Diagnosing English learners’ writing skills: A cognitive diagnostic modeling study

Zahra Shahsavar

Abstract: This study aims to clearly diagnose EFL students’ writing strengths and weaknesses. I adapted a diagnostic framework which comprises multiple writing components associated with three main dimensions of second language writing. The essays of 304 English learners who enrolled in an English language program were double-rated by 13 experienced English teachers. The data were submitted to cognitive diagnostic assessment (CDA). The provided evidence supports the validity of 21 writing descriptors tapping into three academic writing skills: content, organization, and language. The skill mastery pattern across writing shows that language was the easiest skill to master while content was the most difficult. In mastering language, the highest mastery levels were found in using capitalization, spelling, articles, pronouns, verb tense, subject-verb agreement, singular and plural nouns, and prepositions while some sub-skills such as sophisticated or advanced vocabulary, collocation, redundant ideas or linguistic expressions generated a hierarchy of difficulty. Researchers, teachers, and language learners may benefit from using the 21-item checklist in the EFL context to assess students’ writing. This method can have additional benefits for educators to know how to adapt other instruments for various purposes.

Subjects: Statistics & Probability; Educational Research; Higher Education; Classroom Practice

ABOUT THE AUTHOR
Zahra Shahsavar is an assistant professor at Shiraz University of Medical Sciences in Iran. She obtained her PhD in English Language from University Putra Malaysia (UPM). Her current research focuses on critical thinking in education, argumentative writing, writing assessment, online learning, and the use of technology for teaching and learning.

PUBLIC INTEREST STATEMENT
The purpose of this study was to provide the checklist to clearly diagnose EFL students’ writing strengths and weaknesses. Kim’s (2011) checklist that has been used in the ESL context was adapted. To identify the extent to which Kim’s checklist can diagnose EFL students’ writing, three writing experts reviewed and refined the checklist. Then, 13 experienced English teachers used the checklist to assess the writing of 304 female English learners. Our results support the validity of 21 writing components in the checklist for the EFL context. The skill mastery pattern across writing shows that language was the easiest skill to master while content was the most difficult one. Researchers, teachers, and language learners may benefit from using the 21-item checklist in the EFL context to distinguish writing skill masters from non-masters. This method can have additional benefits for educators to know how to adapt other instruments for various purposes.
Keywords: cognitive diagnostic assessment; EFL writing; Q-matrix validation; writing assessment

1. Introduction

Writing is a fundamental language skill which requires most assistance from teachers to develop (Vakili & Ebadi, 2019; Xie, 2017). In writing assessment, holistic and analytic rating approaches are commonly used. The holistic approach is helpful to perform a summative evaluation of students’ writing strengths (Bacha, 2001; Lee, Gentile, & Kantor, 2010), whereas the analytic approach serves to assess students’ writing based on pre-specified writing criteria and provide information concerning individual’s writing ability (Weigle, 2002).

Despite the advantages of the holistic and analytic approaches, the available instruments have rarely been used to diagnose students’ writing deficiencies and problems. One reason could be the conventional definitions of these scales: holistic scales typically sum up and reduce the performance of learners to a single mark, without any requirements to provide information on students’ areas of lack or strengths. Analytical scales, on the other hand, do provide information about coarse-grained components of language learners’ writing skills, but they do not provide sufficient discrimination across different criteria. For example, a significant number of such scales pull together multiple textual features such as spelling, punctuation, grammar, syntax, and vocabulary under one main major component or dimension such as language use (Xie, 2017). While these approaches offer the advantage of ease of assessment administration and rating as well as the saving of time, a potential problem is that they have disadvantages for the students who wish to receive more fine-grained information about their writing skills.

To address the shortcomings of the aforementioned approaches in writing assessment, some researchers (e.g. Banerjee & Wall, 2006; Kim, 2011; Struthers, Lapadat, & MacMillan, 2013; Xie, 2017) have shifted their attention to cognitive diagnostic assessment (CDA) to develop fine-grained evaluation systems that enable them to assess the areas where students need help to solve their problems in writing. These evaluation systems break down students’ second language (L2) writing performance into a wide range of descriptors which are related to different aspects of writing. For example, Banerjee and Wall’s (2006) checklist assesses the performance of English as a second language (ESL) students who studied at UK universities. The checklist contains a list of binary choices, can (yes) and do (must pay attention to) descriptors which can assess students’ academic writing as well as reading, listening and speaking abilities. The validity of the checklist was investigated in two stages: content relevance and convergence and making judgments about using the checklist by each student.

In another study which was inspired by Crocker and Algina’s (1986) steps to developing checklists, Struthers et al. (2013) evaluated various types of cohesion in ESL students’ writing. To assess the checklist reliability and validity, the panel of reviewers judged the items after generating an item pool based on the literature. After piloting and computing the inter-rater reliability, a 13-descriptor checklist which evaluated cohesion in writing was developed.

Likewise, Kim (2011) developed an empirically derived descriptor-based diagnostic (EDD) checklist that evaluates ESL students’ writing. The checklist consists of 35 descriptors in five distinct writing skills including content fulfillment, organizational effectiveness, grammatical knowledge, vocabulary use, and mechanics. There are some overlaps among the items of the writing skills (for more information readers are referred to Kim, 2011, p. 519).

Kim trained 10 English teachers to use the checklist to mark 480 TOEFL iBT (Internet-based Test) essays, which was then validated through CDA modeling. Recently, Xie (2017) adapted Kim’s (2011) instrument into diagnose ESL students’ writing in Hong Kong. Xie found Kim’s checklist more accurate to diagnose students’ strengths and weaknesses in L2 writing than using traditional scoring. However, she modified and shortened the list to make it fit for her context of research. The
main concern was that only the core components of Kim’s checklists are generalizable across different ESL contexts, which renders it a suitable device amenable to diagnosing the basic components of writing ability such as coherence, grammar, and vocabulary (Ariyadoust & Shahsavar, 2016).

Despite the encouraging results of adapting Kim's (2011) checklist, there are main concerns about using it. First, due to logistical problems, Kim’s EDD checklist was not applied in a real classroom setting where diagnostic assessment is highly demanded and most useful (Alderson, Haapakangas, Huhta, Nieminen, & Ullakonoja, 2015). Second, as Xie (2017) has shown, to gain highly reliable scores, the surface features of some checklist items need to be changed. That is, as programs in academic writing are different, Kim's checklist components may not function likewise in other contexts. Finally, the checklist was developed in an ESL setting once it was applied to the English as a foreign language (EFL) setting in Hong Kong in Xie’s (2017) study, the extent to which its components reliably assessed students’ writing in classrooms was not as certain as that in Kim’s study.

