) (Part I), CDF can provide additional information about his or her mastery status of latent attributes (Part I and Part II). The former is a simple item- or task-level feedback, while the latter introduces additional attribute-level feedback; namely, the latter provides more information than the former.

In the field of learning science, feedback can be divided into different levels according to the complexity of the information it provides (Whyte et al., 1995), such as (a) knowledge of response (KOR), which informs the respondent whether his/her response was correct; (b) knowledge of correct response (KCR), which informs the respondent of the correct response; and (c) knowledge of correct response plus

additional information (KCRI), which not only informs the respondent of the correct response but also details why the correct answer was correct or why the incorrect answer was not correct. At present, most feedback-related studies consider KCR to be the most basic feedback mode to promote learning (Bangert-Drowns et al., 1991; Butler et al., 2013; Kluger & DeNisi, 1996). Correspondingly, in this study, the CIRF (Part I in Figure 1) and the CDF (Part I and Part II in Figure 1) belong to KCR and KCRI, respectively.

However, adding more information to feedback does not necessarily promote learning, as we subconsciously assume. There is debate as to whether informational feedback (i.e., KCRI) is more effective than simple feedback (i.e., KOR or KCR) in promoting learning. Some researchers believe that informational feedback can promote students' learning (e.g., Corbett et al., 1997; Gibbons & Fairweather, 1998). Some studies even suggest that the more information feedback provides, the greater the effect on promoting learning (Dunn et al., 2012; Maddox et al., 2008). For example, Shute et al. (2008) compared the effect of KOR and KCRI in promoting learning and found that KCRI was more effective for student learning than KOR. By contrast, some researchers argue that increasing the complexity of feedback is detrimental to learners (e.g., Bangert-Drowns et al., 1991). Some studies on how feedback affects learning have found that KCRI has no significant advantage over KOR or KCR (e.g., Andre & Thieman, 1988; Kulhavy et al., 1985; Peeck, 1979; Whyte et al., 1995). For example, Butler et al. (2013) found no significant difference between KCR and KCRI in promoting students' performance on repeated questions, but that the latter is more beneficial in promoting students' performance on new inference questions.

For the existing studies related to feedback, in addition to the lack of a unified conclusion on whether increased information in the feedback is beneficial to promoting learning, some other issues might affect their conclusions. First, in most studies, the dependent variable used to reflect learning performance is the observed raw score rather than latent constructs, such as latent ability or cognitive attributes. The raw scores contain a lot of noise information, and the latent variable models can estimate the true abilities, which might be more informative. Second, some studies used tests with repeated measure design or discrete, unlinked tests. The former cannot exclude the practice/memory effect, and the latter cannot ensure the comparability of test results. Third, most studies used a pretest–posttest design involving only two time points, which may not reflect the interaction effect between feedback modes and test times. Studies that include more test time points can better map how the effect of feedback on learning changes over time. To this end, it may be better to use a longitudinal CDM to analyze the data from a longitudinal CDA involving more than two time points to explore the impact of feedback on promoting learning.

In contrast to cross-sectional CDAs, which fail to assess students' learning development (e.g., Chen, 2012; Liu et al., 2013), longitudinal CDAs evaluate students' latent attributes and identify their strengths and weaknesses over a period of time (Zhan, 2020a). The data collected from longitudinal CDAs have provided researchers with opportunities to develop models for learning tracks, which can be used to diagnose individual developmental trajectories over time and evaluate the effectiveness of CDF and corresponding remedial teaching. In recent years, several longitudinal CDMs have been proposed (for reviews, see Zhan, 2020b), such as the longitudinal higher order DINA (Long-DINA) model (Zhan et al., 2019), the latent transition analysis-DINA model (F. Li et al., 2016), and some extensions (e.g., Huang, 2017; Kaya & Leite, 2017; Madison & Bradshaw, 2018; Pan et al., 2020; S. Wang et al., 2018; Zhang & Chang, 2020). The results of these studies suggest that longitudinal CDMs as a methodology can diagnose each student's developmental trajectory.

Several studies have attempted to explore the effectiveness of CDF in promoting learning; however, their results do not seem to be satisfactory. For example, Wu (2019) compared the learning facilitation effects of online individualized remedial teaching and traditional group-based remedial teaching; the former utilizing the attribute-level feedback without KOR (i.e., only Part II in Figure 1), the latter utilizing the class-level information provided to teachers to illustrate students' mastery proportion of each attribute. The results indicated that, in remedial teaching, individualized remedial teaching is more effective than group-based remedial teaching. L. Wang et al. (2020) compared the learning facilitation effects of two types of individualized remedial teaching: explaining the meaning of students' nonmastered attributes (i.e., CDF) and narrating the correct problem-solving procedures (i.e., KCRI with task-level information). The results indicated that utilizing CDF in individualized remedial teaching is more effective in promoting learning than utilizing KCRI with task-level information. Ren et al. (2021) compared the learning facilitation effects of targeted and nontargeted intervention materials; the former included students' poorly mastered attributes (i.e., only Part II in Figure 1), and the latter contained all the attributes involved in the tests (i.e., control group with untargeted feedback). The results of this study indicated that providing attribute-level feedback is more effective in promoting learning than providing untargeted feedback. However, the findings cannot answer whether attribute-level feedback is more effective than any other traditional feedback (e.g., KOR or KCR). Overall, these existing studies have some commonalities. First, they all adopted the pretest–posttest design with parallel tests. Second, they all repeatedly used the cross-sectional CDM for data analysis in pretest and posttest. Third, they all performed statistical hypothesis testing only on the raw scores and not the latent variables. Fourth, since CDF (or attribute-level feedback without KCR) is a necessary but not sufficient condition

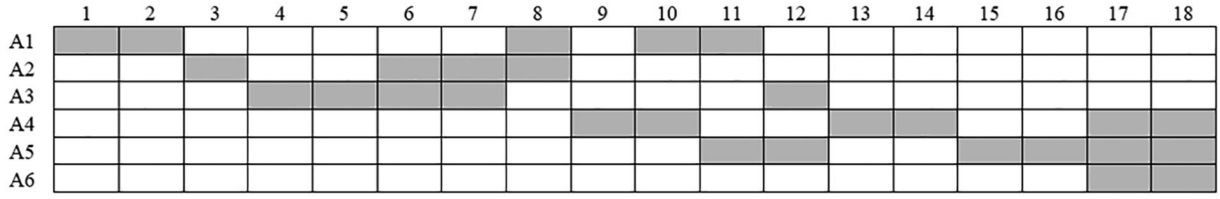

FIGURE 2. *Q-matrix of the instrument of rational number operations.*
Note. Blank means "0" and gray means "1"; A1 = rational numbers; A2 = related concepts of rational numbers; A3 = axis; A4 = addition and subtraction of rational numbers; A5 = multiplication and division of rational numbers; A6 = mixed operation of rational numbers.

for remedial teaching, Wu's (2019) and L. Wang et al.'s (2020) studies confounded the effects of CDF and remedial teaching on promoting learning. It can be argued that there is still a lack of research to explore the effectiveness of CDF (without remedial teaching) in promoting learning; more importantly, there is also a lack of research that has attempted to use longitudinal CDA to explore the effectiveness of feedback in promoting learning.

