Effatpanah et al. Language Testing in Asia (2019) 9:12
https://doi.org/10.1186/s40468-019-0090-y

Language Testing in Asia

RESEARCH Open Access

# Diagnosing EFL learners' writing ability: a diagnostic classification modeling analysis

Farshad Effatpanah, Purya Baghaei* and Ali Akbar Boori

* Correspondence: pbaghaei@mshdiau.ac.ir
English Department, Mashhad Branch, Islamic Azad University, Mashhad, Iran

## Abstract

Cognitive diagnostic models (CDMs) have recently received a surge of interest in the field of second language assessment due to their promise for providing fine-grained information about strengths and weaknesses of test takers. For the same reason, the present study used the additive CDM (ACDM) as a compensatory and additive model to diagnose Iranian English as a foreign language (EFL) university students' L2 writing ability. To this end, the performance of 500 university students on a writing task was marked by four EFL teachers using the Empirically derived Descriptor-based Diagnostic (EDD) checklist. Teachers, as content experts, also specified the relationships among the checklist items and five writing sub-skills. The initial Q-matrix was empirically refined and validated by the GDINA package. Then, the resultant ratings were analyzed by the ACDM in the CDM package. The estimation of the skill profiles of the test takers showed that vocabulary use and content fulfillment are the most difficult attributes for the students. Finally, the study found that the skills diagnosis approach can provide informative and valid information about the learning status of students.

**Keywords:** Cognitive Diagnostic Assessment (CDA), ACDM, Q-matrix, L2 Writing Sub-skills

## Introduction

Written language is only one of the fundamental ways of communication and linguistic expression which allows individuals from different cultures and backgrounds to participate in various aspects of today's global community. Given the rapid advances in technology and communication throughout the world, the ability to write a second or foreign language has become a crucial skill more than ever. Learning to write well is thus a need for all students in academic and second/foreign language programs to not only generate new information but also transfer their knowledge (Weigle, Boldt, & Valsecchi, 2003).

Many researchers have argued that writing is a highly complex and multidimensional process with many underlying cognitive components which play vital roles in communicating, thinking, planning, and learning (Dunsmuir & Clifford, 2003; Williams & Larkin, 2013). To write effectively, writers are expected to have a mastery on a set of lower-level (namely, spelling, punctuation, sentence construction) and higher-level writing skills (namely, textual coherence, logical development of relevant arguments) (McCutchen, 2011; Wilson, Olinghouse, McCoach, Santangelo, & Andrada, 2016).

Difficulty in each of these micro-skills can impede the development of L2 writing. Therefore, it is essential to accurately respond to students' specific writing skills where lack of mastery or faulty strategy is indicated. If difficult and problematic areas of writing are identified, students can receive sufficient and immediate feedback on their performance. As a result, they can adopt some strategies to eliminate or remedy their weaknesses during their learning process.

Along the same lines, the last few decades have witnessed a growing interest in the role of feedback on L2 writing. Feedback conventionally refers to "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (Hattie & Timperley, 2007, p. 81). Many researchers have accentuated that effective feedback should close the gap between the actual level and desired level of performance (Kumar & Stracke, 2011; Ramaprasad, 1983; Sadler, 1989). Appropriate feedback has been also shown to have a major role in motivating students, encouraging self-regulated learning, and improving subsequent learning and teaching (Black & Wiliam, 1998; Hyland, 2013; Kuo, Chen, Yang, & Mok, 2016; Zimmerman, 2000). Consequently, there has been a proliferation of research on the effect of various types of feedback on the performance of students; including corrective (Bitchener, 2008; Bitchener & Knoch, 2008), peer (Yu & Hu, 2017), oral (Erlam, Ellis, & Batstone, 2013), coded (Buckingham & Aktuğ-Ekinci, 2017; Sampson, 2012), content (Ashwell, 2000), form-focused (Truscott, 1996), teacher (Yang, Badger, & Yu, 2006), and computer-mediated (Shintani, 2016). Such large number of studies on feedback reflects this fact that student development as writers entails a supportive environment in which they receive valuable feedback (Blikstad-Balas, Roe, & Klette, 2018; Ferris, 2003). However, the current practices of providing feedback fail to identify the exact nature of students' problems. What is needed are approaches to assessment which are able to provide diagnostic and fine-grained information about the strengths and weaknesses of students in various targeted language domains (Llosa, Beck, & Zhao, 2011).

Diagnostic assessment, as an alternative approach to the existing methods of feedback, has recently received more attention among educational experts because of its capability in providing diagnostic information about students' strengths and weaknesses (Lee & Sawaki, 2009a). Information obtained from diagnostic tests can be useful for identifying areas where students need more assistance and designing appropriate materials for further learning or instruction (ALTE, 1998). According to Nation and Macalister (2010), diagnostic assessment is of paramount importance in needs analysis both prior and during a course. The findings of diagnostic assessments specifically pin down what goes into a course to be in accordance with students' needs. Particularly, diagnostic tests can be considered as needs assessment to identify students' gaps, causes of problems, priorities, and possible solutions. It is thus evident that diagnostic tests are in line with assessment for learning (AFL) or formative assessment. In other words, such tests are able to integrate assessment with instruction and curriculum which results in enhancing learning (Pellegrino & Chudowsky, 2003).

In addition to developing tests that are suitable for diagnostic purposes, there is another trend in developing and conducting diagnostic assessment which allows experts to model statistically the test takers' cognitive operations. This approach is known as Cognitive Diagnostic Assessment (CDA) which is able to yield more fine-grained information about learners' performance. CDA is a merger of cognitive psychology and

educational measurement that can provide informative and detailed information about the learning status of students on a set of multiple fine-grained (sub)skills (Rupp, Templin, & Henson, 2010). Unlike traditional psychometric frameworks, such as classical test theory (CTT) and item response theory (IRT), which provide single overall scores along a proficiency continuum, CDA offers categorical mastery/non-mastery diagnostic information about strengths and weaknesses of the examinee's skills. Because CDA is basically diagnostic, complex statistical models, known as cognitive diagnostic models (CDMs), are used to measure to what extent students have mastered a set of sub-skills required for successful performance on a test (de la Torre & Minchen, 2014).

Technically speaking, CDMs decompose tasks into multiple strategies, processes, and knowledge that a student must possess in order to respond correctly to a given test item or task (Birenbaum, Kelly, & Tatsuoka, 1993). This characteristic allows CDMs to generate "multidimensional diagnostic profiles based on statistically-driven multivariate classifications" (Kunina-Habenicht, Rupp, & Wilhelm, 2009, p. 64) of students according to the degree mastery on each of the requisite traits. Such information can maximize opportunities to learn by giving diagnostic feedback to all stakeholders and ultimately enhance language learning and teaching.

Cognitive diagnostic models have primarily been used to serve two main purposes: (a) to classify examinees into similar skill mastery profiles on the basis of their observed response patterns and (b) to identify whether there is a compensatory or non-compensatory interaction between the postulated attributes underlying a given skill (Ravand & Robitzsch, 2018). A wide array of CDMs with different theories or assumptions about the way of interaction between attributes (see Ravand & Baghaei, 2019, for a review) have been proposed. Most of these models have been applied in language assessment contexts and demonstrated to be useful for providing diagnostic feedback in service of instruction and learning (Nichols, 1994). The models include rule space methodology (RSM) (Tatsuoka, 1983, 1995), the attribute hierarchy method (AHM) (Leighton, Gierl, & Hunka, 2004), the higher-order DINA model (HO-DINA) (; de la Torre & Douglas, 2004), the multi-strategy DINA (MS-DINA) (de la Torre & Douglas, 2008), the DINO and NIDO models (Templin & Henson, 2006), the full non-compensatory reparameterized unifed model (full NC-RUM)/fusion model (Hartz, 2002; Roussos et al., 2007), the compensatory RUM (C-RUM) (de la Torre, 2011), GDINA (de la Torre, 2011), the general diagnostic model (GDM) (von Davier, 2008; Xu & von Davier, 2008), the log-linear cognitive diagnosis model (LCDM) (Henson, Templin, & Willse, 2008), and the additive CDM (de la Torre, 2011).

