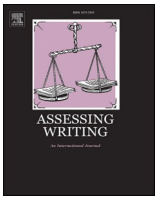
Assessing Writing 59 (2024) 100803

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw



A mixed Rasch model analysis of multiple profiles in L2 writing

Farshad Effatpanah [a,*,1], Purya Baghaei [b,2], Mohammad N. Karimi [c,3]

[a] Research Unit of Psychological Diagnostics, Faculty of Rehabilitation Sciences, TU Dortmund University, Dortmund, Germany
[b] English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran
[c] Department of Foreign Languages, Kharazmi University, Tehran, Iran

ARTICLE INFO

Keywords:
L2 writing
Linguistic features
Multiple profiles
Item response theory
Mixed Rasch Model

ABSTRACT

The present study used the Mixed Rasch Model (MRM) to identify multiple profiles in L2 students' writing with regard to several linguistic features, including content, organization, grammar, vocabulary, and mechanics. To this end, a pool of 500 essays written by English as a foreign language (EFL) students were rated by four experienced EFL teachers using the Empirically-derived Descriptor-based Diagnostic (EDD) checklist. The ratings were subjected to MRM analysis. Two distinct profiles of L2 writers emerged from the sample analyzed including: (a) Sentence-Oriented and (b) Paragraph-Oriented L2 Writers. Sentence-Oriented L2 Writers tend to focus more on linguistic features, such as grammar, vocabulary, and mechanics, at the sentence level and try to utilize these subskills to generate a written text. However, Paragraph-Oriented Writers are inclined to move beyond the boundaries of a sentence and attend to the structure of a whole paragraph using higher-order features such as content and organization subskills. The two profiles were further examined to capture their unique features. Finally, the theoretical and pedagogical implications of the identification of L2 writing profiles and suggestions for further research are discussed.

1. Introduction

In recent years, there has been a growing research interest in characterizing second/foreign language (L2) writing and exploring multiple profiles of (successful) L2 writers with respect to the distribution and frequency of several linguistic features that are assumed to affect L2 writing quality and distinguish between various levels of L2 writing proficiency (e.g., Crossley et al., 2014; Jarvis et al.,

Abbreviations: 2PL, Two-parameter logistic; 3PL, Three-parameter logistic; AIC, Akaike's Information Criterion; ANOVA, Analysis of Variance; AWE, Automated Writing Evaluation; BIC, Bayesian Information Criterion; CA, Cluster Analysis; CAIC, Consistent AIC; CON, Content Fulfillment; EDD, Empirically-Derived Descriptor-Based Diagnostic; EFL, English as a Foreign Language; ELT, English Language Teaching; GRM, Grammatical Knowledge; L1, First Language; L2, Second/Foreign Language; LCA, Latent Class Analysis; LCM, Latent Class Model; LPA, Latent Profile Analysis; MA, Multidimensional Analysis; MANOVA, Multivariate Analysis of Variance; MCH, Mechanics; MFRM, Many-Facet Rasch Measurement; MIRT, Mixture Item Response Theory; MNSQ, Mean Square; MRM, Mixed Rasch Model; ORG, Organizational Effectiveness; PCA, Principal Component Analysis; RM, Rasch Model; VOC, Vocabulary Use.

* Correspondence to: Research Unit of Psychological Diagnostics, Faculty of Rehabilitation Sciences, TU Dortmund University, Emil-Figge Street 50, 44227 Dortmund, Germany.
E-mail address: farshad.effatpanah@tu-dortmund.de (F. Effatpanah).
[1] ORCID ID: https://orcid.org/0000-0003-3970-5588
[2] ORCID ID: https://orcid.org/0000-0002-5765-0413
[3] ORCID ID: https://orcid.org/0000-0001-7834-6368

https://doi.org/10.1016/j.asw.2023.100803
Received 20 June 2023; Received in revised form 8 November 2023; Accepted 2 December 2023
Available online 15 December 2023

2003; Kim, 2020). A large number of studies in the past twenty years have investigated different variables likely to predict L2 writing performance, including L2 linguistic knowledge, writing strategies, instructional background, the purpose of writing, audience and topic, cultural expectations, text characteristics, and the writing medium. Among the wide array of variables that affect writing, linguistic knowledge has been documented to be the most noticeable factor in L2 writing impacting writing quality and raters' judgment (Kim & Crossley, 2018; Lee et al., 2021). Consequently, a number of studies, which will be reviewed in the following section, have examined the relationship between the distribution of different linguistic features (e.g., content, vocabulary, grammar, organization, and mechanics) and L2 writing quality, and have found significant moderate-to-strong correlations between them. Jarvis et al. (2003), however, argue that this research has primarily used correlational analysis and relied on a linear modeling approach between linguistic features and writing quality. Put differently, previous research has assumed that all writers adopt similar processes and patterns in developing a written text, and low- and high-quality writings follow a uniform or single pattern. However, there exist multiple configurations of a successful essay, and writers are likely to employ different patterns throughout the composing process (Crossley et al., 2014; Kim, 2020). Within this line of research, numerous studies have been conducted in the past few years to explore multiple profiles of (successful) writers using traditional approaches, such as latent class analysis (LCA), cluster analysis (CA), latent profile analysis (LPA), principal component analysis (PCA), and multidimensional analysis (MA). Although these methods are useful for classifying writers into different profiles, the use of Mixture Item Response Theory (MIRT) models, especially Mixed Rasch model (MRM; Rost, 1990) to identify profiles of writers in developing an essay has been neglected in writing research. The MRM conflates latent class models (LCM; Lazarsfeld & Henry, 1968) and the Rasch model (RM; Rasch, 1960/1980) to identify subpopulations of participants that differ substantially in their patterns of item difficulties. In other words, MRM is premised on the assumption that individuals from the same population are likely to use different solution patterns or strategies to complete a set of items/tasks, or employ the same solution patterns in different ways. Thus, it allows researchers to detect differential patterns across the respondents in different latent subpopulations or classes.

Against this background, the present study aims to use the MRM to explore multiple profiles of L2 writers who are likely to use linguistic features in different ways to produce a text. Finding patterns in the use of linguistic features and categorizing individuals into distinct classes can provide valuable information on variations in educational contexts (Hickendorff et al., 2018). Specifically, a better understanding of individuals' writing profiles provides information that can help advance theories of writing development, facilitate instruction and learning, and help teachers design remedial instructional materials for individual learners (Hickendorff et al., 2018).

2. Literature review

2.1. Linguistic features and writing quality

Writing in both first language (L1) and L2 is a complex cognitive process that relies on the integration of many (meta)cognitive processes and linguistic resources (Huang & Zhang, 2022; Zhao & Liao, 2021). Over the past few decades, numerous theoretical models have been developed to present a comprehensive representation of writing processes and developmental aspects of writing proficiency. Seminal models of writing (e.g., Chenoweth & Hayes, 2003; Flower & Hayes, 1981; Galbraith, 2009; Hayes, 2012; Kellogg, 1996) generally recognize that writing performance requires a writer to coordinate multiple cognitive (sub)processes and linguistic knowledge, which places a substantial demand on the central executive function of working memory capacity (Kellogg, 1996). According to Flower and Hayes (1981) model, which shifted the attention from simple linear sequence models to cognitive processes, the act of writing consists of three dynamic and hierarchical processes: planning, translating, and reviewing. They argue that L1 and L2 writers make plans, generate ideas, organize them, and set goals to create an internal representation of the knowledge that they intend to present in their writing; translate their ideas into texts through retrieving linguistic resources and relevant knowledge from their memory; and finally, systematically review and revise their texts to rectify errors. These processes tap into various skills, knowledge types, such as knowledge of linguistic resources and sociolinguistic knowledge, metacognitive, and neurodevelopmental processes (Schoonen et al., 2011).

Among the writing processes, translation is the major process through which propositional content or ideas are transformed into appropriate linguistic forms (van Gelderen et al., 2011). The process calls for sufficient knowledge of linguistic features, including content, grammar and sentence construction, word choice, organization or textual connection, and mechanics (e.g., spelling and punctuation). Generally, content and organization are considered as higher-level, and mechanics, grammar, and vocabulary as lower-level writing skills (Schoonen et al., 2011). The synchronization of lower- and higher-level writing skills constrains working memory capacity and impacts the quality of written texts (Vasylets & Marín, 2021). Previous studies have indicated that fluent or automatic use of lower-level skills in a text reduces the cognitive processing load and allows a writer to give more attentional capacity to higher-level skills (Effatpanah & Baghaei, 2021; Schoonen et al., 2011). Since linguistic knowledge is the foremost predictor of writing performance, a large number of studies have been conducted to examine the role of linguistic features in L2 writing performance (e.g., Abdi Tabari & Wang, 2022; Lan et al., 2022; Lee et al., 2021). The analysis of linguistic features has proven effective in assessing writing quality and differentiating between the performance of L2 writers with different proficiency levels. The research has established that writers with greater language proficiency are more likely to generate higher-quality texts (Weigle, 2002). Given the major role that linguistic knowledge plays in language learning, in general, and L2 writing, in particular, Bachman and Palmer (1996), in their model of language competence, maintain that language knowledge, in both L1 and L2, comprises two specific components: (a) organizational knowledge and (b) pragmatic knowledge. Organizational knowledge refers to the ability to generate or discern grammatically correct sentences, comprehend their propositional content, and order them to construct a text. Organizational knowledge consists of grammatical knowledge, which includes the knowledge of phonology, vocabulary, syntax, and graphology

involved in producing accurate sentences or utterances; and textual knowledge, which includes the knowledge of rhetorical organization and cohesion involved in composing. Pragmatic knowledge, however, is concerned with the relationship between utterances or texts and communicative goals or intentions of writers (or speakers).

