ASIAN JOURNAL OF THE SCHOLARSHIP OF TEACHING AND LEARNING VOLUME 4 ■ NUMBER 3 ■ SEP 2014

ORIGINAL ARTICLE

Examining the development of paragraph writing ability of tertiary ESL students: A continuous assessment study

Vahid ARYADOUST[1]

[1] Centre for English Language Communication
National University of Singapore

Address for Correspondence: Dr Vahid ARYADOUST, Centre for English Language Communication, National University of Singapore, 10 Architecture Drive, Singapore 117511.
Email: vahid.aryadoust@nus.edu.sg

Recommended citation:
Aryadoust, V. (2014). Examining the development of paragraph writing ability of tertiary ESL students: A continuous assessment study. *Asian Journal of the Scholarship of Teaching and Learning, 4*(3), 153-179.
https://doi.org/10.24112/ajsotl.43313

Examining the development of paragraph writing ability of tertiary ESL students: A continuous assessment study

Abstract

This study investigates the development in paragraph writing ability of 116 undergraduate English as a second language (ESL) students enrolled in a paragraph writing course. Students wrote sample paragraphs before, during, and after the course, and these were marked on an analytical scale by multiple expert raters. The results were first subjected to many-facet Rasch model (MFRM) analysis to measure differences in rater severity and identify rater misfits; raters' scores were anchored to these initial results to generate fair scores for students. Next, a curve-of-factors latent growth model was fitted to the scores. The results showed that students' ability in multiple writing skills grew gradually and linearly from the beginning of the course. This progress was found to be independent of the writing prompts. Students' development is attributed to a variety of facilitative factors, including explicit lessons and frequent practice, regular feedback through a continuous assessment (CA) approach and various opportunities to engage with class tutors, and the use of online technology in the course.

Recent years have seen a growing scholarly interest in the assessment of writing skills. Research into both first-language (L1) and second-language (L2) writing has shown that students' writing skills grow as their neural connections, linguistic awareness (i.e., awareness of syntax, morphology, and orthography), and cognitive functions develop (Abbott, Berninger & Fayol, 2010; Berninger et al., 2010). At the early stages of writing, children have established few functional neural connections, and have difficulty mastering the alphabets, orthography, and spelling (Marr & Cernak, 2003). As they mature, their neural networks develop, which helps them master primary linguistic functions such as syntax, vocabulary, and grammar (Abbott & Berninger, 1993). With proper instruction, mature learners may also master more cognitively complex writing functions such as planning, drafting, and revising (Long, 2007).

Typically, L1 writers possess relatively rich linguistic resources; for example, they usually require little to no explicit vocabulary instruction (Nation, 2006). What they do need is to acquire cognitive strategies necessary to develop coherent paragraphs. By contrast, L2 writers may struggle both with poor cognitive strategies and with limited linguistic resources (Hodges, Whitten, Horner, Webb & Miller, 1990). To help these learners develop their writing skills,

L2 researchers and practitioners have developed and implemented a variety of "pedagogical interventions" in L2 studies (Benevento & Storch, 2011).

Among the interventions attempted for L2 learners, a substantial body of research suggests that ongoing instructor feedback can play a major role in L2 writing pedagogy (Ferris & Roberts, 2001). Ferris and Hedgcock (1998) argued that the provision of both direct and indirect feedback can help L2 learners achieve higher levels of accuracy over time, and Ferris's (2003) empirical study supported this presumption. Evidence also suggests that providing "full, explicit written feedback, together with individual [learner-teacher] conference feedback" results in (non-linear) development in the structure and lexico-grammatical accuracy of L2 student texts, and helps significantly reduce linguistic errors, such as errors in vocabulary and sentence structure (Bitchener, Young & Cameron, 2005, p. 201).

More recently, Storch and Tapper (2009) examined the effect of a 12-week writing module on the fluency, accuracy, coherence, and lexico-grammatical proficiency of L2 writers. Most students' writing components improved significantly over time; Storch and Tapper postulated that continuous feedback would be the most important cause of this improvement (Bitchener et al., 2005; Storch & Tapper, 2000), and found that the provision of explicit grammar and vocabulary lessons and student exposure to academic texts also played a role (Ferris, 2003; Hinkel, 2004; Leki, 2006). These findings concur with Polio, Fleck, and Leder's (1998) 15-week study of an academic writing module in which teachers provided continuous feedback to L2 students, leading to significant improvements in the accuracy of their writing.

Storch's (2009) examination of L2 university students enrolled in an Australian university also appears to highlight the importance of active "pedagogical intervention" to L2 writing development. The students received no writing lessons and no instructor feedback during the 10-week study, and made no noticeable improvement in any writing skill, which partially resonates with Hinkel's (2003) study of college-bound L2 learners studying in the US. Storch argued that semester-long L2 writing programs alongside content courses can help improve multiple aspects of L2 writing, including structure, knowledge of genre, and lexico-grammatical knowledge, but that educational programs lacking L2 writing support might not be equally effective (DeKeyser, 2007).

Pedagogical intervention has also been shown effective by portfolio studies, where students write multiple drafts and continuously revise their texts according to their teachers' feedback (Andrew & Romova, 2011; Adler-Kassner & O'Neill, 2010). For example, students in Benevento and Storch's (2011) L2 portfolio study received repeated feedback on their writing, and experienced improvements in their grammar, phraseology, coherence, and overall writing quality; Polio et al. (1998) and Chandler (2003) found similar results. Benevento and Storch (2011, p. 107) argued that feedback is effective to the extent that learners "notice" it

and "implement corrections in their own work." The efficiency of feedback and instructions also depends on the type and requirements of a given task, because different tasks "[necessitate] the use of vocabulary and grammatical structures of varying difficulty and familiarity" (Benevento & Storch, 2011, p. 107).

Possible Gaps in Learning and Assessment

Different writing skills generally do not develop evenly, and there is some evidence that important skills may remain both underassessed and relatively undeveloped among L2 writers. Bae and Lee (2012) found a non-linear growth pattern in coherence, grammar, and punctuation skills among young Korean learners of English, but no discernable growth in task fulfillment—writers' ability to provide relevant or "on-topic" responses to a task prompt. To our knowledge, except Bae and Lee's study, no research has examined task fulfillment in developmental studies of writing. Task fulfillment has obvious significance in overall writing quality, as it indicates whether the writer is actually producing content that is relevant to the requirements of a given writing task (Hinkel, 2004).

Furthermore, L2 writing assessment itself may be biased to an unknown extent by the influence of different writing tasks and task types on performance. Abbott et al. (2010, p. 283) state that "assessing [L2 writing over time] requires repeated testing of the same individual." To perform repeated assessments, developmental researchers have either developed parallel writing tasks (Weigle, 2002) or used the same tasks across multiple time points at a fixed interval (Storch & Tapper, 2009). However, there is yet no consensus over the potential role that different writing task types may play in assessment. Some evidence exists that young L2 learners perform significantly differently on different types of writing tasks, but the amount and type of impact on adult L2 learners is inconclusive (Benevento & Storch, 2011). For example, Koda (1993) compared the difficulty of descriptive and narrative tasks, and found descriptive tasks to be less cognitively demanding for young Japanese learners of English; Way, Joiner, and Seaman (2000) found similar results among secondary school French learners. Comparing the difficulty of several tasks, Way et al. reported a lower mean score for expository tasks, indicating that it taxed more cognitive resources than descriptive and narrative tasks, and argued that task difficulty might influence low-ability writers more than mature and competent writers. By contrast, Aryadoust (2012) found no task effect on adult learners of English.

Finally, most L2 developmental writing research has explored the effect of educational interventions on word-, sentence-, and essay-level cognitive and linguistic processes. Too little attention has been given to paragraphs—discourse units comprising a group of sentences which develop a main idea (Oshima & Hogue, 1991). The degree to which task effects may influence paragraph-writing skills is largely unexplored.