These issues have motivated us to re-validating Kim’s (2011) checklist in our study in order to diagnose EFL students' writing skills. Like Xie’s (2017) study, I set out to inspect the psychometric features of the instrument using the same quantitative data analysis techniques used by Kim (2011), while being cognizant of the fact that the initial and post-hoc editing of the instrument would be necessary to make it suitable for our context. Such modifications and updates in adapting measurement devices are required steps in validation research, which can render the end product more useful from the original instrument(s) for the context under investigation (Aryadoust & Shahsavar, 2016).

In the following section, I review the relevant literature and some of the issues that can arise in re-validating adapted instruments such as construct-irrelevant variance induced by raters (Messick, 1989). The review starts by defining CDA and covers the importance of CDA models particularly in writing an assessment.

1.1. Cognitive diagnostic assessment (CDA)
CDA is a relatively new development in the assessment that provides fine-grained information concerning learners’ mastery and non-mastery of tested language skills. CDA has its roots in item response theory (IRT) which is designed to find the probability of test takers answering items. However, unlike IRT models which assess one primary attribute and discard the items that do not fit the measured attribute, CDA models assess and measure multidimensional constructs which are influenced by both test-taker ability and item parameters (i.e. difficulty, discrimination, and pseudo-guessing). In CDA, the successful performance of each item depends on the particular IRT model and the mastery of multiple subskills being applied.

Recently, the demand for CDA has been increasing among educators and assessment developers (e.g. Kato, 2009; Lee, Park, & Taylan, 2011), as it pragmatically combines cognitive psychology and statistical techniques to evaluate learners’ language skills (e.g. Jang, 2008; Lee & Sawaki, 2009). The main advantage of CDA is its generic applicability to any assessment context (North, 2003) which allows the researchers to use it in language assessment not only in diagnosing the receptive skills such as reading/listening (e.g. Aryadoust & Shahsavar, 2016; Kim, 2015; Li & Suen, 2013) but also in the productive skills such as writing (Kim, 2011; Xie, 2017).

The first step in developing CDA models in language assessment is constructed definition and operationalization. Early studies in language assessment resulted in, for example, the Four Skills Model of writing comprising phonology, orthography, lexicon, and grammar (Carroll, 1968; Lado, 1961), or Madsen’s (1983) model comprising mechanics, vocabulary choice, grammar and usage, and organization. According to North (2003), such models suffer from three major drawbacks, namely, lack of differentiation between range or accuracy of vocabulary and grammar, lack of...
communication quality and content assessment, and overlooking fluency. To overcome these drawbacks in construct definition and operationalization, CDA researchers have emphasized the use of communicative-competence-based models. For example, Kim’s (2011) diagnostic checklist contains 35 binary descriptors (yes or no) associated with five major writing subskills (content fulfillment, organizational effectiveness, grammatical knowledge, vocabulary use and mechanics) which are largely based on communicative competence.

These representations of communicative competence have been validated through a class of CDA models called the reparameterized unified model (RUM) by Xie (2017) and Kim (2011). Like other CDA models, RUM also needs to have a Q-matrix, an incident matrix to specify the conceptual relationships between items and target subskills (Tatsuoka, 1983). Q-matrix can be defined as $Q = \{q_{ik}\}$ when a test containing $i$ test items measures $k$ sub-skills. If sub-skill $k$ is applied in answering item 1, then $q_{1k} = 1$, and if sub-skill $k$ is not assessed by the item, then $q_{ik} = 0$. Figure 1 presents a truncated version of a Q-matrix. It demonstrates four test items engaging four hypothetical sub-skills. For example, item 1 assesses sub-skills $c$ and $d$, whereas item 4 merely taps into sub-skill $a$. The accurate measurement of an identified sub-skill should be assessed by at least two or three test items. Each test item should be limited to testing a small number of subskills but the researcher is free to define the sub-skills and the generated Q-matrix accordingly (Hartz, Roussos, & Stout, 2002).

![Figure 1. Illustration of a Q-matrix of 4 items by 4 sub-skills.](image)

One of the concerns in CDA-based writing research is the role of construct-irrelevant factors such as raters’ erratic marking. The available CDA models do not provide any option to control the effect of rater severity as a source of variation in data (Kim, 2011). Hence, conventional latent variable modeling such as many-facet Rasch measurement (MFRM; Aryadoust, 2018; Linacre, 1994) have been used by many scholars to investigate and iron out such sources. Likewise, Kim (2011) recommends that MFRM should be used as a pre-CDA analysis to investigate whether/which raters have marked erratically. In addition, if several writing prompts/tasks are used, it is important to investigate their difficulty levels in CDA modeling. In our context, I have used three writing prompts which would lend themselves to CDA analysis should they have the same cognitive demands for the students (Engelhard, 2013).

In this study, I adapted Kim’s (2010, 2011) checklist and its construct in XXX [anonymized location] where diagnostic writing assessment is least-researched, and teachers are less familiar with the provision of diagnostic feedback with or without CDA checklists. The study particularly examines three research questions:

1. To what extent is the adapted writing assessment framework useful to diagnose EFL students’ writing?
2. To what extent can CDA identify the EFL students’ mastery of the pre-specified writing sub-skills?
3. What are the easiest and most difficult sub-skills for the EFL students to master in the present context?

2. Methodology

2.1. Participants

2.1.1. Raters
All participants took part in the study after receiving ethics committee approval. Thirteen female EFL teachers aged between 33 and 51 ($M = 39.85$; $SD = 5.20$). Among these teachers, 10 held
master's degrees and the rest bachelor's degrees in teaching English or English literature. They had teaching experience between 8 and 22 years (M = 14.69; SD = 4.90) in a Language School (LS; anonymized). LS is a language institute which was founded in 1979 in [anonymized location]. It is the largest language school in [anonymized location] with multiple branches throughout XX cities.

2.1.2. Students
In this study, 304 female EFL students from (Anonymized location) aged between 16 and 46 (M = 20.80; SD = 5.07) took the writing test in LS. Among these students, 128 (42%) was high school students, 21 (7%) held diploma, 137 (45%) got bachelor’s degrees, and 18 (6%) holds master’s degrees in different fields of study such as teaching English, English literature, information technology (IT), medicine, etc. Although the majority of students enroll in other language institutes before coming to LS, to assess the approximate level of students’ knowledge of English and place them in an appropriate level, they have to take the placement test designed by LS.

In this study, students were in advanced levels who had been studying English for approximately four years at LS. They had English class for two sessions per week. Each session took 115 minutes. During each session, 15 minutes were allotted for teaching and practicing various writing skills; students learned about the topic sentence, supporting sentence, body, unity, coherence, and completeness as essential parts of a good paragraph. They learned about a definition paragraph such as a concept, an idiom or an expression. They also learned various kinds of paragraphs such as narration, classification, comparison, contrast, exemplification, cause-effect, definition, process, and description.