Research has demonstrated that linguistic features can exert a great impact on writing quality and raters' judgments, such that L2 writers with high levels of proficiency have higher knowledge of content, grammar, vocabulary, organization, rhetorical structures, spelling, and punctuation (McNamara et al., 2010; Weigle, 2002). For instance, good writers have extensive domain-specific and discourse knowledge for the topic they are writing about. Consequently, they can use that knowledge in sophisticated ways and develop more relevant ideas in their composition (Olinghouse & Graham, 2009), including more arguments, contrasting ideas, supporting ideas, and exemplifications (McNamara et al., 2013). Studies have shown that higher levels of knowledge about the topic make significant contributions to writing development (Wen & Coker, 2020). Topical knowledge influences the time and cognitive workload allocated to writing processes (Kessler et al., 2022), and enables writers to produce well-structured texts (Yang et al., 2015). Discourse knowledge also helps writers to easily change discourse modes, based on the differences between narrative, comparative, descriptive, and argumentative essays (Crossley et al., 2014).

Vocabulary knowledge can be considered the other major factor in successful writing. It is broadly defined as the use of diverse, sophisticated, and advanced lexical items in spoken and written language (Laufer & Nation, 1995). A great deal of research has examined the relations between vocabulary knowledge and L2 writing quality, and reported positive moderate to strong correlations (e.g., Kim et al., 2022; Zhang, 2022). Research has shown that more proficient L2 writers use a wide variety of appropriate lexical items. Higher-proficiency L2 writers show a high level of lexical diversity and avoid lexical repetition (Kim & Crossley, 2018) indicating less word overlap and fewer referential links (Kim et al., 2022). More proficient L2 writers, furthermore, use less frequent sophisticated words more often than lower-level L2 writers (Zhang, 2022). More specifically, several researchers have found that higher-proficiency L2 writers tend to use cohesive devices or conjunctions (e.g., however, in addition to) or logical connectives/connectors to show logical links between sentences or clauses and direct readers to detect the organization of the text (Crossley et al., 2011; Kyle & Crossley, 2016). Higher-proficiency L2 writers are also more inclined to use more *collocations* (Haswell, 2000), *hedges* (e.g., likely, sort of), as a kind of metadiscourse, to show the writer's uncertainty (Biber et al., 2020), and *demonstratives* (these, those, this, and that) to enhance the text cohesion (Biber et al., 2020).

Aside from vocabulary knowledge, grammatical knowledge has been shown to be an influential predictor of successful L2 writing. Grammatical knowledge refers to the ability to use varied and complex grammatical structures (Ortega, 2015). Much research has examined the link between grammatical knowledge and L2 writing quality, and the findings indicate that syntactic structures and L2 writing quality are significantly associated, with reported correlations ranging from moderate to strong (Crossley et al., 2011; Lan et al., 2022; Lee et al., 2021). Numerous studies have found that higher-proficiency L2 writers typically produce longer clauses and sentences (Crossley & McNamara, 2014; Yang et al., 2015) and even more complicated syntactic structures (Crossley & McNamara, 2014; Kyle & Crossley, 2018) than less proficient L2 writers. In similar studies that investigated the level of syntactic complexity, a number of researchers have found that high-scored L2 essays include more prepositions, passives, transitional devices or conjunctions, adverbs and adjectives, pronouns and referents, and articles (Grant & Ginther, 2000; Hyland, 2002). Studies have also indicated that higher-rated texts use less sentence fragments and comma splices or run-on sentences (Crossley et al., 2011) and contain correct word forms, word order, singular and plural nouns, subject-verb agreement, and verb tenses (Lee et al., 2021).

Another important factor in successful writing is text organization. Organization concerns the cohesive and coherent development of ideas and supporting sentences within and between paragraphs, also known as discourse features (e.g., links between text elements). A great deal of research has examined the role of discourse features in explaining human ratings of L2 writing quality, and found controversial results. Although early studies on the relationship between L2 writing quality and discourse features reported positive correlations (e.g., Yang & Sun, 2012), several studies have argued that discourse features of L2 writing quality are likely to differ based on discourse types (Crossley et al., 2016; Pu et al., 2023). Studies have also indicated that higher-proficiency L2 writers have a tendency to generate less cohesive texts as measured by word overlap, although their texts are linguistically sophisticated (Crossley & McNamara, 2014). Numerous studies, however, have reported that higher-proficiency L2 writers use more explicit cohesive devices or connectives to develop coherence in their texts compared to lower-proficiency L2 writers and thus produce more cohesive texts (Haswell, 2000). Other studies have demonstrated that higher-rated L2 writing contains accurate and logically developed topic and supporting sentences (Pu et al., 2023).

Finally, successful L2 writing can be attributed to the use of mechanics, which is characterized as the use of conventions of English writing including spaces, indentation, spelling, punctuation, capitalization, and handwriting. Studies have shown that there is a direct relationship between mechanics and L2 writing quality (Harrison et al., 2016; Vögelin et al., 2018). Higher-rated L2 essays contain more accurate spelling, punctuation, capital letters, indentation, and margins (Limpo et al., 2017). Spelling has also been shown to tap several skills and knowledge sources including sound-letter correspondence, or phonological, morphological, and orthographic awareness (Limpo et al., 2017).

Taken together, the studies reviewed above support the fact that more proficient L2 writers can produce higher-quality written texts with regard to different linguistic features (e.g., content, vocabulary, grammar, organization, and mechanics). While these studies have provided valuable insights into the relations between linguistic features and L2 writing quality as well as characteristics of higher- and lower-proficiency L2 writers, this line of research has mostly addressed the issue of successful writing on the basis of linear modeling approaches, through the use of traditional analytical approaches such as Analysis of Variance (ANOVA), correlation, and regression-based techniques (Crossley et al., 2014). Previous studies have assumed that more and less proficient L2 writers use the same strategies, processes, and patterns in composing a text, acknowledging a certain level of homogeneity among individuals. However, L2 writers are likely to not adopt linguistic features in similar ways (Crossley et al., 2014). For that reason, Jarvis et al.

(2003) suggest that the analysis of multiple profiles of (successful) writers can offer more accurate information about the features of writers.

2.2. Writing profiles

Numerous research studies have focused on finding patterns and categorizing writers into multiple profiles. Two lines of research can be identified in the relevant literature. Some of this research has investigated profiles of writers based on writing strategies and (meta)cognitive factors. For instance, Torrance et al. (1994), using CA, recognized three distinct groups of writers (e.g., planners, revisers, and mixed-strategy writers) based on the responses of social sciences undergraduate students to a questionnaire, measuring writing strategies, experiences of writing, and productivity. Torrance et al. (2000) also used PCA on 715 essays produced by undergraduate psychology students and identified four profiles of writing behavior (minimal-drafting, detailed-planning, outline-and-develop, and think-then-do). In another study, Kim (2020) identified four profiles (e.g., Search-based, Plan-based, Revision-based, and Correction-based) among 260 Korean undergraduate students on the basis of some features of planning, revision, and internet searching activities during digital writing. She used LPA and a one-way Multivariate Analysis of Variance (MANOVA) to specify differences in text quality across the profiles. Cruz Cordero et al. (2023) further conducted a study to explore writing motivation and ability profiles in U.S. middle-school students participating in an automated writing evaluation (AWE) intervention using MI Write. They also aimed to discover transition paths between profiles as a result of the intervention. Using latent profile and latent transition analysis, four motivation and ability profiles with self-reported writing self-efficacy, attitude towards writing, and a measure of writing emerged: Low, Low/Mid, Mid/High, and High. Troia et al. (2023) recently used LPA to capture the underlying distinct profiles of 335 upper elementary students in order to investigate aspects of individual knowledge, motivation, and cognitive processes. The analysis yielded five interpretable and relatively evenly distributed performance patterns: Globally Weak, At Risk, Average Motivated, Average Unmotivated, and Globally Proficient. The authors concluded that the presence of five distinct profiles highlights the necessity of comprehensive but differentiated instruction to address knowledge, skills, and motivation to produce desirable outcomes.