Present Study

The current study seeks to investigate the growth of English as a second language (ESL) university students' paragraph writing in three areas: "Organisation," the structure and coherence of the written texts; "Content," (task fulfilment) the relevance and originality of their written ideas to the task as assessed by human raters; and "Language," the accuracy and fluency of the language used (Hayes & Flower, 1980; Manchón, 2009).

Unlike previous research, this study uses paragraphs as the unit of investigation. The institute that hosted the present study offers writing courses that start by teaching paragraph writing. The expectation is that students who master paragraph writing skills will be prepared to construct lengthier units of discourse, such as essays and theses (Hinkel, 2004).

Specific research questions addressed in this study include:

> a) How do first-year university learners enrolled in an academic paragraph writing course improve in their paragraph writing skills (assessed as Organisation, Content, and Language), as well as in their overall writing ability, over a period of one academic semester?
>
> b) Does the rate of writing growth differ across writing tasks?

Similar to Benevento and Storch's (2011), Storch's (2009), and Storch and Tapper's (2009) research, this study does not include a control group, as it has not been designed to compare various interventions; rather, it adheres to the curriculum and explores trends in students' development over 12 weeks of lessons.

The present study has several important advantages over previous research. First, the assessment of the learners in this study is based on six fairly lengthy texts produced at the beginning, in the middle, and at the end of an academic semester, thereby reducing measurement error and enabling more precise measurement (Aryadoust, 2012); most previous research used two or three texts for evaluation. Second, in response to the call by Abbott et al. (2010) for developing reliable assessment tasks, the present study applied an anchor-based many-facet Rasch model (MFRM) to the data collected in the three time points. It also benefits from a large sample size, improving the generalizability of the results.

Writing Models

The most influential model of cognitive strategies in writing is the model proposed by Hayes and Flower (1980), which comprises three recursive cognitive processes: (a) planning—the process of brainstorming, setting elaborate goals, and organising ideas; (b) translating—the process of converting ideas into written text; and (c) reviewing—the process of evaluating and, if necessary,

revising the written text (see also Manchón, 2009).

During planning, writers retrieve relevant information from their working memory and structure it coherently. This retrieval can be either top-down (resulting from a pre-planned macrostructure) or bottom-up (planned spontaneously during writing) (Elbow, 1981). The writers then evaluate the structured information against their goals and adjust it according to their understanding of the writing task.

During translation, writers match their ideas with linguistic units (words, sentences, and paragraphs) in their working memory. Research offers strong evidence that an individual writer often constructs these linguistic units quite differently: for example, the ability to generate words is uncorrelated with the ability to produce sentences, which is in turn uncorrelated with the ability to write paragraphs (Crossley, Weston, McLain Sullivan & McNamara, 2011; Whitaker, Berninger, Johnston & Swanson, 1994). Finally, during reviewing, writers work to rectify orthographic mistakes and grammatically ill-formed structures, rarely changing the written text's overall meaning (MacArthur, Graham & Schwartz, 1991).

Hayes and Flower's (1980) process-oriented model is not necessarily in conflict with post-process paradigms of teaching and learning in writing (e.g., Kent, 1999). It has been argued that process-oriented models give too much weightage to learners' cognitive mechanisms, thereby ignoring the social dimension of writing (Kent, 1999). However, this assertion can be easily refuted, because students who are engaged in writing process do not write in a vacuum; rather, they typically write in response to a task or have an audience or a social situation in mind. Additionally, post-process and similar frameworks do not seem to offer a significant advantage over Hayes and Flower's model. They assert that writing processes are not generalisable across individuals, which is at odd with most well-researched theories of mind and comprehension/production (Kintsch & Mangalath, 2011). Matsuda (2003, p. 74) accuses post-process theorists of "discursive construction" of the history of writing pedagogy, arguing that "the notion of post-process needs to be understood not as the rejection of process but as the recognition of the multiplicity of L2 writing theories and pedagogies" (Matsuda, 2003, p. 65).

In the present study, the three chosen writing skills correspond to the components of the writing models proposed by Hayes and Flower (1980) and Manchón (2009): the students set goals and generate ideas during planning, a process which corresponds to the Organisation component of the employed writing scale. Also during planning, they evaluate their own ideas against their understanding of the writing task and adjust accordingly; that is, they attempt to fulfill the requirements of the task by describing, evaluating, discussing, and arguing (Hinkel, 2004). This process is represented and measured by the

Content component. Finally, they translate their thoughts into sentences and paragraphs by drawing on their linguistic skills, a process which corresponds with the Language component. Throughout the writing process, students revise their texts (Abbot et al., 2010; MacArthur et al., 1991; Adler-Kassner & O'Neill, 2010).

METHODOLOGY

Participants

The study comprised 116 first-year university students (both males and females) aged between 18 and 22 years, enrolled in a basic academic English course at the Centre for English Language Communication (CELC) of the National University of Singapore (NUS). They come from various countries, including China, Malaysia, Indonesia, India, Myanmar, Singapore, and Vietnam. Undergraduate students admitted to NUS are required to take a qualifying English test, a source-based writing placement test which must be taken by the students who are not competent in academic English language. The qualifying test will determine which students should take the basic English module—which focuses on paragraph writing skills—and English for academic purposes, and which students may be exempted from taking supplementary academic English programs. Following a placement test, they were placed in the basic English module and taught in 11 classes by four tutors.

The students are from a range of disciplines, including business, computer and electrical engineering, geography, social science, and real estate.

The Writing Module

The paragraph writing classes met twice per week, each session taking approximately two hours. The main elements[1] of the schedule were, as follows:

(a) *Explicit writing lessons*. The tutors explain the new lesson; illustrate a model for the lesson; involve students in practice by creating links between reading and writing; and provide scaffolding and guidance until students attain independence in writing. For example, the tutors first teach accurate use of English grammar, and cohesion (e.g., connectors and anaphora). Next, students are asked to read a number of texts of varying lengths, and to identify and repair poorly structured or incoherent passages. Other important concepts stressed include unity and organisation, as well as the structure of topic, support, and conclusion statements. Teachers provide multiple example paragraphs and discuss their underlying structure in class, and students practice analysing paragraph structure with peers. Students also perform systematic text-editing exercises.

1 This module includes other components such as producing effective outlines, research, and oral presentations, which for space constraints are not discussed.

(b) *Grammar and academic vocabulary lessons*. The lessons contain explicit (and sometimes implicit) instruction in grammar and vocabulary, two major components of writing. Students are given numerous sentences and paragraphs containing target academic vocabulary. Improvements in these areas can significantly improve students' writing skills (Coxhead, 2012), specifically in the translation stage of writing (Abbott et al., 2010). The grammar and vocabulary lessons in the course are designed to improve the degree of organisation and clarity with which students can express their thoughts.

(c) *Continuous assessment*. The class makes use of ongoing assessments to maintain a supportive classroom environment, and to aid the tutors in pitching their teaching at a suitable level for students (Kramer, 1999). The continuous assessment (CA) approach allows teachers to assess students' strengths and weaknesses, and to adjust their teaching methodology and materials based on these assessments periodically (Le Grange & Reddy, 1998). Like formative assessment, CA is intended to provide an efficient method for assessing students' writing skills; encourage interactions between tutors and students; and strike a balance between assessment and learning (Hamp-Lyons & Condon, 2000).

(d) *Consultation and teacher feedback*. For various affective and motivational reasons, students occasionally shy away from engaging in dialogues with their peers and tutors (Pajares & Valiante, 1997). Therefore, the tutors convene multiple 20- to 30-minute conferences with each student, and provide written and oral feedback on stylistic and language issues across several drafts of the student's writing; the tutors assess Content, Organisation, and Language. This study uses the students' graded first drafts, because the texts students independently produce best represent their "actual development level." (Later drafts, which incorporate the tutors' facilitative feedback, better represent the "level of potential development" or "zone of proximal development": forms of language expression which learners have not yet mastered, but have the potential to develop with proper scaffolding and support [Vygotsky, 1978, p. 86].)