2.2. Instrument
2.2.1. Prompts
After consulting with the LS teachers, based on students’ proficiency level and areas of interest, three appropriate argumentative writing prompts were chosen by the researcher (see Appendix A). Students were tasked with writing three essays. Time intervals of two weeks was considered for writing each essay. The average passage length was between 250 and 300 words.

2.2.2. The EDD checklist
In this study, a diagnostic assessment scheme, the EDD checklist, was adapted from Kim (2011) to diagnose students’ writing strengths and weaknesses in the EFL context. To adapt Kim’s checklist accurately, three writing experts including the researcher reviewed and refined the checklist to make it amenable to assessment in the context of the study. The review panelists holding postgraduate degrees in applied linguistics evaluated the clarity, readability, usefulness, and content of Kim’s checklist independently. The results of the review indicated that the conceptual representation of some descriptors overlapped. For example, I identified an overlap between 16 descriptors in MCH—a descriptor for evaluating the degree to which the writer follows English academic conventions—and in GRM—a descriptor for evaluating the degree to which the writer follows English grammar rules appropriately.

To remove the overlapping traits, GRM, VOC, and MCH were merged to form a unitary “language” dimension. Also, an abridged version of the checklist was created by rephrasing some items. For example, item 1 (i.e. This essay answers the question) was rephrased as “This essay answers the topic (prompt)”. Some double-barreled items were removed or broken down. For example, item 2 (i.e. This essay is written clearly enough to be read without having to guess what the writer is trying to say) was removed as it seemed repetitive and broke down item 3 (i.e. This essay is concisely written and contains few redundant ideas or linguistic expressions) into two independent items (i.e. items 2 and 25 in Appendix B). Based on this, a modified checklist comprising 25 descriptors were created tapping into three academic writing skills: content (7 items), organization (5 items), and language (13 items) (Appendix B).
2.3. Data collection

Thirteen LS teachers were tasked to use the checklist to assess students' essays. First, the researcher trained each rater individually, instructing them on how to use the checklist and apply binary response options. After training, nine randomly selected essays were given to all raters and the checklist inter-rater reliability was assessed by using the correlation between a single rater and the rest of the raters. After that, the raters evaluated 300 pieces of writing that were written by the students. The number of essays given to the raters differed based on their availability.

2.4. Data analysis

2.4.1. Considerations in developing the Q-matrix

In this study, RUM was applied to analyze the collected data. As noted earlier, the initial stage of the RUM analysis is the development of the Q-matrix. Aryadoust & Shahsavar's (2016) guidelines were followed for developing the Q-matrix, which include: no identified sub-skill should be measured by all test items since this would result in low discrimination between test takers who have and have not mastered the sub-skill. The proper diagnostic practice would be to compare students' performance on items that do and do not measure the particular sub-skill. Associating multiple sub-skills with exactly the same items were avoided, because, again, the diagnostic model would be unable to discriminate between test takers' mastery of the respective sub-skills, resulting in collinearity. It is generally impossible to conduct a reliable diagnostic assessment for a Q-matrix with collinear sub-skills.

In the present study, a Q-matrix is an incidence matrix that shows the relationship between skills and test items or descriptors. The results showed that RUM is a fruitful method to distinguish writing skill masters from non-masters and that EDD can provide accurate fine-grained information to EFL learners.

2.4.2. Choosing the items with sufficient fit

Two rounds of RUM analysis were conducted. In the first round, four malfunctioning items were removed from the analysis. The second round was performed to re-examine the psychometric features of the remaining items. Multiple fit and item indicators were considered to identify the best fitting items. Then, $r^n$, $p_*$, and $c_i$ parameters which inform the researcher about test takers, the Q-matrix, and the misspecifications observed were estimated. A reliable Q-matrix would produce $r^n$ coefficients less than .90, indicating the effectiveness of the test items at distinguishing between skill mastery and non-mastery; values less than 0.50 indicate that the particular sub-skill is required in order to answer the item accurately (Roussos, Xueli, & Stout, 2003). By contrast, high $p_*$ values are desirable; the higher the $p_*$ estimate, the higher the chance of test takers who have mastered the relevant sub-skill successfully employing the sub-skills tapped by that item. The parameter $c_i$, also known as the "completeness index," has a range between zero and three (Montero, Monfils, Wang, Yen, & Julian, 2003) and values closer to three indicate that the sub-skills required to accurately answer the test item are correctly specified in the Q-matrix, whereas zero indicates they are not fully specified (Zhang, DiBello, Puhan, Henson, & Templin, 2006).

2.4.3. Estimating the posterior probability of mastery and learner classification consistency

For each learner, a posterior probability of mastery (ppm) index for each sub-skill was estimated— ppm is the likelihood that the learner has mastered the sub-skill. We treated learners with $ppm > 0.60$ as masters, those with $ppm < 0.40$ as non-masters, and those between 0.40 and 0.60 as unknown (Roussos, DiBello, Henson, & Jong, 2010). The 0.40–0.60 region is called the indifference region and helps improve the reliability of mastery-level specifications, though it leaves the learners in that region unspecified. I estimated the proportion of the learners whose correct patterns were misclassified.
I further estimated learner classification consistency by using a set of statistics which indicate correspondence between real and simulated data. RUM generates multiple simulated samples to estimate master/non-mastery profiles for each simulated sample. For each sub-skill, the model estimates the Correct Classification Reliability (CCR)—which is the proportion of times that the learners are classified accurately—and the associated Test–retest Classification Reliability (TCR)—which is the rate by which the learner is classified identically on the two simulated parallel tests. CCR and TCR are estimated and reported individually for masters and non-masters of each skill (Roussos et al., 2010).

I further estimated learner sub-skills mastery levels ($\alpha_j$) as well as a composite score for all mastery estimations or the supplemental ability ($\eta_j$).