Some of this research has also explored multiple profiles of writers with regard to the distribution of several linguistic features. For instance, Crossley et al. (2014), using CA approach, found four distinct profiles of successful essays: accessible style, action and depiction style, lexical style, and academic style. Their study provided empirical evidence that mixtures of linguistic features are used by successful writers to produce a quality text. In the context of L2 writing research, Jarvis et al. (2003) examined multiple profiles of highly rated L2 learner compositions and compared them in terms of 21 linguistic features. Performing CA on two datasets of timed compositions (one corpus of 160 University ESL placement compositions composed by 40 learners and a corpus of 178 Test of Written English (TWE) written by 178 language learners), they explored five clusters for the first dataset and three clusters for the second dataset. Their results showed that the clusters for both datasets were unique and different with regard to word length, adverbs, nouns, pronouns, prepositions, stative verbs, and present tense verbs. Extending Jarvis et al.'s (2003) study, Friginal et al. (2014) identified six writing profiles based on 23 linguistic features for highly-rated essays written by non-native speaking (NNS) and native English speaking (NS) university students. The profiles differed in text length, type/token ratio (e.g., a measure of vocabulary variation), nouns and adjectives, pronouns, nominalizations, prepositions, and pronouns. The authors concluded that there are multiple profiles even in a homogeneous group of highly rated essays. In a similar vein, Friginal and Weigle (2014) applied MA to detect the functional dimensions of L2 academic essays. They explored four functional dimensions or clusters for over 80 linguistic features across 209 essays: Addressee-Focused Description vs. Personal Narrative, Personal Opinion vs. Impersonal Evaluation/Assessment, Involved vs. Informational Focus, and Simplified vs. Elaborated Description. Although these studies have successfully demonstrated the existence of multiple profiles of writing, they primarily employed traditional methods such as CA, LCA, LPA, and MA for classifying writers. The major problem with these latent transition models is that they are dependent on mean scores and observed variables (Tabachnick & Fidell, 2013) and fail to model item responding (Aryadoust & Zhang, 2016). As argued by Hickendorff et al. (2018, p. 9), although latent transition models provide a clear representation of the underlaying structure of the data, they bring along with several assumptions. For example, latent class models explain all relations between the observed variables (Weller et al., 2020). They also typically make the assumptions that the indicators conform to a particular distribution, often the normal distribution, within each class (Hickendorff et al., 2018). Considering the shortcomings of latent transition models, this study uses MRM, as a psychometric analytic technique, to explore multiple profiles of L2 writers with regard to a set of linguistic features.

2.3. Mixed Rasch Model (MRM)

Mixed Rach Model (MRM; Rost, 1990) is an integration of the RM (Rasch, 1960/1980) and LCM (Lazarsfeld & Henry, 1968). Under the RM, the probability of a correct response to a particular item is modeled as a function of the ability of the person and the difficulty of the item. An important assumption of the RM is parameter invariance, e.g., the item difficulties should be constant between individuals and all members of the population. However, this assumption is unlikely to be satisfied if there are qualitative differences, like the use of different cognitive processes, between different groups of individuals. MRM relaxes this assumption and allows item parameters to differ across latent classes of a population, when the unidimensional RM does not hold for the entire population, although the RM holds within each latent class.

A distinguishing feature of the MRM is that the number of latent classes are specified without a priori identification of group membership, and each person must belong to only one latent class with the highest probability (Rost, 1990). Through the comparison of the patterns of item parameters across classes, characteristics of latent classes can be described, indicating structural and qualitative

differences between the sub-populations (Rost, 1990). The latent ability distribution and item parameters are further dependent on latent classes (Preinerstorfer & Formann, 2012). Understanding how individuals in different classes give different responses to each item of a given test or a task is totally in accordance with the concept of construct validity. According to Messick (1989), the main feature of construct validity is to grasp what mental processes and strategies individuals use to answer and/or solve test items or tasks. Due to these features, the MRM has already been used in a variety of research studies to examine the fit of the unidimensional RM, differential item functioning (DIF), test calibration, standard setting, response styles and faking personality, problem-solving strategies, and test speededness (see Sen & Cohen, 2019, for a comprehensive review of the applications of IRT mixture models).

3. Method

3.1. Data

The data for the present study were obtained from the performance of five hundred Iranian English as a foreign language (EFL) students on the following writing prompt (at least 350 words) in their writing courses: "*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 data came from a research poject conducted by Effatpanah et al. (2019) to use cognitive diagnostic models to diagnose L2 writing ability. The participants consisted of 212 junior, 152 senior, and 136 postgraduate students. A demographic questionnaire was used to collect participants' background information, including their age, gender, English learning background, and the amount of English use. Findings from the questionnaire survey showed that the participants' age ranged from 19 to 58 years (M = 24.89 years, SD = 6.30). There were 151 male (30.2%) and 349 female (69.8%) students. The gender imbalance comes from the heavy dominance of females in English departments in many Iranian universities. The participants reported their English learning background from 3 to more than 10 years, and their amount of English use from cannot say to almost every day. All students were native speakers of Persian and were studying English as an academic major. Informed consent was obtained from all individual participants included in the study. They were reassured that their information would remain confidential and anonymous.

Four experienced raters were also recruited to score the essays. All of the raters knew. English as their foreign language and Persian as their first language and, but all had native-like English language proficiency. Three held a master's degree in English language teaching and had scored 8 (overall) in IELTS exam, while one rater was a Ph.D. candidate of English Language Teaching (ELT) and a lecturer in university. The group of raters consisted of one female and three males aged between 28 and 39 years old (M = 31.75; SD =

Table 1
Writing sub-skills and their linguistic features.

Writing Skills	Description	Linguistic Features
Content Fulfillment (CON)	Content fulfillment assesses a writer's ability to address a given question by presenting unity and relevance of supporting sentences, information, and examples	• Thesis statement • Few redundant ideas • Supporting ideas and examples • Main arguments
Organizational Effectiveness (ORG)	Organizational effectiveness assesses a student's ability to develop and organize ideas and supporting sentences cohesively and coherently within and between paragraphs	• Topic sentence • Supporting sentences • Parapgraph
Grammatical Knowledge (GRM)	Grammatical knowledge assesses a student's ability to demonstrate syntactic variety and complexity accurately	• Syntactic variety • Word order • Sentence fragments • Run-on sentences or comma splices • Verb tenses • Sunject-verb agreem • Singular and plural nouns • Articles • Prepositions • Prnouns and referents • Word order • Word forms • Transitional devices
Vocabulary Use (VOC)	Vocabulary use assesses a student's ability to use a wide range of lexical items accurately and appropriately	• Sophisticated or advanced vocabulary • A wide range of vocabulary • Choice of vocabulary • Appropriate collocations • Appropriate tone and register
Mechanics (MCH)	Mechanics assesses a student's ability to follow the conventions of English writing such as margins and indentation, punctuation, spelling, and capitalization	• Punctuation • Capital letters • Indentation and margins • Spelling

4.99) with an average 14.3 years of experience in teaching and assessing L2 writing.

The raters evaluated the essays based on a diagnostic assessment scheme called the Empirically-derived Descriptor-based Diagnostic (EDD) checklist (Kim, 2010, see Appendix). The scale was originally devised for the independent essay section of Test of English as a Foreign Language ™ Internet-based Test (TOEFL iBT) to assess and describe non-native English-speaking students' writing in an academic setting. The EDD checklist included 35 dichotomous (Yes, No) descriptors measuring five writing sub-skills, including content fulfillment (CON), vocabulary use (VOC), grammatical knowledge (GRM), organizational effectiveness (ORG), and mechanics (MCH). All of the sub-skills relate to some linguistic features, and each descriptor can measure more than one sub-skill. The sub-skills and the associated linguistic features are illustrated in Table 1.

Because variation across raters in the rating procedure is a major threat to construct validity (Wind & Peterson, 2018), all raters underwent a 2-hour training session to decrease potential inconsistencies (construct-irrelevant variance) before rating the essays. During the session, the raters were instructed how to use the scale, and discussions were moderated to clarify the content of each descriptor separately. The raters were trained to choose the *yes* option for descriptors if writers had generally fulfilled the descriptors' standards; otherwise, the *no* option would be thought as appropriate. As suggested by Weigle (2002), a small sample of essays were then given to raters to acquaint themselves with the scale and represent its specific characteristics. The essays were randomized and compiled in four packages, and assigned to the judges after training. Each rater received 125 essays and copies of the checklist. Thirty five essays were inserted in each package to be scored by all the raters. The total score in the 35-descriptor checklist ranged from 0 and 35 with a mean of 17.62 and standard deviation of 7.59. The Cronbach's alpha reliability of the checklist was 0.89, which is highly satisfactory. Pearson correlation analysis was used to assess inter-rater reliability between the raters. The correlation coefficient across all raters was 0.82, indicating a satisfactory agreement in the two ratings. Cohen's Kappa was also computed to be 0.62, suggesting a high agreement.

3.2. Data analysis

3.2.1. Many-Facet Rasch Measurement (MFRM)

To ensure that the performance of raters did not exert any influence on the ratings and the class membership of writers, we used many-facet Rasch measurement (MFRM; Linacre, 1989). The MFRM is an elaboration of the Rasch family of models that allows researchers to include additional variables of interest (referred to as facets) besides item and person parameters that impact test scores. As defined by Bond et al. (2020, p. 145), facets are aspects of the measurement process that "routinely and systematically interpose themselves between the difficulty of the test and the ability of person". The most commonly used facets are raters in a constructed-response assessment, demographic characteristics (e.g., gender, race/ethnicity, etc.), domains in an analytic rating scale, and item/prompt type. The interposition of these facets may contaminate the measurement and lead to construct-irrelevant variance (Messick, 1989). For instance, if there is any indication that the performance of raters (e.g., rater severity or leniency) has affected the scores of students, it is essential to make appropriate adjustments on the scores before using them for any intended purposes (Eckes, 2011). For the purpose of this study, a three-faceted MFRM analysis was performed using the FACETS computer package, Version 3.71 (Linacre, 2014a) to examine the psychometric quality of the writers' scores prior to applying the MRM to the data.