Despite the increasing interest for diagnostic purposes and potential of CDMs in providing detailed information about test takers' strengths and weaknesses, too little research has been devoted to L2 writing ability in a foreign/second language-learning context (Knoch, 2011; Ranjbaran & Alavi, 2017). In response to this call, this study seeks to apply a cognitive diagnostic model to identify strengths and weaknesses of Iranian English as a foreign language (EFL) university students' L2 writing ability. Before ending this section, it must be noted that unlike previous studies in which researchers selected a CDM based on their intuition about the interaction between postulated attributes or common application of a particular CDM in language testing, the Additive CDM (ACDM), as a compensatory model, was utilized in this study in light of the results obtained from the previous study carried out by (Effatpanah, Baghaei, and Ravand, A comparison of different cognitive diagnostic model

for EFL writing, submitted) who found that the ACDM fits better than other CDM models for EFL writing performance. In this regard, the current study aims to investigate strengths and weaknesses of Iranian EFL students in L2 writing ability.

## Review of the literature

### Cognitive diagnostic models

Cognitive diagnostic models (CDMs) are discrete and multidimensional latent variable models developed mainly for diagnosing students' mastery profiles on a set of skills or attributes based on their observed item response patterns. According to Rupp and Templin (2008), CDMs are "probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and non-compensatory ways to generate latent classes" (p. 226).

Similar to item response theory models, CDMs are probabilistic models. They model the likelihood of executing a correct response with regard to a number of latent traits. In unidimensional item response theory models, the probability of producing a correct answer relies on a single latent trait, θ, in such a way that those examinees with higher ability have a higher probability of success. On the contrary, CDMs express a given student's performance level in terms of the probability of mastery of each attribute separately, or the probability of belonging to each latent class with a particular skill-mastery profile (Lee & Sawaki, 2009a).

CDMs are also inherently confirmatory. Like confirmatory factor analysis, latent traits in CDMs are defined a priori through an incidence matrix called Q-matrix (Tatsuoka, 1983), which is considered as the loading structure of CDMs. It specifies a substantive hypothesis about the underlying response processes of students. The Q-matrix indicates the association between each item (rows) and its target cognitive sub-skills (columns) through a pattern of "1 s" and "0 s". If an item requires sub-skill $k$, $qik$ =1; otherwise, $qik$ =0. Additionally, Rupp and Templin (2008) state that another manifestation of confirmatory nature of CDMs is the priori specification of the way different attributes interact in the response process, that is, whether there exists a compensatory (disjunctive) or non-compensatory (conjunctive) relationship among the required attributes.

Furthermore, CDMs belong to multidimensional item response theory models. CDMs contain multiple latent traits inasmuch as the successful performance on an item (or a task) requires mastery of numerous sub-skills. Because each item is related to multiple attributes, CDMs have a complex loading structure. However, compared to multidimensional IRT and factor analysis (FA) in which latent traits are continuous, CDMs possess discrete or categorical latent variables.

In regard to assuming varying inter-skill relationships among the predictor latent attributes, CDMs are classified into different categorizations. One way is to distinguish between disjunctive/conjunctive and compensatory/non-compensatory. The assumption in compensatory models is that inadequacy of one attribute can be made up for by the presence of other required attributes. In these models, mastery of more attributes does not increase the probability of success in a given item. In contrast, in non-

compensatory models, all the attributes are required to get an item right, that is, non-mastery of one attribute cannot be compensated for by the mastery of other attributes. More recently, additive CDMs have been proposed as a new category of CDMs which assume that presence of any one of the attributes increases the probability of a correct response independent of the presence or absence of other attributes (Ma, de la Torre, & Sorrel, 2018).

Another categorization of CDMs is specific vs. general. Specific CDMs are models which allow for only one type of relationship in the same test: conjunctive, disjunctive, and additive. On the other hand, general CDMs allow each item to select its own model that best fits it rather than imposing a specific model to all the items. de la Torre (2011) showed that several specific CDMs can be derived from general models if appropriate constraints are applied in the parameterization of general models. For instance, the generalized deterministic inputs, noisy "and" gate (GDINA) (de la Torre, 2011), as a general model, can be turned into DINA, DINO, ACDM, NC-RUM, and C-RUM by changing the link function into *log* and *logit* and setting the interaction effects to zero.

### ACDM model

Additive CDM (ACDM; de la Torre, 2011) is a compensatory cognitive diagnostic model. It is similar to the generalized deterministic inputs, noisy "and" gate (GDINA) (de la Torre, 2011), as a general model, which allows both compensatory and non-compensatory relationships between attributes within the same test. The modeling approach adopted by the GDINA is similar to analysis of variance (ANOVA) (Ravand, 2015). In this saturated model, all possible interaction and main effects are used. If some main or interaction effects are removed, several specific CDMs can be derived from the GDINA model. For instance, by setting all interaction effects to zero, the GDINA model turns into the ACDM. The difference between the two models is that the GDINA model has a multiplicative impact, but the ACDM has an additive impact on the probability of a correct response (de la Torre, 2011). Also, unlike the GDINA model which partitions test takers into $(2^{k^*_j})$, the ACDM has $K^*_j + 1$ parameters for item $j$. The ACDM postulates that (1) the probability of success increases by mastering each of the required attributes and lack of one attribute can be made up for by the presence of other attributes; and (2) the contribution of each attribute is independent from the other attributes. The item response function (IRF) for the ACDM is:

$$P\left(\alpha^*_{lj}\right) = \delta_{j0} + \sum_{k=1}^{k^*_j} \delta_{jk} a_{lk}$$

where $P\ (\alpha^*_{lj})$ is the probability of success, $k$ is the number of attributes required for item $j$, $\delta_{j0}$ is the intercept which represents the probability of a correct response when none of the required skills is present, and $\delta_{jk}$ is the main effect due to attribute $\alpha_k/$.

### Previous applications of CDMs

CDMs are largely used in two ways: (a) to develop "true diagnostic tests" (Ravand & Baghaei, 2019, p. 4) for the purpose of providing fine-grained information about the strengths and weaknesses of test takers and (b) to retrofit (post hoc analysis) non-diagnostic tests in order to extract richer information about their underlying

latent traits. It might be somewhat odd to say that although diagnostic tests are so compelling, very few true diagnostic tests have been designed (e.g., DIALANG by Alderson, 2005; Alderson & Huhta, 2005; DELNA (www.delna. auckland.ac.nz/uoa); and DELTA by Urmston, Raquel, & Tsang, 2013; Ranjbaran & Alavi, 2017) due to the difficult and time-consuming process of test construction (Alderson, 2005). Conversely, most applications of CDMs in educational measurement in general and second/foreign language testing in particular are cases of retrofitting to existing achievement or proficiency tests. Many researchers noted that this practice of CDMs application may raise various problems concerning the validity of any inferences about the test takers' skill competencies (DiBello, Roussos, & Stout, 2007; Jang, 2009). Nevertheless, Lee and Sawaki (2009a) argued that "retrofitting efforts could serve as an important step in advancing diagnostic language assessment research... it is worth examining the extent to which useful diagnostic information could be extracted from existing assessments before delving into an expensive, time-consuming process of designing a new cognitive diagnostic test" (p. 174).

In addition, the results derived from retrofitting assessments can be useful for universities and non-tertiary education institutions. A single score from standardized language tests such as IELTS and test of English as a foreign language (TOEFL) indicates only whether or not an applicant has a sufficient command of English to satisfy the requirements of academic studies in that language. Nonetheless, fine-grained diagnostic information about students' ability can provide further information for stakeholders and university admission offices. According to Cumming (2015), "The coordinator or faculty advisor for an academic program at that same university, however, should expect more detailed information than a single score from a language test, relevant to the expectations of that academic program. Can a student applicant read, write, and interact orally in a certain language with sufficient proficiency to complete course assignments effectively, conduct required research tasks successfully, and perhaps work as teaching or research assistant? To make such decisions requires diagnostic information from a language assessment" (p. 414).