This study aims to explore the effectiveness of CDF in promoting learning by using longitudinal CDA. To this end, this study conducts a quasi-experiment (Hacker et al., 2000) by utilizing a longitudinal CDA to compare the effect of three feedback modes on promoting learning, including CDF, CIRF, and no feedback. As a quasi-experiment study, this study attempts to find evidence that CDF can promote learning and has a more significant facilitation effect on learning than CIRF. The results of this study help us understand whether adding attribute-level information to KCR helps promote learning, namely, whether KCRI with attribute-level information is more effective than KCR in promoting learning.

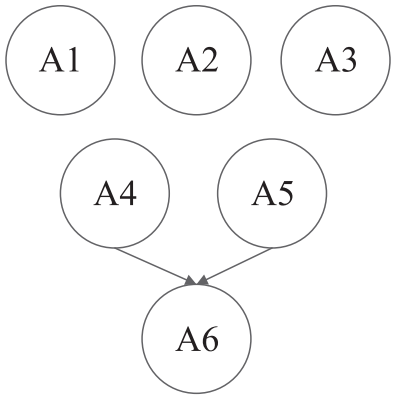

FIGURE 3. *Attribute hierarchy of the rational number operations.*
Note. A1 = rational numbers; A2 = related concepts of rational numbers; A3 = axis; A4 = addition and subtraction of rational numbers; A5 = multiplication and division of rational numbers; A6 = mixed operation of rational numbers.

Method

Instrument

A developed longitudinal CDA of rational number operations (Tang & Zhan, 2020) is used to carry out this study. As the repeated measure design is not always feasible in longitudinal educational measurement, especially for high-stakes tests, the developed instrument in this study used the design of the parallel test, which consists of three parallel tests, namely Formal Test A, Formal Test B, and Formal Test C. A part of the data underlying this longitudinal instrument has been used as an empirical example in previous methodological studies (e.g., Zhan, 2021; Zhan & He, 2021).

Since the development process is not the focus of this study, we only give a brief introduction to it, and more details about it can be found in Tang and Zhan (2020). The development process contains three main phases: (a) the Q-matrix (Tatsuoka, 1983) construction and item development, (b) the pilot test for item quality monitoring (the preliminary screening of the items is conducted mainly according to item difficulty and discrimination based on classical test theory), and (c) the formal test for test quality control (including the Q-matrix validation [de la Torre, 2008], reliability and validity testing [W. Wang et al., 2015], differential item functioning checking [Hou et al., 2014], and parallel tests checking). The results indicated that the reliability and validity of the developed instrument are high and that the three tests included in it meet the requirements of parallel tests. Overall, the developed instrument can be used for longitudinal CDA to track students' learning.

The three parallel tests have the same Q-matrix (see Figure 2), which means they contain the same number of items and the same number of attributes. Each test contains 18 dichotomous items, including 12 multiple-choice items and six calculation items. Note that, as parallel tests, each test contains a different set of items than the others. Six attributes of rational number operations are required: (A1) rational numbers, (A2) related concepts of rational numbers, (A3) axis, (A4) addition and subtraction of rational numbers, (A5) multiplication and division of rational numbers, and (A6) mixed operation of rational numbers. These six attributes followed a hierarchical structure (Leighton et al., 2004; see Figure 3), in which A1 to A3 are structurally independent, and A4 and A5 are both needed to master A6. A

reachability matrix[1] was contained in the Q-matrix, and at least two items assessed each attribute to make the Q-matrix complete and make the model identifiable (Ding et al., 2010; Gu & Xu, 2020).

Research Design and Analysis

As a classroom study or field experiment, in order not to interfere with the regular teaching progress and to ensure that the results have high ecological validity, this study adopted a quasi-experimental design. Researchers cannot and should not strictly control some extraneous variables that may affect the research results in the actual teaching situation. For example, students cannot be forbidden to learn other relevant knowledge outside the research time, to ensure minimal interruption of students' learning and for equity and humanitarian reasons. Hence, we assume that everything the students do outside of the research time is naturally occurring and in line with real-life situations. Although the quasi-experimental design cannot strictly control extraneous variables, it may be the best choice when studying behavior and cognition in the natural environment (Hacker et al., 2000). We need to be cautious when making causal inferences based on the results from quasi-experiments (Shadish et al., 2002).

This study followed a 3 × 3 design. The independent variables were the feedback mode (with three levels, namely the CDF, CIRF, and no feedback) and the test time (with three levels, namely the first time, the second time, and the third time), with the former as the between-subject variable and the latter as the within-subject variable. The dependent variable was the mathematical performance, reflected in the raw scores and model-based diagnostic results on the three parallel tests.

To select a suitable model, three longitudinal CDMs with different condensation rules (i.e., conjunctive, disjunctive, and compensatory condensation rules [Maris, 1995, 1999]) were used to fit this data (see Section A1 in the Supplemental Appendix, available in the online version of this article, for details). The results indicated that the simplified Long-DINA (sLong-DINA) model (Zhan et al., 2019) was the best fitting model among those evaluated.[2] The sLong-DINA model assumes that the correct item response requires that the attributes obey the conjunctive condensation rule, which was verified during the instrument development phase (Tang & Zhan, 2020).

There are two main data analysis strategies in longitudinal assessments: simultaneous estimation strategy using the longitudinal model and separated estimation strategy by repeatedly using the cross-sectional model (Zhan, 2020a). The former requires students to wait until all the tests are complete before an analysis of the results becomes available, whereas the latter can provide students with timely feedback after each test. Timely feedback on students' performance has generally been shown to support student learning (Kluger & DeNisi, 1996; Shute et al., 2008). However, considering that the simultaneous estimation strategy provides more accurate parameter estimation than the separated estimation strategy (Zhan, 2020a), in this study, the former is used, together with the sLong-DINA model, to obtain the final model-based diagnostic results. In contrast, the latter is used with the cross-sectional DINA model to provide students with timely feedback after each test (e.g., Wu, 2019).

TABLE 1
Assignment of Six Classes in Different Groups

Group	Teacher 1	Teacher 2
Diagnosis group	Class 1	Class 2
Traditional group	Class 3	Class 4
Control group	Class 5	Class 6

Note. The class number in the table is not the original class number.

Participants

The participants are students in the first semester of junior high school, Grade 7 (around 13 years old). By adopting convenience sampling, 289 students from six parallel classes of a similar academic level (according to students' admission scores) participated in this quasi-experiment.[3] The six classes were divided into three groups (i.e., two classes in each group), including the group with CDF (denoted as the diagnosis group), the group with CIRF (denoted as the traditional group), and the group with no feedback (denoted as the control group). These six classes have two math teachers (each teacher teaches three classes). To balance the impact of teachers' teaching styles on students, the three classes taught by each teacher were randomly assigned into three different groups. It should be noted that, as a classroom study, students were randomly assigned at the class level rather than the individual level. For ease of exposition, we renumbered the six classes, as shown in Table 1.