3. Results

3.1. Descriptive statistics

Table 1 presents the descriptive statistics of the items. The mean score indicates the difficulty with which that item was rated by the raters. Accordingly, the easiest item to score highly on is Item 24 ($M=0.795$; $SD=0.404$) and the most difficult is Item 20 ($M=0.20$; $SD=0.325$). Item 20 also has the highest skewness and kurtosis values (2.344 and 3.513, respectively) likely due to the uneven distribution of zero and one scores which would skew the tails of the distributions (high skewness) and increase the height of the distribution (high kurtosis). High skewness and kurtosis coefficients would weaken the normal distribution, but because RUM makes no

| Item  | Mean  | SD    | Skewness | Kurtosis |
|-------|-------|-------|----------|----------|
| Item1 | 0.756 | 0.429 | -1.200   | -0.562   |
| Item2 | 0.372 | 0.484 | 0.529    | -1.729   |
| Item3 | 0.540 | 0.498 | -0.165   | -1.982   |
| Item4 | 0.346 | 0.476 | 0.649    | -1.586   |
| Item5 | 0.392 | 0.488 | 0.441    | -1.815   |
| Item6 | 0.492 | 0.500 | 0.029    | -2.009   |
| Item7 | 0.302 | 0.460 | 0.861    | -1.265   |
| Item8 | 0.504 | 0.500 | -0.019   | -2.009   |
| Item9 | 0.401 | 0.490 | -0.015   | -2.009   |
| Item10| 0.503 | 0.500 | 0.199    | -1.970   |
| Item11| 0.450 | 0.498 | 0.194    | -1.972   |
| Item12| 0.451 | 0.498 | 0.800    | -1.367   |
| Item13| 0.314 | 0.465 | -0.694   | -1.525   |
| Item14| 0.663 | 0.473 | -0.694   | -1.525   |
| Item15| 0.663 | 0.473 | -0.349   | -1.888   |
| Item16| 0.585 | 0.493 | -0.209   | -1.966   |
| Item17| 0.551 | 0.497 | -0.924   | -1.151   |
| Item18| 0.689 | 0.463 | -0.820   | -1.333   |
| Item19| 0.120 | 0.325 | 2.344    | 3.513    |
| Item20| 0.477 | 0.500 | 0.092    | -2.001   |
| Item21| 0.238 | 0.426 | 1.235    | -0.477   |
| Item22| 0.718 | 0.450 | -0.977   | -1.051   |
| Item23| 0.795 | 0.404 | -1.468   | 0.156    |
| Item24| 0.346 | 0.476 | 0.649    | -1.586   |
assumptions regarding the normality of the distribution, the large kurtosis and skewness coefficients would not jeopardize the validity of the data.

3.1.1. MFRM

MFRM analysis was performed to investigate the performance of raters as well as the difficulty level of the three prompts. Raters’ performance was evaluated in terms of their fit statistics and spread of their severity levels, and task difficulty was evaluated in terms of the observed differences between their logits as well as the reliability; in this particular case, a low reliability was aimed since reliability in MFRM indicates the degree of dispersion. Figure 2 provides an overview of

| Measure | Students | Rater | Task | Items |
|---------|----------|-------|------|-------|
| 4 +     | +        | +     | +    |       |
| .       |          |       |      |       |
| .       |          |       |      |       |
| .       |          |       |      |       |
| 3 +     | .        | +     | +    |       |
| .       |          |       |      |       |
| .       |          |       |      |       |
| .       |          |       |      |       |
|          | *        |       |      |       |
| 2 +     | *        | +     | +    |       |
| .       |          |       |      |       |
| .       |          |       |      |       |
| .       |          |       |      |       |
|          | *        |       |      |       |
| 1 +     | ***      | +     | +    | *     |
| .       |          |       |      |       |
| *       |          |       |      |       |
| 0 +     | *****    | 13 3  | *    | 2 3   |
| .       |          |       |      |       |
| .       |          |       |      |       |
| .       |          |       |      |       |
| 9 +     | 1 11 7 8 | *     |      |       |
| .       |          |       |      |       |
| 8 +     | 10 2 9   | *     |      |       |
| .       |          |       |      |       |
| .       |          |       |      |       |
| 7 +     | *        | +     | +    | **    |
| .       |          |       |      |       |
| 6 +     | *        | +     | +    | **    |
| .       |          |       |      |       |
| 5 +     | *        | +     | +    | **    |
| .       |          |       |      |       |
| 4 +     | *        | +     | +    | **    |
| .       |          |       |      |       |
| 3 +     | *        | +     | +    | **    |
| .       |          |       |      |       |
| 2 +     | *        | +     | +    | **    |
| .       |          |       |      |       |
| 1 +     | *        | +     | +    | **    |
| .       |          |       |      |       |
| 0 +     | *        | +     | +    | **    |
| .       |          |       |      |       |

Figure 2. Variable maps of the data.
the spread of the four facets in the analysis: students, raters, tasks, and items. Students’ ability constitutes a widespread from −3.01 to 3.86 logits with the reliability and separation of 0.84 and 2.31 ($p < 0.05$), indicating that the assessment has successfully discriminated approximately two groups of writing proficiency. The level of rater severity ranges from −0.35 to 0.72 logits the reliability and separation of 0.90 and 3.00 ($p < 0.05$), indicating the presence of three different rater severity levels. Although raters are different in their severity levels, their infit and outfit MNSQ values (between 0.81 and 1.22) indicate that they rated consistently across tasks and students. Finally, the three tasks are virtually equally difficult (reliability/separation = 0.00), suggesting that the data generated can be submitted to RUM analysis.

3.1.2. The reparametrized unified model (RUM)

3.1.2.1. Round 1. RUM estimated the proportion of masters and non-masters of content, language, and organization subskills. Table 2 presents the results of the first analysis. The far left column gives the identifying number of the items, followed by a $p_i^*$ column, three $r^*n$ columns, and a $c_i$ (completeness index) column. Problem indices have been emboldened. The $p_i^*$ values are greater than zero, with the majority of them falling above 0.600, indicating a high likelihood that the test takers who have mastered the subskills executed them successfully. As $r^*n$ columns present, items one to seven and items eight to 12 discriminate between high and low ability students in content and organization, respectively. Items 13 to 25 tap language skills; of these items 13, 21, and 25 have $r^3$ indices greater than 0.900, indicating poor discrimination power. The completeness index for eight items is low, indicating that the subskills required to accurately