Considering the advantages of retrospective CDMs, these models have been applied in different studies on different language skills (Aryadoust, 2018; Buck et al., 1998; Buck & Tatsuoka, 1998; Buck, Tatsuoka, & Kostin, 1997; Chen & Chen, 2016; Jang, 2009; Kasai, 1997; Kim, 2014; Lee & Sawaki, 2009b; Li, 2011; Li & Suen, 2013; Ranjbaran & Alavi, 2017; Sawaki, Kim, & Gentile, 2009; Scott, 1998; Sheehan, 1997; von Davier, 2008). Compared to reading and listening, too little studies employed CDMs for the productive skills, e.g., writing and speaking. The reason for this lies in the limited number of writing items and inadequacy of existing rating scales for providing detailed information about the strengths and weaknesses of students (Kim, 2011). Alderson (2005) suggests that there should be a difference between diagnostic tests of writing and other tests of writing (e.g., placement or proficiency). Of course, he accentuated the use of indirect tests instead of performance tests in assessing writing. He argued that "in the case of diagnostic tests, which seek to identify relevant components of writing ability, and assess writers' strengths and weaknesses in terms of such components, the justification for indirect tests of writing is more compelling' (pp. 155–156). However, indirect tests are not adequately advocated in the era of performance assessment because

of their invalid measures of the multi-faceted nature of writing (Weigle, 2002). Knoch (2011, p. 82) also maintained that "rating scales developed for purposes other than diagnostic testing are not appropriate for diagnostic purposes." In light of this call, several studies have attempted to develop diagnostic checklist to identify the problematic areas of writers (Banerjee & Wall, 2006; Kim, 2011; Struthers, Lapadat, & MacMillan, 2013). In their study, Banerjee and Wall (2006) developed and tailored a new instrument assessment for different skills of English. Struthers et al. (2013) also designed a checklist, containing 13 items, to assess only cohesion in the writing of children in grades 4–7. In another study, Kim (2011) developed an Empirically derived Descriptor-based Diagnostic (EDD) checklist, which contains 35 descriptors measuring five sub-skills of academic writing in English. She then applied the reduced reparameterized unified model (RRUM) to study the extent to which information obtained from CDMs are accurate, discriminant, and reliable. Recently, Xie (2016) examined the validity of the EDD checklist for formative assessment. Like Kim, she used the RRUM to assess first-year undergraduates in Hong Kong and evaluated the usefulness of the checklist for diagnosing academic writing in English. As it is apparent, the number of CDM studies on L2 writing is still scant and more research is required to apply different CDMs in order to determine which model can better reflect the underlying L2 writing attributes.

The present study investigated the extent to which the diagnostic information of the EDD checklist, which was estimated using the ACDM model, was accurate for specifying Iranian EFL university students' L2 writing ability. To accomplish this purpose, the following research questions are posed:

1. Does the ACDM model produce reliable and useful diagnostic information about Iranian EFL university students in terms of mastery/non-mastery of the L2 writing attributes?
2. What are the strengths and weaknesses of Iranian EFL students' L2 writing ability?
3. To what extent is the final Q-matrix valid for diagnosing the students' L2 writing ability?

## Methodology

### Participants

A total number of 500 English students participated in this research. The participants produced an essay in courses related to L2 Writing. They were selected from 21 universities of Iran according to their convenient accessibility. They were taught by 32 non-native full-time English teachers who followed different syllabi and materials for instruction. All teachers had a wide range of experience, between 9 and 24 years, in teaching writing. The sample consisted of 349 female (69.8%) and 151 male (30.2%) students who ranged in ages between 19 and 58 years (M = 24.89 years, SD = 6.30). This relative gender disparity is due to the typical distribution of students in English departments in many Iranian universities. Of the total sample, 268 (53.6%) studied English Teaching, 128 (25.6%) English Literature, and 104 (20.8%) English Translation. Also, 212 participants were Junior, 152 Senior undergraduate, and 136 M.A. students. Their length of time learning English varied from

1 to 2 years to more than 10 years. Similarly, they reported their frequency of use of English from cannot say to almost every day.

Moreover, four raters were recruited to mark the essays. The sample of raters included one female and three male raters between the ages of 28 to 39 years old (M = 31.75; SD = 4.99). They were all non-native speakers of English, knowing Persian as their first language and English as their foreign language. Each rater was provided with an assessment package containing 125 essays along with marking guidelines to score them based on the EDD checklist (Kim, 2011).

### Instruments

There were two measures employed in this study. One was a writing prompt, and the other was a diagnostic checklist for rating the essays (Additional file 1).

#### *Writing sample*

The students were asked to complete a writing task. The task consisted of two sections. The first part included demographic information of participants such as name (optional), age, gender, major, educational level, length of time learning English, and frequency of use of English. The second part comprised a writing task. The respondents were asked to write at least a 350-word descriptive essay in class within an hour in response to the following prompt:

How to be a first-year college student? Write about the experience you have had. Make a guide for students who might be in a similar situation.

*Describe* how to make new friends, how to escape homesickness, how to be successful in studying, etc.

#### *The Empirically derived Descriptor-based Diagnostic (EDD) checklist*

The essays were marked using a diagnostic assessment scale called the Empirically derived Descriptor-based Diagnostic (EDD) checklist (Kim, 2011; see appendix) accompanied by detailed guidelines. It was developed for scoring and describing the writing of non-native English-speaking students in an academic context. The EDD checklist consists of 35 dichotomous (Yes, No) descriptors assessing five writing attributes. The sub-skills are content fulfilment (CON), organizational effectiveness (ORG), grammatical knowledge (GRM), vocabulary use (VOC), and mechanics (MCH). The five subskills were defined as (1) a student's ability to address a given question by presenting unity and relevance of supporting sentences, information, and examples (CON); (2) a student's ability to develop and organize ideas and supporting sentences cohesively and coherently within and between paragraphs (ORG); (3) a student's ability to demonstrate syntactic complexity and variety accurately (GRM); (4) a student's ability to use a wide range of lexical items accurately and appropriately (VOC); and (5) a student's ability to follow the conventions of English writing such as margins and indentation, punctuation, and spelling and capitalization.

## Procedure

Data collection procedure began in September 2018, just at the commencement of the academic year. A consent form was sent to 25 English departments of different

universities across the country to allow us to give the test to their students. Only 21 universities consented to be involved in the study. From each university, two or three intact classes at different levels (junior and senior B.A. and M.A.) were selected. In total, 500 English students participated in this study who were taught by 32 teachers. Some teachers taught two classes. The students were asked to write at least a 350-word descriptive essay in class in response to the prompt. They were encouraged to rely on themselves and do not use any dictionaries, books, and internet during the writing to give the researchers and their teachers the opportunity to diagnose their strengths and weaknesses in L2 writing. They were reassured that their information would remain confidential and anonymous. Then, four raters were recruited to rate these papers on the basis of the EDD checklist (Kim, 2011). Reliability coefficients of the checklist were estimated using Cronbach alpha and a value of 0.89 was obtained which is acceptable. This value indicates that there is a high internal consistency for overall ratings of the checklist items.

### Rater training

Rater variability is a serious source of construct-irrelevant variance, which may threaten construct validity (Barrett, 2001; Lumley & McNamara, 1995; Weigle, 1999). Rater training is an effective way for reducing such variations. It has been shown that training can maximize the self-consistency of individual raters (Eckes, 2011; McIntyre, 1993; Weigle, 1998) and allows raters to interact, ask questions, review different dimensions of writing prompt and scoring rubrics, and get feedback on their scoring. Hence, a two and a half-hour session was held to introduce raters with the purpose of the study, writing attributes, Q-matrix construction, and the EDD checklist, as the rating scale. Then, all the raters received the 35-item checklist and a matrix form to specify their idea about each item and its associated attributes. All the descriptors and their attribute associations were discussed one by one. A master form was collectively prepared to combine all the raters' ideas into a comprehensive matrix. Next, following the suggestion of Weigle (2007), raters were asked to try out the rating scale on a set of essays to compare their answers with each other. Each rater scored three essays. It helped us to discuss different points where less agreement was achieved.