To measure fit of the data to the MFRM, infit and outfit mean square (MNSQ) fit indices were computed (Linacre, 2002). Infit MNSQ is inlier-sensitive or information-weighted fit which is more sensitive to the response patterns of examinees to items targeted on the person, and vice versa. However, outfit MNSQ is outlier-sensitive fit which is more sensitive to unexpected responses to items with difficulty far from a person, and vice versa. The measurement is more reliable when the MNSQ fit statistics closely approach a value of 1.0. A reasonable range for outfit MNSQ and infit MNSQ is 0.5–1.50 (Linacre, 2014b). Values lower than 0.50 are called overfit and show less variation than expected by the model, which are generally benign; values higher than 1.50 are considered misfit and show aberrant response patterns than the expectations of the model, which distort the measurement and cause construct-irrelevant variance. To visually display the extent of variation and targeting in each facet, MFRM vertical ruler of all facets was also checked. The vertical ruler, as an interval scale, allows comparing the location of items/descriptors, students, and raters. In fact, it presents the relationships between items, persons, and raters on an item-person-rater map, known as Wright map, which expresses person ability estimates, item difficulty estimates, and rater severity measures on the same metric calibrated in logits. The Wright map shows how items disperse with respect to the ability of examinees and rater's performance.

We further examined rater severity, exact agreement, rater reliability, and rater separation. Rater severity, expressed in logits (or log odd-units), indicates to what extent raters have been severe or lenient in scoring the essays. Lower scores indicate higher difficulty measures or higher severity measures. According to Linacre (2014b), observed exact agreement is the proportion of times one rater gives the exact same score as another rater for all examinees. When there is unanimous agreement among raters, the observed exact agreement is 0; however, when all raters are in complete agreement, the observed exact agreement is 1. In cases where only some raters agree or raters agree on some examinees, then the observed exact agreement is calculated as the fraction of all opportunities to agree with other raters on all examinees. For individual raters, exact agreement indicates the percentage of times a given rater assigns exactly the same score as another rater when rating the same examinees (Linacre, 2014b). Reliability also shows to what extent the estimated parameters (e.g., severity measures, examinee ability, and additional facets) would be reproducible if the test were administered to a different sample with similar characteristics from the same population. Reliability varies from zero to one, with a high coefficient indicating a high degree of measurement consistency and a strong ability to differentiate between examinees. Finally, separation denotes the number of statistically different strata in the data (Linacre, 2014b).

3.2.2. Mixed Rasch Model analysis

The computer program WINMIRA 2001 (von Davier, 2001a) was used to apply the MRM to the data. A number of researchers (e.g., Alexeev et al., 2011; Sen, 2018) have shown that when MRM is applied to a test initially designed to conform to a two-parameter logistic (2PL) IRT model and three-parameter logistic (3PL) IRT model, it would result in detection of spurious latent classes, which, in turn, can lead to erroneous or ambiguous conclusions that have considerable effects on practitioners. It is thus necessary to initially investigate whether the tests conform to the 2PL and 3PL IRT models. The estimation of the 3PL IRT model with an additional guessing parameter appeared not reasonable for the data used in this study because it does not contain any multiple-choice items where guessing behavior is likely to occur. Therefore, a 2PL IRT model was fitted to the data and its results were compared against the

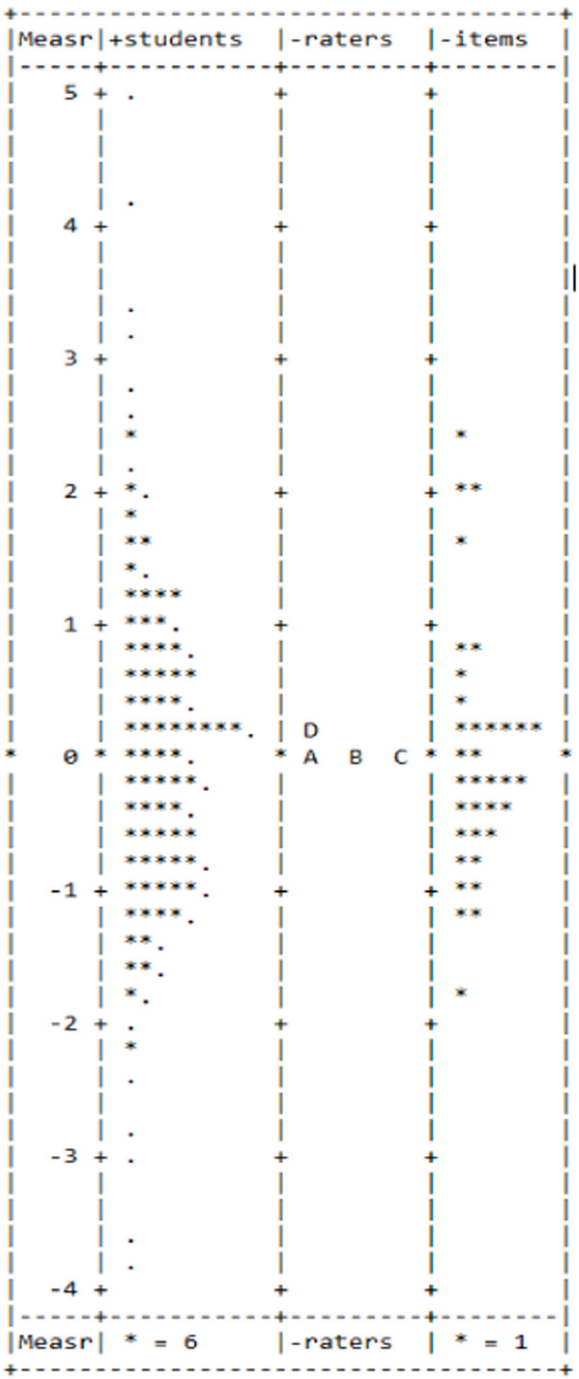


Fig. 1. : *All facets vertical summary.*

MRM.

There are generally two common approaches to MRM analysis: (1) exploratory analysis approach in which the number of latent classes is firstly identified, and then the relationship between a set of manifest variables (covariates) and the latent class membership is examined to explain the qualitative differences between the latent classes, and (2) confirmatory analysis approach in which several manifest variables (covariates) are directly incorporated into the model during the estimation of the latent classes (De Ayala & Santiago, 2016). In this approach, the existence of substantive a priori evidence helps to verify whether the covariates moderate person and item parameters (Sen & Cohen, 2019).

The present study took a two-stage exploratory approach to data analysis. First, because the number of classes is not a model parameter to be estimated, the fit of alternative models with one to four latent classes were compared to explore the optimal number of classes, and their results were compared against the 2PL IRT model. The models were evaluated by comparing three relative fit statistics: Akaike's Information Criterion (AIC) $= -2\ log\ L + 2\ P$, where L is the maximum likelihood function value, P is the number of parameters; Bayesian Information Criterion (BIC) $= -2\ log\ L + P\ \ln[n]$, where ln[n] is the natural log of sample size; and Bozdogan's Consistent AIC (CAIC) $= -2\ log\ L + p\ [\ln(n) + 1]$. The model with the lowest fit values indicated parsimony and better fit. Among these relative fit statistics, AIC has been shown to be less accurate and inconsistent unless the true model is among the competing models (Sen et al., 2019). According to Burnham and Anderson (2002), AIC tends to asymptotically choose the model that diminishes the mean square error prediction. However, as AIC does not impose any penalty for sample sizes, the more complex or highly parameterized model is selected with an increase in sample size. In contrast, with simulation studies, BIC and CAIC were found to have fairly more capacity in discerning the true model because they penalize more for sample sizes and the number of parameters, favoring the most parsimonious model (Preinerstorfer & Formann, 2012; Sen et al., 2019).

After detecting the number of classes, item fit statistics and item difficulty parameters were estimated for each resultant latent profile. Item difficulty parameters with logit units or log odds units for items represent the location of each item on the latent trait continuum (e.g., L2 writing ability). To thoroughly recognize the items that bring about qualitative differences between classes, within-class item difficulty parameters were compared in terms of graphical demonstration, reliability coefficients, and the correlation of item difficulties across the classes. For item fit analysis, furthermore, the item Q-index and its asymptotically standardized form, ZQ with zero mean and variance of one (Rost & von Davier, 1994; von Davier, 2001b), were used to examine the fit of descriptors/items. The index utilizes separability and conditional inference of descriptors/items to the latent variable, and is considered as an item discrimination index because it shows the extent to which a single item fits to the RM (Rost & von Davier, 1994). The Q-index is within the range of 0–1, where 0 shows perfect fit, 0.50 indicates random response behavior or the lack of relation between the item and the latent variable, and 1 reflects perfect misfit for the model. The ZQ also has a range between − 1.96 and + 1.96 (95% confidence interval), suggesting acceptable fit. A p-value is used to test the statistical significance of the observed misfit.