(e) *Online follow-up practices and discussions*. The class uses online technology in several ways. First, the tutors set up an online chat room for students to converse on defined topics. This is intended to foster a sense of rapport between tutors and students, and to assure students that their attempts to make progress are recognised and appreciated (Deden & Carter, 1996). In addition, tutors keep students updated through emails and online messages. Finally, students submit the drafts of their individually written paragraphs via a web service called *Turnitin*.

Writing Prompts

Students were tested four times: once before and once after the course, and twice during it. The two mid-course assessments were separated by only one week, so they have been combined into one "mid-course" assessment result.

Pre- and post-course prompts. The pre- and post-course writing prompts were not originally part of the curriculum, but were added to track students' development. The researcher and course coordinator initially prepared four major prompts based on the course objectives, and on Kroll and Reid's (1994) guidelines for developing clear, unbiased, and unambiguous prompts. These four prompts were submitted to five writing experts with extensive teaching and research experience in academic writing. The experts rated the degree of clarity, cultural bias, and ambiguity of the prompts (Appendix 1), and agreed that an expository and a comparison-and-contrast prompt would be well-suited for the pre- and post-course assessments and parallel with the mid-course prompts (Table 1).

Mid-course prompts. The mid-course assessments have been used during the class as a significant element of CA. The syllabus advocates administering multiple prompts to students at each assessment point; students may choose to respond to any prompt. The first mid-course assessment comprised an expository and a comparison-and-contrast prompt, with each student responding to both prompts at each assessment. Combining the two mid-course assessments allowed for a longitudinal data set and control for the effect of genre (Table 1).

The pre- and post-course data collection stages used identical prompts. Most existing research (Benevento & Storch, 2011; Storch, 2009; Storch & Tapper, 2009) supports the expectation that, given the 12-week interval between the two measurements, students' prior exposure to the prompts would be unlikely to affect their performance (Anderson, 2001).

Data Generation and Analysis

The paragraphs were initially marked using a pre-established analytical rating scale consisting of three criteria: Language (the ability to use accurate grammar, vocabulary, and mechanics), Organisation (use of the rhetorical

Table 1. Pre- and post-course, and multiple mid-course prompts

Assessment	Prompt	Task type
Pre- and post-course prompt 1	Explore one or more reasons why teenagers are hooked on computer games.	Expository
Pre- and post-course prompt 2	Compare and contrast classroom learning with and without the aid of computers.	Comparison and contrast
Mid-course prompt 1	Compare and contrast your learning experience in two different modules at the university.	Comparison and contrast
Mid-course prompt 2	Explore the factors leading to a growing income gap between the wealthy and the poor.	Expository

conventions of paragraph writing, including the clarity of topic and conclusion statements, and the organisation and coherence of ideas), and Content (the ability to generate content that is relevant and fulfills the task requirements by using proper description, evaluation, and argumentation techniques). Students' assigned scores were aggregated into a single mark between 1 and 100, with the Language criterion receiving a 60% overall weighting, and Organisation and Content 20% each. This aggregate mark was then translated into grades F through A+. Language was more heavily weighted because the course is an academic proficiency module that concentrates on raising language accuracy in academic writing, as students placed in the module are found to have language errors that affect comprehensibility. The development and validation of the rating scale and weights was informed by the ongoing writing research and modeling conducted by the Assessment Committee, which uses psychometric modeling to examine tests' validity and reliability (Tan & Wu, 2011).

Because of the large number of scripts (n = 232 X 3 = 696), the marking load was divided among multiple raters. Three experienced raters who had taught various academic writing modules were contracted to mark the pre- and post-course paragraphs. One rater was unable to continue with the post-course marking, and another rater was contracted to replace her. The mid-course assignments were marked by the class tutors. Table 2 shows the raters' background information.

Table 2. Raters' marking information

Rater	Gender	# of pre-course paragraphs marked	# of mid-course paragraphs marked	# of post-course paragraphs marked
1	F	57 SM; 60 CS		58 SM; 59 CS
2	F	57 SM; 60 CS		57 SM; 59 CS
3	F	58 SM; 60 CS		
4	F			58 SM; 59 CS
5	M		58 SM; 36 CS	
6	M		56 SM; 36 CS	
7	F		54 SM; 36 CS	
8	F		28 SM; 36 CS	
Total		232	232	232

Notes. SM = single-marked papers; CS = common scripts which were double-marked by all raters.

To adjust for rater effects such as rater inconsistency and differences in rater severity, and to control for any other potential source of construct-irrelevant variance, I subjected the data to MFRM analysis on the FACETS computer package (Linacre, 2012a). MFRM analysis offers several important benefits over traditional methods of determining data reliability and validity: it adjusts observed marks for raters' severity and other facets affecting it, thereby yielding precise measures; it yields quality control or fit indices including mean square (MNSQ) and standardised fit statistics; and, notably, it helps the researcher achieve invariance, or specific objectivity (Engelhard, 2012). In other words, when data fit the MFRM sufficiently, students' estimated ability measures are

invariant across raters, test items, and tasks (Kubinger, 2005), or are rater-, item-, and task-independent (Engelhard, 2012).

According to Bond and Fox (2007), given a sample size of between 30 and 250, the smallest and largest MNSQ indices that can be confidently described as fitting the MFRM are 0.5 and 1.5, respectively. Erratic fit indices (misfits) indicate the possibility that raters are making errors in rating (Engelhard, 1994; Linacre, 2012b).

Rasch model reliability and separation indices were computed for both students and raters. These indicate "statistically distinguishable levels of performance" for every facet in the analysis (Linacre, 2012b, p. 293). High student reliability and separation statistics correspond to more precise measurement, as do low rater reliability statistics (which indicate that raters do not have significantly different severity measures) (Linacre, 2012b).

MFRM is a technique for establishing the reliability and precision of writing marks. It can detect perturbations and construct-irrelevant variables that have affected students' marks; estimate rater severity and adjust marks for it; and detect bias affecting students' marks [see Engelhard (2012) for a discussion of benefits of MFRM]. Another important advantage of the model is that it can measure and adjust for the effect of any "facet" that would presumably exert an impact on students' marks, such as students' or raters' gender. The present study, however, does not factor in gender or other potential facets because that information was unavailable when the study was conducted (Bond & Fox, 2007).

Modelling the Level and Rate of Growth

To explore the students' development of writing skills over time, and the effect of class tasks on this development (Duncan, Duncan and Strychker[2], 2006), I estimated a curve-of-factors and a conditional curve-of-factors latent growth model (LGM) (McArdle, 1988) on AMOS computer package, Version 21. LGM is a special case of confirmatory factor analysis which has commonly been used in language assessment (e.g., Phakiti, 2008; Purpura, 1999; Sawaki, 2012).

Curve-of-factors LGMs fit growth trajectories into factor scores, which represent the similarity between repeated measures across time points (McArdle, 1988). Curve-of-factors is basically a higher-order model, which subjects all observed variables to higher-level confirmatory factor analysis and generated factor scores. These factor scores are then used for growth modeling (Duncan et al., 2006). Figure 1 presents the present study's curve-of-factors model for writing

2 The precondition of fitting curve-of-factors LGMs is estimating an associative model where each writing skill and fair score has a separate slope and intercept factor (Duncan et al., 2006, p. 67). This model determines "whether the behaviors are related." I estimated the fit of the associative model which was sufficient [$\chi^2 = 55.05$ ($p < 0.05$); χ^2 / df = 2.62; NNFI = 0.927; CFI = 0.944; AIC = 209.05; and RMSEA = 0.119]. It was not discussed further due to space constraints (Appendix 2).

