Educational Methods & Practice

https://emp-open.de
Volume 2 (2024)

Research Article

Examining the Dimensionality of Linguistic Features in L2 Writing Using the Rasch Measurement Model

Farshad Effatpanah*
Technische Universität Dortmund, Dortmund, Germany

Purya Baghaei
International Association for the Evaluation of Educational Achievement (IEA), Hamburg, Germany

Abstract

It has been acknowledged that second/foreign language (L2) writing is a complex and multi-dimensional cognitive process, and linguistic knowledge is the foremost predictor of L2 writing. Previous research on developing models and orientations for characterizing L2 writing and its linguistic features are based on methods rooted in classical test theory (CTT) which mostly tend to overlook qualitative differences among writers. The use of item response theory (IRT) and Rasch models has been disregarded in L2 writing research. This study aimed to psychometrically investigate the dimensionality of linguistic features in L2 writing using the Rasch model. To achieve this, 500 Iranian English as a foreign language (EFL) students wrote an essay marked by four experienced raters using an empirically-derived descriptor-based diagnostic checklist. The scores derived from the marking of the essays were subjected to Rasch model analysis. Individual item/descriptor fit, separation and reliability, unidimensionality, and local item dependency (LID) were examined. The results provided evidence for the multidimensionality of linguistic features in L2 writing. The analysis of the positive and negative item loadings on Factor 1, extracted from the Rasch model residuals, revealed the presence of two sets of descriptors that contribute to the definition of two groups of L2 writers. The first set comprises descriptors with positive loadings mostly related to higher-level linguistic features of L2 writing, including content fulfillment (CON) and organizational effectiveness (ORG). However, the second set includes descriptors with negative loadings chiefly related to lower-level linguistic features, such as vocabulary use (VOC), grammatical knowledge (GRM), and mechanics (MCH). Implications and suggestions for further research are discussed.

Keywords

L2 writing, linguistic features, dimensionality, rating scales, Rasch model

Corresponding author:
*Farshad Effatpanah, Research Unit of Psychological Diagnostics, Faculty of Rehabilitation Sciences, Technische Universität Dortmund, Emil-Figge Street 50, 44227 Dortmund, Germany
Email: farshad.effatpanah@tu-dortmund.de
Farshad Effatpanah: https://orcid.org/0000-0003-3970-5588
Purya Baghaei: https://orcid.org/0000-0002-5765-0413

Received 25 October 2023; Received in revised form 6 January 2024; Accepted 28 January 2024
Available online 8 February 2024

1 | Introduction

Written communication in both first language (L1) and second/foreign language (L2) is an intricate cognitive process that is influenced by several factors and multifold underlying cognitive processes which are integral to learning, thinking, planning, and communicating (Williams & Larkin, 2013). It requires the combination of various linguistic knowledge and (meta)cognitive processes (Huang & Zhang, 2022). To better understand the nature of writing, a variety of theoretical models (e.g., Abbott & Berninger, 1993; Alamargot & Chanquoy, 2001; Bereiter & Scardamalia, 1987; Chenoweth & Hayes, 2003; Fayol et al., 2012; Flower & Hayes, 1980, 1981, 1983; Galbraith, 2009; Hayes, 2012; Grabe & Kaplan, 1996; Kellogg, 1996; Schoonen et al., 2003, 2009) and orientations (Manchón, 2011) have been proposed to gain a deeper insight into the nature of writing and its factors, dimensions, processes, and issues as well as their interplay in the writing process. Researchers have come to the consensus that L2 writing is a multi-dimensional process in such a way that each dimension is related to other dimensions. Linguistic knowledge has also been shown to be the most important predictor in L2 writing which encompasses various linguistic features including content, vocabulary, grammar, mechanics, and discourse or organization. These linguistic features interact in complex ways to shape the overall quality of a written text (Lee et al., 2021). Research has indicated that different levels of linguistic features can serve as indicators of L2 writing proficiency and affect raters' judgments. For example, high quality L2 essays tend to exhibit greater complexity in both lexical items (e.g., Zhang, 2022) and grammatical structures (e.g., Kyle & Crossley, 2018). They also contain more relevant ideas and arguments (Olinghouse & Graham, 2009), more explicit cohesive devices for text cohesion and coherence (e.g., Pu et al., 2023), and more appropriate mechanics (e.g., Vögelin et al., 2018). Although previous studies have provided valuable insights into the nature of L2 writing, its linguistic features, and its dimensionality which have significant implications for research, assessment, and pedagogy, they are dependent on methods rooted in classical test theory (CTT) and assume that writers adopt similar (meta)cognitive processes and linguistic features while producing a text. However, individual writers may differ in their writing processes and the use of linguistic features. In effect, the use of linguistic features may not be the same across L2 writers, and there might be multiple configurations through which writers can produce a quality text (Effatpanah et al., 2024; Jarvis et al., 2003). This, in turn, is more likely to cause multidimensionality of linguistic features in L2 writing.

Along the same lines, Hickendorff et al. (2018) argue that researchers commonly use total raw scores to develop models in statistical analyses. The use of total scores presupposes uniformity in the pattern of individuals' performance, treating any variability within and among individuals mainly as statistical noise. In this way, the proposed model is a general model that describes the average behavior observed in a sample. However, when distinct subgroups with qualitative differences exist within a population, the general model may not accurately represent their characteristics (Hickendorff et al., 2018).

To the best of our knowledge, too little attention has been paid to psychometrically examining the dimensionality of linguistic features in L2 writing. A neglected approach for analyzing the dimensionality of linguistic features in L2 writing is the use of item response theory (IRT) and Rasch (Rasch, 1960/1980) models. To address this gap, this study aims to use Rasch model to analyze the dimensionality of linguistic features in L2 writing based on an empirically-derived descriptor-based checklist for scoring L2 writing. Examining the dimensionality of L2 writing regarding its linguistic features can privilege researchers to revise theories and models of L2 writing, design more robust empirical research, and facilitate pedagogy.

2 | Background

2.1 | L2 Writing Dimensionality

According to Cumming (2016, p. 65), "L2 writing is inherently multi-faceted, involving multiple issues and orientations that may not even be commensurable with each other". The dimensionality of L2 writing refers to a variety of dimensions and factors involved in the process of L2 writing (Hirvela et al., 2016). It encompasses a wide range of dimensions and factors including linguistic (e.g., content, grammar, vocabulary, organization or sentence structure, and mechanics),

cognitive (e.g., planning, organizing, and translating ideas), cultural (e.g., audience expectations, cultural norms, communicative purposes), textual (e.g., coherence and structure of the written text), affective (e.g., attitudinal and emotional aspects of writing), and strategic (e.g., paraphrasing, cognates, and language resources) dimensions. Successful L2 writing requires the integration of these factors and dimensions to produce a quality text. These myriad factors overshadow the assessment of L2 writing, that is, the dimensions are more likely to vary depending on the specific approach, orientation, or framework used for assessing L2 writing. Typical dimensions considered for assessing L2 writing are content, language use, organization, conventions, style, genre knowledge, writing processes (e.g., planning, drafting, revising, and editing), audience and topic awareness, cultural awareness and expectations, and task response.

Over the past few decades, L2 writing scholars have presented numerous orientations and approaches toward L2 writing. Manchón (2011) states that there are three complementary perspectives in the study of the nature of L2 writing in the literature: learning-to-write (LW), writing-to-learn-content (WLC), and writing-to-learn-language (WLL). Each of these orientations highlights distinct aspects of writing and is intricately connected to the purposes of learning and teaching writing as well as the diverse contexts in which L2 writing is acquired and instructed. Under the LW, researchers concentrate on three kinds of approaches to L2 writing. The first sort of approaches is concerned with the writer and the cognitive processes used to create texts. In this writing-oriented approach, numerous models have been developed to explain cognitive processes involved in writing (e.g., Abbott & Berninger, 1993; Alamargot & Chanquoy, 2001; Bereiter & Scardamalia, 1987; Chenoweth & Hayes, 2003; Fayol et al., 2012; Flower & Hayes, 1980, 1981, 1983; Galbraith, 2009; Hayes, 2012; Grabe & Kaplan, 1996; Kellogg, 1996; Schoonen et al., 2003, 2009). The second sort of approaches focus on the products of writing by investigating texts as autonomous objects and discourse (e.g., Feez, 2001; Hyland, 2004; Johns, 1997). The third sort of approaches concerns what writers do incorporate a sense of audience or address the readers based on their expectations (e.g., Hyland, 2009).

The second perspective, i.e., WLC, focuses on the examination of how the act of writing can serve as a tool for learning other disciplinary subject-matter in the content areas. The core idea of the WLC revolves around the concept of transfer which "occurs when people make use of prior experiences to address new challenges" (Cleary, 2013, p. 62). The WLC states that students utilize writing not only to showcase their acquired knowledge in a written form but also to enhance their learning by leveraging the resources that writing offers (James, 2009).