### Q-matrix validation

Correct identification of attributes underlying performance and their associations with test items improve the quality of a cognitive diagnostic assessment. If a Q-matrix is misspecified, diagnostic information may result in invalid inferences (Rupp & Templin, 2008). In the present study, raters, along with the authors, were considered as content experts to collectively indicate the major attributes required to perform correctly on each item. The raters were trained how to code the attributes measured by each item. On the basis of the consensus among the experts on the item–sub-skill associations, an initial Q-matrix was developed. The initial Q-matrix included 48 "1 s" which was less denser than the final Q-matrix (50 "1 s") from Kim (2011). The manipulated parameters related to 7 items (2, 3, 11, 12, 14, 31, 35). For example, item 3 was associated with CON, ORG, and GRM in the present study while it was specified to be related to CON, ORG, and VOC in the Kim (2011).

In the second step, the Q-matrix was empirically revised and validated by the procedure suggested by de la Torre and Chiu (2016) using "G-DIINA" package (Ma et al., 2018). In the first run of the analysis, some suggestions for the Q-matrix revision were provided. For example, it was suggested that ORG is not involved for Descriptor 7 (The supporting ideas and examples in this essay are appropriate and logical). In this case, only CON remains for the item. Accepting that statistical analysis should not be considered as the mere driving force for Q-matrix revision, the experts inspected the content of the descriptor and agreed that ORG is a requisite attribute for the item, so we kept it in the Q-matrix. Also, for Descriptors 1 and 10, it was suggested that ORG and GRM are respectively involved for the items; hence, we added them to the Q-matrix. However, for Descriptors 3, 12, and 18, the suggestions were to respectively remove ORG, CON, and GRM from the Q-matrix. After several rounds of revisions and rerunning the GDINA package, the final Q-matrix presented in Table 1 was developed. Of the total descriptors, nine of them were affiliated with CON, 12 items with ORG, 15 with GRM, five with VOC, and seven with MCH. In Table 1, 1 s indicate that the successful performance of students on each item depends on the mastery of the attributes whereas 0 s indicate that the item does not require the attributes. For example, if a test taker wants to perform successfully on item 2 (the essay is written clearly enough to be read without having to guess what the writer is trying to say), he/she should have the mastery of ORG and GRM. In this item, content (CON), vocabulary (VOC), and mechanics (MCH) are not necessary.

## Results

### Inter-rater reliability

As a preliminary check, the degree of inter-rater reliability was investigated using Pearson correlation coefficients. Each rater randomly received 37 scripts from the other raters to score for the second round. Totally, 147 out of 500 papers were remarked. The correlation coefficient between the total scores of the remarked papers from the two ratings was 0.82 ($n$ = 147, $p < 0.001$) which shows a satisfactory agreement in the two ratings. In addition, a paired-samples $t$-test was conducted to compare the means of the sample in the two ratings. The results showed that there was not a statistically significant difference between the two means of raw scores across the two different ratings ($df$ = 146, $p$ = 0.06, $t$ = 1.87, mean difference = 0.63).

### Model fit

As a fundamental step in any statistical modeling, estimated parameters in CDMs are interpretable to what extent the model fits the data. Data were analyzed using the CDM package version 6.1-10 (Robitzsch, Kiefer, George, & Uenlue, 2018) in the R statistical software (R core Team, 2013). The CDM package employs marginal maximum likelihood estimation using the EM (Expectation-Maximization) algorithm for estimating and fitting the models (George, Robitzsch, Kiefer, Uenlue, & Grosz, 2016). The CDM package also produces a set of relative and absolute fit statistics which can be used for checking the fit of a model to the data and comparing multiple models to select the most appropriate model. As mentioned earlier, authors (A comparison of different cognitive diagnostic model for EFL writing, submitted) showed that the ACDM has

**Table 1** Final Q-matrix

| | Content | Organization | Grammar | Vocabulary | Mechanics |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 |
| 2 | 0 | 1 | 1 | 0 | 0 |
| 3 | 1 | 0 | 1 | 0 | 0 |
| 4 | 1 | 1 | 0 | 0 | 0 |
| 5 | 1 | 1 | 0 | 0 | 0 |
| 6 | 1 | 0 | 0 | 0 | 0 |
| 7 | 1 | 1 | 0 | 0 | 0 |
| 8 | 1 | 0 | 0 | 0 | 0 |
| 9 | 0 | 1 | 0 | 0 | 0 |
| 10 | 0 | 1 | 1 | 0 | 0 |
| 11 | 1 | 1 | 0 | 0 | 0 |
| 12 | 0 | 1 | 0 | 0 | 0 |
| 13 | 1 | 1 | 0 | 0 | 0 |
| 14 | 0 | 1 | 0 | 0 | 1 |
| 15 | 0 | 0 | 1 | 0 | 0 |
| 16 | 0 | 0 | 1 | 0 | 0 |
| 17 | 0 | 0 | 1 | 0 | 1 |
| 18 | 0 | 0 | 0 | 0 | 1 |
| 19 | 0 | 0 | 1 | 0 | 0 |
| 20 | 0 | 0 | 1 | 0 | 0 |
| 21 | 0 | 0 | 1 | 0 | 0 |
| 22 | 0 | 0 | 1 | 0 | 0 |
| 23 | 0 | 0 | 1 | 0 | 0 |
| 24 | 0 | 0 | 1 | 0 | 0 |
| 25 | 0 | 0 | 1 | 0 | 0 |
| 26 | 0 | 0 | 0 | 1 | 0 |
| 27 | 0 | 0 | 0 | 1 | 0 |
| 28 | 0 | 0 | 0 | 1 | 0 |
| 29 | 0 | 0 | 1 | 1 | 0 |
| 30 | 0 | 0 | 1 | 0 | 0 |
| 31 | 0 | 0 | 0 | 0 | 1 |
| 32 | 0 | 0 | 0 | 0 | 1 |
| 33 | 0 | 0 | 0 | 0 | 1 |
| 34 | 0 | 1 | 0 | 0 | 1 |
| 35 | 0 | 0 | 0 | 1 | 0 |

the best fit in terms of absolute and relative fit indices compared to other CDMs for L2 writing ability.

As can be seen from Table 2, there is a nonsignificant value for MX2 ($p > 0$), that is, the value of 31.2 indicates good fit. It must be noted that there is not a clear cut-off score or criterion for fit indices in CDMs. However, as a rule of thumb, the closer the value to zero, the better the model fits. Considering this, the value of 0.058 shows a relative good fit of the model to the data in terms of the MADcor (DiBello et al., 2007; Roussos et al., 2007). As to the MADRES, the value of 1.27 indicates a relatively poor

**Table 2** ACDM fit statistics

| Fit index | Estimate |
|---|---|
| MX2 | 31.2 |
| MADcor | 0.058 |
| MADRESIDCOV | 1.27 |
| MADQ3 | 0.056 |
| SRMSR | 0.074 |
| RMSEA | 0.083 |

fit of the model to the data. For MADQ3 and SRMSR, the values of 0.056 and 0.074 can be considered as good fit of the model. In terms of RMSEA, following Maydeu-Olivares and Joe (2014), the value of 0.083 is satisfactory. Overall, there are convincing evidence for supporting the fit of the ACDM model to the data. The large values of fit indices in this study could be due to the small sample size ($N$ = 500) and the large number of test items ($J$ = 35) (Kang, Yang, & Zeng, 2018; Kunina-Habenicht et al., 2009; Lei & Li, 2016; Lin & Weng, 2014).