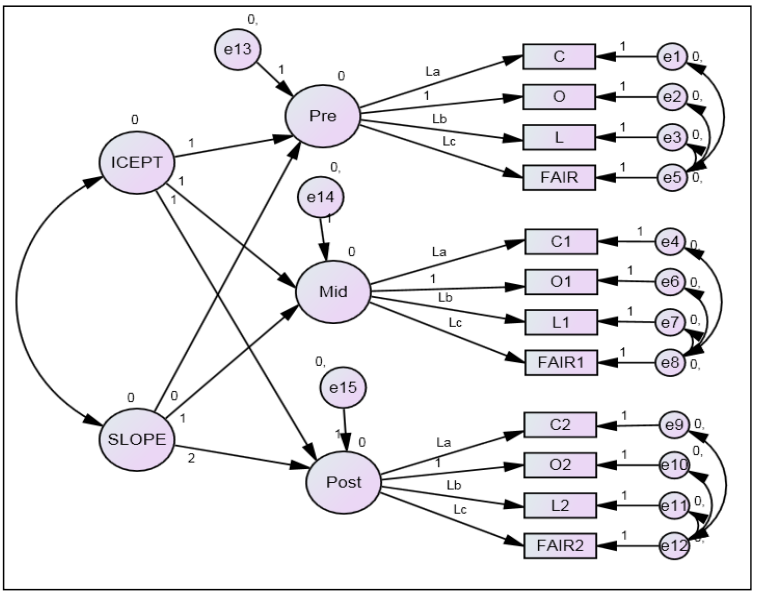

Figure 1. Illustration of the curve-of-factors LGM for writing skills. Legend: ICEPT = Intercept; Pre = Pre-Course; Mid = Mid-Course; Post = Post-Course; C = Content; O = Organization; L = Language; e = error of measurement.

skills and fair scores. This model seeks to answer the first research question, by ascertaining the extent of development in paragraph writing skills over one semester.

As in structural equation modelling (SEM) (e.g., Phakiti, 2008), in this model, large circles represent factors or latent variables; rectangles represent observed variables; and small circles represent measurement error. Bidirectional arrows indicate correlations between factors or items, and unidirectional arrows indicate causality; that is, the latent variables which the test measures are believed to account for observed variance in the items. Because Fair scores are dependent on Content, Language, and Organisation scores, their error terms must be freed (correlated) in the model (Schumacker & Lomax, 2010). Duncan and Duncan (1996, p. 339) state that "[i]n fitting the curve-of-factors LGM, unique factor covariances for each variable over time are allowed to covary, and are included mainly to improve the goodness of fit of the model."

Organisation (Pre_O, Mid_O, and Post_O) was used as the "reference scaling" (for identification purposes) for the lower-order factors (Pre, Mid, and Post), and constrained to unity. In addition, the path coefficients between the lower-order factors and their observed variables were fixed to be equal for Content (L_a), Language (L_b), and Fair scores (L_c) (Duncan et al., 2006). To examine the statistical significance of the coefficients, I computed their p values and critical ratios (path coefficient divided by standard measurement error).

Two primary components of LGMs are slope and intercept. Slope is the rate or pattern of growth or development of the latent trait under assessment, which can be linear or non-linear. Slopes are estimated by constraining the loading coefficients of the time points, as previously discussed. Intercept is the initial level of the latent trait under assessment or the expected score at the initial time point, which is often greater than zero. Intercepts are estimated by constraining their loading coefficient to unity. Since the study examined three time points, trait growth (growth in Content, Language, Organisation, and Fair scores) was modelled to be linear. Therefore, the slopes' higher-order regression weights were constrained to 0, 1, and 2, and the intercept parameters were restricted to unity (Duncan et al., 2006). Finally, to compare the rate and level of growth of the two task types (comparison-and-contrast and cause-and-effect), I regressed intercepts and slope factors on the variable Task.

Figure 2 presents the model which is a conditional curve-of-factors LGM (see Duncan et al., 2006), a curve-of-factors model that incorporates observed variables to predict intercepts and slope factors. In this study, the observed variable that predicts the intercept and slope is the dummy variable Task with two levels: 0 (expository) and 1 (comparison-and-contrast). The conditional LGM model tests whether the variance in the rate of growth (the higher-order slope factor) and initial level of writing skills (intercept) is attributable merely to students' writing skills, or to both students' writing skills and task type.

To assess model fit, I used multiple fit statistics: (a) Chi-square (χ^2) test, an index of the difference between the observed and implied covariance matrices; (b) χ^2/df (normed χ^2), the ratio of χ^2 to the degrees of freedom (*df*), with values below

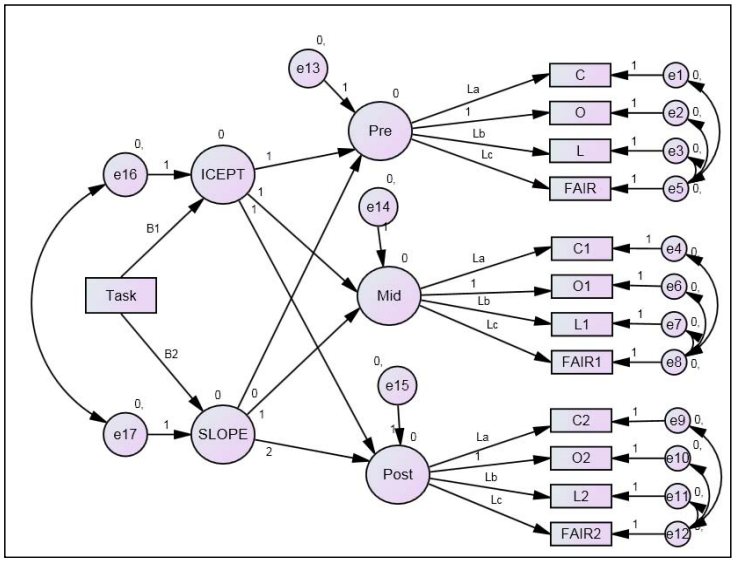

Figure 2. Illustration of the conditional curve-of-factors LGM for writing skills. Legend: ICEPT = Intercept; Pre = Pre-Course; Mid = Mid-Course; Post = Post-Course; C = Content; O = Organization; L = Language; e = error of measurement.

3 indicating good fit; (c) RMSEA (Root Mean Square Error of Approximation), which indicates the fit of the model to the covariance matrix of the population, with values below 0.6 indicating good fit; (d) GFI (Goodness-of-Fit Index), NNFI (Non-Normed Fit Index), and CFI (Comparative Fit Index), which evaluate fit relative to a baseline model and range between 0 and 1, with values greater than 0.90 indicating good fit; and (e) Akaike Information Criterion (AIC) a parsimony index, with smaller values suggesting good fit (Schumacker & Lomax, 2010).

RESULTS

Many-Facet Rasch Model (MFRM)

To examine the reliability of the writing scores, I subjected the data to MFRM analysis. As previously discussed, I initially used raters' performance on a number of paragraphs to construct individual anchored files for the three data sets. Next, I constrained raters' severity level and scale categories to the anchor statistics, and fit the data to the MFRM.

Raters' Severity and Reliability Across Time

Table 3 shows MFRM estimates, after anchoring, of raters' observed and Fair severity measures, standard error, fit statistics, and reliability and separation indices.

The two raters (1 and 2) who were able to participate in both the pre-course and post-course sessions experienced a large shift toward a neutral severity score: from -0.62 and 0.53, respectively, to -0.01 for both.