As a recent orientation to L2 writing studies, Cumming (1990) argues that the act of writing in an L2 has the potential to improve the acquisition of L2 linguistic knowledge. This improvement arises both from the act of writing and the reflection upon written corrective feedback provided for one's own writing. Writing could encourage learners to "analyze and consolidate second language knowledge that they have previously (but not fully) acquired" (Cumming, 1990, p. 483). When L2 writers engage in L2 writing, they are likely to need to "monitor their language production in a way that is not necessary or feasible under time constraints of comprehending or conversing in a second language" (p. 483). Consequently, it is essential to examine the way L2 writers learn how to write and the importance of understanding the instrumental role that writing plays in the development of an L2 in educational settings (Harklau, 2002).

Among the orientations and models proposed for describing writing, Flower and Hayes's (1981) model is the most influential model of writing. Flower and Hayes argue that writing comprises three hierarchical and dynamic processes: (1) planning where both L1 and L2 writers engage in idea generation, organization, and goal-setting to form an internal representation of the information or knowledge they aim to convey through their writing; (2) translating where writers translate or convert their ideas into written texts by drawing on pertinent knowledge and linguistic resources stored in their memory; and (3) reviewing where writers assess and revise their texts in a systematic way to correct errors.

Translation serves as the primary mechanism among the various writing processes, facilitating the conversion of ideas or propositional content into suitable linguistic expressions (van Gelderen et al., 2011). This intricate process demands a thorough understanding of linguistic components, such as content, sentence structure, grammar, word selection, textual coherence, organization, and mechanics (e.g., punctuation and spelling). Writing subskills such as mechanics, grammar,

and vocabulary are typically regarded as lower-level, whereas higher-level subskills encompass organization and content (Schoonen et al., 2011). To produce a well-constructed text, writers should coordinate these higher- and lower-level writing subskills which impose constraints on working memory capacity and influence the overall quality of written texts (Güvendir & Uzun, 2023). The impediment or any difficulty in the mastery of these subskills can override the improvement of L2 writing. It is thus crucial for teachers and educational experts to accurately assess and identify specific writing weaknesses or flawed strategies among students. Once problematic areas are pinpointed, students can receive appropriate and timely feedback and subsequently pursue some strategies to remedy and develop their writing skills during their learning process.

2.2 | Diagnostic Assessment

Diagnostic assessment of L2 writing ability has received a great deal of attention among L2 researchers (Alderson, 2005) due to its potential to yield rich and informative diagnostic information about writers' specific weaknesses and strengths in writing (Knoch, 2011). Diagnostic assessment identifies what students have already mastered in their learning process and where they need more assistance to eliminate their deficiencies. The information obtained from diagnostic assessment can be used to support inferences about students' knowledge and help teachers tailor remedial instruction based on individual students' needs (Jang, 2009). More specifically, diagnostic assessment can be viewed as a form of needs assessment aimed at pinpointing students' gaps, understanding the root causes of issues, determining priorities, and exploring potential solutions (Nation & Macalister, 2010). Diagnostic information can also be useful for providing diagnostic feedback to all stakeholders about students' learning status and ultimately enhancing the quality of learning by maximizing the opportunities to learn (Jang, 2009). Many researchers have noted that providing diagnostic information can raise students' awareness of their own learning process, encouraging self-regulated learning, and motivating them (Kuo et al., 2016; Zimmerman, 2000). In this regard, diagnostic assessment aligns with formative assessment or assessment for learning (AFL). In fact, it effectively combines assessment with curriculum and instruction (Pellegrino & Chudowsky, 2003).

Alderson (2005) contends that there is a confusion between diagnostic tests and placement or proficiency tests. However, according to Bachman (1990, p. 60), when we speak of a diagnostic test,

> we are generally referring to a test that has been designed and developed specifically to provide detailed information about the specific content domains that are covered in a given program or that are part of a general theory of language proficiency. Thus, diagnostic tests may be either theory or syllabus-based.

In the context of writing assessment, Alderson (2005) advocates the use of indirect tests for assessing writing ability of students rather than performance tests. However, the current trend in writing assessment favors performance tests over indirect ones. This shift is attributed to the perception that indirect tests cannot adequately and validly measure the multi-faceted aspects of writing (Weigle, 2002).

An important aspect in the performance assessment of writing is rating scales. Within the context of performance rater-mediated assessments, raters are usually provided with some kinds of multiple-category or ordinal rating scales to evaluate students' performance in a writing task. They utilize these scales to convey their interpretation of the quality of students' performance. However, LIosa et al. (2011) have argued that

> most existing tools for assessing writing [including analytic and holistic scales] are not sufficiently informative to identify the exact nature of students' problems with composing nor to provide instructionally useful feedback. Most existing assessments of writing are designed to evaluate writing achievement by eliciting an extended response to a writing prompt and evaluating the response with a holistic score. While useful for determining whether students have met performance goals in composition, such assessments do not provide sufficiently specific information to be useful for instructional purposes (p. 258).

What is needed are rating scales that fulfill diagnostic purposes. In this regard, a large number of diagnostic rating scales, which will be reviewed in the following sections, have been developed to identify strengths and weaknesses in students' writing ability. Scoring of students' written performance using rating scales holds significant importance in the

conceptualization of validity (Kane, 2013) and is essential in the practical aspects of language testing and assessment (Knoch & Chapelle, 2018). This is due to the fact that scoring establishes a critical connection between a performance and a proficiency claim (Knoch et al., 2021). It is thus necessary to provide evidence for the efficiency of rating scales used to assess writing ability of students.

2.3 | Rating Scales for Scoring Writing

A number of researchers have suggested numerous classifications of rating scales. The most widely cited categorization is the distinction between holistic and analytic scales (Hamp-Lyons, 1991; Lee et al., 2010; Weigle, 2002). In holistic scoring (sometimes referred to as 'impressionistic' scoring), a single score to a piece of writing is assigned based on raters' overall impression of it. Holistic scores have been shown to be rapid and appropriate for writings characterized by internal coherence or congruency, but not for texts exhibiting internal complexity, such as those that are highly developed but frequently interrupted by grammatical errors (Weigle, 2002). Raters further mostly tend to provide a global score based on the strengths of a text not its weaknesses (Jarvis et al., 2003). On the other hand, analytic scoring (sometimes referred to as 'multiple trait assessment') involves the assignment of a separate score for each of a number of aspects or criteria of a task. This kind of scoring has the advantage of being highly reliable, providing higher construct validity for L2 writers, and offering more diagnostic information about the performance of students, because they measure writing across multiple criteria. However, empirical research has indicated that raters fail to conceptually discriminate between different criteria of a scale (Hughes & Hughes, 2020). That is to say, raters tend to judge each of the criteria independently of the others, known as 'halo effect' (Cooper, 1981; Engelhard, 2013). This tendency of raters has been shown to produce 'flat' profiles (e.g., where students are often assigned ratings that are uniformly negative or uniformly positive across various criteria), compromise the diagnostic capacity of the analytic approach for assessing writing, and finally provide misleading information about the performance of writers (Hamp-Lyons, 1991; Knoch, 2009).

Another well-known categorization of rating scales indicates their construction methodology, or the way they are developed. In his dyadic typology of rating scales, Fulcher (2003) makes a distinction between intuitive methods and empirical methods. Scales constructed intuitively rely on existing scales or the intuition of scale developers for specifying common features at different proficiency levels. Fulcher et al. (2011) contends that such scales are more likely to depend on theory indirectly, as the developers' beliefs are influenced by theory, but they do not originate from an examination of real performance data. However, scales developed based on empirical methods are developed based on real world and corpora and take into consideration real performance. Fulcher et al. (2011) extend his typology and distinguishes between '*measurement-driven'* scales and *'performance-driven'* scales, which are relatively equivalent to empirical and intuitive methods, respectively. Fulcher (2012) argues that well-designed empirical or performance-driven scales will better reflect the way language is naturally used in real-world situations. As Fulcher et al. (2011) argue, what is important in scale construction is that rating scales should be empirically developed and establish a meaningful connection with real world language use. In response to Fulcher et al.'s call for developing rating scales empirically, a number of researchers have developed scales and checklists for assessing writing ability of students. For a comprehensive review of approaches to scale development, refer to Knoch, 2009 and Knoch et al., 2021.