Students were excluded according to the following rules: missing more than five items in three tests, missing any of three tests, and not answering carefully (e.g., off-topic). After data cleaning, 90, 92, and 94 valid students (276 students in total) were collected in the diagnosis, traditional, and control groups, respectively. According to the results of previous studies (e.g., Zhan, 2020a; Zhan et al., 2019), such numbers of students meet the parameter estimation requirements of the sLong-DINA model.

Procedure

Figure 4 illustrates the research flowchart. We began by handing out the learning materials on knowledge of rational

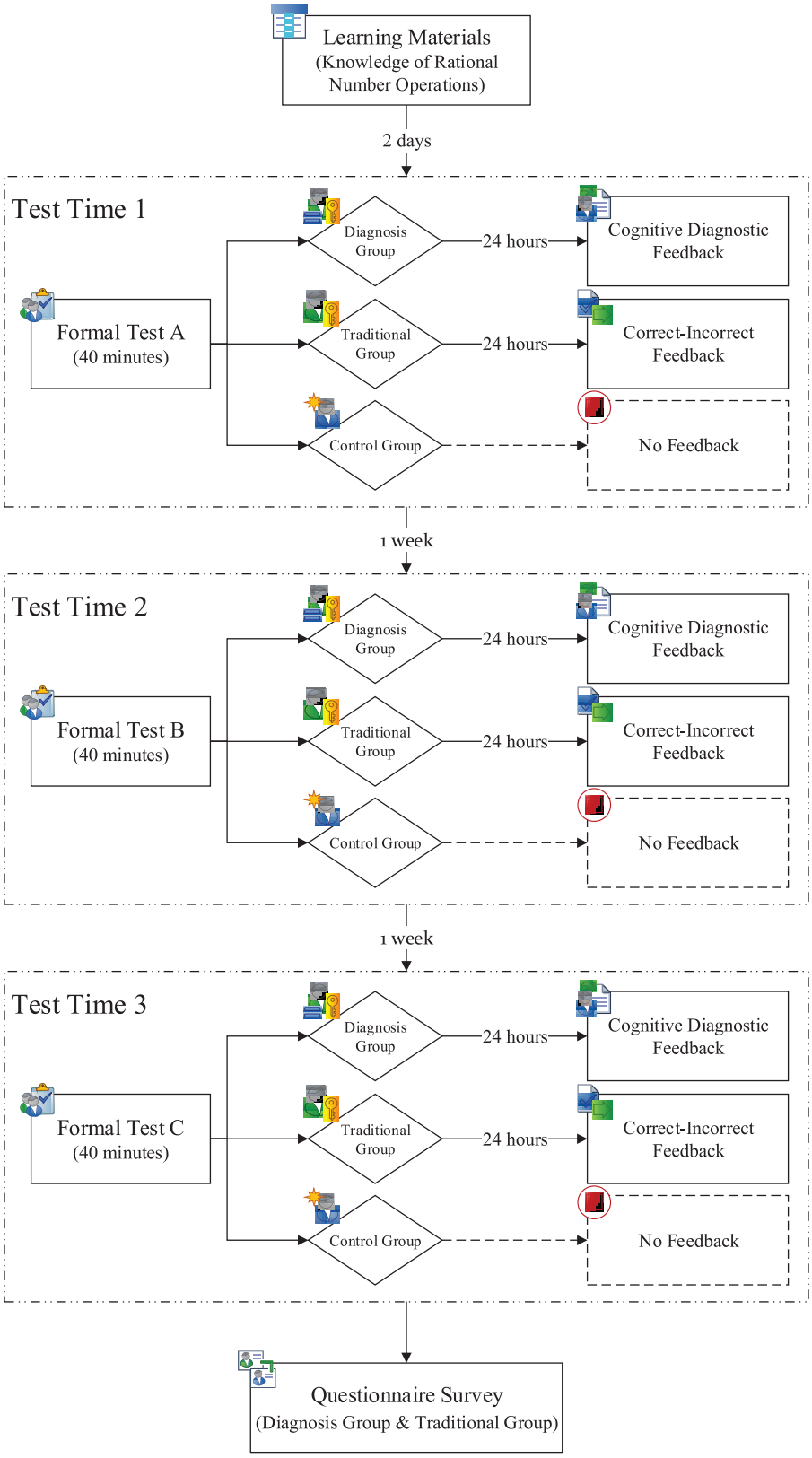

FIGURE 4. *Research flowchart.*

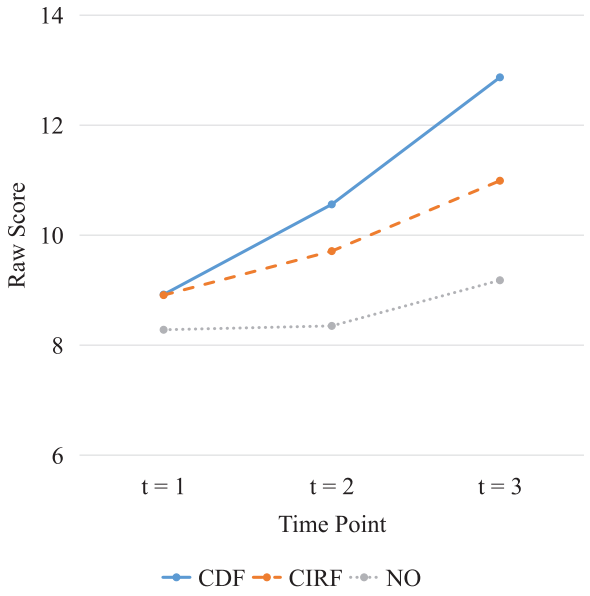

FIGURE 5. *Developmental trend of average raw scores in different groups over time.*
Note. t = test time point; CDF = cognitive diagnostic feedback; CIRF = correct–incorrect response feedback; NO = no feedback.

number operations to all three groups of students (a sample of the learning material for attribute one can be found in the online Supplemental Figure S1). Students were asked to study the learning materials and take the first test 2 days later. Then, at the first test time point, all three groups of students took Formal Test A. Twenty-four hours after the end of the first test (Butler et al., 2007), CDF (i.e., Part I and Part II in Figure 1) and CIRF (i.e., Part I in Figure 1) were provided to the students in the diagnosis group and the students in the traditional group, respectively. After receiving the feedback report, students voluntarily pursued self-remediation by using the initial learning materials. They did not receive any targeted instruction (e.g., remedial teaching) other than the feedback reports. In addition, to prevent students in the control group from proofreading their answers, they did not receive feedback or instruction until all three tests were completed.

Subsequent tests were performed 1 week after the initial test. The procedure for the second test time point and the third test time point was the same as that for the first test time point. All classes administered the three parallel tests in the same order. Finally, to understand how often students use feedback reports and their subjective feelings about whether the feedback was helpful to their learning, students in the diagnosis group and the traditional group were required to complete a questionnaire survey after all tests (see Section A2 in the online Supplemental Appendix). The results show that most students used their feedback reports and believed that feedback helped their learning.

In addition, for humanitarian reasons, the students in the control group and the traditional group also received their CDF after the experiment. No financial compensation was provided for completing the experiment.