Another finding is that most raters fit the MFRM, as evidenced by MNSQ fit statistics between 0.50 and 1.5. Rater 5 underfits the model; therefore, her performance data, standardised MFRM residuals, and the fit statistics of those students marked by her were reexamined by the coordinator and researcher carefully. In all, however, MNSQ squares indicate generally high intra-rater consistency. The reliability and separation indices do testify to inter-rater variation in pre- and mid-course assessments, although this variation disappears in the post-course assessment. Raters' severity levels do vary, but the MFRM adjusts for this difference, and the generated fair scores are not contaminated by discrepancies in rater severity (Linacre, 2012b; Engelhard, 2012).

Student Rasch Model Reliability and Fit Statistics

Table 4 shows students' MFRM reliability indices and mean infit and outfit MNSQ statistics. Higher MFRM reliability and separation statistics are desirable in this set of statistics. (Lower indices are desirable for raters, since a test should distinguish between students but not raters.)

Table 3. Raters' severity measures and fit across time

Observed average	Fair score	Severity measure	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	Rater
Pre-course								
21.36	21.17	-0.62	0.04	1.25	2.00	1.10	0.89	1
16.66	16.57	0.53	0.04	0.98	-0.08	1.05	0.41	2
18.57	18.37	0.09	0.04	0.87	-1.06	0.80	-1.81	3
Reliability = .79		Separation = 2.18						
Mid-course								
19.20	19.73	0.37	0.07	1.15	1.0	1.19	1.30	4
20.60	20.22	0.09	0.09	1.88	2.5	1. 48	2.30	5
20.30	20.38	0.01	0.07	0.88	-0.8	0.91	-0.70	6
21.10	21.32	-0.47	0.06	0.68	-2.70	0.71	-2.90	7
Reliability = .94		Separation = 4.11						
Post-course								
18.80	19.02	-0.01	0.05	1.10	0.10	0.88	-1.10	1
18.80	19.02	-0.01	0.05	1.17	0.90	1.21	1.30	2
18.90	18.93	0.03	0.05	0.99	0.00	0.92	-0.4	8
Reliability = 0.00		Separation = 0.00						

Table 4. Rasch model reliability statistics of students' scores after anchoring across time

	Pre-course	Mid-course	Post-course
Rasch model reliability	.95	.90	.91
Rasch model separation	4.19	2.98	3.10
Overall infit MNSQ	1.02	1.03	1.09
Overall outfit MNSQ	0.95	0.93	1.04

Longitudinal reliability indices are at least .90 across time, suggesting negligible measurement error. In addition, separation indices indicate four consistently distinguishable levels in the pre-course writing performance data and three levels in mid- and post-course data, suggesting a gradual decline of student heterogeneity across time.

On average, students' performance data generated at each time point fit the Rasch model well, with MNSQ fit statistics between 0.9 and 1.1 (Linacre, 2012b). However, a close examination of students' individual data revealed several misfitting trends across time, as well as large linearised Rasch model residuals. These papers were re-marked by either the researcher, coordinator, or both. The finalised writing skill and fair scores were used in the LGM analysis.

Exploring the Growth in Students' Writing Performance

To explore the growth in students' writing skills, I initially assessed the fit of the curve-of-factors LGM. Table 5 shows the results.

Table 5. Fit statistics of the estimated latent growth curve models

Model	χ^2	p	df	χ^2/df	NNFI	CFI	AIC	RMSEA
Curve-of-factors LGM	70.49	0.02	48	1.46	0.979	0.984	186.78	0.045
Conditional curve-of-factors LGM	238.61	0.00	58	4.10	0.913	0.944	330.06	0.122

Table 6. Mean and variance of the intercept and slope factors of the curve-of-factor LGM

		Effect	SE	p value
Means				
	Intercept	0.924	0.024	< 0.05
	Slope	0.924	0.204	< 0.05
Variance				
	Intercept	0.115	0.074	< 0.05
	Slope	0.078	0.034	< 0.05
Covariance		0.103	0.046	< 0.05

The curve-of-factors model fit the data reasonably well [$\chi^2 = 70.49$ ($p < 0.01$); $\chi^2/df = 1.46$; NNFI = 0.979; CFI = 0.984; AIC = 154.50; RMSEA = 0.045). In all, the curve-of-factors LGM proved to be a well-fitting model and was later reparameterised to include the effect of task types as the predictor of the higher-order factors (see next page). Due to space constraints, I present only the parameter estimates of the curve-of-factors LGM.

Table 6 shows that the curve-of-factors LGM intercept and slope had significant means and variances (intercept: $M = 0.924$; variance = 0.115; $p < 0.05$; slope: $M = 0.924$; variance = 0.078; $p < 0.05$), suggesting that the initial degree of writing skills was significantly different from zero and that the students' paragraph writing skills improved over time. The intercept and slope factors covaried significantly (0.103, $p < 0.05$) and the differences among students in the higher-order intercept and slope factors were significant at $p < 0.05$, indicating substantial variation.

Table 7 shows the lower-order standardised and non-standardised path coefficients (or factor loadings), squared multiple correlations, standard errors (SE), and *p* values of the curve-of-factors LGM.

The non-standardised paths, which were constrained to be equal, were statistically significant: $L_a = .928$ ($SE = 0.042$; critical ratio = 20.076; $p < 0.001$), $L_b = .733$ ($SE = 0.037$; critical ratio = 19.697; $p < 0.001$), $L_c = .739$ ($SE = 0.086$; critical ratio = 8.576; $p < 0.001$). Organisation (Pre_O, Mid_O, and Post_O) was set as the reference scaling for the lower-order factors and constrained to unity.

Squared multiple correlations indicate the amount of variance in the observed variable which the factors explained, and range between .364 (36.4% of the observed variance of Post_O) to .964 (96.4% of the observed variance of Post_ Fair). Overall, the factors explain a significant amount of variance, indicating precise measurement across time.

Table 7. Parameter estimates, standard error, critical ratio, and p values of the curve-of-factors model

Observed variable	*SE*	Label	Non-standardised path coefficient	Standardised path coefficient	Squared multiple correlation	*CR*	*p* value
Pre_C	0.042	L_a	0.928	.897	.804	22.076	***
Pre_O			1.000	.885	.783		
Pre_L	0.037	L_b	0.733	.787	.619	19.697	***
Pre_Fair	0.086	L_c	0.739	.982	.502	8.576	***
Mid_C	0.042	L_a	0.928	.830	.689	22.076	***
Mid_O			1.000	.844	.713		
Mid_L	0.037	L_b	0.733	.732	.536	19.697	***
Mid_Fair	0.086	L_c	0.739	.733	.537	8.576	***
Post_C	0.042	L_a	0.928	.739	.546	22.076	***
Post_O			1.000	.751	.364		
Post_L	0.037	L_b	0.733	.604	.564	19.697	***
Post_Fair	0.086	L_c	0.739	.709	.964	8.576	***

*** $p < 0.001$

Finally, the conditional curve-of-factors LGM, in which slope and intercept factors are statistically regressed on the exogenous variable Task, was fitted to the data. This model did not fit the data well [$\chi^2 = 238.61$ ($p < 0.01$); $\chi^2/df = 4.10$; NNFI = 0.913; CFI = 0.944; AIC = 330.06; RMSEA = 0.122] and the Task variable did not predict either intercept or slope factors, indicating that the type of task did not strongly influence students' degree and rate of development; the non-standardised Task-slope coefficient was 0.013 (SE = 0.279; $p > 0.05$), and the non-standardised Task-intercept coefficient was -.032 (SE = 0.560; $p > 0.05$). Other parameter estimates of this model are approximately equal to the estimates of the curve-of-factors LGM.

DISCUSSION

This study set out with the main aim of investigating first year university learners' improvement in paragraph writing skills assessed as Content, Organisation, and Language, and overall writing ability over a period of one academic semester. It further explored whether the improvement would differ across tasks.

Reliability of scores. MFRM was applied to provide a reliable measurement dimension whose results would be free of construct-irrelevant variance. This measurement incorporated a rater anchoring method: raters first marked multiple papers, generating estimated severity measures that were then anchored to their severity parameters in the subsequent MFRM analysis.