2.4 | Diagnostic Checklists for Writing Assessment

Over the past few decades, a number of researchers have focused on diagnostic assessment of writing. Two lines of research emerge in the relevant literature. The first line comprises studies developing diagnostic rating scales for assessing writing ability of students (e.g., Fulcher, 1996; He et al., 2021; Kim, 2010; Knoch, 2007; Lukácsi, 2021; Ma et al., 2022; North & Schneider, 1998; Safari & Ahmadi, 2023; Struthers et al., 2013; Turner & Upshur, 2002; Upshur & Turner, 1999). For example, Knoch (2007) developed a theoretically-based and empirically-developed rating scale for a diagnostic writing con-

text. The scale assesses accuracy, fluency, complexity, mechanics, reader-writer interaction, content, coherence, and cohesion. Kim (2010) also developed an empirically derived descriptor-based diagnostic (EDD) checklist using think-aloud protocols from raters. The original purpose of constructing the scale was to assess and characterize the writing of non-native English-speaking students in an academic context, specifically for the independent essay section of the Test of English as a Foreign Language™ Internet-based Test (TOEFL iBT). This checklist contains 35 descriptors assessing five sub-skills of academic writing in English, including content fulfillment, organizational effectiveness, vocabulary use, grammatical knowledge, and mechanics. Following the procedure suggested by Crocker and Algina (1986), in another study, Struthers et al. (2013) developed a 13-item checklist to only assess cohesion in the writing of children in Grades 4–7 in order to inform instructional practices. Furthermore, Lukácsi (2021) designed a level-specific checklist for assessing the English as a Foreign Language (EFL) writing proficiency at the B2 level. Based on a mixed-methods strategy of inquiry (e.g., think-aloud protocols, qualitative data from stimulated recall, and semi-structured interviews), 35 binary items were developed based on CEFR (Common European Framework of Reference for Languages) to allow researchers and teachers to use the rating scale for level testing regardless of the test purpose. He et al. (2021) also developed a diagnostic checklist using the descriptors of China's Standards of English Language Ability (CSE). The scale consists of 15 descriptors and measures four attributes including grammatical knowledge, textual knowledge, functional knowledge, and sociolinguistic knowledge. Ma et al. (2022) further developed an English writing EDD checklist by using raters' think-aloud protocol and expert judgment. The scale includes 30 descriptors measuring five attributes such as content, organization, vocabulary, syntax, and mechanics. Most recently, based on L2 students' verbalizations of their challenges in integrated writing tasks, Safari and Ahmadi (2023) developed and validated a binary 30-descriptor empirically-based diagnostic checklist for L2 reading-listening-writing integrated tasks. The traits are source use concerns, fulfilling task demands, organization and structure, grammatical knowledge, vocabulary knowledge, and mechanics.

The second line of research comprises studies modifying and applying the above-mentioned rating scales in different educational contexts for different purposes. An advantage of the development of binary diagnostic checklists is that they increase the number of writing items which, in turn, allow researchers and practitioners to apply complex statistical methods and measurement models to writing. For that reason, a large number of researchers have applied cognitive diagnostic models (CDMs; Ravand & Baghaei, 2019) to diagnose writing ability of students (e.g., Effatpanah et al., 2019; Shahsavar, 2019; Shi et al., 2024; Xie, 2017); Shi et al. (2024) used the checklist developed by Ma et al. (2022), whereas other studies either utilized the Kim's (2010) EDD checklist or modified it. Numerous studies have also used the EDD checklist to apply advanced quantitative methods to L2 writing such as the linear logistic test model (LLTM; Fischer, 1973), as an IRT-based cognitive processing model, for examining the underlying cognitive operations of L2 writing performance (Effatpanah & Baghaei, 2021) and the mixed Rasch model (Rost, 1990) for exploring multiple profiles of L2 writers (Effatpanah et al., 2024). All of the reviewed studies provide empirical evidence for the benefits of using descriptor-based diagnostic checklists for assessing L2 writing performance of students and offering fine-grained diagnostic feedback.

3 | The Present Study

Using the Rasch model (Rasch, 1960/1980), this study seeks to examine the dimensionality of L2 writing based on an empirically-derived descriptor-based checklist. Rasch model is a probabilistic model that is used to predict the outcome of encounters between persons and a set of items. The Rasch model assumes that the probability of giving a correct response to an item or successfully completing a task is a function of the ability of a person and the difficulty of a given item/task. The model yields a logistic function based on the differences between person ability and item difficulty. The higher is the probability of success when a person's ability is greater than an item's difficulty. Under the Rasch model, the probability that person v gets an item j right, given his/her ability θ_v and the item difficulty β_j is expressed as:

$$P(X_{vj} = 1 \mid \theta_v, \beta_j) = \frac{exp(\theta_v - \beta_j)}{1 + exp(\theta_v - \beta_j)} \qquad (1)$$

where $P(X_{vj})$ is the probability of success. A distinguishing characteristic of the Rasch model is its ability to transform item and person raw measures into interval-scaled measures. This transformation involves placing items and examinees on a calibrated scale, where their positions correspond to their difficulty and ability measures. In the present study, the following research questions were addressed:

RQ1: Are the linguistic features of L2 writing as measured by the empirically-derived descriptor-based diagnostic (EDD) checklist unidimensional?

RQ2: What are the underlying psychometric linguistic dimensions of L2 writing as measured by the empirically-derived descriptor-based diagnostic (EDD) checklist?

4 | Method

4.1 | Setting and Participants

The study conducted secondary analysis of data obtained from a research project carried out by Effatpanah et al. (2019) to diagnose Iranian EFL students' writing performance using cognitive diagnostic models (CDMs). The data were collected from 500 Iranian EFL students aged between 19 to 58 (M = 24.89 years, SD = 6.30) from different classes and universities in Iran. The students were taught by 32 non-native full-time English teachers following various materials and syllabi for instruction. Their L2 writing teaching experience to adults varied from 9 to 24 years. There were 349 female (69.8%) and 151 male (30.2%) participants. This gender imbalance is due to the typical distribution of students in English departments in many Iranian universities, which are heavily dominated by females. The sample included 212 junior, 152 senior, and 136 postgraduate students. Of the total sample, 104 (20.8%) studied English Translation, 128 (25.6%) English Literature, and 268 (53.6%) Teaching English as a Foreign Language (TEFL). The students reported their English learning background from three to more than ten years and their amount of English use from cannot say to almost every day. All participants were native speakers of Persian and were studying English as an academic major. They were informed that the retrieved data would remain anonymous and confidential, and informed consent was obtained from all students.

Four experienced raters (1 female and 3 male) whose ages ranged from 28 to 39 years old (M = 31.75; SD = 4.99) were also recruited to assess the essays. They had an average 14.3 years of experience in teaching and assessing L2 writing. One rater was a Ph.D. candidate of TEFL and a lecturer in university, while the remaining three raters held master's degrees in TEFL and had achieved an overall score of 8 in the IELTS exam. All of the raters were non-native English speakers (with Persian as their first language and English as their foreign language), but all had native or near-native English language proficiency.

4.2 | Instrument

4.2.1 | Writing Prompt

The students were required to write at least a 350-word essay in response to the following writing 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 writing prompt was administered as a class activity in L2 writing courses. The students were urged to depend on their own abilities and refrain from utilizing any books, dictionaries, or internet resources while writing to provide the researchers and their teachers with the opportunity to assess their strengths and weaknesses in L2 writing.

4.2.2 | The Empirically-derived Descriptor-based Diagnostic (EDD) Checklist

To operationalize the measurement of writing, a diagnostic assessment scale called the empirically derived descriptor-based diagnostic (EDD) checklist (Kim, 2010, 2011) was used. The checklist comprises 35 binary (Yes, No) descriptors measuring five writing sub-skills. The sub-skills are: (1) content fulfillment (CON) assessing a student's ability to address a given

question by ensuring unity and relevance in supporting sentences, information, and examples; (2) vocabulary use (VOC) assessing a student's ability to accurately and appropriately use a broad array of lexical items; (3) grammatical knowledge (GRM) assessing a student's ability to accurately show syntactic variety and complexity; (4) organizational effectiveness (ORG) assessing a student's ability to cohesively and coherently develop and organize ideas and supporting sentences within and between paragraphs; and (5) mechanics (MCH) assessing a student's ability to adhere to English writing conventions, including indentation and margins, capitalization, spelling, and punctuation.

4.3 | Rater Training and Formal Scoring

Within the context of rater-mediated performance assessment, rater variations are a major threat to construct validity in the rating procedure (Wind & Peterson, 2018). A 2-hour training session prior to scoring the essays was held to mitigate potential inconsistencies (construct-irrelevant variance) or variations across the raters in this study. This training session involved instructing raters on the use of the scale and moderating discussions to separately elucidate the content of each descriptor. They were also trained how to interpret the *yes* or *no* response option; when writers generally met the standards, a *yes* response was recommended; otherwise, the *no* option was chosen. Following the guidelines of Weigle (2002), a small subset of essays was provided to raters to familiarize themselves with the scale and its specific properties and compare their performance with each other. After training, the essays were randomized and grouped into four packages. Each rater received 125 essays and copies of the checklist. Thirty-five essays were included in each package to be scored by all the raters. The total score on the checklist ranged from 0 to 35, with a mean of 17.62 and a standard deviation of 7.59. As a preliminary check, the Cronbach's alpha reliability of the checklist was investigated, and a value of 0.89 was obtained, which is highly satisfactory. The degree of inter-rater reliability among the raters was also examined using Pearson correlation coefficients analysis. The correlation coefficient across the raters was 0.82. According to Dancey and Reidy's (2004) criteria, coefficients smaller than 0.30 are considered weak, between 0.30 and 0.60 moderate, and above 0.60 strong. Cohen's Kappa was also computed, and a value of 0.62 was obtained, indicating a high agreement among the raters.