In the second stage of the analysis, a set of background variables including age, gender, English learning background, and the amount of English use were used to further explore the nature of each latent class. Previous studies have employed different covariates such as gender, ethnicity, age, metacognitive ability, intervention condition, and overall language ability (see Sen & Cohen, 2019, for a comprehensive review).

4. Results

4.1. Many-Facet Rasch Model (MFRM) analysis

The MFRM analysis showed that rater severity measures ranged between − 0.09 and 0.16, indicating no large differences in rater severity tendencies. The observed exact agreement and expected exact agreement rates, as indicators of inter-rater agreement, ranged from 87.6 to 91.2 and 60.9–63.8, respectively. This shows that the observed exact agreement statistics were well above the model expected agreement statistics, suggesting that in 87% of the cases, the raters assigned the same ratings to the writers being assessed. The infit and outfit MNSQ statistics were further examined to identify any misfitting raters. Infit and outfit MNSQ values ranging from 0.91 to 1.06 and 0.90–1.10, respectively, fell within the acceptable boundary of 0.5–1.5 (Linacre, 2014b), indicating that all the raters achieved satisfactory intra-rater consistency in scoring essays. The separation and reliability statistics of the rater facet were also 2.82 and 0.89, respectively.

The measures for persons, raters, and items/descriptors are graphically shown in Fig. 1. As can be seen, the variability in severity among raters is considerably narrower compared to that of the students and items. This indicates well-behaved performance of the raters. Considering the results of the MFRM, it can be concluded that the rater effect is negligible, and the results of the MRM analysis would not be affected by rater bias. The consistent behavior of the raters in this study could be due to their rater training background and the two training sessions they underwent before scoring the essays.

4.2. Mixed Rasch Model (MRM) analysis

The data was first subjected to the 2PL IRT model. The values of AIC (19546.21), BIC (19841.62), and CAIC (19911.10) revealed that this model had a poor fit compared to the MRM, except for one-latent class or Rasch model. This suggests that the 2PL item response functions do not accurately capture the underlying structure of the data. To find the optimal number of profiles, as explained earlier, MRM with different number of latent classes (one to four classes) were then fitted to the data. As presented in Table 2, the results of information criteria (AIC, BIC, and CAIC) were not consistent across the four models. According to the AIC, a four-class model was the best fitting model due to its higher complexity. However, the BIC and CAIC selected the model with two latent classes. Such

discrepancies between information criteria have been frequently reported in previous studies for model comparisons (Sen & Cohen, 2019). Since a model with two latent classes produced smaller values on the BIC and CAIC compared to the rival models, it was chosen as the best model and used for further analyses.

Table 3 provides essential class-specific information about the expected and observed class size, and mean class assignment probabilities of the two latent classes for the chosen two-class model. As can be seen, there was a significant conformity between the model expectations and the observed data patterns because approximately 76% of the sample were assigned to Class 1 and about 23% were assigned to Class 2, which amount to the expected sizes (e.g., 76% and 23%). The mean probability also indicated the average hypothetical class assignment probabilities of the two latent classes. The values of mean probability across the two classes showed that the diagonal values are not smaller than the off-diagonal elements, indicating high classification accuracy (Baghaei & Carstensen, 2013; Baghaei et al., 2019). The classification precision showed that test takers with a high probability of assignment to Class 1 (0.977), according to the two-latent class model, had a low probability of being classified in Class 2 (0.023).

Since a two-class model was selected, the parameters of the model were examined to explore distinct nature of the two profiles. Table 4 presents the difficulty parameters for the 35 descriptors, Q-indices, the standardized form of Q-index (ZQ), and *p* values for item/descriptor fit. As shown in Table 4, for Class 1, all descriptors yielded Q indices smaller than 0.50, indicating that the descriptors have sufficient discriminant power to distinguish well between low-ability students who possess a lower chance of performing successfully on each descriptor and high-ability students who possess a higher probability. However, ZQ statistics of six descriptors (e.g., 2, 3, 22, 26, 30, and 34) were beyond the boundary of ± 1.96 at 95% confidence interval, and their *p* values were significant. For Class 2, on the other hand, except for descriptor 9 whose Q index is equal to 0.50, all descriptors had adequate fit and discriminated well between high- and low-level students. With regard to ZQ index, four descriptors (e.g., 5, 11, 15, and 28) had values greater than the cut-off points and produced significant *p* values. The misfit descriptors in both classes may be due to the multidimensional nature of the checklist. As noted earlier, a couple of descriptors require more than one writing sub-skill.

The patterns of item difficulty parameters across the two latent classes are presented in Fig. 2. The *x*-axis indicates 35 descriptors of the checklist and *y*-axis indicates difficulty estimates of the descriptors in logit scale. Class 1 is illustrated with a solid line and Class 2 with a dotted line. As can be seen, the patterns of difficulty parameters showed an inconsistent pattern across the two classes, suggesting different cognitive structures for students from both classes (Rost, 1990), that is, there are substantial qualitative differences between students with regard to the use of linguistic features in writing. For instance, difficulty parameter estimates for Class 2 showed greater variability (ranging from −3.94 to 6.99) than Class 1 in which difficulty parameter estimates ranged from − 1.36 to − 2.56. For Class 1, almost the first half of descriptors (e.g., 1–15) was mostly easier than the rest of descriptors (e.g., 16–35) for which Class 2 showed a better performance. For Class 1, the easiest descriptors were 16, 17, 21, and 23, whereas descriptors 16, 19, 25, and 31 were the easiest descriptors for Class 2. The most difficult descriptors for Class 1 were 35, 34, 26, and 27, whereas descriptors 9, 10, 34, and 26 were the most difficult descriptors for Class 2. As demonstrated in Fig. 2, descriptors 4, 8, 9, 10, 11, 12, 13, 20, 29, and 32 contributed to variability between the two latent classes.

The considerable difference across the classes was supported by a moderate Spearman rank-order correlation ($r = 0.557$, $p = 0.001$) between difficulty parameter estimates from the two classes, suggesting slight agreement between the item parameter estimates across the two classes. An independent-samples *t*-test was also used to compare the mean of raw scores for the two classes. Results revealed a significant difference in means for Class 2 ($M = 11.07$, $SD = 5.83$) and Class 1 ($M = 19.62$, $SD = 6.92$, t (498) $= 12.12$), $P < 0.000$), with a better performance of Class 1 compared to Class 2. However, no substantial differences in reliability were observable across the two classes: reliability for Class 1 was 0.864, and for Class 2, it was 0.869. Females were further dominant in both classes (269 females in Class 1 and 80 females in Class 2).

Finally, multiple Chi-square tests were conducted to compare the two classes with regard to a set of background variables. A Chi-square test for independence indicated that the latent classes had a significant association with age χ^2 (4, n = 500) = 10.32, $p = 0.03$, phi = 0.14), English learning background χ^2 (5, n = 500) = 11.97, $p = 0.03$, phi = 0.15), and the amount of English use χ^2 (4, n = 500) = 30.51, $p = 0.00$, phi = 0.24), whereas there was no significant association between the classes and gender χ^2 (1, n = 500) = 0.14, $p = 0.70$, phi = −0.01).

5. Discussion

This study used MRM to explore multiple profiles of L2 writers who are likely to be different in using a set of linguistic features to produce a well-written text. To ensure that the students' marks and their class membership were not affected by the performance of the raters, MFRM was first applied before running the MRM. The analysis of MFRM showed that all of the raters achieved adequate fit to the model, indicating that the rater effect was negligible and would not affect the results of the MRM analysis. Models with one to four

Table 2
Mode-data fit information for the four estimated MRM.

Models	AIC	BIC	CAIC
2PL One-Latent Class	19546.21	19841.62	19911.10
Rasch One-Latent Class	19714.10	19865.83	19901.83
Rasch Two-Latent Class	19336.39	19644.05	19717.05
Rasch Three-Latent Class	19195.85	19659.46	19769.46
Rasch Four-Latent Class	19144.45	19764.00	19911.00

Table 3
Class-specific statistics across the two-latent class model.

Class	Expected Size	Observed Size	Mean Probability Class 1	Mean Probability Class 2
Class 1	0.762	0.765	0.977	0.023
Class 2	0.234	0.234	0.071	0.929

Table 4
Difficulty parameters of the descriptors, item fit statistics, and *p*-values for the two latent classes.