In addition to checking the absolute fit of the model to the data, the fit of the ACDM was testified by estimating classification consistency $P_c$ and classification accuracy $P_a$. As demonstrated in Table 3, the classification accuracy (*Pa*) and consistency (*Pc*) for the whole latent class pattern is 0.74 and 0.63, respectively, indicating that the test possesses a 74% probability of accurately classifying a randomly selected respondent into his/her correct latent class from a single-test administration. It also has a 63% probability of classifying a randomly selected respondent consistently on two administrations of the test. The other rows of the table show the consistency and accuracy of classifying examinees according to the mastery or non-mastery of each attribute. Similar to absolute fit statistics, there is not a definite criterion for *P*α and *Pc* values. In the light of the results obtained by Cui, Gierl, and Chang (2012), Wang, Song, Chen, Meng, and Ding (2015), and Johnson and Sinharay (2018), the values of accuracy and consistency are fairly high and acceptable in the current study. These moderate values of *P*α and *Pc* at the whole-pattern can be due to the presence of moderate correlations among the attributes (Cui et al., 2012).

**Table 3** Classification consistency $P_c$ and accuracy $P_a$

| Classification accuracy and consistency | ACDM |
|---|---|
| P_a | 0.74 |
| P_c | 0.63 |
| P_a CON | 0.94 |
| P_c CON | 0.90 |
| P_a ORG | 0.95 |
| P_c ORG | 0.93 |
| P_a GRM | 0.93 |
| P_c GRM | 0.91 |
| P_a VOC | 0.94 |
| P_c VOC | 0.91 |
| P_a MCH | 0.91 |
| P_c MCH | 0.87 |

### ACDM analysis

As the first analysis of the ACDM model, the attributes mastery status of the whole group show that (Table 4), of the five sub-skills, the most difficult attributes to master were VOC and CON with probabilities of 32% and 43%, respectively, indicating that only 32% of students have mastered vocabulary use and 43% have mastered content fulfillment. However, ORG and MCH were identified as the easiest attributes with 61% and 56% probability, respectively followed by GRM with 52%. It suggests that 61% of the students mastered organizational effectiveness, 56% mastered mechanics, and only 52% had a mastery of grammatical knowledge.

As noted earlier, CDMs classify examinees into different latent classes according to the total number of attributes, e.g., $2^k$. In the present study, there are 32 possible latent classes (five sub-skills with $2^5 = 32$ latent classes) with respect to the Q-matrix configuration. To save space, data for only the first and last four latent classes are presented in the first column of Table 5. The second column shows the possible attribute pattern or profile for all the 32 latent classes.

As can be seen from the third column of the table, the attribute profile $\alpha_1$= [00000] had the highest class probability. It indicates that 22% of students, containing about 113 respondents (as shown in the last column), were classified as belonging to this latent class; therefore, these students are expected to have mastered none of the attributes. Skill profile of $\alpha_{32}$ = [11111] was the second most populated latent class with 19% probability, indicating that approximately 96 students belonged to this class who were expected to have mastered all of the attributes. The remaining profiles relate to students who mastered one of the attributes to four of the attributes.

Table 6 shows, for space considerations, the ACDM parameters for only the first three items or descriptors of the checklist. The first column gives the item number, the second column shows the requisite attributes for each item, the third column displays the attribute mastery patterns, and the last column represents the probability of a successful performance on each item with respect to the mastery of the required attributes by any given test item. As an illustration, successful performance on item 1 requires the presence of CON and ORG. Those students who have mastered none of the required attributes have only 6% probability of guessing to get the item right (e.g., item intercept). However, those students who have mastered CON have 37% chance to respond correctly to the item. Therefore, masters of CON had 0.06 + 0.37 = 0.43 probability of success (not slipping) on the item. Similarly, for those who have mastered ORG, there is 20% probability to give a correct response to the item. Thus, those who have mastered ORG have 0.06 + 0.20 = 0.26 probability of success (not slipping) on the item. It can be inferred that the CON discriminates more among masters and non-masters compared to ORG. In fact, mastery of CON increments the probability of a

**Table 4** Attribute difficulty

| Attributes | Attribute probability 1 |
|---|---|
| CON | 0.43 |
| ORG | 0.61 |
| GRM | 0.52 |
| VOC | 0.32 |
| MCH | 0.56 |

**Table 5** Class probabilities

| Latent class | Attribute pattern | Class probability | Class expected frequency |
|---|---|---|---|
| 1 | 00000 | 0.226 | 113.11 |
| 2 | 10000 | 0.019 | 9.69 |
| 3 | 01000 | 0.083 | 41.93 |
| 4 | 11000 | 0.023 | 11.85 |
| … | … | … | … |
| 29 | 00111 | 0.011 | 5.92 |
| 30 | 10111 | 0.020 | 10.23 |
| 31 | 01111 | 0.033 | 16.88 |
| 32 | 11111 | 0.192 | 96.15 |

successful performance on the given item. By mastering the two attributes, the probability of getting the item right increases to 63% (0.06 + 0.37 + 0.20). This shows the nature of the ACDM as an additive model in which each attribute additively contributes to the increase in the probability of a correct response.

Table 7 demonstrates the probabilities (posterior probability) that each student belonged to each of the 32 latent classes. For space considerations, the class probability for the first four students in the data are presented. The first column gives the latent classes ($2^5$ = 32) and the second column represents response patterns of the four students. As an illustration, there are 38% and 26% probabilities for student number 1 to belong to latent classes 19 and 3, respectively. By checking the latent classes 20 and 19, it reveals that he/she has 38% chance to have mastered ORG and MCH and 26% chance for mastering only ORG. For student number 4, as another example, there is 89% probability of belonging to latent class 1, indicating that he/she has mastered none of the attributes by checking the latent class.

Table 8 further provides the mastery probability of each student on any of the requisite attributes for a given test item or task. Due to the space limitation, the attribute mastery probability of only four randomly selected students are presented. The first column shows the student ID, followed by response pattern, attribute profile, the probability of belonging to this profile, and the attribute mastery probabilities. For instance, the probabilities that student 267 with the skill profile of [11011] has mastered the attributes CON to MCH are 0.96, 0.99, 0.00, 0.65, and .097, respectively. Put differently, there is a probability of 96% that he/she has mastered CON and 0% probability for mastering GRM. On the other hand, student number 101 with the profile mastery [00011] has a probability of mastery of 3%, 35%, 20%, 67%, and 63%, indicating that he/she has mastered VOC and MCH, but not CON, ORG, and GRM. The values above 0.50 show a high confidence for the mastery status of different sub-skills for each student (Hu, Miller, Huggins-Manley, & Chen, 2016).

Finally, an inspection of the tetrachoric correlation among the attributes revealed that there exist a moderate to strong correlations between the sub-skills. Overall, the values larger than 0.70 are considered as strong, 0.50 and 0.70 as moderate, and less than 0.50 as weak. Empirical studies showed that 0.50 is a logical value for correlation among attributes (e.g., Henson, Templin, & Douglas, 2007; Kunina-Habenicht, Rupp, & Wilhelm, 2012). As demonstrated in Table 9, there was a strong correlation coefficients between GRM and MCH (0.87) followed by GRM and VOC (0.78). It suggests that if a student

**Table 6** ACDM parameters

| Item Number | Required Attributes | Mastery Patterns | Probability |
|---|---|---|---|
| I1 | CON-ORG | A00 | 0.06 |
| I1 | CON-ORG | A10 | 0.37 |
| I1 | CON-ORG | A01 | 0.20 |
| I1 | CON-ORG | A11 | 0.63 |
| I2 | ORG-GRM | A00 | 0.12 |
| I2 | ORG-GRM | A10 | 0.30 |
| I2 | ORG-GRM | A01 | 0.45 |
| I2 | ORG-GRM | A11 | 0.87 |
| I3 | CON-GRM | A00 | 0.17 |
| I3 | CON-GRM | A10 | 0.32 |
| I3 | CON-GRM | A01 | 0.42 |
| I3 | CON-GRM | A11 | 0.91 |

performed well on GRM, he/she was likely to perform well on MCH and VOC as well, and vice versa. In fact, cognitive operations required to perform successfully on a test item or task were similar to each other. On the contrary, there existed a moderate correlation between VOC and ORG (0.50), and VOC and MCH (0.63); indicating that a good performance on VOC was not necessarily associated with a good performance on ORG and MCH.