Results

Figure 5 illustrates the developmental trend of the average raw scores of the three groups (more details can be found in the online Supplemental Table A2). The average raw scores of all three groups increased over time. The largest increase was seen in the diagnosis group, followed by the traditional and control groups.

A two-factor mixed-design analysis of variance (ANOVA) was performed, using feedback mode and test time as the independent variables and the raw score as the dependent variable. The results of the normality test, the homogeneity test of variance, and Mauchly's test of sphericity on the raw scores are reported in Section A3 in the online Supplemental Appendix. As shown in the results presented in Table 2, a significant interaction was found between feedback mode and test time. After performing a simple effect analysis (Tables 3 and 4), no significant difference was found between the three groups of students at the first test time; however, a significant difference between the diagnosis group and the control group was found at the second and third test times, and a significant difference between the three groups was found at the third test time. In addition, the diagnosis group and traditional group did not show significant differences until the third test time (i.e., after two times of feedback), indicating that the pretest–posttest design in most previous studies that included only one intervention may not fully demonstrate the advantages of CDF over CIRF.

Meanwhile, there were significant differences between the three time points for the cognitive and traditional groups, indicating that CDF and CIRF may have a continuous promoting effect on learning. By contrast, there were significant differences between the third test time and previous test times for the control group. Overall, it was clear that the learning progress of the traditional and control groups was not as good as that of the students in the diagnosis group.

Figure 6 illustrates the developmental trend of the average general latent abilities of the three groups (more details can be found in online Supplemental Table A3). Overall, the increase was the greatest in the diagnosis group, followed by the traditional group and control group. Compared with Figure 5, it can be seen that the developmental trend of the average general latent ability of each of the three groups of students was roughly but not the same as that of their average raw score.

A two-factor mixed-design ANOVA was performed, using feedback mode and test time as the independent variables and general latent ability as the dependent variable. The normality test, the homogeneity test of variance, and Mauchly's test of sphericity on the general latent ability are reported in Section A3 in the online supplemental appendix. As shown in the results in Table 5, a significant interaction

TABLE 2
Mixed-Design Analysis of Variance of Different Feedback Modes and Test Times for Raw Score

	df	*F*	*p*	η^2	Partial η^2	BF_{10}
Feedback mode	2	5.69	.004	0.04	0.04	9.141
Test time	1.84	493.58	<.001	0.54	0.64	>100
Feedback mode * Test time	3.67	70.74	<.001	0.16	0.34	>100

Note. The assumption of sphericity had been violated, the Greenhouse–Geisser correction adjusted the degrees of freedom (*df*) for the within-subject effect; the Bayes factor (BF_{10}) was calculated using the JASP software based on the Bayesian estimation. For example, BF_{10} = 9.141 means that the current data are 9.141 times more likely to occur under the alternative hypothesis (H1) being true than under the null hypothesis (H0) being true. Dienes (2014) suggested that BF_{10} less than 1, 1/3, and 1/10 represents weak, moderate, and substantial evidence for the H0, respectively. By contrast, BF_{10} greater than 1, 3, and 10 represents weak, moderate, and substantial evidence for the H1, respectively.

TABLE 3
Simple Effect Analysis of Feedback Modes for Raw Score

	t = 1				*t* = 2				*t* = 3			
	B	BF_{10}	T	D	B	BF_{10}	T	D	B	BF_{10}	T	D
CDF–CIRF	.989	0.161	1.000	1.000	.209	0.216	.455	.454	.003	17.759	.005	.005
CDF–NO	.344	0.243	.718	.717	.001	12.654	.004	.004	<.001	>100	<.001	<.001
CIRF–NO	.349	0.237	.723	.731	.043	0.537	.152	.151	.004	4.958	.020	.020

Note. CDF = cognitive diagnostic feedback; CIRF = correct–incorrect response feedback; NO = no feedback; *t* = test time point; B = *p* value of Bonferroni test; BF_{10} = Bayes factor; T = *p* value of Tamhane's T2 test; D = *p* value of Dunnett's T3 test.

TABLE 4
Simple Effect Analysis of Test Times for Raw Score

	CDF				CIRF				NO			
	B	BF_{10}	T	D	B	BF_{10}	T	D	B	BF_{10}	T	D
(*t* = 1)–(*t* = 2)	<.001	>100	.034	.034	<.001	>100	.560	.559	.466	0.148	.999	.999
(*t* = 1)–(*t* = 3)	<.001	>100	<.001	<.001	<.001	>100	.006	.006	<.001	>100	.469	.468
(*t* = 2)–(*t* = 3)	<.001	>100	<.001	<.001	<.001	>100	.143	.143	<.001	>100	.564	.563

Note. CDF = cognitive diagnostic feedback; CIRF = correct–incorrect response feedback; NO = no feedback; *t* = test time point; B = *p* value of Bonferroni test; BF_{10} = Bayes factor; T = *p* value of Tamhane's T2 test; D = *p* value of Dunnett's T3 test.

was found between feedback mode and test time. After performing a simple effect analysis (Tables 6 and 7), it can be seen that although most of the results were consistent with the raw score results, there were a few inconsistent results. For example, no significant difference was found between the traditional and control groups across all three test times, indicating that providing CIRF may not be effective in promoting learning compared with no feedback. Meanwhile, for the control group, there was no significant difference between the three test times, indicating that the general latent abilities of students do not show significant changes over time.

As stated earlier, students were randomly assigned at the class level. To investigate whether the effects of different feedback modes on two classes within the same group are similar, Figure 7 shows the developmental trend of the average raw scores and general latent abilities of the six classes (more details can be found in online Supplemental Table A4). It can be seen that the two classes within the same group have roughly the same developmental trend in both raw scores and general latent abilities; from another perspective, the difference between the effects of the different feedback modes is robust to the classes taught by the two teachers.

Figure 8 shows the developmental trend of the mastery proportion of the six attributes of the three groups over time. According to the mastery proportion on the first test time, it can be seen that the six attributes differ in difficulty, which is related to the hierarchical structure between them (see Figure 3). For example, attribute A1 is the easiest,

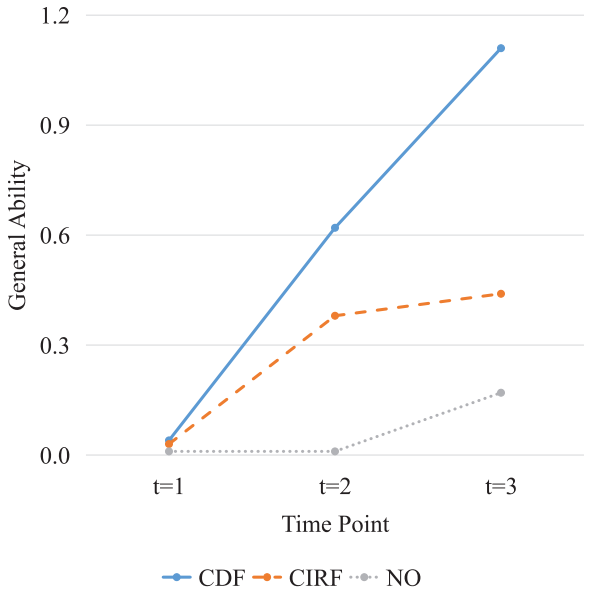

FIGURE 6. *Developmental trend of average general latent ability in different groups over time.*
Note. t = test time point; CDF = cognitive diagnostic feedback; CIRF = correct–incorrect response feedback; NO = no feedback.

while attribute A6 is, comparatively, the most difficult. In the diagnosis group, the mastery proportion of each attribute improved over time; in the traditional group, although the mastery proportion of each attribute also improved over time, the improvement was smaller than that in the diagnosis group. By contrast, there was a slight improvement in the mastery proportion of each attribute in the control group, especially for the relatively tricky attributes.