| Item | $p_i^*$ | $r^1$ (content) | $r^2$ (organization) | $r^3$ (language) | $c_i$ |
|------|---------|----------------|----------------------|----------------|------|
| 1    | 0.984   | 0.719          | 0                    | 0              | 1.871|
| 2    | 0.905   | 0.762          | 0.675                | 0              | 0.113|
| 3    | 0.899   | 0.662          | 0                    | 0              | 0.868|
| 4    | 0.960   | 0.231          | 0                    | 0              | 0.330|
| 5    | 0.937   | 0.362          | 0                    | 0              | 0.655|
| 6    | 0.961   | 0.611          | 0                    | 0              | 0.499|
| 7    | 0.661   | 0.578          | 0                    | 0              | 0.338|
| 8    | 0.854   | 0              | 0.123                | 0              | 2.296|
| 9    | 0.802   | 0              | 0.061                | 0              | 1.557|
| 10   | 0.888   | 0              | 0.186                | 0              | 1.600|
| 11   | 0.949   | 0              | 0.517                | 0              | 0.333|
| 12   | 0.915   | 0              | 0.662                | 0              | 0.256|
| 13   | 0.485   | 0              | 0                    | 0.912          | 0.635|
| 14   | 0.907   | 0              | 0                    | 0.534          | 2.473|
| 15   | 0.966   | 0              | 0                    | 0.435          | 2.366|
| 16   | 0.940   | 0              | 0                    | 0.244          | 2.701|
| 17   | 0.760   | 0              | 0                    | 0.568          | 2.196|
| 18   | 0.914   | 0              | 0                    | 0.583          | 2.587|
| 19   | 0.918   | 0              | 0                    | 0.543          | 2.555|
| 20   | 0.306   | 0              | 0                    | 0.432          | 0.271|
| 21   | 0.925   | 0              | 0                    | 0.948          | 0.132|
| 22   | 0.664   | 0              | 0                    | 0.327          | 0.084|
| 23   | 0.947   | 0              | 0                    | 0.684          | 1.803|
| 24   | 0.936   | 0              | 0                    | 0.846          | 2.010|
| 25   | 0.372   | 0              | 0                    | 0.980          | 2.503|
answer the test item might need to be respecified in the Q-matrix. These findings indicate that although most of the items functioned reasonably well, the Q-matrix would benefit from some respecification, and some items would need to be deleted from the analysis. It was decided to delete items 13, 20, 21, and 25 to re-specify the Q-matrix and re-run the analysis.

3.1.2.2. Research question 1.
To what extent is the adapted writing assessment framework useful to diagnose EFL students’ writing? In the second Round, we initially examined the $r^*$, $p^*$, and $c_i$ parameters to answer the first research question, since test takers’ estimated mastery levels are reliable to the extent that item parameters fall within the constraints tenable. Table 3 presents the results of the second round of RUM analysis. This analysis comprises 21 items, which function adequately, as testified by their $r^*$, $p^*$, and $c_i$ parameters (see Appendix C). For example, the index of item 1 is 0.986 indicating a high chance that the learners mastering the sub-skill tapped by item 1 executed that sub-skill successfully. The majority of items discriminate masters from non-masters, as indicated by their $r^*$ statistics which either falls below 0.50 or is slightly greater than 0.50. I made a decision to keep the item on the checklist because it is an important descriptor whose discrimination power might improve if it is administered to a larger sample. Finally, a general improvement in $c_i$ statistics was found, suggesting that the sub-skills tested by the items have been specified successfully. In addition, the average observed and modeled item difficulty values were 0.495 and 0.502, respectively (correlation = 0.999), suggesting a significantly high degree of model-data fit.

| Item | $p^*$ | $r^*$ (content) | $r^*$ (organization) | $r^*$ (language) | $c_i$ |
|------|-------|----------------|----------------------|-----------------|------|
| 1    | 0.986 | 0.822         | 0                    | 0               | 1.492|
| 2    | 0.875 | 0             | 0.597                | 0               | 0.100|
| 3    | 0.946 | 0             | 0.685                | 0               | 0.571|
| 4    | 0.906 | 0.138         | 0                    | 0               | 0.901|
| 5    | 0.860 | 0.183         | 0                    | 0               | 1.911|
| 6    | 0.970 | 0.546         | 0                    | 0               | 0.652|
| 7    | 0.609 | 0.367         | 0                    | 0               | 1.204|
| 8    | 0.861 | 0             | 0.174                | 0               | 2.414|
| 9    | 0.771 | 0             | 0.057                | 0               | 2.295|
| 10   | 0.896 | 0             | 0.202                | 0               | 1.875|
| 11   | 0.914 | 0             | 0.485                | 0               | 0.518|
| 12   | 0.860 | 0             | 0.559                | 0               | 0.533|
| 14   | 0.905 | 0             | 0                    | 0.512           | 2.493|
| 15   | 0.974 | 0             | 0                    | 0.429           | 2.169|
| 16   | 0.944 | 0             | 0                    | 0.248           | 2.581|
| 17   | 0.753 | 0             | 0                    | 0.521           | 2.436|
| 18   | 0.911 | 0             | 0                    | 0.571           | 2.698|
| 19   | 0.918 | 0             | 0                    | 0.573           | 2.363|
| 22   | 0.657 | 0             | 0                    | 0.372           | 0.070|
| 23   | 0.956 | 0             | 0                    | 0.696           | 1.671|
| 24   | 0.946 | 0             | 0                    | 0.752           | 1.902|
| pk   |       |               |                      |                 |      |
| Num  | 1     | 2             | 3                    |                 |      |
| Value| 0.402 | 0.515         | 0.532                |                 |      |
Another fit estimation analysis was the correlation between estimated general ability against their model-expected values which is presented in Figure 3. The correlation of these estimates is .950 ($R^2 = 0.90$), which indicates a high level of global fit.

I also investigated the diagnostic capacity of the estimated skill mastery profiles by comparing the proportion-correct values for the masters and non-masters across the items. As Figure 4 presents, the item masters performed significantly better than the item non-masters. The average scores for the masters ranged from 0.336 to 0.97 and average scores for the non-masters ranged from 0.005 to 0.643, suggesting that the RUM analysis has achieved significantly high precision and that on average masters have achieved higher scores. The results indicate that the CDA approach can distinguish master writers from non-masters.
3.1.2.3. Research question 2. To what extent can CDA identify the EFL students’ mastery of the pre-specified writing sub-skills? Table 4 presents learner classification consistency represented by CCR and TCR statistics for masters and non-masters and Cohen’s Kappa. CCR and TCR statistics for masters/non-mastered across content, organization, and language are greater than 0.900, indicating significantly high accuracy in determining the mastery level of learners. The Cohen’s kappa values which report the correspondence between expected and observed classifications also testify to the high accuracy of classifications. The far right-hand side column gives the proportion of learners in the indifference region, which is negligible across the three sub-skills.

| Sub-skills | CCR | Master % CCR | Non-Master % CCR | Cohen’s kappa | TCR | Master TCR | Non-Master TCR | Proportion |
|------------|-----|--------------|------------------|---------------|-----|------------|----------------|------------|
| content    | 0.949 | 0.948       | 0.949            | 0.893         | 0.903 | 0.902      | 0.903          | 0.046      |
| organization | 0.963 | 0.969       | 0.956            | 0.925         | 0.926 | 0.940      | 0.915          | 0.017      |
| language   | 0.949 | 0.968       | 0.928            | 0.897         | 0.904 | 0.937      | 0.866          | 0.029      |
| Average    | 0.954 | 0.963       | 0.945            | 0.905         | 0.912 | 0.927      | 0.895          |            |

I estimated the proportion of the learners whose correct patterns were misclassified; Table 5 shows that there was almost 0.00% of misclassification out of 100,000 simulated learners’ responses; almost 100% of learners had no error for three sub-skills classifications. In addition, only 0.6% and 12.2% of master learners were misclassified as non-masters for two and three sub-skills, respectively. The estimated accuracy of classifications is 87.2%, suggesting that the classification achieved high reliability.