Misfitting cases and large residuals were subjected to close scrutiny. Raters were in most cases found to have correctly graded multiple cases, though several misfits were flagged which were treated by re-rating the papers (Engelhard, 2012; Linacre, 2012b; VanPatten, 1990).

Research Question One

The results of the curve-of-factors LGM showed that students' writing skills and overall writing ability improved considerably over time. Students' writing skills were assessed on Content, Organisation, and Language, and on overall writing ability, which is the sum of these three components. Overall, this study supports the finding by Storch and Tapper (2009) and Andrew and Romova (2011) that writing programs consisting of grammatical and vocabulary lessons, writing practices, and continuous feedback might result in improvement in L2 writers' improvement over one or two academic semesters. Results are further discussed below.

Content (Task fulfillment). In the planning stage of Hayes and Flower's (1980) model, writers generate ideas based on the task requirements, or the goals set by the writing tasks. The cognitive process during planning can be captured by think-aloud protocols or questionnaires, but the output of this and successive stages is merely captured by examining the texts written by students; examining this output was the major goal of the present study (see, for example, Bae & Lee, 2012). To measure task fulfillment, Content was defined as the relevance of students' paragraphs to the prompt and their ability to analyse, describe, evaluate and argue efficiently. Examining Content required particular care, since little is known about task fulfillment and its growth in L2 writing (Aryadoust, 2012).

As previously discussed, the mean score of the validated scores increased gradually across time. The linear LGM fit the data well with moderate slope variance and mean; students gradually progressed from writing off-track paragraphs loosely connected to the prompts to paragraphs giving fairly effective and relevant responses, indicating that they increasingly attended to the task requirements and applied efficient analysis, description, evaluation, and argumentation techniques (Aryadoust, 2012).

This finding is inconsistent with Bae and Lee's (2012) findings that Korean young learners did not make any significant progress in content during the first few months of instruction, although they made significant progress after the few months. Bae and Lee's finding might be due to the limited world knowledge of students, their weakness in generating ideas pertinent to the task (Abbot et al., 2010), a lack of efficient analysis, evaluation, or argumentation techniques, and a failure to understand the tasks (Hayes, 1996), as well as of "learned writing schema" (Deane, Odendahl, Quinlan, Fowles, Welsh, & Bivens-Tatum, 2008, p. 4). These weaknesses adversely affect the planning stage, and are reflected in poor textual content (Alamargot & Chanquoy, 2001; Hayes, 1996). Due to their age, the students in Bae and Lee's study make progress in terms of text content after a fairly lengthy period of initial instruction. By contrast, the university students in the present study had relatively rich and mature world knowledge, and the tutors' instructions and in-class activities apparently helped them learn

to recognise and fulfill task requirements. Age and maturity seem to be two important differentiating factors between the present study and Bae and Lee's research (Hayes and Flower, 1980).

For students whose text content is slightly irrelevant to the prompt, teachers should stress reviewing the prompt's requirements, pointing students to the statements which are incongruent with the prompt, and guiding students to rectify their mistakes (Hinkel, 2004). Students in this group are close to achieving their level of potential development, and so a minimum amount of feedback and scaffolding should be sufficient for them (Vygotsky, 1978).

Completely irrelevant texts, on the other hand, might signal serious issues surrounding students' comprehension, writing techniques (such as evaluation and argumentation), and task complexity. The class tutors find that these students require substantially more scaffolding, support, and written and oral corrective feedback (Ellis, 2010), as their level of potential development appears to be distant from their actual proficiency level. Mature students who fully miscomprehend the main requirements of the writing prompts may be unaware of their lack of comprehension, and may not be engaging in dialogue with their peers and tutors (Pajares & Valiante, 1997). If they receive teacher scaffolding and constructive peer feedback, they might achieve accurate comprehension of the prompts' requirements and provide relevant responses (Pfingstag, 1998). As previous L2 writing research has not treated task fulfillment in much detail, assessing its development over time should be a priority.

Ability to organise ideas (organisation). The mean index of Organisation marks increased rather linearly through the semester. This finding is consistent with Bae and Lee (2012), who also found a developmental pattern in students' ability to write coherent paragraphs. Bae and Lee (2012, p. 364) argued that the ability to organise thoughts and ideas "is definitely teachable, is learnable, and can be improved." This further resonates with the findings of Benevento and Storch (2011) and Polio et al. (1998) that students' ability to write well-structured paragraphs improved over time, likely due to teacher instruction and feedback.

According to Hayes and Flower (1980), students achieve successful organisation if they have sufficient linguistic resources to translate their thoughts into words. A reasonable conclusion is that the numerous course lessons and exercises helped develop students' skills in organising their thoughts at the paragraph level (Bea & Lee, 2012). A number of studies have found that engaging students in repeated practice is important to achieving coherence and unity in writing (Cumming, 2003; Silva, 1990). Some practitioners have suggested that students be formally helped to transition from a "dependent" stage of writing—in which they require scaffolding, illustration, and example texts—to a stage where they attain independence in applying their language skills successfully (Swales & Feak, 1994).

Linguistic skills (language). Language comprises the ability to use grammar, vocabulary, mechanics, and a range of sentence types; these abilities are activated during the translation stage of writing effectively (Bourdin & Fayol, 1994). Students' Language scores developed linearly across time, indicating a gradual but significant improvement in sentence variety, length, and clarity. This development may be largely attributed to explicit grammar and academic vocabulary lessons, although other elements, such as consultation and feedback, likely had a significant influence (Bae & Lee, 2012; Bitchener et al., 2005; Storch & Tapper, 2000). Bitchener et al. found one-on-one teacher-learner consultation to be the most helpful technique for providing feedback. It is plausible to presume that the systematic teacher-learner consultations throughout an academic semester can help students substantially. Notably, the consultation sessions in the present study were longer than in the study performed by Bitchener et al., suggesting that the students might have benefitted more from the meetings. Previous research has also suggested that grammar and vocabulary are two underlying components of learning (Coxhead, 2012; Nation, 2006) and the key features that determine text quality (Walters & Wolf, 1996). Improving these elements is key to improving students' overall ability to read and write (Kobayashi & Rinnert, 2013; Widdowson, 1990).

On an analytical writing assessment scale, students' total score is the sum of their marks on the writing components, and represents their overall writing ability. LGM analysis showed that a linear growth trend can best explain students' progress in both the components and overall fair scores. A significant advantage of analytical scales is the amount of information concerning students' ability level in multiple skills as well as their overall ability level (Hamp-Lyons & Condon, 2000).

Research Question Two

The study also found that students' progress is task-invariant, meaning that students were able to transfer their skills across writing tasks, and that students' writing quality did not vary significantly across tasks. Carlman (1986) and Skehan (1998) found similarly consistent performance across writing tasks; however, this finding appears to contradict Kuiken and Vedder (2008), as well as Robinson's (2005, p. 29) suggestion that writing performance should vary with "the attentional, memory, reasoning, and other information processing demands imposed by the structure of the task." A possible explanation might be that the tasks in the present study had the same rhetorical structure (expository), and were chosen by content specialists on the basis of their clarity, similar cognitive demands, and relevance to first-year university students' experience (Ellis, 2010). This similarity would render the tasks almost equally demanding, and partial out the effect of any task-related influence on student performance (Robinson, 2005).

This finding is also inconsistent with those of Koda (1993) and Way et al. (2000). This inconsistency might be attributed to the maturity and age of the participants in the present study. Participants in the studies conducted by Koda and Way et al. were young learners whose lexico-grammatical resources and writing techniques (for example, evaluation and argumentation) might have not been as advanced as those of the tertiary students in the present study. The students in this study are required to participate in many writing and reading activities, which might have facilitated the transfer of their writing skills across tasks. This presumption might be assessable by knowledge transfer theory, which seeks to explain the formation and distribution of knowledge (Lobato, 2003). Further research should investigate the transfer of writing techniques across tasks, and could model knowledge transfer mechanisms and students' reservation or willingness to apply acquired linguistic skills to other tasks and contexts.