To ensure that students' scores have not been affected by construct irrelevant factors such as rater severity that can confound measurement, a three-faceted many-facet Rasch measurement (MFRM; Linacre, 1989) was performed using the software FACETS, Version 3.71 (Linacre, 2014a). For intra-rater consistency, infit and outfit mean squares (MNSQs), as two rater variability indices, were computed. Linacre (2014b) recommended the range of 0.5–1.50 for both fit indices. With the range of outfit between 0.90 and 1.10, infit between 0.91 and 1.06, all the raters achieved sufficient fit to the model, indicating satisfactory self-consistency in scoring the essays using the descriptors. Rater severity measures further varied from -0.09 to 0.16, suggesting negligible differences in rating severity. Two measures of the variance of rater severity, i.e., rater separation reliability ($R = 0.89$) and rater separation ratio ($G = 2.82$), also indicated that raters' severity does not vary considerably. Finally, checking the vertical summary (e.g., an interval scale which compares the location of items/descriptors, persons, and raters on the same scale calibrated in logits) showed that rater severity has not affected the students' marks.

4.4 | Data Analysis

The ratings awarded by the raters were analyzed by the WINSTEPS computer package version 3.73 (Linacre, 2009a) to examine the fit of the data to the Rasch model. For the purpose of this study, individual item/descriptor fit, separation and reliability, unidimensionality, and local item dependency (LID) were examined. Item difficulty parameters are estimated based on the proportion of examinees who get an item right or successfully complete a task regardless of those examinees' ability levels. They indicate the locations of items/descriptors on the latent trait continuum and are expressed in log-odd units or logits. Unlike the CTT, an error of measurement index is provided for each item. The error of measurement indicates the accuracy of estimated item difficulty parameters.

Furthermore, item and person reliability and separation indices were examined. The reliability of person and item

indices serves as an indicator of the scale's precision in measuring examinee ability and item difficulty (Linacre, 2009b). Separation reliability refers to the ratio of item or person true standard deviation to error standard deviation (e.g., root mean square error (RMSE)). It signifies the degree to which person and item parameters are distinct on the latent trait. In fact, item separation is used to verify the hierarchy of items of a scale. In cases where separation values are low (< 3 for high, medium, low item difficulties, item reliability < 0.9), it suggests that the sample size is not sufficient to validate the hierarchy of item difficulties within the scale (Linacre, 2009b). However, person separation is used to classify examinees. Low person separation values (< 2, person reliability < 0.8) imply that the scale may lack sensitivity to differentiate between examinees at varying proficiency levels (Linacre, 2009b). The range of separation reliability values extends from zero to infinity. A higher separation value for persons/items indicates a greater probability that persons/items with higher ability/difficulty estimates possess higher estimates compared to those with lower estimates (Linacre, 2009b).

To investigate the fit of the data to the model, outfit and infit mean square (MNSQ) fit statistics were calculated (Linacre, 2002). Both statistics indicate to what extent items of a scale align with the underlying latent variable being measured and involve the division of chi-square values for items by their degrees of freedom (Linacre, 2009b). As Linacre (2002) argued, outfit MNSQ serves as an outlier-sensitive fit statistic which detects erratic response patterns from examinees on items that are either relatively very easy or very difficult for them, and vice versa. On the other hand, infit MNSQ serves as an inlier-sensitive fit statistic, identifying unexpected response patterns to items that are specifically targeted to examinees, and vice versa. Linacre (2009b) considered a value between 0.50–1.50 as an acceptable range for fit indices. Moreover, point-measure correlations were estimated for all items/descriptors to measure the extent to which observed scores align with the expected latent trait. Point-biserial (or point-measure) correlations indicate to what extent the responses to each item within a scale are correlated with the overall measure.

An important feature of the Rasch model is its ability to generate a ruler-like device, functioning as an interval scale. This facilitates the comparison of item/descriptor and person locations. Specifically, it creates an item-person map, often referred to as the Wright map, which represents item difficulty and person ability estimates on the same metric calibrated in logits. The Wright map visually illustrates how items/descriptors are distributed in relation to the abilities of examinees.

Moreover, the unidimensionality of the scale was checked. Unidimensionality is an important assumption in Rasch and IRT models which posits that all items of a scale should measure a single unidimensional latent trait, that is, only one latent trait should explain variability in the observed responses. This requirement is fundamental for measurement and, more importantly, item construction because a scale claims to measure different levels of a latent trait should be affected by different levels of another latent trait (Ziegler & Hagemann, 2015). The Rasch model is a parametric model which imposes certain assumptions for unidimensionality. When the assumptions of the model hold (e.g., the data fit the Rasch model), it is an indication that all the items of a scale measure a single unidimensional latent trait, and that persons and items can be located on a common interval continuum. It must be noted that "unidimensionality does not imply that performance on items is due to a single psychological process. In fact, a variety of psychological processes are involved in responding to a set of items. However, as long as they are involved in unison-that is, performance on each item is affected by the same process and in the same form-unidimensionality will hold" (Bejar, 1983, p. 31). Items can still be regarded as unidimensional if they measure the same processes to the same degree (Fischer, 1997). Because infit and outfit MNSQ fit statistics exhibit minimal susceptibility to systematic threats against unidimensionality (Smith & Plackner, 2009, p. 424), the unidimensionality of the scale was checked using the principal component analysis of linearized Rasch residuals (PCAR). Because items usually do not adhere to the expectations of the Rasch model, there remain some residuals after data-model fit. In fact, residuals denote the disparities between Rasch model predictions and actual observations (e.g., observed data) and are regarded as differences between predictions of the Rasch model and the actual observations. As unexpected part of the data, residuals are expected to be uncorrelated and randomly distributed (Linacre, 2009b). Smaller residual values indicate a closer fit to the model. It must be noted that the expected latent trait is excluded from the analysis when PCAR is conducted based on standardized residuals. When the data have adequate fit to the Rasch model, the latent trait is expected

to account for all information in the data, with the residuals reflecting random noise. Any component taken from the residuals is thus identified as a second dimension, indicating the breach of the unidimensionality assumption (Linacre, 2009b). The strength of the emergent component (e.g., the capacity of the component for explaining the common variance in data) is compared with the strength of the target dimension. According to Linacre (2009b), eigenvalues smaller than 2 verify the unidimensionality of the scale.

In the PCAR, loadings indicate the correlation between the items and an off-target secondary component extracted from the residuals that is unrelated to the primary target dimension (Baghaei & Cassady, 2014). Items/Descriptors with positive and negative loadings form two distinct sets that are orthogonal to the target dimension. Items with a correlation close to zero do not contribute to this secondary component. A high loading on the secondary component indicates that the item is associated with the off-target dimension and is more likely to have a weaker correlation with the target Rasch dimension or the latent trait (Baghaei & Cassady, 2014). A scrutiny of the content of contrasting clusters of items with high negative and positive loadings (exceeding ±0.40) helps identify meaningful interpretations of the secondary components as additional dimensions (Linacre, 2009b).

To assess the unidimensionality of a scale, Smith (2002) suggested the estimation of person parameters based on two subsets of a scale. Unidimensionality implies that person parameter estimates should remain consistent regardless of the subset of items encountered by examinees. If the ability estimate of an examinee varies between different subsets of the scale, it indicates that the data may reflect more than one dimension, threatening the construct validity of the scale (Baghaei & Cassady, 2014). The equality of estimates is investigated using *t*-tests. Statistically significant results suggest a lack of equality across the subsets with regard to person parameters and the presence of additional dimensions in the scale.

The results of item loadings on the first factor extracted from the residuals are also illustrated as a map by WINSTEPS. The plot visually illustrates the distribution of item loadings on the off-target dimension, with higher loadings at the two extremes. Items/Descriptors at the upper end exhibit positive loadings, while those at the lower end display negative loadings. If there is no discernible pattern in the residuals of items, they are expected to disperse across various regions of the map without clustering in either the positive or negative loading regions (Linacre, 2009b). Conversely, notable dimensions lead to clusters of items referred to as "contrasts," which emerge in opposing regions of the plot based on their loading values. These contrasts reflect structural differences among examinees with regard to their performance in a test.

Finally, LID was also evaluated using Pearson correlation analysis of linearized Rasch residuals. As stated by Wright (1994), residuals indicate the extent to which the items of a scale are locally easier or more difficult than the expectation of the model. Substantial correlations between the residuals of two items suggest local dependency, signifying potential shared dimensions or replicated features (Linacre, 2009b). After removing the Rasch dimension, locally dependent item pairs show notably high positive or negative residual correlations. Consistent with Christensen et al.'s (2017) recommendation, correlations exceeding 0.20 indicate the presence of local dependency.