	Class 1				Class 2			
Items	Estimate	Q-Index	ZQ-Index	*p*-value	Estimate	Q-Index	ZQ-Index	*p*-value
1	0.647	0.216	0.185	0.426	1.634	0.257	0.596	0.275
2	-0.629	0.149	-2.008	0.977 * **	-2.067	0.181	0.087	0.465
3	-0.332	0.129	-2.671	0.996 * ** *	-1.653	0.186	0.231	0.408
4	-0.363	0.185	-0.904	0.817	0.608	0.256	0.812	0.208
5	0.059	0.194	-0.576	0.717	0.724	0.097	-1.837	0.966 * **
6	0.399	0.181	-0.918	0.820	0.646	0.209	0.289	0.386
7	0.130	0.178	-1.108	0.866	0.753	0.116	-1.527	0.936
8	0.009	0.213	0.054	0.478	0.482	0.154	-0.694	0.756
9	-0.030	0.213	0.049	0.480	6.999	0.500	0.064	0.474
10	-0.662	0.167	-1.455	0.927	3.034	0.098	-0.665	0.747
11	-0.574	0.198	-0.528	0.701	1.299	0.437	2.644	0.004 * *
12	-0.756	0.227	0.285	0.387	0.957	0.308	1.297	0.097
13	-0.118	0.210	-0.106	0.542	1.395	0.215	-0.152	0.560
14	0.811	0.203	-0.259	0.602	0.908	0.121	-1.334	0.908
15	0.209	0.197	-0.425	0.664	0.358	0.314	2.161	0.015 *
16	-1.361	0.176	-1.084	0.860	-3.947	0.100	-0.368	0.643
17	-1.097	0.178	-1.149	0.874	-1.773	0.150	-0.184	0.573
18	-0.667	0.210	-0.209	0.582	-1.464	0.222	0.590	0.277
19	-0.617	0.212	-0.126	0.550	-3.897	0.236	0.416	0.338
20	-0.424	0.197	-0.536	0.704	0.094	0.123	-1.023	0.846
21	-0.902	0.261	1.173	0.120	-1.780	0.137	-0.377	0.647
22	-0.711	0.280	1.837	0.033 *	-0.595	0.156	-0.453	0.674
23	-0.824	0.264	1.280	0.100	-1.610	0.245	0.828	0.203
24	-0.540	0.264	1.473	0.070	-0.232	0.166	-0.556	0.711
25	-0.219	0.257	1.447	0.073	-2.706	0.245	0.854	0.196
26	1.898	0.310	2.315	0.010 *	2.811	0.291	0.570	0.284
27	1.682	0.220	0.178	0.429	1.242	0.113	-1.461	0.928
28	0.862	0.204	-0.232	0.591	0.444	0.040	-2.461	0.993 * ** *
29	0.033	0.221	0.322	0.373	-1.138	0.151	-0.339	0.632
30	-0.307	0.281	2.151	0.015 *	-2.001	0.092	-0.882	0.811
31	-0.228	0.215	0.070	0.472	-2.516	0.266	1.051	0.146
32	0.065	0.233	0.742	0.228	-0.343	0.219	0.641	0.260
33	-0.094	0.186	-0.840	0.799	-1.269	0.187	0.164	0.434
34	2.092	0.305	2.027	0.021 *	2.917	0.290	0.560	0.287
35	2.562	0.211	-0.365	0.642	1.684	0.142	-1.008	0.843

Note. * $p < 0.05$ * * $p < 0.01$ * ** $p > 0.95$ * ** * $p > 0.99$

classes were then fitted to the data, and the model with two latent classes proved to fit better to the data. Class 1 consisted of approximately 76% and Class 2 about 23% of the sample. The analysis of qualitative differences between the two classes showed that they are significantly different with respect to the cognitive processes writers adopt to successfully complete a writing task.

The inspection of class differences indicated that Class 1 had a higher mean in writing performance relative to Class 2. Most of the first fifteen descriptors, related to content fulfillment (CON) and organizational effectiveness (ORG), were mainly easier for Class 1, and the remaining descriptors, associated with grammatical knowledge (GRM), vocabulary use (VOC), and mechanics (MCH), were easier for Class 2. This finding suggests that Class 2 includes writers who tend to focus more on linguistic features, such as GRM, VOC, and MCH, at the sentence level and try to utilize these subskills to produce a text. However, Class 1 consists of writers who go beyond the boundary of a sentence and devote attention to the structure of a whole paragraph by making use of CON and ORG subskills. These two subtypes of L2 writers can be labeled *'Sentence-Oriented'* and *'Paragraph-Oriented'* L2 writers, respectively.

A closer scrutiny of the patterns of difficulty parameters of the descriptors across the classes, presented in Fig. 2, reflects that Sentence-Oriented L2 writers tend to exhibit more correct English word orders, consistent subject-verb agreement, appropriate prepositions, correct pronouns and referents, fewer sentence fragments, and fewer run-on sentences and comma splices. With regard to vocabulary, Sentence-Oriented writers are more likely to utilize a variety of vocabulary items, appropriate collocations and word choices, and appropriate word forms. Concerning mechanics, they also exhibit more use of correct spelling, proper punctuation marks, suitable capital letters, and appropriate tone and register. On the other hand, the performance of Paragraph-Oriented writers show that they depend on higher-order subskills. They can present a clearer thesis statement, develop more appropriate and logical arguments,

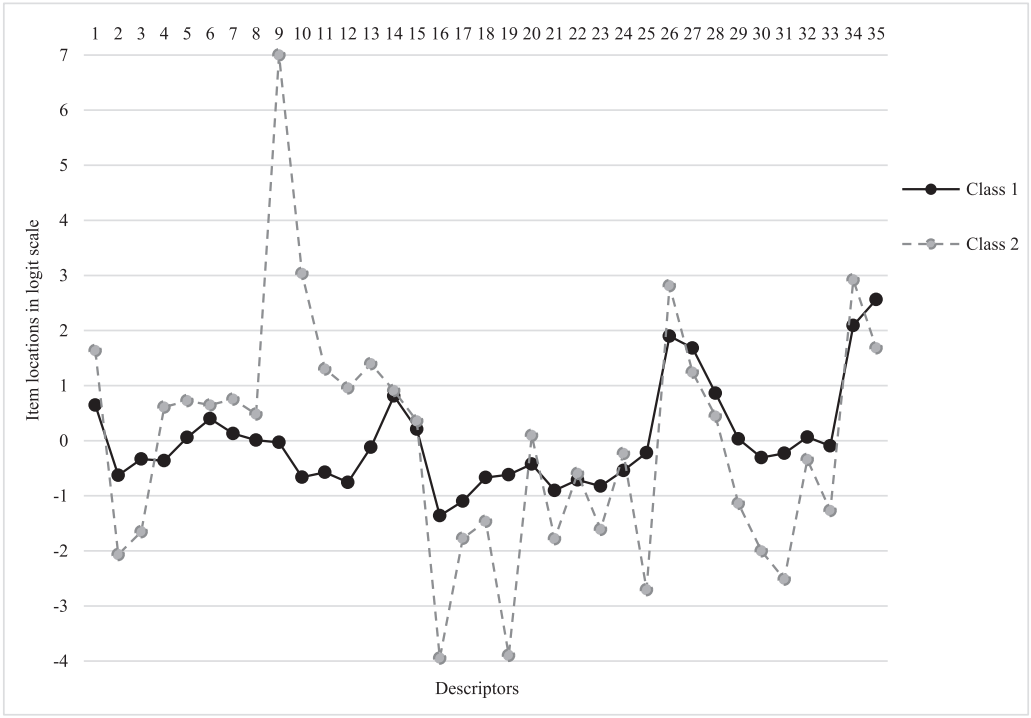


Fig. 2. : Difficulty parameters across the two latent classes.

supporting ideas, and examples. They also exhibit a higher ability to organize ideas into paragraphs, develop or expand ideas more explicitly within and across paragraphs, cohesively connect paragraphs to each other, and provide a clear topic sentence tied to supporting sentences to present one distinct idea within each paragraph. They further appear to be able to use more sophisticated or advanced vocabulary items, appropriate articles, and proper verb tenses than Sentence-Oriented writers. This finding agrees with the study conducted by Schoonen and De Glopper (1996) who argue that writers at varying proficiency levels have a different perception of what a good characteristic of a written text is; lower-proficiency writers prefer to focus on mechanics and layout, including vocabulary and grammar, whereas higher-proficiency writers focus on higher order features such as content and text organization. Notwithstanding the differences across the classes, the two classes are convergent in terms of using effective transition devices, syntactic variety, and appropriate singular and plural nouns.

The results from the analysis of the classes further highlight the importance of working memory capacity in the cognitive processes that writers need to handle to write successfully. Automatizing or being fluent in retrieving lower-level linguistic features (e.g., vocabulary, grammar, and mechanics) may occupy little of writers' attention and thus allow writers to leave more cognitive capacity for higher-order features, including content and organization (Schoonen et al., 2003). More specifically, rapid access to knowledge of lexis, grammar, and mechanics has a great impact on the efficient use of working memory capacity. Studies have shown a significant correlation between working memory measures (both in terms of efficiency and accessibility), writing fluency, and writing quality (Vasylets & Marín, 2021). In the current study, the performance of Sentence-Oriented L2 writers suggests that they do not appear to have fluent access to lower-level linguistic features (e.g., vocabulary, grammar, and mechanics). On the contrary, the performance of Paragraph-Oriented L2 writers indicate that they have efficient access to lower-level linguistic features which reduce the cognitive processing load and leave little of writers' attention. Consequently, this increases the cognitive capacity of the writers to pay attention to higher-order linguistic features (e.g., content and organization), and finally enhance writing quality.