| # out of 100,000 | 0 sub-skills | 1 sub-skill | 2 sub-skills | 3 sub-skills |
|------------------|--------------|-------------|--------------|-------------|
| Proportion (%)   | 0.872 (87.2%) | 0.122 (12.2%) | 0.006 (0.6%) | 0 (0.00%) |

3.1.2.4. Research question 3. What are the easiest and most difficult sub-skills for the EFL students to master in the present context? Table 6 demonstrates the average ability scores of learners on each of the sub-skills alongside their standard deviation indices. Content has the lowest average score (M = 0.401; SD = 0.031) and language the highest (M = 0.532; SD = 0.040), indicating that content was the most difficult skill to gain mastery of and language was the easiest.

| Sub-skill | Mean | SD |
|-----------|------|----|
| content   | 0.401 | 0.031 |
| organization | 0.514 | 0.033 |
| language | 0.532 | 0.040 |

4. Discussion
This study investigates the diagnostic capacity of the adapted writing framework developed by Kim (2010) in an EFL context. The results indicated that the CDA approach can distinguish EFL master
writers from non-masters reasonably well, which is in line with the previous research where CAD was found to be a practical and reliable method to discriminate students’ writing skills (Imbos, 2007; Kim, 2011; Xie, 2017). The results also confirm the findings of Alderson (2005) who argued that diagnostic assessment can be applied to assess students’ writing analytically and to make a more useful basis for providing feedback to language learners. In response to the limitations of teacher feedback which does not provide reasonable interpretations for students (Roberts & Gierl, 2010), diagnostic assessment integrates diagnostic feedback to L2 writing and helps teachers identify students’ weaknesses and strengths. Indeed, the CDA technique can assist teachers in communicating with curriculum developers, school administrators, and parents to enhance students’ learning (Jang, 2008).

Nevertheless, one of the caveats of CDA is that it requires expertise in psychometric modeling. While achieving expertise in CDA can be a challenge for many teachers, if not all, such a goal is not implausible. Kim (2010) argues that in multiple Asian environments including Japan, Hong Kong, and Taiwan, many local teachers have received training on psychometrics and developed localized language tests, thus abandoning the “high-stakes” tests of English. Despite the financial hurdles that can exert negative impacts on teacher education programs in low-resource environments, teacher education and psychometric training should be considered by language schools and institutions and teachers should be spurred to pursue psychometric knowledge.

Another finding emerging from the study is that RUM provided precise and correct mastery profiles for the writing skills of EFL learners. The proportion of the learners whose correct patterns were misclassified was almost 0.00% out of 100,000 simulated learners’ responses. This useful statistics can be used to judge how likely it is that an individual examinee has a certain number of errors in their estimated skill profile. For example, in a study conducted by Henson, Templin, and Willse (2009), 99% of the simulated examinees had one or no errors. Such statistics are easy for score users to interpret. When they get an estimated score profile, they can compare the results with their own records. Thus, if score users find one skill classification that does not seem quite right, they should feel comfortable that they can disregard it without spending too much time and effort trying to figure out the discrepancy. This is similar to the reliability estimation that teachers may be able to understand and use in their diagnostic score interpretations for a given skills profile.

In this study, the skill mastery pattern across writing shows that language was the easiest skill to master while content was the most difficult. There are several possible implications for these results.

First, in mastering language, the highest mastery was found in using capitalization, spelling, articles, pronouns, verb tense, subject-verb agreement, singular and plural nouns, and prepositions. The result is consistent with Kim’s (2011) study where the learners tended to master spelling, punctuation, and capitalization before they mastered other skills. It is also confirms Nacira’s (2010) belief that although a lot of students’ writing mistakes related to capitalization, spelling, and punctuation, mastering these skills is easier than mastering other skills such as content and organization (Nacira, 2010).

Second, in this study, the language was the easiest skill to master, while some sub-skills (e.g. sophisticated or advanced vocabulary, collocation, redundant ideas or linguistic expressions) generated a hierarchy of difficulty. This problem has been detected by Kim (2011) who noted that mastering vocabulary use was highly demanding yet essential for students to improve their writing. According to Kim, the learners need advanced cognitive skills along with exact, rich, and flexible vocabulary knowledge in L2 writing. This result matches the findings of other researchers who found mastering vocabulary as the most difficult skill for students. They believe that
improving students’ vocabulary has a leading role in developing their language skills such as writing (Kobayashi & Rinnert, 2013).

Third, students’ lack of skill in mastering collocation supports the previous research which indicates that it is hard for students to master collocation in improving their writing skill because of the native culture influence (Sofi, 2013). Moreover, students’ lack of skill in mastering vocabulary likely affects their collocation mastery. This speculation is based on Bergström’s (2008) study that shows a strong relationship between vocabulary knowledge and collocation knowledge.

Finally, Students’ lack of skill in linguistic expressions may imply that L2 writers try to interpret their ideas by indicating the meaning rather than applying an accurate linguistic expressions (Ransdell & Barbier, 2002).

5. Conclusions
Announcing students’ writing scores without sufficient descriptions on scores can be useless to test takers who are eager to improve their writing (Knoch, 2011). Although rating scales function as the main test construct in a writing assessment, there is little research which identifies how rating scales are constructed for diagnostic assessment.

English researchers would benefit from diagnostic assessment of writing due to the heightened need to respond to EFL students’ writing products; they should raise their awareness of their strengths and weaknesses. Knoch (2011) argues that most available writing tests are inappropriate for diagnosing EFL/ESL learners’ writing issues and proposes multiple guidelines for developing diagnostic writing instruments. Unlike diagnostic assessment, many writing assessment models only focus on surface level textual features such as grammar, syntax, and vocabulary.

This study supported the validity of 21 writing descriptors tapping into three academic writing skills: content, organization, and language (see Appendix C). RUM applied as a fruitful method to distinguish writing skill masters from non-masters and that the validated framework can provide accurate fine-grained information to EFL learners. In fact, CDA applied in this study may provide valuable information for researchers seeking to assess EFL students’ writing by using the framework. The findings may carry significant implications for teachers who complain about the quality of students’ writing or those who do not have substantial assessment practices in evaluating EFL students’ writing. As reporting on simple scores may not lead to improvements in their writing (Knoch, 2011), it seems reasonable to provide a detailed description of test takers’ writing behavior by using diagnostic assessment. In addition, the report on strengths and weaknesses of test takers’ writing may help students monitor their own learning (Mak & Lee, 2014). The results may also draw interest from researchers, curriculum developers, and test takers who are research-minded and eager to promote EFL students’ writing skill.