## Discussion

This study set out to examine the utility of the ACDM, as a compensatory model, in generating reliable and useful diagnostic information about strengths and weaknesses of Iranian EFL university students' L2 writing ability. As a fundamental step in CDMs, an initial Q-matrix was designed based on the consensus among the raters, as content experts, on the association between the checklist items and the five writing sub-skills. The Q-matrix was then empirically validated through the procedure suggested by de la Torre and Chiu (2016) in GDINA package. In the next step, the data were subjected to the ACDM model for checking fit of the data to the model. The results of absolute fit statistics yielded adequate evidence supporting the fit of the ACDM model to the data based on almost all the criteria. Then, the fit of the ACDM was further supported by checking the classification consistency and accuracy and tetrachoric correlations among the attributes. The analysis of the attribute-level and profile-level *Pc* and *P*α indices indicated acceptable values for pattern-level and high values for sub-skill–level. In addition, the writing sub-skills generally produced moderate to high correlations with each other which could be considered as a sort of evidence for claiming that there exists a compensatory relationship among the L2 writing attributes. This finding is in line with the previous study conducted by authors (A comparison of different cognitive diagnostic model for EFL writing, submitted) who found that compensatory models outperform non-compensatory models in explaining L2 writing ability.

The results of the study also showed that the CDM-based checklist approach is very useful in both identifying the major patterns of skill mastery of Iranian EFL students in L2 writing ability and providing finer-grained information about their learning status.

**Table 7** Class probabilities for respondents

| Latent Class | Response patterns | | | |
|---|---|---|---|---|
| | 10110001011010101101110010000011000 | 01010001010100111000100000100000000 | 00001000010001001100111110000100110 | 00011001000100010000110010001100000 |
| Class 1 | 0.00 | 0.36 | 0.10 | 0.89 |
| Class 2 | 0.01 | 00 | 00 | 0.05 |
| Class 3 | 0.26 | 0.63 | 0.13 | 0.05 |
| Class 4 | 0.03 | 00 | 00 | 00 |
| Class 5 | 00 | 00 | 00 | 00 |
| Class 6 | 00 | 00 | 00 | 00 |
| Class 7 | 0.01 | 00 | 00 | 00 |
| Class 8 | 00 | 00 | 00 | 00 |
| Class 9 | 00 | 00 | 00 | 00 |
| Class 10 | 00 | 00 | 00 | 00 |
| Class 11 | 00 | 00 | 00 | 00 |
| Class 12 | 00 | 00 | 00 | 00 |
| Class 13 | 00 | 00 | 00 | 00 |
| Class 14 | 00 | 00 | 00 | 00 |
| Class 15 | 00 | 00 | 00 | 00 |
| Class 16 | 00 | 00 | 00 | 00 |
| Class 17 | 0.03 | 00 | 0.45 | 00 |
| Class 18 | 0.01 | 00 | 00 | 00 |
| Class 19 | 0.38 | 00 | 0.21 | 00 |
| Class 20 | 0.15 | 00 | 00 | 00 |
| Class 21 | 00 | 00 | 0.05 | 00 |
| Class 22 | 00 | 00 | 00 | 00 |
| Class 23 | 0.10 | 00 | 0.06 | 00 |

**Table 7** Class probabilities for respondents *(Continued)*

| Latent Class | Response patterns | | | |
|---|---|---|---|---|
| | 10110001011010101101110010000011000 | 01010001010100111000100000100000000 | 00001000010001001100111110000100110 | 00011001000100010000110010001100000 |
| Class 24 | 0.05 | 00 | 00 | 00 |
| Class 25 | 00 | 00 | 00 | 00 |
| Class 26 | 00 | 00 | 00 | 00 |
| Class 27 | 00 | 00 | 00 | 00 |
| Class 28 | 00 | 00 | 00 | 00 |
| Class 29 | 00 | 00 | 00 | 00 |
| Class 30 | 00 | 00 | 00 | 00 |
| Class 31 | 00 | 00 | 00 | 00 |
| Class 32 | 00 | 00 | 00 | 00 |

**Table 8** Skill-mastery probabilities

| Test takers | Response pattern | Attribute profile | Probability | CON | ORG | GRM | VOC | MCH |
|---|---|---|---|---|---|---|---|---|
| 7 | 00011110011110111100101110000001000 | 11001 | 0.43 | 0.90 | 0.99 | 0.08 | 0.00 | 0.54 |
| 101 | 10000001100100111100011111010011001 | 00011 | 0.17 | 0.03 | 0.35 | 0.20 | 0.67 | 0.63 |
| 267 | 11000111011111001100010010011001100 | 11011 | 0.62 | 0.96 | 0.99 | 0.00 | 0.65 | 0.97 |
| 487 | 11111111111111111111111111111111111 | 11111 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

As the results of the study showed, the two "flat" skill-mastery profiles, namely "non-master of all attributes" $\alpha_1$= [00000] and "master of all attributes" $\alpha_{32}$= [11111], were the most prevalent skill profiles. This is in contrast to the previous applications of CDMs to L2 writing by Kim (2011) and Xie (2016) who reported a small portion of test takers fell into the flat categories. They noted that generating various types of skill-mastery profiles is an indication of the model's high discriminant function. However, it should be noted that they reported a weak to moderate correlation among the writing attributes which affect the classification of examinees. The existence of flat skills profile can be due to the unidimensionality of the measured scale and the high correlations between the attributes (Lee & Sawaki, 2009a; Rupp et al., 2010). Lee and Sawaki (2009a) rightly noted that "… a CDA analysis may classify most of the examinees into flat profiles. This makes additional scores reported redundant, suggesting that reporting separate attribute scores provides little additional information over and above what a total score or overall proficiency score can offer. This can happen, for example, when a CDA is applied to a non-diagnostic test that was designed to be an essentially psychometrically unidimensional test for a target population (e.g., Luecht, Gierl, Tan, & Huff, 2006). When this happens, one can say that the utility of profile scoring is questionable from the psychometric point of view. This is a likely scenario in a domain such as language assessment where constructs are often found to be highly correlated among themselves" (p. 185).

As stated above, moderate to high correlations between the writing attributes were observed in the current study which can be considered as the reason for classifying most students into the flat profiles.

Furthermore, the results of the study showed that VOC and CON were the most difficult attributes among the five sub-skills of writing, followed by GRM, MCH, and ORG. This finding is in agreement with studies of Kim (2011) and Xie (2016), but compared with them, the students in the present study appeared to have a worse dominance on different attributes of L2 writing, with a higher mastery rates for GRM, MCH, and ORG (52%, 56%, and 61%, respectively) and a lower mastery rates for VOC and CON (43% and 32%, respectively). This finding substantiates English as a second language (ESL)/EFL writing research indicating that content and vocabulary are the most important attributes in the process of producing a high-level essay (Cheng, Klinger, & Zheng, 2007; Milanovic, Saville, & Shuhong, 1996; Schoonen, Snellings, Stevenson, & van Gelderen, 2009). According to Wilson et al. (2016), mechanics and grammar are low-level writing skills and vocabulary, organization, and content are higher-level writing skills. Therefore, it is evident that since vocabulary and content are higher-level skills, they require more cognitively advanced operations and therefore are difficult to master.