5 | Results

5.1 | Item/Descriptor Characteristics and Fit Statistics

Table 1 shows descriptive statistics of the data, including mean, standard deviation (SD), skewness, and kurtosis, computed on SPSS for Windows, Version 25, item difficulty measures in logits, standard errors of measurement, infit and outfit MNSQ statistics, and point-measure correlations. As can be seen, Item/Descriptor 16 had the highest mean score (M = 0.82, SD = 0.385), and Item/Descriptor 35 (M = 0.12, SD = 0.323) had the lowest. The skewness and kurtosis values of some descriptors were high, possibly due to their difficulty measures causing an imbalance in the distribution tails (Bachman, 2004). Because the data is dichotomous, a descriptor being either easy or difficult results in a high prevalence of 1.00 or 0.00 (correct or incorrect) cases, respectively. This creates an asymmetry in the data shape, with a skew towards the side with lower frequency, leading to a long-tailed distribution and a high skewness value.

Table 1

Descriptive Statistics, Item Measures, Fit Statistics, and Point-Measure Correlations

Items	Mean	SD	Skewness	Kurtosis	Measures	Model S.E.	Infit MNSQ	Outfit MNSQ	PT-Measures
1	0.34	0.476	0.659	-1.572	0.83	0.11	0.97	0.91	0.49
2	0.67	0.472	-0.706	-1.508	-0.91	0.11	0.99	0.87	0.46
3	0.60	0.490	-0.427	-1.825	-0.56	0.10	0.94	0.90	0.50
4	0.52	0.500	-0.088	-2.000	-0.13	0.10	0.88	0.97	0.54
5	0.45	0.498	0.185	-1.974	0.23	0.10	0.88	0.97	0.54
6	0.40	0.491	0.401	-1.847	0.50	0.10	0.91	0.91	0.52
7	0.44	0.497	0.234	-1.953	0.29	0.10	0.87	0.80	0.56
8	0.47	0.499	0.129	-1.991	0.15	0.10	0.94	0.90	0.51
9	0.43	0.496	0.275	-1.932	0.34	0.10	0.93	0.89	0.52
10	0.53	0.499	-0.129	-1.991	-0.18	0.10	0.84	0.76	0.58
11	0.54	0.499	-0.153	-1.985	-0.21	0.10	0.93	0.85	0.52
12	0.57	0.496	-0.284	-1.927	-0.38	0.10	0.95	0.87	0.50
13	0.47	0.499	0.129	-1.991	0.15	0.10	0.92	0.86	0.52
14	0.33	0.471	0.716	-1.494	0.90	0.11	0.94	0.89	0.50
15	0.44	0.497	0.242	-1.949	0.30	0.10	0.99	0.94	0.48
16	0.82	0.385	-1.671	0.795	-1.94	0.13	1.07	1.06	0.33
17	0.71	0.452	-0.950	-1.102	-1.19	0.11	0.92	0.92	0.47
18	0.64	0.479	-0.603	-1.643	-0.78	0.11	1.02	0.98	0.44
19	0.73	0.446	-1.016	-0.971	-1.27	0.11	1.27	1.23	0.26
20	0.55	0.498	-0.185	-1.974	-0.25	0.10	0.90	0.82	0.54
21	0.69	0.463	-0.824	-1.326	-1.05	0.11	1.06	1.06	0.39
22	0.61	0.488	-0.461	-1.795	-0.61	0.10	0.99	1.50	0.44
23	0.67	0.470	-0.735	-1.466	-0.94	0.11	1.08	1.44	0.37
24	0.57	0.495	-0.300	-1.918	-0.40	0.10	0.99	1.06	0.46
25	0.63	0.483	-0.549	-1.705	-0.71	0.11	1.29	1.34	0.26
26	0.17	0.372	1.801	1.248	2.07	0.13	1.11	1.31	0.33
27	0.21	0.406	1.443	0.083	1.71	0.12	0.99	0.95	0.44
28	0.34	0.473	0.697	-1.521	0.87	0.11	0.95	0.87	0.50
29	0.52	0.500	-0.096	-1.999	-0.14	0.10	1.05	1.05	0.43
30	0.62	0.487	-0.478	-1.778	-0.63	0.10	1.17	1.26	0.34
31	0.63	0.484	-0.522	-1.734	-0.68	0.10	1.20	1.29	0.31
32	0.49	0.500	0.056	-2.005	0.06	0.10	1.03	1.11	0.44
33	0.55	0.498	-0.202	-1.967	-0.27	0.10	1.01	0.99	0.46
34	0.15	0.353	2.011	2.053	2.26	0.14	1.11	1.26	0.33
35	0.12	0.323	2.375	3.657	2.56	0.15	0.98	0.93	0.40

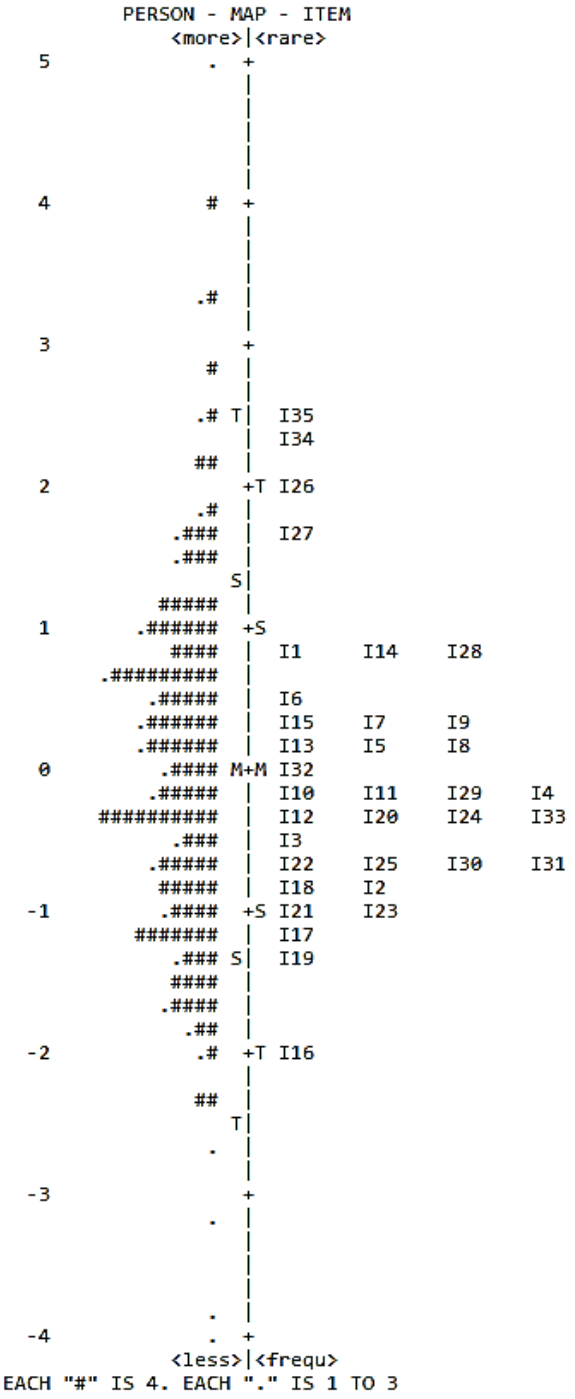
Note. M = Mean; SD = Standard Deviation; S.E. = Standard Error of Measurement; MNSQ = Mean Square; PT-Measures = Point-Measure Correlations.

The results of item difficulty parameters also showed that they ranged from -1.94 to 2.56 logits with item reliability coefficients and separation values of 0.99 and 8.78, respectively. Descriptors 19 and 16 were the easiest descriptors, and Descriptors 35 and 34 were the most difficult. Person estimates ranged from -3.89 to 4.08, with item reliability coefficients and separation values of 0.87 and 2.62, respectively, suggesting the wide range of examinees' abilities. With regard to infit and outfit MNSQ fit statistics, except for Descriptor 22, all the fit values were within the acceptable boundary of 0.50–1.50. Descriptor 22 is a poor descriptor for inclusion in the EDD checklist and can be excluded in future studies. The point-measure correlations also indicated that all correlations are positive and mostly medium. These values indicate a substantial agreement between the patterns of item difficulties in the data and the Rasch model (Linacre, 2009b).

A Wright map of the concurrent distribution of persons and the difficulty of the descriptors of the EDD checklist is shown in Figure 1. The first column shows the interval scale in logits. The dotted dividing line represents the hypothetical latent trait continuum. On the left side of the line, Hash marks (#) indicate the distribution of examinees. On the right side of the line, numbers show the difficulty of the descriptors. On both side of the dotted line, M represents the mean, S and T are one and two standard deviation(s) from the mean. Descriptors and examinees ranged from the easiest and least able at the bottom to the most difficult and most able at the top of the scale, respectively. Descriptors/Items are expected to be located across the whole scale to effectively measure the 'ability' of all examinees (Bond et al., 2020). As can be seen in Figure 1, descriptors cover a wide range of difficulty, although few difficult descriptors should be added to capture ability of most able examinees at the top of the scale. This provides evidence for the representativeness of the descriptors.