The results of this study are subject to some limitations. First, students were not trained on how to apply the EDD checklist items. Sufficient context-dependent training and supervision in applying EDD checklist may help students improve their writing in various situations in their future performance. Second, the number of prompts and also the ordering of the essays may affect the result of CDA, too. Third, the students did not collaborate with researchers in developing the checklist; the results could have been different had I had students’ assistance by conducting interviews and using think-alouds in developing the framework. Fourth, students’ mastery in different subscales is another concern; longitudinal studies should be conducted to confirm if students retain their mastery in their future writing. Fifth, although I applied the CDA model to assess students’ writing analytically, it is not evident that a discrete-point method alone can evaluate students’ writing in the best way. There is a need for further research to combine the holistic and analytical evaluation in this area. Finally, the content has been considered the main element characterizing good writing by Milanovic, Saville, and Shuhong (1996), and similarly, it was the most difficult skill to master in this study. The lowest proportion
of masters in the content may accrue from students’ lack of deep thinking in writing clear thesis statements that support their ideas and provide logical and appropriate examples to support their writing. How to apply various sub-skills to improve students’ writing is another crucial issue to investigate in the future research.

Taken together, these findings substantiate the claim that students’ attribute mastery profiles inform them of their strengths and weaknesses and aid them in achieving mastery of the skills targeted in the curriculum. Indeed, CDA of language skills can engage students in different learning and assessment activities if teachers help them understand the information provided and use it to plan new learning (Jang, 2008).

Acknowledgements
I would like to thank Dr. Vahid Aryadoust of the National Institute of Education of Nanyang technological University, Singapore, for his mentorship during the design of the study, data analysis, writing the results, and editing the manuscript. I also wish to thank the raters for taking their valuable time and effort in scoring students’ writing.

Funding
This work was supported by Shiraz Medical School of Sciences, followed by the grant number 98-01-10-19547.

Author details
Zahra Shahsavar E-mail: shahsavarzahra@gmail.com
1 English Language Department, Shiraz University of Medical Sciences, Shiraz, Iran.

Citation information
Cite this article as: Diagnosing English learners’ writing skills: A cognitive diagnostic modeling study, Zahra Shahsavar, Cogent Education (2019), 6: 1608007.

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

Alderson, J. C., Haapakangas, E. L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). The diagnosis of reading in a second or foreign language. New York, NY: Routledge.

Aryadoust, V. (2018). Using recursive partitioning Rasch trees to investigate differential item functioning in second language reading tests. Studies in Educational Evaluation, 56, 197–204. doi:10.1016/j.stueduc.2018.01.003

Aryadoust, V., & Shahsavar, Z. (2016). Investigating the validity of the Persian blog attitude questionnaire: An evidence-based approach. Journal of Modern Applied Statistical Methods, 15(1), 417–451. doi:10.22237/jmasm/1462074660

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? System, 29, 371–383. doi:10.1016/S0346-251X(01)00025-2

Bonnerjee, 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. doi:10.1016/j.jeap.2005.11.003

Bergström, K. (2008) Vocabulary and receptive knowledge of English collocations among Swedish upper secondary school students (Unpublished doctoral dissertation), Stockholm University, Sweden.

Carrol, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), Language Testing Symposium: A Psycholinguistic Approach (pp. 46–69). Oxford, UK: Oxford University Press.

Crocken, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Engelhard, G. (2013). Invariant measurement: Using Rasch models in the social, behavioral, and health sciences. New York: Routledge.

Hartz, S., Roussos, L., & Stout, W. (2002). Skill Diagnosis: Theory and practice [Computer software user manual for Arpeggio software]. Princeton, NJ: Educational Testing Service.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. Psychometrika, 74 (2), 191–210. doi:10.1007/s11159-008-9089-5

Imbos, T. (2007). Thoughts about the development of tools for cognitive diagnosis of students’ writings in an e-learning environment. IAISE/ISI Satellite Retrieved from http://iase-web.org/documents/papers/sat2007/Imbos.pdf

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y. R. Chung, & J. Xu (Eds.), Towards adaptive CALL: Natural Language Processing for Diagnostic Language Assessment (pp. 117–131). Ames: Iowa State University.

Kato, K. (2009) Improving efficiency of cognitive diagnosis by using diagnostic items and adaptive testing (Unpublished doctoral dissertation), Minnesota University, USA.

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. Language Testing, 32(2), 227–258. doi:10.1177/0265532214538457

Kim, Y. H. (2010). An argument-based validity inquiry into the empirically-derived descriptor-based diagnostic (EDD) assessment in ESL academic writing (Unpublished doctoral dissertation), University of Toronto, Canada.

Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified. Language Testing, 28(4), 509–541. doi:10.1177/0265532214008860

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. doi:10.1016/j.asw.2011.02.003

Kobayashi, H., & Rinner, C. (2013). L1/L2 writing development: Longitudinal case study of a Japanese multicompetent writer. Journal of Second Language Writing, 22(1), 4–33. doi:10.1016/j.jslw.2012.11.001

Lado, R. (1961). Language testing. New York: McGraw-Hill.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. International Journal of Testing, 11(2), 144–177. doi:10.1080/15305058.2010.534571
Lee, Y. W., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. Applied Linguistics, 31(3), 391–417. doi:10.1093/applin/amp040

Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnostic approaches to language assessment: An overview. Language Assessment Quarterly, 6(3), 72–189.

Li, H., & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. Educational Assessment, 18(1), 1–25. doi:10.1080/01443410.2013.761522

Linacre, J. M. (1994). Many-facet Rasch measurement. Chicago: MESA Press.

Madsen, H. S. (1983). Techniques in Testing. Oxford, UK: Oxford University Press.

Mak, P., & Lee, I. (2014). Implementing assessment for learning in L2 writing: An activity theory perspective. System, 47, 73–87. doi:10.1016/j.system.2014.09.018

Messick, S. (1989). Validity. In R. Linn (Ed.), Educational Measurement (3rd ed., pp. 13–103). Washington, DC: American Council on Education/Macmillan.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), Studies in language testing 3: Performance testing, cognition and assessment (pp. 92–111). Cambridge: Cambridge University Press.

Montero, D. H., Monfils, L., Wang, J., Yen, W. M., & Julian, M. W. (2003, April). Investigation of the application of cognitive diagnostic testing to an end-of-course high school examination. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Chicago, IL.

Nacira, G. (2010). Identification and analysis of some factors behind students Poor writing productions the case study of 3rd year students at the English department (Unpublished master thesis), Batna University, Algeria.

North, B. (2003). Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. TOEFL Monograph 24. Princeton NJ: Educational Testing Service.