CONCLUSION AND FUTURE RESEARCH

This study shows that the students made significant progress throughout an L2 academic paragraph writing course, and attributes this progress largely to the structure of the course itself (Ferris & Roberts, 2001). Student writing skills are argued to develop most quickly in educational programs designed to bring multiple facilitative factors (e.g., multiple kinds of lessons and tasks, and multiple ways of engaging with tutors) to bear on various aspects of students' writing (Benevento & Storch, 2011; Cho & MacArthur, 2011; Storch, 2009).

Academic writing curricula would benefit from incorporating researchable elements—techniques and strategies such as explicit grammar and vocabulary lessons—to enhance the skills that are significantly correlated with writing. Student development is best nurtured by regular consultations with tutors, and by establishing a stress-free environment where students perceive that their efforts are appreciated and that they will not be left behind; to this end, continuous assessment (CA) is recommended. Incorporating CA into class curricula allows teachers to monitor students' progress and response to their teaching methodology, and to quickly identify and help students who are not progressing at the necessary pace.

As Slomp (2012) has recently noted, improvement in writing skills depends on a wide range of cognitive, interpersonal, instructional, and institutional factors. This study primarily examined students' growth in paragraph writing skills as a function of instructional factors; future research can further address the effect of students' interpersonal and cognitive attributes on their development of writing skills (Kobayashi & Rinnert, 2013). It is important to note that the resources and the supportive role of the experienced lecturers and raters at CELC have been a major contributing factor in students' development. Examining the role of these elements does merit further investigation.

Another important line of research has focused on the textual features of student texts, as measured by Coh-Metrix (e.g., Graesser, McNamara, & Kulikowich, 2011). Very little research has been conducted on the development of these textual features across time. Research could fruitfully examine the relationship between the development of textual features such as semantic and syntactic complexity indices, and human raters' assessment of development in students' writing skills. Finally, as Canagarajah (2006) has argued, assessing English as lingua franca is becoming an emerging field. Future research can investigate the developmental patterns of students who learn English as lingua franca and compare them against ESL learners. This comparison would permit researchers to (re)assess and validate their theories which are primarily developed in ESL contexts.

ACKNOWLEDGEMENTS

This study was funded by the Centre for English Language Communication of the National University of Singapore. I am thankful to Professors Wu Siew Mei, Susan Tan, and Richard Seow for their support of the project and their valuable comments; Teck Kiang Tan for his comments on the latent growth models; and Professors Irene Tan and Maliga Jeganathan and the raters and experts who assisted me throughout the project. Any errors are solely my responsibility.

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*, 478-508.

Abbott, R., Berninger, V., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology, 102*, 281-298.

Adler-Kassner, L., & O'Neill, P. (2010). *Reframing writing assessment to improve teaching and learning.* Logan: Utah State University Press.

Alamargot, D., & Chanquoy, L. (2001). *Studies in writing series: Vol. 9. Through the models of writing.* Dordrecht, the Netherlands: Kluwer Academic.

Anderson, N. H. (2001). *Empirical direction in design and analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Andrew, M., & Romova, Z. (2011). Teaching and assessing academic writing via the portfolio: benefits for learners of English as an additional language. *Assessing Writing, 16*(2), 112-122.

Aryadoust, V. (2012). How does "sentence structure and vocabulary" function as a scoring criterion alongside others in writing assessment? *Iranian Journal of Language Testing, 2*(1), 28-58.

Bae, J., & Lee, Y.-S. (2012). Evaluating the development of children's writing ability in an EFL context. *Language Assessment Quarterly, 9*(4), 348-374

Benevento, C., & Storch, N. (2011) Investigating writing development in secondary school learners of French. *Assessing Writing, 16*(1) 97-110.

Berninger, V., Abbott, R., Swanson, H. L., Lovitt, D., Trivedi, P., Lin, Gould, L., et al. (2010). Relationship of word- and sentence-level working memory to reading and writing in second, fourth, and sixth grade. *Language, Speech, and Hearing Services in Schools, 41*, 179-193.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL students. *Journal of Second Language Writing, 12*(3), 191-205.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Bourdin, B., & Fayol, M. (1994). Is written language production really more difficult than oral language production? *International Journal of Psychology, 29*, 591-620.

Canagarajah, A. S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an International Language. *Language Assessment Quarterly, 3*, 229-242.

Carlman, N. (1986). Topic differences on writing tests: How much do they matter? *English Quarterly, 19*, 39-47.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267-296.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103*(1), 73.

Coxhead, A. (2012). Academic vocabulary, writing and English for academic purposes: Perspectives from second language learners. *RELC Journal, 43*(1), 137-145.

Crossley, S. A., Weston, J., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*(3) 282-311.

Cumming, A. (2003). Experienced ESL/EFL writing instructors' conceptualizations of their teaching: Curriculum options and implications. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 71-92). New York: Cambridge University Press.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Cognitive models of writing: Writing proficiency as a complex integrated skill*. ETS RR-08-55. New Jersey: Educational Testing Service.

Deden, A., & Carter, V. K. (1996). Using technology to enhance students' skills. In E. Jones (ed.), *Preparing competent college graduates: Setting new and higher expectations for student learning. New directions for higher education* (No. 96, pp. 81-92). San Francisco: Jossey-Bass.

DeKeyser, R. M. (2007). Introduction: Situating the concept of practice. In R. M. DeKeyser (Ed.), *Practice in a second language* (pp. 1-18). New York: Cambridge University Press.

Duncan, S. C., & Duncan, T. E. (1996). A multivariate latent growth curve analysis of adolescent substance use. *Structural Equation Modeling: A Multidisciplinary Journal, 3*(4), 323-347.

Duncan, T. E., Duncan, S. C., & Strychker, L. A. (2006). *An introduction to latent variable growth curve modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Elbow, P. (1981). *Writing with power*. Oxford, England: Oxford University Press.

Ellis, R. (2010). Epilogue: A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition, 32*, 335-349.

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 33*, 93-112.

Engelhard, G., Jr. (2012). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.

Ferris, D. (2003). *Response to student writing. Implications for second language students*. Mahwah, New Jersey: Lawrence Erlbaum.

Ferris, D. R., & Hedgcock, J. S. (1998). *Teaching ESL composition: Purpose, process, and practice*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ferris, D. R., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*, 161-184.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223-234.

Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory, and research*. Cresskill, NJ: Hampton Press.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1-27). Mahwah, NJ: Lawrence Erlbaum.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum.

Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly, 37*, 275-301.

Hinkel, E. (2004). *Teaching academic ESL Writing: Practical techniques in vocabulary and grammar*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hodge, J. G., Whitten, M. E., Horner, W. B., Webb, S. S., & Miller, R. K. (1990). *Harbrace college handbook*. Orland, Florida: Harcourt Brace Jovanovich.

Kent, T. (Ed.). (1999). *Post-process theory: Beyond the writing-process paradigm*. Carbondale, IL: Southern Illinois University Press.

Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science, 3*, 346-370.

Kobayashi, H., & Rinnert, C. (2013). L1/L2/L3 writing development: Longitudinal case study of a Japanese multicompetent writer. *Journal of Second Language Writing, 22*(1), 4-33.

Koda, K. (1993). Task-induced variability in FL composition: Language-specific perspectives. *Foreign Language Annals, 26*, 332-346.

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology, 50*, 569-598.

Kroll, B., & Reid, J. (1994). Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing, 3*(3), 231-255.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model: Some critical suggestions on traditional approaches. *International Journal of Testing, 5*, 377-394.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17*, 48-60.