Figure 1

Wright Map of the Distribution of Persons and Descriptors of the EDD Checklist on the latent variable



5.2 | Unidimensionality and Local Independence

PCAR was carried out to check the unidimensionality of the checklist. The results of PCAR revealed that the model explains 30.9% of the observed variance; 14% are explained by item measures, and 16.9% are explained by person measures. The observed model variance was higher than the model expectation of 30.7%, although 69.1% of the variance are still unexplained. The first factor (contrast) accounted for 4.9% of the unexplained variance, with the eigenvalue equals to 2.5, which was higher than the critical value of 2, indicating the multidimensionality of the checklist.

Table 2 gives the loadings of the 35 descriptors on the first factor identified in the PCAR, excluding the first component. The first factor showed that there are two sets of descriptors that contribute to the definition of the EDD checklist. Descriptors with positive loadings on Factor 1 are mostly associated with higher-level linguistic features of L2 writing, such as content fulfillment (CON) and organizational effectiveness (ORG). However, descriptors with negative loadings are chiefly associated with lower-level linguistic features, including vocabulary use (VOC), grammatical knowledge (GRM), and Mechanics (MCH).

Table 2

Item Loadings for the Descriptors of the EDD Checklist on the First Factor in PCA of Residuals

Items	Loadings	Items	Ladings
2	-0.20	**1**	0.34
3	-0.03	**4**	0.27
16	-0.27	**5**	0.31
17	-0.17	**6**	0.40
18	-0.16	**7**	0.39
19	-0.33	**8**	0.30
20	-0.07	**9**	0.30
21	-0.22	**10**	0.52
23	-0.11	**11**	0.46
24	-0.10	**12**	0.37
25	-0.27	**13**	0.24
27	-0.02	**14**	0.10
28	-0.14	**15**	0.05
29	-0.33	**22**	0.00
30	-0.43	**26**	0.14
31	-0.41	**34**	0.04
32	-0.12		
33	-0.26		
35	-0.02		

To assess whether the two sets produce equivalent person parameter estimates, five hundred *t*-tests were run to compare the person parameters derived from the two sets of the descriptors for all the examinees. The analysis showed that, out of five hundred, 81 cases (16.2%) of the *t*-tests were statistically significant. As argued by Smith (2002), less than 5% of *t*-tests should be significant if a unidimensional scale is desired. The findings of the present study strongly support a lack of equality across the two sets of descriptors, indicating the multidimensionality of the EDD checklist.

Loading patterns of descriptors on the first hypothesized factor in the linearized residuals are presented in Figure 2. Descriptors with negative loadings are represented by small letters and located on the bottom end, while descriptors with positive loadings are represented by capital letters and located on the top. As can be seen, the residuals of the descriptors have formed two distinguishable clusters. In fact, descriptors have not scattered across the map, so the EDD checklist is multidimensional.

Finally, the largest standardized residual correlations for identifying dependent items are provided in Table 3. As can be seen, all the correlations of item residual pairs are higher than 0.20, suggesting that the items share a large proportion of their random variance which is an indication of dependency among item pairs.

Figure 2

Plot of Item Loadings on the First Factor in PCAR

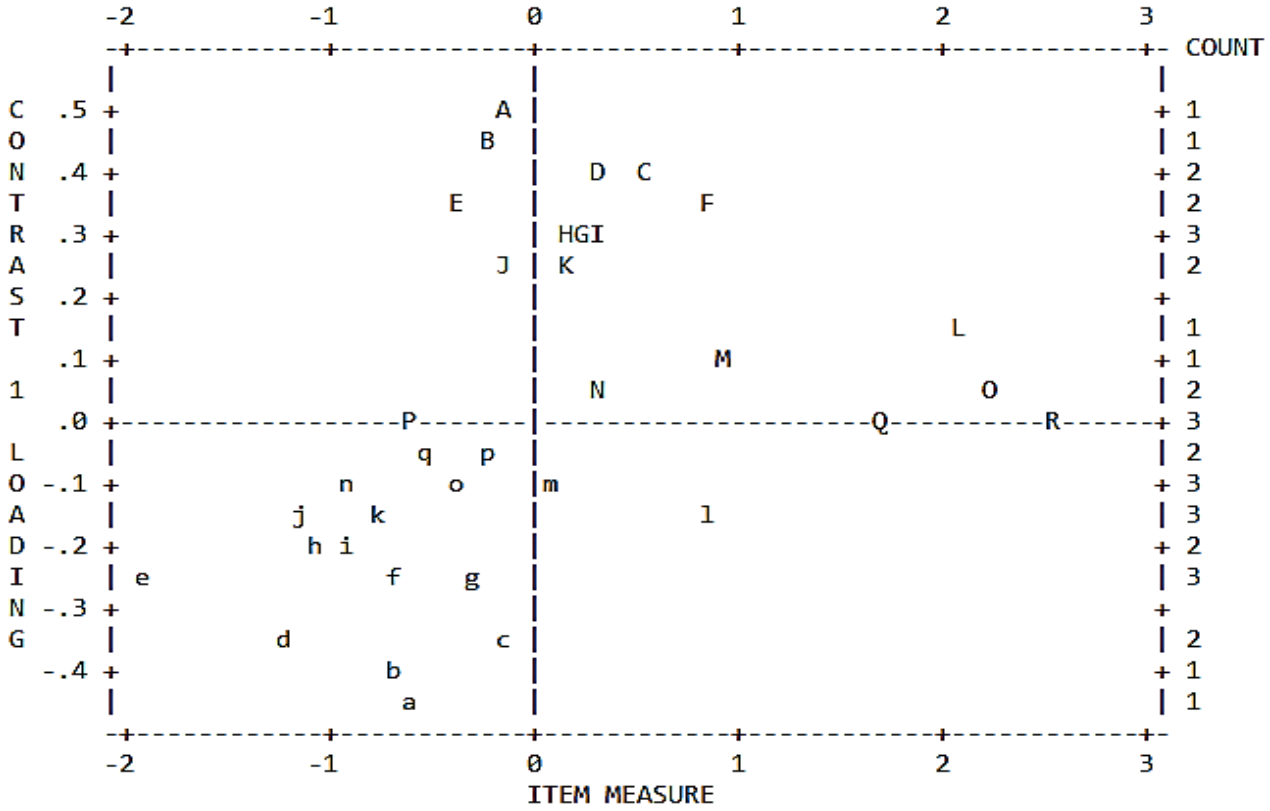


Table 3

The Largest Standardized Residual Correlations

Correlation	Entry Number Item	Entry Number Item
0.36	10	11
0.30	11	12
0.27	6	7
0.27	29	30
0.25	28	29
0.23	6	8
0.22	2	3
0.22	5	6
0.22	7	8
0.22	9	10

6 | Discussion

This study set out to use the Rasch model to psychometrically investigate the dimensionality of linguistic features in L2 writing. To operationalize the measurement of L2 writing, an empirically-derived descriptor-based diagnostic (EDD) checklist (Kim, 2010, 2011) commonly used within the context of L2 writing research was utilized. The analysis of item/descriptor difficulty parameters, fit statistics, and point-measure correlations showed the conformity between the

observed data and the expectations of the Rasch model, although one descriptor (e.g., Descriptor 22) had a poor fit. This descriptor needs to be carefully analyzed before being used in future studies. Item and person separation and reliability coefficients were also evaluated. The results of LID analysis also indicated the existence of dependency between some item pairs of the checklist. This could be due to the replicated features or potential shared dimensions across descriptors of the checklist. However, according to Schroeders et al. (2014), such dependency represents linguistic knowledge and "lie at the core of the construct; they are by no means irrelevant for the measurement of language abilities" (p. 414). It is highly recommended for future studies which intend to develop new rating scales for assessing L2 writing to avoid including descriptors with many replicated features.

Furthermore, the unidimensionality of the checklist was analyzed using PCAR. The results showed that eigenvalue of the first factor (contrast) was greater than 2, suggesting the multidimensionality of the EDD checklist. The analysis of item loadings further revealed the involvement of two sets of descriptors in defining the checklist. The comparison of person estimates across the two sets of descriptors showed a lack of equality across the two sets and provided further evidence for the multidimensionality of the checklist and L2 writing.

A closer analysis of negative and positive item loadings showed that descriptors that loaded negatively on Factor 1 were mainly related to lower-level linguistic features, such as vocabulary use (VOC), grammatical knowledge (GRM), and Mechanics (MCH). However, descriptors that loaded positively on Factor 1 were primarily associated to higher-level linguistic features of L2 writing, including content fulfillment (CON) and organizational effectiveness (ORG). This can be considered as an empirical evidence that there are two groups of L2 writers who are qualitatively different with respect to using linguistic features in their writing. This finding converges with Effatpanah et al. (2024) who recently conducted a research study to detect multiple classes of L2 writers using the mixed Rasch model (Rost, 1990). The authors recognized two classes for L2 writers. The first class consists of L2 writers who are inclined to attend more to lower-level linguistic features, including grammar, vocabulary, and mechanics, at the sentence level in order to produce a written text. On the other hand, the second class is characterized by L2 writers who prioritize higher-level linguistic features, such as content and organization, aiming to move beyond the boundaries of a sentence and pay more attention to the whole structure of a paragraph. This finding also supports Schoonen and De Glopper's (1996) contention that writers of different proficiency levels perceive the qualities of a well-written text differently. Writers with lower proficiency tend to prioritize layout and mechanics, including grammar, vocabulary, and mechanics, while those with a higher level of proficiency concentrate on more advanced or higher-order features including text organization and content.