Given that Sentence-Oriented writers have limited knowledge of lower-order linguistic features and do not possess adequate cognitive capacity for retrieving higher-order features, which may impact the quality of writing, they are more likely to produce a text that includes many shortcomings in terms of intelligibility and organization. Still, writers may have recourse to various compensatory strategies to overcome the negative impacts of working memory constraints (van Gelderen et al., 2011). According to the compensation assumption proposed by van Gelderen et al. (2011, p. 283), when writers do not have automatic access to lower-level features circumventing their monitoring of higher-level features, "[they] 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". Van Gelderen et al. further clarify that instead of simultaneous processing, working memory can be spent on sequential processing of various facets of the writing task. Moreover, Schoonen et al. (2009) contend that a rich repertoire of vocabulary knowledge may provide writers with opportunities to make up for their deficiency in generating content elements and expressing their ideas. Schoonen et al. (2011) also argue that a good knowledge of grammar and vocabulary helps writers to formulate their ideas more clearly, when they cannot link pieces of text logically. Therefore, the better performance of Sentence-Oriented writers on lower-level linguistic features in this study might be due

to the execution of some compensation strategies developed by writers to make up for their lack of competence in higher-order linguistic features.

Finally, the present study found significant relationships between class membership, age, English learning background, and the amount of English use. This finding indicates that the differences across the two classes can be due to the effect of these variables. A closer inspection of the covariates suggested that writers with lower age, less learning background, and smaller amount of English use belong to Class 1 (e.g., Sentence-Oriented writers). This finding is expected because Class 1 includes lower-level writers who are more likely to have not reached a level at which they can simply retrieve higher-level linguistic features. In contrast, Class 2 consists of writers with higher age, much learning background, and larger amount of English use. On the other hand, no significant relationship between class membership and gender of writers was observed, suggesting that gender does not play a role in assigning students into classes. This finding accords with previous studies which reported the absence of a correspondence between classes and gender (e.g., Aryadoust & Zhang, 2016). Nevertheless, it must be noted that females generally outperformed males in both classes which substantiates the results of previous studies on the relationship between L2 writing performance and gender (e.g., Cheong et al., 2022).

6. Implications, limitations, and directions for future research

The present study has several implications for research and practice related to exploring multiple profiles of L2 writers and using MRM in language testing and assessment. With regard to research, this study builds upon and extends previous applications of MRM to language skills. Unlike previous studies which applied the MRM to receptive skills (e.g., reading and listening), this study aimed to use the MRM within the context of performance-assessment in general and L2 writing in particular. The procedure illustrated in this study provides insights into the practical effectiveness of MRM for identifying qualitative differences across L2 writers' subgroups.

The research findings also have a set of theoretical and pedagogical significance. From a theoretical perspective, understanding multiple profiles of L2 writers and how they adopt different (meta)cognitive processes, or even linguistic features, would allow scholars to identify the exact nature of L2 writing, develop a logical and comprehensive model of L2 writing, and revise current theories of L2 writing performance. From a pedagogical perspective, exploring writing patterns and classifying L2 writers into multiple classes can be an effective way to address individual differences (Hickendorff et al., 2018). The analysis of each profile and its specific characteristics helps teachers and researchers to diagnose problematic areas of writing proficiency, provide sufficient and accurate feedback to individual students, design better tasks, activities, and remedial (authentic) materials for individual students, improve writing instruction, assess L2 writing, and train better raters. Teachers can also inform students of the results of investigating their writing patterns. Students can then take several actions or develop strategies to remove their deficiencies during the process of learning and improve their writing proficiency.

When considering the results from this study, several limitations and suggestions for future research should be noted. Although writing researchers are aware that writing proficiency cannot be treated as dichotomous, the raters in this study were required to make binary decisions while utilizing the EDD checklist. As a reviewer of an earlier draft of this paper rightly pointed out, the results of the present study might have been in part the function of how writing performance has been rated. An integrative skill like writing may not lend itself to dichotomous scoring. The results of the study might have been influenced by bending the construct of writing out of its natural shape. A polytomous scoring of the descriptors might have appeased the concern. Therefore, future studies can use rating scales which include a scale (e.g., a 5-point scale) to make multi-level choices about students' writing performance. Over the past few decades, researchers have developed different models as unidimensional (e.g., Sen et al., 2019; Tseng & Wang, 2021), polytomous (von Davier & Rost, 1995), and multidimensional extensions (e.g., von Davier, 2008) of MRM and successfully applied the models to language data. Future studies can utilize these models to examine their usefulness in language testing and educational measurement. Another interesting area for further investigation is to examine what patterns and profiles would emerge if essays of different genres or across different academic tasks were used. Such analyses would provide additional insights into different (meta)cognitive processes, or even linguistic features, L2 writers adopt to develop a written text. Future studies can also consider various factors and/or covariates to capture a thorough understanding of L2 writers' profiles. Factors such as writers' L1, general language proficiency level of the writer, context of writing, topic, task, etc. are important predictors for the profile membership of a writer.

Funding

The author(s) received no specific funding for this work from any funding agencies.

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability

The datasets generated during and/or analyzed during the current study are freely available in figshare at the following link: https://figshare.com/s/221cc358fa1c4e26cf9.

Appendix. The EDD Checklist (Kim, 2010)

		Yes	No
1	This essay answers the question.	☐	☐
2	This essay is written clearly enough to be read without having to guess what the writer is trying to say.	☐	☐
3	This essay is concisely written and contains few redundant ideas or linguistic expressions.	☐	☐
4	This essay contains a clear thesis statement.	☐	☐
5	The main arguments of this essay are strong.	☐	☐
6	There are enough supporting ideas and examples in this essay.	☐	☐
7	The supporting ideas and examples in this essay are appropriate and logical.	☐	☐
8	The supporting ideas and examples in this essay are specific and detailed.	☐	☐
9	The ideas are organized into paragraphs and include an introduction, a body, and a conclusion.	☐	☐
10	Each body paragraph has a clear topic sentence tied to supporting sentences.	☐	☐
11	Each paragraph presents one distinct and unified idea.	☐	☐
12	Each paragraph is connected to the rest of the essay.	☐	☐
13	Ideas are developed or expanded well throughout each paragraph.	☐	☐
14	Transition devices are used effectively.	☐	☐
15	This essay demonstrates syntactic variety, including simple, compound, and complex sentence structures.	☐	☐
16	This essay demonstrates an understanding of English word order.	☐	☐
17	This essay contains few sentence fragments.	☐	☐
18	This essay contains few run-on sentences or comma splices.	☐	☐
19	Grammatical or linguistic errors in this essay do not impede comprehension.	☐	☐
20	Verb tenses are used appropriately.	☐	☐
21	There is consistent subject-verb agreement.	☐	☐
22	Singular and plural nouns are used appropriately.	☐	☐
23	Prepositions are used appropriately.	☐	☐
24	Articles are used appropriately.	☐	☐
25	Pronouns agree with referents.	☐	☐
26	Sophisticated or advanced vocabulary is used.	☐	☐
27	A wide range of vocabulary is used.	☐	☐
28	Vocabulary choices are appropriate for conveying the intended meaning.	☐	☐
29	This essay demonstrates facility with appropriate collocations.	☐	☐
30	Word forms (noun, verb, adjective, adverb, etc.) are used appropriately.	☐	☐
31	Words are spelled correctly.	☐	☐
32	Punctuation marks are used appropriately.	☐	☐
33	Capital letters are used appropriately.	☐	☐
34	This essay contains appropriate indentation.	☐	☐
35	Appropriate tone and register are used throughout the essay.	☐	☐

References

Abdi Tabari, M., & Wang, Y. (2022). Assessing linguistic complexity features in L2 writing: Understanding effects of topic familiarity and strategic planning within the realm of task readiness. *Assessing Writing, 52*, 1–14. https://doi.org/10.1016/j.asw.2022.100605

Alexeev, N., Templin, J. L., & Cohen, A. S. (2011). Spurious latent classes in the mixture Rasch model. *Journal of Educational Measurement, 48*(3), 313–332. http://www.jstor.org/stable/23018149.

Aryadoust, V., & Zhang, L. (2016). Fitting the mixed Rasch model to a reading comprehension test: Exploring individual difference profiles in L2 reading. *Language Testing, 33*(4), 529–553. https://doi.org/10.1177/0265532215594640

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press,.

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research, & Evaluation, 18*(5), 1–13. https://doi.org/10.7275/n191-pt86

Baghaei, P., Kemper, C. J., Reichert, M., & Greiff, S. (2019). Applying the mixed Rasch model in assessing reading comprehension. In V. Aryadoust, & M. Raquel (Eds.), *Quantitative data analysis for language assessment Volume II: Advanced methods* (pp. 15–32). Routledge.

Biber, D., Gray, B., Staples, Sh, & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes, 46*, 1–11. https://doi.org/10.1016/j.jeap.2020.100869

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach (2nd Ed.)*. Springer.