Le Grange, L., & Reddy, C. (1998). *Continuous assessment: an introduction and guidelines to implementation*. Capetown, South Africa: Juta & Co.

Leki, I. (2006). ‘‘You cannot ignore’’: L2 graduate students’ response to discipline-based written feedback. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing*

(pp. 266-286). Cambridge: Cambridge University Press.

Linacre, J. M. (2012a). *Facets Rasch measurement computer program*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2012b). *A user's guide to FACETS Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Lobato, J. (2003). How design experiments can inform a rethinking of transfer and vice versa. *Educational Researcher, 32*(1), 17-20.

Long, M. (2007). Series editor's preface. In R. M. DeKeyser (Ed.), *Practice in a second language* (p. xi). Cambridge, UK: Cambridge University Press.

MacArthur, C. A., Graham, S., & Schwartz, S. (1991). Knowledge of revision and revising behaviors among students with learning disabilities. *Learning Disability Quarterly, 14*, 61-73.

Manchón, R. M. (Ed.) (2009). *Writing in foreign language contexts: Learning, teaching, and research*. Bristol: Multilingual Matters.

Marr, D., & Cernak, S. (2003). Consistency of handwriting in early elementary students. *American Journal of Occupational Therapy, 57*, 161-167.

Matsuda, P. K. (2003). Process and post-process: A discursive history. *Journal of Second Language Writing, 12*, 65-83.

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In R.B. Cattel& J. Nesselroade (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 561-614). New York: Plennum.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*, 59-82.

Oshima, A., & Hogue, A. (1991). *Writing academic English*. NY: Addison Wesley Longman.

Pajares, F., & Valiante, G. (1997). Influence of writing self-efficacy beliefs on the writing performance of upper elementary students. *Journal of Educational Research, 90*, 353-360.

Pfingstag, N. (1998). The neglected lesson: Teaching L2 writers to decipher writing prompts. *The neglected lesson*, 1-8. Retrieved, from Eric (1225630).

Phakiti, A. (2008). Strategic competence as a fourth-order factor model: A structural equation modeling approach. *Language Assessment Quarterly, 5*(1), 20-42.

Polio, C., Fleck, C., & Leder, N. (1998). ''If I only had more time:'' ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing, 7*, 43-68.

Purpura, J. (1999). *Strategy use and second language test performance: A structural equation modeling approach*. Cambridge: Cambridge University Press.

Robinson, P. (2005). Cognitive complexity and task sequencing: studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43*, 1-32.

Sawaki, Y. (2012). *Structural equation modeling in language assessment*. The Encyclopedia of Applied Linguistics.

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modeling*. New York: Routledge.

Silva, T. (1990). Second language composition instruction: Developments, issues, and directions in ESL. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 11-23). New York: Cambridge University Press.

Skehan, P. (1998). Task-based instruction. *Annual Review of Applied Linguistics, 18*, 268-286.

Slomp, D. (2012). Challenges in assessing the development of writing ability: Theories,

constructs and methods. *Assessing Writing, 17*(4), 81-91.

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing, 18*(2), 103-118.

Storch, N., & Tapper, J. (2000). Discipline specific academic writing: what content teachers comment on. *Higher Education Research and Development, 19*, 337-356.

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes, 8*(3), 207-223.

Swales, J., & Feak, C. (1994). *Academic writing for graduate students*. Ann Arbor: University of Michigan Press.

Tan, S., & Wu, S. M. (2011). *Enhancing test validity: Using the many-facets Rasch model for analyzing rater reliability and rater errors*. Internal report submitted to the Centre for Development of Teaching and Learning of the National University of Singapore.

VanPatten, B. (1990). Attention to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition, 12*, 287-301.

Vygotsky, L. (1978). Interaction between learning and development. In T. M. Cole, *Mind in society* (pp. 79-91). Cambridge, MA: Harvard University Press.

Walters, J., & Wolf, Y. (1996). Language awareness in non-native writers: metalinguistic judgments of need for revision. *Language Awareness 5*(1), 3-25.

Way, D. P., Joiner, E. G., & Seaman, M. A. (2000). Writing in the secondary foreign language classroom: The effects of prompts and tasks on novice learners of French. *The Modern Language Journal, 84*, 171-184.

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

Whitaker, D., Berninger, V., Johnston, J., & Swanson, H. L. (1994). Intraindividual differences in levels of language in intermediate grade writers: Implications for the translating process. *Learning and Individual Differences, 6*, 107-130.

Widdowson, H. G. (1990). *Aspects of language teaching*. Oxford: Oxford University Press.

Appendix 1: The pool of four tasks and the questionnaire administered to the writing experts

Prompt 1: Compare and contrast classroom learning with and without the aid of computers.
Prompt 2: Compare and contrast a means of communication in the past with one at present.
Prompt 3: Explore one or more reasons why teenagers are hooked to computer games.
Prompt 4: What are three important effects on the wide use of mobile phones in society today?

	Item	Strongly disagree	Disagree	Agree	Strongly agree
1	The tasks is sufficiently challenging to discriminate between high-and low-ability students.				
2	The ideas in the tasks are within the experience of the students.				
3	The tasks is culturally ambiguous.				
4	The tasks leads students to construe the topic differently than intended.				
5	The tasks allows for some degree of freedom to show their background knowledge.				
6	The tasks is understandable to low-ability readers.				
7	Students can address the task in the time frame.				
8	The tasks specifies the rhetorical properties of the response (e.g., comparison & contrast).				

Appendix 2: Associative LGM

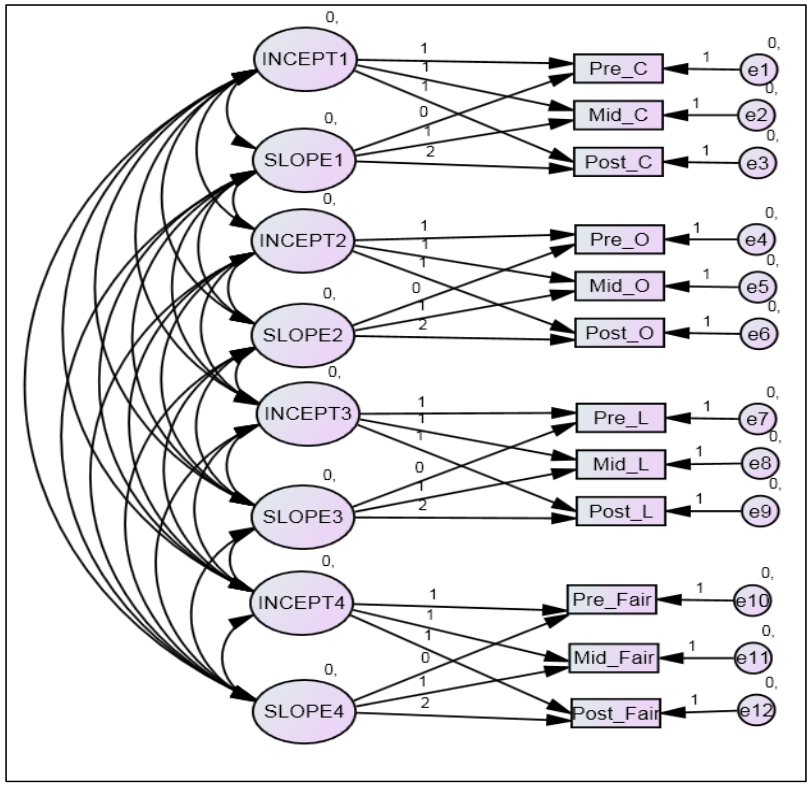

Illustration of the associative LGM for writing skills. Associative LGM allows for estimating multivariate representation of measurements which are regressed on slope and intercept factors. I tested the fit of the model initially by estimating the fit of individual repeated measures (i.e., the univariate LGMs) of the associative and subsequently by assessing the fit of the multivariate LGCM.