Another finding of the study is that unlike reading and listening comprehension, the involvement of lower- and higher-level classifications of writing sub-skills may not be legitimate. Rather, each writing sub-skill can involve both higher- and lower-level constituents. For example, those vocabulary items which have more frequency and are widely used can be retrieved more easily compared to those vocabulary items which are sophisticated and require more cognitive processing load to be elicited. Similarly, simple grammatical structures can be easily retrieved and used in a written text, but complex grammatical structures need more cognitive capacity. This could be the possible reason why some descriptors, marked as higher-level linguistic features, were more difficult for the group who tends to focus on lower-level linguistic features at the sentence level.

This account of the hierarchy of linguistic features in L2 writing seems plausible because less proficient L2 writers have limited linguistic knowledge (e.g., content, vocabulary, grammar, organization, and mechanics) and generate written texts marked by many shortcomings, including limited intelligibility, lack of cohesion and coherence, distorted organization, and numerous lower-order issues such as grammar and spelling (Trapman et al., 2018). Consequently, lower level writers are more likely to resort to some strategies to make up for their lack of linguistic knowledge. This process can be justified by the *inhibition* and *compensation* assumptions proposed by van Gelderen et al. (2011). They argue that

> According to the inhibition assumption, inexperienced writers' inefficient use of grammatical and lexical knowledge impedes their monitoring of text production on the level of content and their use of higher

> order strategies for optimizing text quality on a global level … . According to the compensation assumption, … inexperienced writers may well be able to adopt higher order strategies in writing, although their capability for efficient retrieval and production of linguistic elements is still limited. For example, by using efficient strategies, working memory capacity can be spent on sequential processing of different aspects of the writing task, instead of simultaneous processing. (P. 283)

7 | Conclusion

This study has several theoretical, pedagogical, and methodological implications. From a theoretical standpoint, the findings of this study supported the multidimensionality of the linguistic features in L2 writing. Understanding the dimensionality of linguistic features in L2 writing enables L2 writing researchers and scholars to characterize the nature of L2 writing. This understanding facilitates the development of a coherent and comprehensive model for L2 writing, as well as the refinement of existing theories related to L2 writing production.

From a pedagogical standpoint, examining qualitative differences among L2 writers with regard to several linguistic features can prove to be a valuable approach for addressing individual variances. The analysis of qualitative differences among L2 writers allows teachers to pay much more attention to different sets of linguistic features (e.g., higher- and lower-level) in writing classes, identify problematic aspects of writing ability, offer targeted and appropriate feedback to individual students, devise more effective tasks, activities, and remedial materials tailored to each student, enhance writing instruction, and evaluate L2 writing. Students themselves can adopt some strategies to overcome deficiencies in their learning process and enhance their writing proficiency.

From a methodological standpoint, this study extends previous studies on the dimensionality of L2 writing by using the standard dichotomous Rasch model. More importantly, researchers typically use methods rooted in CTT and MFRM to investigate the psychometric qualities of rating scales. However, the use of various group-level inter-rater agreement and inter-rater consistency for a given dataset can generate inconsistent results, and the presence of high agreement and reliability among raters does not necessarily represent the precision of ratings (Wind & Peterson, 2018). Furthermore, studies on the use of MFRM mostly tend to focus on the performance of raters, especially their severity and leniency, and/or the interaction of raters with other factors such as the gender of examinees. Too little attention has been devoted to the analysis of the dimensionality of a scale and its construct. Another important implication of the present study for research in validating rating scales within the context of L2 writing research is that checking item loadings and LID can provide valuable information about the development and validation of rating scales for assessing L2 writing. Examining item loadings can explore qualitative differences that underlie the performance of examinees. Similarly, investigating LID allows researchers to avoid including descriptors and elements with replicated features.

This study had limitations that could be addressed in future research. First, the generalizability of the present results was limited to the current sample under investigation. Similar further studies are needed to shed more light on the generalizability of the findings from this study. Second, as repeatedly argued, a problem with binary diagnostic checklists is the dichotomization of an integrated skill like writing, although all researchers, practitioners, and teachers acknowledge that writing competence cannot be simplified into binary distinction. An intriguing avenue for further research involves examining the dimensionality of linguistic features in L2 writing across various academic tasks and genres. Such analyses would provide more in-depth insights into structural or qualitative differences among L2 writers with regard to their (meta)cognitive processes, especially their linguistic features.

Funding

The authors did not receive support from any organization for the submitted work.

Conflict of Interest

The authors have no competing interests to declare that are relevant to the content of this article.

Data Availability Statements

The datasets generated during and/or analyzed during the current study are freely available in Figshare at the following link: https://doi.org/10.6084/m9.figshare.12425357.v1.

How to Cite:

Effatpanah, F., & Baghaei, P. (2024). Examining the dimensionality of linguistic features in L2 writing using the Rasch measurement model. *Educational Methods & Practice, 2*:9.

References

Abbott, R. D., & Berninger, V. W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology, 85*(3), 478–508. https://doi.org/10.1037/0022-0663.85.3.478

Alamargot, D., & Chanquoy, L. (2001). *Through the models of writing*. Kluwer Academics.

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

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Baghaei, P., & Cassady, J. (2014). Validation of the Persian translation of the Cognitive Test Anxiety scale. *SAGE Open*, *4*(4). 1–11. https://doi.org/10.1177/2158244014555113

Bejar, I. I. (1983). *Achievement testing: Recent advances*. Sage.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum.

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th Ed.). Routledge.

Chenoweth, N. A., & Hayes, J. R. (2003). The inner voice in writing. *Written Communication, 20*(1), 99–118. https://doi.org/10.1177/0741088303253572

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Cleary, M. N. (2013). Flowing and freestyling: Learning from adult students about process knowledge transfer. *College Composition and Communication*, *64*(4), 661–687. https://www.jstor.org/stable/43490784

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218–244. https://doi.org/10.1037/0033-2909.90.2.218

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.

Cumming, A. (1990). Metalinguistic and ideational thinking in second language composing. *Written Communication, 7*(4), 482–511. https://doi.org/10.1177/0741088390007004003

Cumming, A. (2016). 3. Theoretical orientations to L2 writing. In R. Manchón & P. Matsuda (Ed.), *Handbook of second and foreign language writing* (pp. 65-88). De Gruyter Mouton. https://doi.org/10.1515/9781614511335-006

Dancey, C., & Reidy, J. (2004). *Statistics without maths for psychology: Using SPSS for Windows*. Prentice Hall.

Effatpanah, F., & Baghaei, P. (2021). Cognitive components of writing in a second language: An analysis with the linear logistic test model. *Psychological Test and Assessment Modeling*, *63*(1), 13–44. URL:https://www.psychologie-aktuell.com/journale/psychological-test-and-assessment-modeling/currently-available/inhaltlesen/psychological-test-and-assessment-modeling-2021-1.html

Effatpanah, F., Baghaei, P., & Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia, 9*(12), 1–23. https://doi.org/10.1186/s40468-019-0090-y

Effatpanah, F., Baghaei, P., & Karimi, M. N. (2024). A mixed Rasch model analysis of multiple profiles in L2 writing. *Assessing Writing, 59*, 1–15. https://doi.org/10.1016/j.asw.2023.100803

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

Fayol, M., Alamargot, D., & Berninger, V. (2012). From cave writers to elite scribes to professional writers to universal writers, translation is fundamental to writing. In M. Fayol, D. Alamargot, & V. Berninger (Eds.), *Translation of thought to written text while composing: Advancing theory, knowledge, research, methods, tools, and applications* (pp. 3–1). Psychology Press.

Feez, S. (2001). Heritage and innovation in second language education. In A. M. Johns (Ed.), *Genre in the classroom* (pp. 47–68). Lawrence Erlbaum.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Fischer, G. H. (1997). Unidimensional linear logistic Rasch models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 225–243). Springer.

Flower, L. S., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31–50). Lawrence Erlbaum.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387. https://doi.org/10.2307/356600

Flower, L. S., & Hayes, J. R. (1983). *A cognitive model of the writing process in adults* [Final Report]. ERIC Document Reproduction Service No. ED 240608. Carnegie Mellon University.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*(2), 208–238. https://doi.org/10.1177/026553229601300205

Fulcher, G. (2003). *Testing second language speakin*g. Longman/Pearson Education.

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 378–392). Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29. https://doi.org/10.1177/0265532209359514

Galbraith, D. (2009). Writing as discovery. *British Journal of Educational Psychology Monograph Series II, 6*, 5–26. https://doi.org/10.1348/978185409X421129

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspective*. Longman.