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

Cheong, C. M., Zhang, J., Yao, Y., & Zhu, X. (2022). The role of gender differences in the effect of ideal L2 writing self and imagination on continuation writing task performance. *Thinking Skills and Creativity, 46*, 101–129. https://doi.org/10.1016/j.tsc.2022.101129

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication, 31*(2), 184–214. https://doi.org/10.1177/0741088314526354

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1–16. https://doi.org/10.1016/j.jslw.2016.01.003

Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 438–440). Springer.

Cruz Cordero, T., Wilson, J., Myers, M. C., Palermo, C., Eacker, H., Potter, A., & Coles, J. (2023). Writing motivation and ability profiles and transition during a technology-based writing intervention. *Frontiers in Psychology, 14*, 1–15. https://doi.org/10.3389/fpsyg.2023.1196274

De Ayala, R. J., & Santiago, S. Y. (2016). An introduction to mixture item response theory models. *Journal of School Psychology, 60*, 25–40. https://doi.org/10.1016/j.jsp.2016.01.002
Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang,.
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/fileadmin/download/ptam/1-2021/Seiten_aus_PTAM_2021-1_ebook_2.pdf.
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
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
Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi- dimensional analysis. *Journal of Second Language Writing, 26*, 80–95. https://doi.org/10.1016/j.jslw.2014.09.007
Friginal, E., Li, M., & Weigle, S. (2014). Revisiting multiple profiles of learner compositions: A comparison of highly rated NS and NNS essays. *Journal of Second Language Writing, 23*, 1–16. https://doi.org/10.1016/j.jslw.2013.10.001
Galbraith, D. (2009). Writing as discovery. In *British Journal of Educational Psychology Monograph Series II, 6* pp. 5–26). https://doi.org/10.1348/978185409×421129
Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123–145. https://doi.org/10.1016/S1060-3743(00)00019-9
Harrison, G. L., Goegan, L. D., Jalbert, R., McManus, K., Sinclair, K., & Spurling, J. (2016). Predictors of spelling and writing skills in first- and second-language learners. *Reading and Writing: An Interdisciplinary Journal, 29*(1), 69–89. https://doi.org/10.1007/s11145-015-9580-1
Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication, 17*(3), 307–352. https://doi.org/10.1177/0741088300017003001
Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388. https://doi.org/10.1177/0741088312451260
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
Huang, Y., & Zhang, L. J. (2022). Facilitating L2 writers' metacognitive strategy use in argumentative writing using a process-genre approach. *Frontier in Psychology, 13*, 1–17. https://doi.org/10.3389/fpsyg.2022.1036831
Hyland, K. (2002). Options of identity in academic writing. *ELT Journal, 56*(4), 351–358. https://doi.org/10.1093/elt/56.4.351
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.
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.
Kessler, M., Ma, W., & Solheim, I. (2022). The effects of topic familiarity on text quality, complexity, accuracy, and fluency: A conceptual replication. *TESOL Quarterly, 56*(4), 1163–1190. https://doi.org/10.1002/tesq.3096
Kim, H. (2020). Profiles of undergraduate student writers: Differences in writing strategy and impacts on text quality. *Learning and Individual Differences, 78*, 1–12. https://doi.org/10.1016/j.lindif.2020.101823
Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. *Assessing Writing, 37*, 39–56. https://doi.org/10.1016/j.asw.2018.03.002
Kim, M., Crossley, S. A., & Kim, B.-K. (2022). Second language reading and writing in relation to first language, vocabulary knowledge, and learning backgrounds. *International Journal of Bilingual Education and Bilingualism, 25*(6), 1992–2005. https://doi.org/10.1080/13670050.2020.1838434
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.
Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003
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
Lan, G., Li, X., & Zhang, Q. (2022). Revisiting grammatical complexity in L2 writing via exploratory factor analysis. *Frontiers in Psychology, 13*, 1–6. https://doi.org/10.3389/fpsyg.2022.860753
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307
Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton Mifflin.
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
Limpo, T., Alves, R. A., & Connelly, V. (2017). Examining the transcription-writing link: Effects of handwriting fluency and spelling accuracy on writing performance via planning and translating in middle grades. *Learning and Individual Differences, 53*, 26–36. https://doi.org/10.1016/j.lindif.2016.11.004
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. (2014a). Facets Rasch measurement. Chicago, IL: Winsteps.com (Computer program).
Linacre, J.M. (2014b). A user's guide to Facets Rasch-model computer programs. Chicago, IL: Winsteps.com.
McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of quality writing. *Written Communication, 27*(1), 57–86. https://doi.org/10.1177/0741088309351547
McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods, 45*(2), 499–515. https://doi.org/10.3758/s13428-012-0258-1
Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd Ed., pp. 1–103). American Council on Education/Macmillan.
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
Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing, 29*, 82–94. https://doi.org/10.1016/j.jslw.2015.06.008
Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology, 65*(2), 251–262. https://doi.org/10.1111/j.2044-8317.2011.02020.x
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.
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
Rost, J., & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement, 18*(2), 171–182. https://doi.org/10.1177/014662169401800206
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

Sen, S. (2018). Spurious latent class problem in the mixed Rasch model: A comparison of three maximum likelihood estimation methods under different ability distributions. *International Journal of Testing, 18*(1), 71–100. https://doi.org/10.1080/15305058.2017.1312408

Sen, S., & Cohen, A. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives, 17*(4), 177–191. https://doi.org/10.1080/15366367.2019.1583506

Sen, S., Cohen, A. S., & Kim, S. H. (2019). Model selection for multilevel mixture Rasch models. *Applied Psychological Measurement, 43*(4), 272–289. https://doi.org/10.1177/0146621618779990

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics (6th Ed.)*. Pearson Education,.

Torrance, M., Thomas, G. V., & Robinson, E. J. (1994). The writing strategies of graduate research students in the social sciences. In *Higher Education, 27* pp. 379–392). https:// doi.org/10.2307/3448190.

Torrance, M., Thomas, G. V., & Robinson, E. J. (2000). Individual differences in under-graduate essay-writing strategies: A longitudinal study. *Higher Education, 39*(2), 181–200. https://doi.org/10.1023/A:1003990432398

Troia, G. A., Wang, H., & Lawrence, F. R. (2023). Latent profiles of writing-related skills, knowledge, and motivation for elementary students and their relations to writing performance across multiple genres. *Contemporary Educational Psychology, 71*, 1–14. https://doi.org/10.1016/j.cedpsych.2022.102100

Tseng, M. C., & Wang, W. C. (2021). The Q-Matrix anchored mixture Rasch model. *Frontiers in Psychology, 12*, 1–9. https://doi.org/10.3389/fpsyg.2021.564976

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

Vasylets, O., & Marín, J. (2021). The effects of working memory and L2 proficiency on L2 writing. *Journal of Second Language Writing, 52*, 1–14. https://doi.org/10.1016/j.jslw.2020.100786

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

von Davier, M. (2001a). WINMIRA [computer program]. Groningen, the Netherlands: ASC-Assessment Systems Corporation. USA and Science Plus Group.

von Davier, M. (2001b). WINMIRA user manual. Groningen, the Netherlands: ASC- Assessment Systems Corporation. USA and Science Plus Group.

von Davier, M. (2008). The mixture general diagnostic model. In G. R. Hancock, & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1–24). Information Age Publishing.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch models. In G. H. Fischer, & I. W. Molennar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 371–379). Springer Verlag.

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

Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent class analysis: A guide to best practice. *Journal of Black Psychology, 46*(4), 287–311. https://doi.org/10.1177/0095798420930932

Wen, H., & Coker, D. L., Jr. (2020). The role of discourse knowledge in writing among first- graders. *Journal of Writing Research, 12*(2), 453–484. https://doi.org/10.17239/jowr-2020.12.02.05

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

Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education, 23*(1), 31–48. https://doi.org/10.1016/j.linged.2011.09.004

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28*, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002

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

Zhao, C. G., & Liao, L. (2021). Metacognitive strategy use in L2 writing assessment. *System, 98*, 1–11. https://doi.org/10.1016/j.system.2021.102472

Farshad Effatpanah is a Ph.D. student and a research assistant at the Faculty of Rehabilitation Sciences, TU Dortmund University, Germany. He holds an M.A. in TEFL (teaching English as a foreign language) from the Islamic Azad University, Mashhad Branch, Mashhad, Iran. His major research interest is the application of item response theory models in analyzing educational and psychological data as well as test validation and scaling.

Purya Baghaei is an associate professor in the English Department, Islamic Azad University, Mashhad Branch, Mashhad, Iran. He holds a PhD in applied linguistics from Alpen-Adria Universität, Klagenfurt, Austria. He is a scholar of the Alexander von Humboldt Foundation in Germany and has conducted post-doctoral research at universities in Vienna, Berlin, Jena, and Bamberg. His major research interest is in foreign language proficiency testing with a focus on the applications of item response theory models in test validation and scaling. He has published numerous articles on language testing and cognitive components of second language acquisition in international journals.

Mohammad N. Karimi is a professor of Applied Linguistics, Kharazmi University, Tehran, Iran. His areas of interest include reading/writing, teacher education and cognitive aspects of language acquisition. His papers have appeared in international journals such as *System*, *Modern Language Journal*, *Language Awareness*, *Applied Linguistics*, *Language Teaching Research*, etc.