A two-factor mixed-design ANOVA was performed, using feedback mode and test time as the independent variables and the mastery status of each attribute as the dependent variable. The results are presented in Table 8, which allows us to compare the effects of different feedback methods at a deeper and more advanced attribute level. Combining the results of general latent ability, it can be seen that the facilitation effects of the different feedback modes for each attribute were generally consistent with those for general latent abilities, but some differences remained. For example, although for general latent abilities, there were no significant differences between the traditional and control groups at any of the three time points, there were significant differences between the two groups for some attributes (e.g., attributes A2 and A3) at the third test time (i.e., after two times of feedback). In addition, based on the results for raw scores and general latent ability, it was found that the diagnosis and traditional groups did not show significant differences until the third test time; this result is corroborated by the degree of change in each attribute. Specifically, for relatively easy attributes (e.g., attributes A1 and A2), both CDF and CIRF can effectively promote students to master these attributes. At the same time, the relative advantage of CDF gradually emerged as the difficulty of the attributes increased and the amount of feedback increased.

In summary, for the research questions in this study, the results of this quasi-experimental study provided some evidence for the two conclusions:

1. CDF can effectively promote student learning compared with no feedback; and
2. CDF is more effective than CIRF in promoting student learning, especially for relatively tricky attributes.

Discussion

To explore whether CDF can promote learning and whether it is more effective than CIRF in promoting learning, this study conducted a quasi-experiment using a longitudinal CDA of rational number operations to compare the effectiveness of three feedback modes: CDF, CIRF, and no feedback. The results of this quasi-experimental study provided some evidence for the conclusions that CDF can promote students' learning and that it is more effective than CIRF in promoting learning, especially for difficult knowledge areas.

The findings of this study support those of some previous studies. For example, the finding that CDF is significantly better than CIRF in promoting students' mathematical performance supports the idea that the more information the feedback provides, the more helpful it is in promoting learning (Dunn et al., 2012; Maddox et al., 2008). In other words, the results of this study support the view that KCRI is more beneficial than KCR in promoting learning. Therefore, this study's results also help clarify some of the controversies surrounding the effectiveness of informational feedback in promoting learning. In CDAs, increasing the complexity of feedback (i.e., providing additional information about latent attribute mastery status) was more beneficial in promoting students' learning. The conclusions of this study add some insight into how informational feedback affects learning.

In addition, the results of this study cannot fully support the view that CIRF can effectively promote learning. Specifically, the analysis of raw scores showed that CIRF was more effective in promoting student learning than no feedback, which supported the view of some previous studies that CIRF can promote learning (e.g., Butler et al., 2007; Mullet et al., 2014); however, the analysis of general latent ability showed no significant difference between CIRF and no feedback in promoting learning. By looking at the response data of individual students in the traditional group, we found that some students showed an increase in raw scores but almost no change in general latent ability. Further observation revealed that the number of correct responses for these students increased at subsequent test times for

TABLE 5
Mixed-Design Analysis of Variance of Different Feedback Modes and Test Times for General Latent Ability

	df	*F*	*p*	η^2	Partial η^2	BF_{10}
Feedback mode	2	4.470	.012	0.032	0.032	4.411
Test time	1.57	562.274	<.001	0.496	0.673	>100
Feedback mode * Test time	3.14	148.706	<.001	0.263	0.521	>100

Note. The assumption of sphericity had been violated; the Greenhouse–Geisser correction adjusted the degrees of freedom (*df*) for the within-subject effect. The Bayes factor (BF_{10}) was calculated using the JASP software based on the Bayesian estimation. For example, BF_{10} = 4.411 means that the current data are 4.411 times more likely to occur under the alternative hypothesis (H1) being true than under the null hypothesis (H0) being true. Dienes (2014) suggested that BF_{10} less than 1, 1/3, and 1/10 represents weak, moderate, and substantial evidence for the H0, respectively. By contrast, BF_{10} greater than 1, 3, and 10 represents weak, moderate, and substantial evidence for the H1, respectively.

TABLE 6
Simple Effect Analysis of Feedback Modes for General Latent Ability

	t = 1				*t* = 2				*t* = 3			
	B	BF_{10}	T	D	B	BF_{10}	T	D	B	BF_{10}	T	D
CDF–CIRF	1.000	0.161	1.000	1.000	.683	0.311	.551	.550	<.001	>100	<.001	<.001
CDF–NO	1.000	0.164	.994	.994	.008	11.951	.007	.007	<.001	>100	<.001	<.001
CIRF–NO	1.000	0.161	.998	.998	.212	0.732	.200	.199	.344	0.461	.344	.342

Note. CDF = cognitive diagnostic feedback; CIRF = correct–incorrect response feedback; NO = no feedback; *t* = test time point; B = *p* value of Bonferroni test; BF_{10} = Bayes factor; T = *p* value of Tamhane's T2 test; D = *p* value of Dunnett's T3 test.

TABLE 7
Simple Effect Analysis of Test Times for General Latent Ability

	CDF				CIRF				NO			
	B	BF_{10}	T	D	B	BF_{10}	T	D	B	BF_{10}	T	D
(*t* = 1)–(*t* = 2)	.004	19.570	.004	.004	.145	0.816	.176	.175	1.000	0.158	1.000	1.000
(*t* = 1)–(*t* = 3)	<.001	>100	<.001	<.001	.063	3.756	.028	.028	1.000	0.230	.753	.751
(*t* = 2)–(*t* = 3)	.021	3.024	.037	.037	1.000	0.168	.983	.983	1.000	0.211	.818	.817

Note. CDF = cognitive diagnostic feedback; CIRF = correct–incorrect response feedback; NO = no feedback; *t* = test time point; B = *p* value of Bonferroni test; BF_{10} = Bayes factor; T = *p* value of Tamhane's T2 test; D = *p* value of Dunnett's T3 test.

items that tested accessible attributes, but the number of correct responses for items that tested difficult attributes remained almost unchanged. One possible explanation of this phenomenon is that CIRF promotes mastery of relatively easy attributes but has little to no facilitation effect on relatively tricky attributes. It is challenging to promote the development of general latent ability. In short, given that the effectiveness of CIRF in promoting learning is still inconclusive, the results of this study suggest that providing students with informational feedback (e.g., CDF) in subsequent practice teaching may be a superior option.