Güvendir, E., & Uzun, K. (2023). L2 writing anxiety, working memory, and task complexity in L2 written performance. *Journal of Second Language Writing, 60*, 1–14. https://doi.org/10.1016/j.jslw.2023.101016

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts.* Ablex.

Harklau, L. (2002). The role of writing in classroom second language acquisition. *Journal of Second Language Writing, 11*(4), 329–350. https://doi.org/10.1016/S1060-3743(02)00091-7

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388. https://doi.org/10.1177/0741088312451260

He, L., Jiang, Z., & Min, S. (2021). Diagnosing writing ability using China's Standards of English Language Ability: Application of cognitive diagnosis models. *Assessing Writing, 50*, 1–14. https://doi.org/10.1016/j.asw.2021.100565

Hickendorff, M., Edelsbrunner, P. A., McMullen, J., Schneider, M., & Trezise, K. (2018). Informative tools for characterizing individual differences in learning: Latent class, latent profile, and latent transition analysis. *Learning*

and Individual Differences, 66, 4–15. https://doi.org/10.1016/j.lindif.2017.11.001

Hirvela, A., Hyland, K. & Manchón, R. M. (2016). 2. Dimensions in L2 writing theory and research: Learning to write and writing to learn. In R. M. Manchón & P. Matsuda (Ed.), *Handbook of second and foreign language writing* (pp. 45–64). De Gruyter Mouton. https://doi.org/10.1515/9781614511335-005

Huang, Y., & Zhang, L. J. (2022). Facilitating L2 writers' metacognitive strategy use in argumentative writing using a process-genre approach. *Frontiers in Psychology, 13*, 1–17. https://doi.org/10.3389/fpsyg.2022.1036831

Hughes, A., & Hughes, J. (2020). *Testing for language teachers* (3rd Ed.). Cambridge University Press.

Hyland, K. (2004). *Genre and second language writers*. University of Michigan Press.

Hyland, K. (2009). *Academic discourse*. Continuum.

James, M. A. (2009). "Far" transfer of learning outcomes from an ESL writing course: Can the gap be bridged? *Journal of Second Language Writing, 18*(2), 69–84. https://doi.org/10.1016/j.jslw.2009.01.001

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, *26*(1), 31–73. https://doi.org/10.1177/0265532208097336

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*(4), 377–403. https://doi.org/10.1016/j.jslw.2003.09.001

Johns, A. M. (1997). *Text, role, and context: Developing academic literacies*. Cambridge University Press.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.). *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–72). Lawrence Erlbaum Associates.

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]. https://hdl.handle.net/1807/24786

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. (2007). *Diagnostic writing assessment: The development and validation of a rating scale* [Unpublished doctoral dissertation, University of Auckland].

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. Peter Lang.

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

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*, *38*(4), 602–626. https://doi.org/10.1177/0265532221994052

Kuo, B. C., Chen, C. H., Yang, C. W., & Mok, M. C. M. (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

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine grained clausal and phrasal indices. *The Modern Language Journal, 102*(2), 333–349. https://doi.org/10.1111/modl.12468

Lee, C., Ge, H., Chung, E. (2021). What linguistic features distinguish and predict L2 writing quality? A study of examination scripts written by adolescent Chinese learners of English in Hong Kong. *System, 97*, 1–19. https://doi.org/10.1016/j.system.2021.102461

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. https://doi.org/10.1093/applin/amp040

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. URL:https://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2009a). *WINSTEPS Rasch Measurement* (Version 3.68) [Computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2009b). *A user's guide to WINSTEPS*. Winsteps.

Linacre, J. M. (2014a). *Facets Rasch measurement* [Computer program]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2014b). *A user's guide to Facets Rasch-model computer programs*. Winsteps.com.

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

Lukácsi, Z. (2021). Developing a level-specific checklist for assessing EFL writing. *Language Testing*, *38*(1), 86–105. https://doi.org/10.1177/0265532220916703

Ma, X., Shi, X., Lu, C., & Li, R. (2022). Development and validation of a college English writing scale for classroom-based peer assessment. *Journal of Xi'an International Studies University*, *30*(1), 56–62. https://doi.org/10.16362/j.cnki.cn61-1457/h.2022.01.010

Manchón, R. M. (2011). *Learning-to-write and writing-to-learn in an additional language*. John Benjamins.

Manchón, R. M. (2016). Introduction: Past and future of L2 writing research. In R. M. Manchón & P. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 1–16). De Gruyter Mouton. https://doi.org/10.1515/9781614511335-003

Nation, I. S. P., & Macalister, J. (2010). *Language curriculum design*. Taylor & Francis.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*(2), 217–262. https://doi.org/10.1177/026553229801500204

Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology, 101*(1), 37–50. https://doi.org/10.1037/a0013462

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

Pu, L., Heng, R., & Xu, B. (2023). Language development for English-medium instruction: A longitudinal perspective on the use of cohesive devices by Chinese English majors in argumentative writing. *Sustainability, 15*, 1–15. https://doi.org/10.3390/su15010017

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Expanded Ed.). University of Chicago Press Originally published 1960, Pædagogiske Institut, Copenhagen.

Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing, 20*(1), 24–56. https://doi.org/10.1080/15305058.2019.1588278

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282. https://doi.org/10.1177/014662169001400305

Safari, F., & Ahmadi, A. R. (2023). Developing and evaluating an empirically-based diagnostic checklist for assessing second language integrated writing. *Journal of Second Language Writing, 60*, 1–15. https://doi.org/10.1016/j.jslw.2023.101007

Schroeders, U., Robitzsch, A., & Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with C-tests. *Journal of Educational Measurement*, *51*(4), 400–418. https://doi.org/10.1111/jedm.12054

Schoonen, R., & De Glopper, K. (1996). Writing performance and knowledge about writing. In G. Rijlaarsdam, H. van den Bergh, & M. Couzijn (Eds.), *Theories, models, and methodology in writing research* (pp. 87–107). Amsterdam University Press.

Schoonen, R., Snellings, P., Stevenson, M., & van Gelderen, A. (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* (pp. 49–76). Multilingual Matters.

Schoonen, R., van Gelderen, A., Stoel, R. D., Hulstijn, J., & De Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning*, *61*(1), 31–79. https://doi.org/10.1111/j.1467-9922.2010.00590.x

Schoonen, R., van Gelderen, A., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning, 53*(1), 165–202. https://doi.org/10.1111/1467-9922.00213

Shahsavar, Z. (2019). Diagnosing English learners' writing skills: A cognitive diagnostic modeling study. *Cogent Education*, *6*(1), 1–19. https://doi.org/10.1080/2331186X.2019.1608007

Shi, X., Ma, X., Du, W., & Gao, X. (2024). Diagnosing Chinese EFL learners' writing ability using polytomous cognitive diagnostic models. *Language Testing*, *41*(1), 109–134. https://doi.org/10.1177/02655322231162840

Smith E. V. Jr. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231. URL:http://jampress.org/abst.htm

Smith, R. M., & Plackner, C. (2009). The family approach to assessing fit in Rasch measurement. *Journal of Applied Measurement*, *10*, 424–437. URL:http://jampress.org/abst.htm

Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a check list. *Assessing Writing, 18*(3), 187–201. https://doi.org/10.1016/j.asw.2013.05.001

Trapman, M., van Gelderen, A., van Schooten, E., & Hulstijn, J. (2018). Writing proficiency level and writing development of low-achieving adolescents: the roles of linguistic knowledge, fluency, and metacognitive knowledge. *Reading and Writing, 31*, 893–926. https://doi.org/10.1007/s11145-018-9818-9

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49–70. https://doi.org/10.2307/3588360

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, *16*(1), 82–111. https://doi.org/10.1177/026553229901600105

van Gelderen, A., Oostdam, R., & van Schooten, E. (2011). Does foreign language writing benefit from increased lexical fluency? Evidence from a classroom experiment. *Language Learning, 61*(1), 281–321. https://doi.org/10.1111/j.1467-9922.2010.00612.x

Vögelin, C., Jansen, T., Keller, S. D., & Möller, J. (2018). The impact of vocabulary and spelling on judgments of ESL essays: An analysis of teacher comments. *The Language Learning Journal, 49*(6), 631–647. https://doi.org/10.1080/09571736.2018.1522662

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

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

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*(2), 161–192. https://doi.org/10.1177/0265532216686999

Wright, B. D. (1994). Local dependency, correlations, and principal components. *Rasch Measurement Transactions*, *10*(3), 509–511. URL:https://www.rasch.org/rmt/rmt103b.htm

Xie, Q. (2017). 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

Zhang, X. (2022). The relationship between lexical use and L2 writing quality: A case of two genres. *International Journal of Applied Linguistics, 32*(3), 371—396. https://doi.org/10.1111/ijal.12420

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment, 31*(4), 231–237. https://doi.org/10.1027/1015-5759/a000309

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). Academic Press.