**Table 9** Tetrachoric correlations between the sub-skills

| | CON | ORG | GRM | VOC | MCH |
|---|---|---|---|---|---|
| CON | 1.00 | | | | |
| ORG | 0.64 | 1.00 | | | |
| GRM | 0.69 | 0.71 | 1.00 | | |
| VOC | 0.68 | 0.50 | 0.78 | 1.00 | |
| MCH | 0.69 | 0.64 | 0.87 | 0.63 | 1.00 |

## Conclusion

The present study investigated strengths and weaknesses of Iranian EFL university students in writing to determine their learning status. The findings indicated that most of the students are not able to produce a well-structured essay with respect to the interested attributes, especially in terms of vocabulary and content. One solution for this problem can be the use of diagnostic assessment in writing classes to monitor the performance of students over time (e.g., during a term or course). Writing classes would be more effective to students if a theory-based assessment is employed, as opposed to conventional method of providing feedback to students. This endeavor helps both teachers to have a better understanding of their students' weaknesses and strengths and students to receive short- and long-term feedback (Kim, 2011).

The previous studies on applications of CDMs on L2 writing by Kim (2011) and Xie (2016) suggested the use of the EDD checklist for different academic tasks or writing genres. By taking this recommendation into account, a descriptive essay was used in the current study. The obtained information verified the validity and accuracy of the checklist for applying to descriptive essays. As one of the anonymous reviewers noted, it would be interesting for future research to examine what form the EDD checklist would take if cognitive diagnostic tasks or a comparison of different academic tasks were used.

As in any research endeavor, the limitations of the present study should be acknowledged, and the conclusions drawn should be viewed within the constraints imposed on the study. This study was limited in that it applied a CDA approach to a non-diagnostic test which is problematic in terms of the validity of inferences about the test takers' skill-mastery profiles (DiBello et al., 2007; Jang, 2009). One urgent area for further investigation is designing a true diagnostic test (Ravand & Baghaei, 2019) according to a CDA framework. However, as noted earlier, retrofitting can be advantageous by determining the diagnostic capacity of existing achievement and proficiency tests before developing true diagnostic tests which need a big budget and a lot of time (Lee & Sawaki, 2009a). Overall, what is important is that cognitive diagnostic assessment has shown its potential in providing diagnostic information about the learning status of students. Consequently, more attention should be paid to designing and developing educational assessments in second/ foreign language contexts that are based on a CDM framework. Such an endeavor necessitates the cooperation of various experts from different fields of study (e.g., subject matter, measurement, pedagogy).

## Additional file

**Additional file 1:** The EDD Checklist, adapted from Kim, 2011. (DOCX 18 kb)

### Abbreviations

ACDM: Additive CDM; CDA: Cognitive diagnostic assessment; CDMs: Cognitive diagnostic models; CON: Content fulfillment; CTT: Classical test theory; EDD checklist: The Empirically derived Descriptor-based Diagnostic checklist; EFL: English as a foreign language; ESL: English as a second language; G-DINA: Generalized deterministic inputs, noisy "and" gate; GRM: Grammatical knowledge; IELTS: The International English Language Testing System; IRT: Item response theory; MCH: Mechanics; ORG: Organizational effectiveness; TOEFL: Test of English as a foreign language; VOC: Vocabulary use

### Acknowledgements

We thank all the students and teachers who helped us in collecting the data.

### Authors' contributions

FE collected and analyzed the data. PB supervised data analysis and provided the codes. AB read the manuscript and provided feedback. All authors wrote parts of the manuscript and revised it.

### Funding

We received no funding for this research from any funding agencies.

### Availability of data and materials

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

### Competing interests

The authors declare that they have no competing interests.

Received: 10 May 2019 Accepted: 26 June 2019
Published online: 16 August 2019

### References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. A & C Black.
Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301–320. https://doi.org/10.1191/0265532205lt310oa.
ALTE. (1998). *Multilingual glossary of language testing terms* (p. 1998). Cambridge: Cambridge University Press.
Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore–Cambridge general certificate of education O-level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *International Journal of Listening*, 1–24. https://doi.org/10.1080/10904018.2018.1500915.
Ashwell, T. (2000). Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing, 9*(3), 227–257. https://doi.org/10.1016/S1060-3743(00)00027-8.
Banerjee, J., & Wall, D. (2006). Assessing and reporting performances on pre-sessional EAP courses: Developing a final assessment checklist and investigating its validity. *Journal of English for Academic Purposes, 5*(1), 50–69. https://doi.org/10.1016/j.jeap.2005.11.003.
Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49–58 Retrieved from http://iej.cjb.net.
Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in Algebra using the rule-space model. *Journal for Research in Mathematics Education, 24*(5), 442–459. https://doi.org/10.2307/749153.
Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*(2), 102–118. https://doi.org/10.1016/j.jslw.2007.11.004.
Bitchener, J., & Knoch, U. (2008). The value of written corrective feedback for migrant and international students. *Language Teaching Research, 12*(3), 409–431. https://doi.org/10.1177/1362168808089924.
Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. https://doi.org/10.1080/0969595980050102.
Blikstad-Balas, M., Roe, A., & Klette, K. (2018). Opportunities to write: An exploration of student writing during language arts lessons in Norwegian lower secondary classrooms. *Written Communication, 35*(2), 119–154. https://doi.org/10.1177/0741088317751123.
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. https://doi.org/10.1177/026553229801500201.
Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*(3), 423–466. https://doi.org/10.1111/0023-8333.00016.
Buck, G., VanEssen, T., Tatsuoka, K., Kostin, I., Lutz, D., & Phelps, M. (1998). Development, selection and validation of a set of cognitive and linguistic attributes for the SAT I verbal: Analogy section. ETS Research Report Series, 1998: i-25. https://doi.org/10.1002/j.2333-8504.1998.tb01768.x.
Buckingham, L., & Aktuğ-Ekinci, D. (2017). Interpreting coded feedback on writing: Turkish EFL students' approaches to revision. *Journal of English for Academic Purposes, 26*, 1–16. https://doi.org/10.1016/j.jeap.2017.01.001.
Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly, 13*(3), 218–230. https://doi.org/10.1080/15434303.2016.1210610.
Cheng, L., Klinger, D., & Zheng, Y. (2007). The challenges of the Ontario Secondary School Literacy Test for second language students. *Language Testing, 24*, 185–208. https://doi.org/10.1177/0265532207076363.
Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2011.00158.x.