Despite its promising results, this study still has some limitations that may affect the accuracy of our conclusions. First, to examine whether cognitive diagnosis has its claimed function to promote learning, this study explored whether CDF promotes learning and whether it promotes learning more effectively than traditional feedback. In essence, this study explored whether KCRI with attribute-level information is more effective than KCR, rather than other types of KCRI (e.g., KCRI with correct problem-solving procedures), in promoting learning. However, providing model-based feedback is difficult for practitioners, who may still prefer other types of KCRI that are relatively easy to implement. Therefore, it is still necessary to explore whether CDF promotes learning more effectively than other types of KCRI in further studies.

Second, since the quasi-experimental design cannot strictly control extraneous variables, the results of this study can only provide evidence to support the idea that CDF is more effective than CIRF in promoting learning, and no causal inference can be drawn that the use of CDF led to a better academic performance of students in the

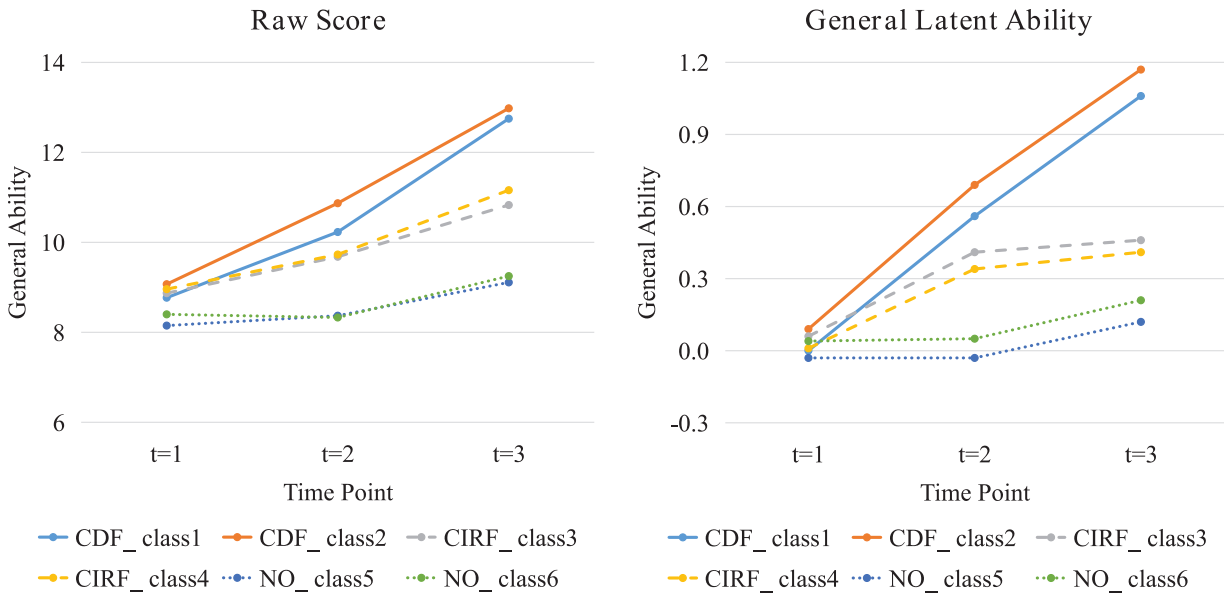

FIGURE 7. *Developmental trend of average raw score and average general latent ability in different classes over time.*
Note. t = test time point; CDF = cognitive diagnostic feedback; CIRF = correct–incorrect response feedback; NO = no feedback; the class number in the figure is not the original class number.

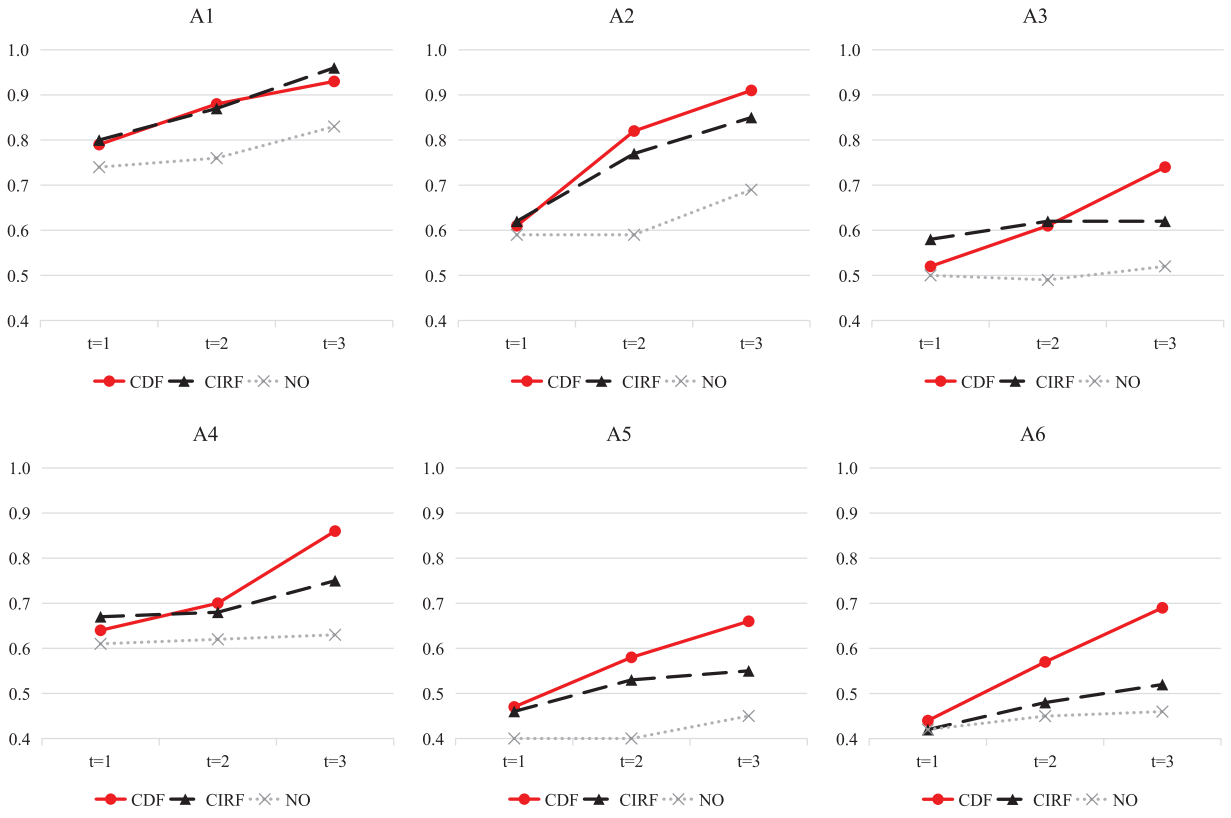

FIGURE 8. *Developmental trend of mastery proportion of attributes over time.*
Note. A1 = rational numbers; A2 = related concepts of rational numbers; A3 = axis; A4 = addition and subtraction of rational numbers; A5 = multiplication and division of rational numbers; A6 = mixed operation of rational numbers; t = test time point; CDF = cognitive diagnostic feedback; CIRF = correct–incorrect response feedback; NO = no feedback.