Ransdell, S., & Barbier, M. L. (2002). New directions for research in L2 writing. The Netherlands: Kluwer Academic Publishers.

Roberts, M. R., & Gierl, M. (2010). Developing score reports for cognitive diagnostic assessments. Educational Measurement: Issues and Practice, 29(3), 25–38. doi:10.1111/emip.2010.29.issue-3

Roussos, L., Xueli, X., & Stout, W. (2003). Equating with the Fusion model using item parameter invariance. Unpublished manuscript, University of Illinois, Urbana-Champaign.

Roussos, L. A., DiBello, L. V., Henson, R. A., & Jang, E. E. (2010). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S. Embretson & J. Roberts (Eds.), New directions in psychological measurement with model-based approaches (pp. 35–69). Washington, DC, US: American Psychological Association.

Sofi, W. (2013). The influence of students’ mastery on collocation toward their writing (Unpublished thesis), State Institute of Islamic Studies, Salatiga, Indonesia.

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. doi: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. doi:10.1111/jedm.1983.20.issue-4

Vakili, S., & Ebadi, S. (2010). Investigating contextual effects on Iranian EFL learners’ mediation and reciprocity in academic writing. Cogent Education, 6(1), 1571289. doi:10.1080/2331186X.2019.1571289

Weigle, S. C. (2002). Assessing writing. UK: Cambridge University Press.

Xie, Q. (2017). Diagnosing university students’ academic writing in English: Is cognitive diagnostic modelling the way forward? Educational Psychology, 37(1), 26–47. doi:10.1080/01443410.2016.1202900

Zhang, Y., DiBello, L., Puhan, G., Henson, R., & Templin, J. (2006, April). Estimating skills classification reliability of student profiles scores for TOEFL Internet-based test. Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, CA.
Appendix A. Selected Prompts

| Prompt | Subject |
|--------|---------|
| 1      | Some people trust their first impressions about a person’s character because they believe these judgments are generally correct. Other people do not judge a person’s character quickly because they believe first impressions are often wrong. Compare these two attitudes. Which attitude do you agree with? Support your choice with specific examples. |
| 2      | Some people believe that university students should be required to attend classes. Others believe that going to classes should be optional for students. Which point of view do you agree with? Use specific reasons and details to explain your answer. |
| 3      | Some people prefer to spend time with one or two close friends. Others choose to spend time with a large number of friends. Compare the advantages of each choice. Which of these two ways of spending time do you prefer? Use specific reasons to support your answer. |

Appendix B. The First EDD Checklist Applied to Assess Students’ Essays

| Instructor’s name: | Prompt no: | Essay number |
|-------------------|------------|--------------|
|                   | Content    | Y  | N  | |
| 1                  | This essay answers the topic (prompt). |       |   |   |
| 2                  | This essay is concisely written. |       |   |   |
| 3                  | This essay contains a clear thesis statement. |       |   |   |
| 4                  | The main arguments of this essay are strong. |       |   |   |
| 5                  | There are enough supporting ideas and examples in this essay. |       |   |   |
| 6                  | The supporting ideas and examples in this essay are appropriate and logical. |       |   |   |
| 7                  | The supporting ideas and examples in this essay are specific and detailed. |       |   |   |
|                   | Organization |       |   |   |
| 8                  | The ideas are organized into paragraphs and include an introduction, a body, and a conclusion. |       |   | |
| 9                  | Each body paragraph has a clear topic sentence explained and elaborated by the supporting sentences. |       |   |   |
| 10                 | Each paragraph presents one distinct and unified idea. |       |   |   |
| 11                 | Each paragraph is connected to the rest of the essay smoothly. |       |   |   |
| 12                 | Transition/coherence devices are used within and between paragraphs to show relational patterns between ideas. |       |   |   |
|                   | Language    |       |   |   |
| 13                 | This essay demonstrates syntactic variety, including simple, compound, and complex sentence structures. |       |   |   |
| 14                 | Verb tenses are used appropriately. |       |   |   |
| 15                 | There is consistent subject-verb agreement. |       |   |   |
| 16                 | Singular and plural nouns are used appropriately. |       |   |   |
| 17                 | Prepositions are used appropriately. |       |   |   |
| 18                 | Articles are used appropriately. |       |   |   |
| 19                 | Pronouns agree with referents. |       |   |   |
| 20                 | Sophisticated or advanced vocabulary is used. |       |   |   |
| 21                 | Vocabulary choices are appropriate for conveying the intended meaning. |       |   |   |
| 22                 | This essay demonstrates facility with appropriate collocations. |       |   |   |
| 23                 | Words are spelled correctly. |       |   |   |
| 24                 | Capital letters are used appropriately. |       |   |   |
| 25                 | This essay contains redundant ideas or linguistic expressions. |       |   |   |

Source: The checklist was the modified version of Kim’s EDD checklist (2011, p.539)
## Appendix C. The modified version of Kim's (2011) EDD Writing Checklist for the EFL context

| # | Item                                                                                                                                   | Y | N |
|---|----------------------------------------------------------------------------------------------------------------------------------------|---|---|
| 1 | This essay answers the topic (prompt).                                                                                                  |   |   |
| 2 | This essay is concisely written.                                                                                                       |   |   |
| 3 | This essay contains a clear thesis statement.                                                                                           |   |   |
| 4 | The main arguments of this essay are strong.                                                                                           |   |   |
| 5 | There are enough supporting ideas and examples in this essay.                                                                            |   |   |
| 6 | The supporting ideas and examples in this essay are appropriate and logical.                                                            |   |   |
| 7 | The supporting ideas and examples in this essay are specific and detailed.                                                             |   |   |
| 8 | The ideas are organized into paragraphs and include an introduction, a body, and a conclusion.                                          |   |   |
| 9 | Each body paragraph has a clear topic sentence explained and elaborated by the supporting sentences.                                     |   |   |
| 10| Each paragraph presents one distinct and unified idea.                                                                                  |   |   |
| 11| Each paragraph is connected to the rest of the essay smoothly.                                                                           |   |   |
| 12| Transition/coherence devices are used within and between paragraphs to show relational patterns between ideas.                          |   |   |
| 13| Verb tenses are used appropriately.                                                                                                     |   |   |
| 14| There is consistent subject-verb agreement.                                                                                                |   |   |
| 15| Singular and plural nouns are used appropriately.                                                                                       |   |   |
| 16| Prepositions are used appropriately.                                                                                                    |   |   |
| 17| Articles are used appropriately.                                                                                                       |   |   |
| 18| Pronouns agree with referents.                                                                                                         |   |   |
| 19| This essay demonstrates facility with appropriate collocations.                                                                         |   |   |
| 20| Words are spelled correctly.                                                                                                           |   |   |
| 21| Capital letters are used appropriately.                                                                                                 |   |   |