Cumming, A. (2015). Design in four diagnostic language assessments. *Language Testing, 32*(3), 407–416. https://doi.org/10.1177/0265532214559115.
de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7.
de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253–273. https://doi.org/10.1007/s11336-015-9467-8.
de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353. https://doi.org/10.1007/bf02295640.
de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595. https://doi.org/10.1007/s11336-008-9063-2.
de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*(2), 89–97. https://doi.org/10.1016/j.pse.2014.11.001.
DiBello, L. V., Roussos, L. A., & Stout, W. (2007). A review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 979–1030). Elsevier.
Dunsmuir, S., & Clifford, V. (2003). Children's writing and the use of information and communications technology. *Educational Psychology in Practice, 19*(3), 170–187. https://doi.org/10.1080/0266736032000109447.
Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Peter Lang.
Erlam, R., Ellis, R., & Batstone, R. (2013). Oral corrective feedback on L2 writing: Two approaches compared. *System, 41*(2), 257–268. https://doi.org/10.1016/j.system.2013.03.004.
Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Mahwah: Lawrence Erlbaum Associates.
George, A. C., Robitzsch, A., Kiefer, T., Uenlue, A., & Grosz, J. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1–24.
Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487.
Henson, R., Templin, J. L., & Douglas, J. 2007. Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement, 44*(4), 361–376. Retrieved from http://www.jstor.org/stable/20461869.
Henson, R. A., Templin, J. L., & Willse, J. T. (2008). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191. https://doi.org/10.1007/s11336-008-9089-5.
Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y.-H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing, 16*(2), 119–141. https://doi.org/10.1080/15305058.2015.1133627.
Hyland, K. (2013). Faculty feedback: Perceptions and practices in L2 disciplinary writing. *Journal of Second Language Writing, 22*(3), 240–253. https://doi.org/10.1016/j.jslw.2013.03.003.
Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing, 26*(1), 031–073. https://doi.org/10.1177/0265532208097336.
Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement, 55*(4), 635–664. https://doi.org/10.1111/jedm.12196.
Kang, C., Yang, Y., & Zeng, P. (2018). Q-Matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621618813104.
Kasai, M. (1997). *Application of the rule space model to the reading comprehension section of the test of English as a foreign language (TOEFL) (Unpublished doctoral dissertation)*. University of Illinois at Urbana Champaign.
Kim, A. Y. (2014). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258. https://doi.org/10.1177/0265532214558457.
Kim, Y.-H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing, 28*(4), 509–541. https://doi.org/10.1177/0265532211400860.
Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81–96. https://doi.org/10.1016/j.asw.2011.02.003.
Kumar, V., & Stracke, E. (2011). Examiners' reports on theses: Feedback or assessment? *Journal of English for Academic Purposes, 10*(4), 211–222. https://doi.org/10.1016/j.jeap.2011.06.001.
Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35*(2), 64–70. https://doi.org/10.1016/j.stueduc.2009.10.003.
Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*(1), 59–81. https://doi.org/10.1111/j.1745-3984.2011.00160.x.
Kuo, B.-C., Chen, C.-H., Yang, C.-W., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple-choice and constructed-response items. *Educational Psychology, 36*(6), 1115–1133. https://doi.org/10.1080/01443410.2016.1166176.
Lee, Y.-W., & Sawaki, Y. (2009a). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172–189. https://doi.org/10.1080/15434300902985108.
Lee, Y.-W., & Sawaki, Y. (2009b). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*(3), 239–263. https://doi.org/10.1080/15434300903079562.
Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement, 40*(6), 405–417. https://doi.org/10.1177/0146621616647954.
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 Retrieved from http://www.jstor.org/stable/1435314.
Li, H. (2011). *Evaluating language group differences in the subskills of reading using a cognitive diagnostic modeling and differential skill functioning approach (Doctoral dissertation)*. State College: Pennsylvania State University.
Li, H., & Suen, H. K. (2013). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing, 30*(2), 273–298. https://doi.org/10.1177/0265532212459031.

Lin, T.-Y., & Weng, L.-J. (2014). Graphical extension of sample size planning with AIPE on RMSEA using R. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 482–490. https://doi.org/10.1080/10705511.2014.915380.
Llosa, L., Beck, S. W., & Zhao, C. G. (2011). An investigation of academic writing in secondary schools to inform the development of diagnostic classroom assessments. *Assessing Writing, 16*(4), 256–273. https://doi.org/10.1016/j.asw.2011.07.001.
Luecht, R.M., Gierl, M.J., Tan, X., & Huff, K. (2006). *Scalability and the development of useful diagnostic scales*. Paper presented at the annual meeting of the NationalCouncil on Measurement in Education, San Francisco.
Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71. https://doi.org/10.1177/026553229501200104.
Ma, W., de la Torre, J., & Sorrel, M. (2018). The Generalized DINA Model Framework (R package version 2.0.8). Retrieved from https://cran.rproject.org/web/packages/GDINA
Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075.
McCutchen, D. (2011). From novice to expert: Language and memory processes in the development of writing skill. *Journal of Writing Research, 3*(1), 51–68. https://doi.org/10.17239/jowr-2011.03.01.3.
McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teachers' assessments of ESL writing samples (Unpublished master's thesis)*. Australia: University of Melbourne.
Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision making behavior of composition markers. In M. M. N. Saville (Ed.), *Studies in language testing 3: Performance testing, cognition and assessment* (pp. 92–111). Cambridge University Press: Cambridge.
Nation, I. S. P., & Macalister, J. (2010). *Language curriculum design*. Taylor & Francis.
Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575–603. https://doi.org/10.3102/00346543064004575.
Pellegrino, J. W., & Chudowsky, N. (2003). FOCUS ARTICLE: The foundations of assessment. *Measurement: Interdisciplinary Research and Perspectives, 1*(2), 103–148. https://doi.org/10.1207/S15366359MEA0102_01.
R Core Team (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.
Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*(1), 4–13. https://doi.org/10.1002/bs.3830280103.
Ranjbaran, F., & Alavi, S. M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation, 55*, 167–179. https://doi.org/10.1016/j.stueduc.2017.10.007.
Ravand, H. (2015). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782–799. https://doi.org/10.1177/0734282915623053.
Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*. https://doi.org/10.1080/15305058.2019.1588278.
Ravand, H., & Robitzsch, A. (2018). Cognitive diagnostic model of best choice: A study of reading comprehension. *Educational Psychology, 38*(10), 1255–1277. https://doi.org/10.1080/01443410.2018.1489524.
Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, C. (2018). CDM: Cognitive diagnosis modeling (Rpackage version 6.1-10). Retrieved from https://cran.rproject.org/web/packages/CDM/index.html
Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). Cambridge: Cambridge University Press.
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.
Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 219–262. https://doi.org/10.1080/15366360802490866.
Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. https://doi.org/10.1007/bf00117714.
Sampson, A. (2012). Coded and uncoded error feedback: Effects on error frequencies in adult Colombian EFL learners' writing. *System, 40*(4), 494–504. https://doi.org/10.1016/j.system.2012.10.001.
Sawaki, Y., Kim, H.-J., & Gentile, C. (2009). Q-Matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*(3), 190–209. https://doi.org/10.1080/15434300902801917.
Schoonen, R., Snellings, P., Stevenson, M., & van Gelderen, V. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchón (Ed.), *Writing in foreign language: Contexts learning, teaching, and research: Multilingual Matters*.
Scott, H. S. (1998). *Cognitive diagnostic perspectives of a second language reading test (Unpublished doctoral dissertation)*. Urbana: University of Illinois, Urbana-Champaign.
Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement, 34*(4), 333–352 Retrieved from http://www.jstor.org/stable/1435113.
Shintani, N. (2016). The effects of computer-mediated synchronous and asynchronous direct corrective feedback on writing: a case study. *Computer Assisted Language Learning, 29*(3), 517–538. https://doi.org/10.1080/09588221.2014.993400.
Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. *Assessing Writing, 18*(3), 187–201. https://doi.org/10.1016/j.asw.2013.05.001.
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 Retrieved from http://www.jstor.org/stable/1434951.
Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale: Lawrence Erlbaum Associates, Inc..
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.
Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327–369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x.

Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency: The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics, 14*(2), 60–82 URL: http://hdl.handle.net/10397/62844.
von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307. https://doi.org/10.1348/000711007x193957.
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.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205.
Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*(2), 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6.
Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*(3), 194–209. https://doi.org/10.1016/j.jslw.2007.0r7.004.
Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly, 37*(2), 345–354. https://doi.org/10.2307/3588510.
Williams, G. J., & Larkin, R. F. (2013). Narrative writing, reading and cognitive processes in middle childhood: What are the links? *Learning and Individual Differences, 28*, 142–150. https://doi.org/10.1016/j.lindif.2012.08.003.
Wilson, J., Olinghouse, N. G., McCoach, D. B., Santangelo, T., & Andrada, G. N. (2016). Comparing the accuracy of different scoring methods for identifying sixth graders at risk of failing a state writing assessment. *Assessing Writing, 27*, 11–23. https://doi.org/10.1016/j.asw.2015.06.003.
Xie, Q. (2016). Diagnosing university students' academic writing in English: Is cognitive diagnostic modelling the way forward? *Educational Psychology, 37*(1), 26–47. https://doi.org/10.1080/01443410.2016.1202900.
Xu, X., & von Davier, M. (2008). Fitting the structured general diagnostic model to NAEP data. *ETS Research Report Series*, i–18. https://doi.org/10.1002/j.2333-8504.2008.tb02113.x.
Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200. https://doi.org/10.1016/j.jslw.2006.09.004.
Yu, S., & Hu, G. (2017). Understanding university students' peer feedback practices in EFL writing: Insights from a case study. *Assessing Writing, 33*, 25–35. https://doi.org/10.1016/j.asw.2017.03.004.
Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). San Diego: Academic Press.

## Publisher's Note