TABLE 8

Mixed-design Analysis of Variance of Different Feedback Modes and Test Times for Latent Attributes and Corresponding Simple Effect Analyses

	Difference between test times			Difference between feedback modes		
Attribute	Feedback mode	*p*	Simple effect analysis	Test time	*p*	Simple effect analysis
A1	CDF	.016	$(t = 1)$–$(t = 3)$*	$t = 1$	.598	
	CIRF	.007	$(t = 1)$–$(t = 3)$**	$t = 2$	.043	
	NO	.314		$t = 3$	.007	CDF–NO* CIRF–NO†
A2	CDF	<.001	$(t = 1)$–$(t = 2)$** $(t = 1)$–$(t = 3)$***	$t = 1$	.185	
	CIRF	.001	$(t = 1)$–$(t = 3)$***	$t = 2$	<.001	CDF–NO*** CIRF–NO***
	NO	.357		$t = 3$	<.001	CDF–NO*** CIRF–NO***
A3	CDF	.008	$(t = 1)$–$(t = 3)$** $(t = 2$–$(t = 3)^{\dagger}$	$t = 1$	.569	
	CIRF	.786		$t = 2$	.134	
	NO	.906		$t = 3$	.007	CDF–NO*** CDF–CIRF† CIRF–NO**
A4	CDF	.004	$(t = 1)$–$(t = 3)$** $(t = 2)$–$(t = 3)$*	$t = 1$	.632	
	CIRF	.477		$t = 2$	.446	
	NO	.956		$t = 3$	.002	CDF–NO*** CDF–CIRF†
A5	CDF	.037	$(t = 1)$–$(t = 3)$*	$t = 1$	.658	
	CIRF	.381		$t = 2$	.050	
	NO	.794		$t = 3$	.017	CDF–NO* CDF–CIRF† CIRF–NO†
A6	CDF	<.001	$(t = 1)$–$(t = 2)^{\dagger}$ $(t = 1)$–$(t = 3)$*** $(t = 2)$–$(t = 3)^{\dagger}$	$t = 1$	.110	
	CIRF	.903		$t = 2$	<.001	CDF–CIRF† CDF–NO***
	NO	1.000		$t = 3$	<.001	CDF–CIRF*** CDF–NO***

Note. CDF = cognitive diagnostic feedback; CIRF = correct–incorrect response feedback; NO = no feedback; *t* = test time point.
$^{\dagger}p < .08$ (marginally significant). *$p < .05$. **$p < .01$. ***$p < .001$.

diagnosis group than in the traditional group. For example, during the experiment, students still receive regular teaching, and other newly learned mathematical knowledge may have an impact on the knowledge of rational number operations; in addition, factors such as learning motivation may also affect the results of this study in the form of a mediating variable: when students get more information, they may become more interested or more motivated in specific knowledge and thus study more. In further studies, researchers can measure some extraneous variables, for example, using surveys or questionnaires to collect extra learning information (not due to the experiment) from participants. If such information can be collected, then a more valid conclusion can be drawn.

Third, to ensure the comparability of the results at each time point, the adopted longitudinal instrument was developed based on the design of the parallel test. However, in practice, perfectly parallel tests do not exist. Thus, the variation in results at different time points may be partly due to the nonparallelism error of the instrument. In addition, alternate-form items in parallel tests may not eliminate practice/memory effects, which may also affect the results of this study. In further studies, a longitudinal instrument with the anchor-item design could be adopted.

Fourth, for ease of operation and to balance the impact of teachers' teaching styles on students, only a class-level random assignment was used in the current study. Even though the results suggested that the difference between the effects of the different feedback modes is robust to the classes taught by the two teachers, the relatively small number of classes may not completely exclude the influence of within-group similarity on the results. From this perspective, the current study's findings are still preliminary findings; thus, a much larger sample and/or a design with individual-level random assignment are still needed in the future to permit meaningful inferences and further verify the findings of the current study.

Fifth, the students in the control group were not given feedback reports until the end of the three tests, which is unusual in a practical test situation. This difference could have affected the ecological validity of the conclusions of this work (fortunately, the comparison between the diagnosis group and the traditional group is not affected).

Sixth, when performing ANOVA on general latent ability and latent attributes, the current study ignores the impact of standard errors of estimates, affecting the analysis results to a certain extent. In further studies, we can try to incorporate the feedback mode as a covariate for change over time and directly test the significance of the coefficient of this covariate. Of course, since the current sLong-DINA model does not have this function, an extended model needs to be developed in the future.

Last but not least, some readers may be concerned about whether the accuracy of the model-based estimated classification will affect the effectiveness of CDF. To our understanding, this influence may indeed exist. First, the classification accuracy is related to the reliability and validity of the measurement instrument and the degree of the model-data fitting. Therefore, it is recommended that practitioners should ensure that the instrument has sufficient reliability and validity and should have some data analysis capabilities (e.g., using specific readily available software) before applying this type of method. Second, it is true that the CDF as reference information still cannot be guaranteed to be 100% correct under the premise of ensuring the instrument has sufficient reliability and validity and the appropriate model. However, students or teachers can refer to the degree of certainty of classification in CDF to make further self-judgments. Third, in practice, we should not be limited to a specific feedback mode but try to combine multiple feedback modes (e.g., KCRI with attribute-level information and correct problem-solving procedures) to promote learning.

Open Practices

The data and analysis files for this article can be found at https://doi.org/10.3886/E153061V1

Acknowledgments

This work was supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Grant No. 19YJC190025) and the National Natural Science Foundation of China (Grant No. 31900795).

ORCID iD

Peida Zhan https://orcid.org/0000-0002-6890-7691

Notes

1. A reachability matrix specifies the direct and indirect relationships among the hierarchical attributes.

2. In contrast to the complete version, the special dimensions used to account for local item dependence (e.g., the practice/memory effect) among anchor/repeated items at different time points are ignored in the simple version to reduce model complexity and computational burden. Since the parallel tests design rather than the anchor-item design was used in this study, this made the simplified model more suitable than the complete model.

3. These students are not the same as those in the instrument development phase.

References

Andre, T., & Thieman, A. (1988). Level of adjunct question, type of feedback, and learning concepts by reading. *Contemporary Educational Psychology*, *13*(3), 296–307. https://doi.org/10.1016/0361-476X(88)90028-8

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213–238. https://doi.org/10.3102/00346543061002213

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology*, *105*(2), 290–298. https://doi.org/10.1037/a0031026

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Chen, Y. H. (2012). Cognitive diagnosis of mathematics performance between rural and urban students in Taiwan. *Assessment in Education: Principles, Policy & Practice*, *19*(2), 193–209. https://doi.org/10.1080/0969594X.2011.560562

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbook of human-computer interaction* (pp. 849–874). Elsevier. https://doi.org/10.1016/B978-044481862-1.50103-5

de la Torre, J. (2008). An empirically based method of q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*(4), 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Ding, S., Yang, S., & Wang, W. (2010). The importance of reachability matrix in constructing cognitively diagnostic testing. *Journal of Jiangxi Normal University*, *34*, 490–494.

Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 840–859. https://doi.org/10.1037/a0027867

Gao, Y., Zhai, X., Andersson, B., Zeng, P., & Xin, T. (2020). Developing a learning progression of buoyancy to model conceptual change: A latent class and rule space model analysis. *Research in Science Education*, *50*(4), 1369–1388. http://doi.org/10.1007/s11165-018-9736-5

Gibbons, A. S., & Fairweather, P. G. (1998). *Computer-based instruction: Design and development*. Educational Technology.

Gu, Y., & Xu, G. (2020). *Identifiability of hierarchical latent attribute models*. arXiv preprint. https://arxiv.org/abs/1906.07869

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*(1), 98–125. https://doi.org/10.1111/jedm.12036

Huang, H.-Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *Journal of Educational Measurement*, *54*(4), 440–480. https://doi.org/10.1111/jedm.12156

Huff, K., & Goodman, D. P. (2007). *The demand for cognitive diagnostic assessment*. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19–60). Cambridge University Press.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Kaya, Y., & Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance. *Educational and Psychological Measurement*, *77*(3), 369–388. https://doi.org/10.1177/0013164416659314

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary Educational Psychology*, *10*(3), 285–291. https://doi.org/10.1016/0361-476X(85)90025-6

Leighton, J. P., & Gierl, M. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*(3), 205–237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x

Li, F., Cohen, A., Bottge, B., & Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement*, *76*(2), 181–204. https://doi.org/10.1177/0013164415588946

Li, H., Kim, M. K., & Xiong, Y. (2020). Individual learning vs. interactive learning: A cognitive diagnostic analysis of MOOC students' learning behaviors. *American Journal of Distance Education*, *34*(2), 121–136. https://doi.org/10.1080/08923647.2019.1697027

Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., & Chang, H. H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification*, *30*(2), 152–172. https://doi.org/10.1007/s00357-013-9128-5

Ma, W., & de la Torre, J. (2020). Choosing between CDM and unidimensional IRT: The proportional reasoning test case. *Measurement: Interdisciplinary Research and Perspectives*, *18*(2), 87–96. https://doi.org/10.1080/15366367.2019.1697122

Maddox, W. T., Love, B. C., Glass, B. D., & Filoteo, J. V. (2008). When more is less: Feedback effects in perceptual category learning. *Cognition*, *108*(2), 578–589. https://doi.org/10.1016/j.cognition.2008.03.010

Madison, M. J., & Bradshaw, L. P. (2018). Assessing growth in a diagnostic classification model framework. *Psychometrika*, *83*(4), 963–990. https://doi.org/10.1007/s11336-018-9638-5

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*(4), 523–547. https://doi.org/10.1007/BF02294327

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212. https://doi.org/10.1007/BF02294535

Mullet, H. G., Butler, A. C., Verdin, B., von Borries, R., & Marsh, E. J. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition*, *3*(3), 222–229. https://doi.org/10.1016/j.jarmac.2014.05.001

Pan, Q., Qin, L., & Kingston, N. M. (2020). Growth modeling in a diagnostic classification model (DCM) framework: A multivariate longitudinal diagnostic classification model. *Frontiers in Psychology*, *11*, 1714. https://doi.org/10.3389/fpsyg.2020.01714

Peeck, J. (1979). Effects of differential feedback on the answering of two types of questions by fifth- and sixth-graders. *British Journal of Educational Psychology*, *49*(1), 87–92. https://doi.org/10.1111/j.2044-8279.1979.tb02401.x

Ren, H., Xu, N., Lin, Y., Zhang, S., & Yang, T. (2021). Remedial teaching and learning from a cognitive diagnostic model perspective: Taking the data distribution characteristics as an example. *Frontiers in Psychology*, *12*, 628607. https://doi.org/10.3389/fpsyg.2021.628607

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence in Education*, *18*(4), 289–316.

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods*, *19*(3), 506–532. https://doi.org/10.1177/1094428116630065

Tang, F., & Zhan, P. (2020). The development of an instrument for longitudinal learning diagnosis of rational number operations based on parallel tests. *Frontiers in Psychology*, *11*, 2246. https://doi.org/10.3389/fpsyg.2020.02246

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*(4), 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Tjoe, H., & de la Torre, J. (2013). Designing cognitively-based proportional reasoning problems as an application of modern psychological measurement models. *Journal of Mathematics Education*, *6*(2), 17–26. https://doi.org/10.1007/s13394-013-0090-7

von Davier, M., & Lee, Y.-S. (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. Springer. https://doi.org/10.1007/978-3-030-05584-4

Wang, L., Tang, F., & Zhan, P. (2020). Effect analysis of individualized remedial teaching based on cognitive diagnostic assessment: Taking "linear equation with one unknown" as an example. *Journal of Psychological Science*, *43*(6), 1490–1497.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*, *43*(1), 57–87. https://doi.org/10.3102/1076998617719727

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. https://doi.org/10.1111/jedm.12096

Whyte, M. M., Karolick, D. M., Nielsen, M. C., Elder, G. D., & Hawley, W. T. (1995). Cognitive styles and feedback in computer-assisted instruction. *Journal of Educational Computing Research*, *12*(2), 195–203. https://doi.org/10.2190/M2AV-GEHE-CM9G-J9P7

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

Wu, H.-M. (2019). Online individualised tutor for improving mathematics learning: A cognitive diagnostic model approach. *Educational Psychology*, *39*(10), 1218–1232. https://doi.org/10.1080/01443410.2018.1494819

Zhan, P. (2020a). A Markov estimation strategy for longitudinal learning diagnosis: Providing timely diagnostic feedback. *Educational and Psychological Measurement*, *80*(6), 1145–1167. https://doi.org/10.1177/0013164420912318

Zhan, P. (2020b). Longitudinal learning diagnosis: Minireview and future research directions. *Frontiers in Psychology*, *11*, 1185. https:///doi.org/10.3389/fpsyg.2020.01185

Zhan, P. (2021). Refined learning tracking with a longitudinal probabilistic diagnostic model. *Educational Measurement: Issues and Practice*, *40*(1), 44–58. https://doi.org/10.1111/emip.12397

Zhan, P., & He, K. (2021). A longitudinal diagnostic model with hierarchical learning trajectories. *Educational Measurement: Issues and Practice*, *40*(3), 18–30. https://doi.org/10.1111/emip.12422

Zhan, P., Jiao, H., Liao, D., & Li, F. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics*, *44*(3), 251–281. https://doi.org/10.3102/1076998619827593

Zhan, P., & Qiao, X. (2020). *A diagnostic classification analysis of problem-solving competence using process data*. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/wtyae

Zhang, S., & Chang, H. (2020). A multilevel logistic hidden Markov model for learning under cognitive diagnosis. *Behavior Research Methods*, *52*(1), 408–421. https://doi.org/10.3758/s13428-019-01238-w

Authors

FANG TANG is a graduate student at the Department of Psychology, Zhejiang Normal University, Jinhua, China. Her research interests are practical applications of cognitive diagnosis.

PEIDA ZHAN is an associate professor at the Department of Psychology, Zhejiang Normal University, and the Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, China. His research interests involve theoretical and applied development in latent variable modeling, such as cognitive diagnosis modeling, item response theory modeling, and response times modeling.