The reliability of single task assessment in longitudinal L2 writing research

Wu, May Y.; Steinkrauss, Rasmus; Lowie, Wander

Published in:
Journal of Second Language Writing

DOI:
10.1016/j.jslw.2022.100950

Publication date:
2023

Journal of Second Language Writing 59 (2023) 100950

Contents lists available at ScienceDirect

Journal of Second Language Writing

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

The reliability of single task assessment in longitudinal L2 writing research

May Y. Wu [*,1], Rasmus Steinkrauss [2], Wander Lowie [3]

Department of Applied Linguistics Faculty of Arts, University of Groningen, The Netherlands

ARTICLE INFO

Keywords:
L2 writing assessment
Generalizability theory
CAF
Complex dynamic systems theory
Task topic

ABSTRACT

Single task writing assessments used in longitudinal studies have raised concerns regarding their reliability. By means of Generalizability Theory (GT), this study investigated the reliability of L2 writing assessments scored on different CAF measures, focusing on a) the reliability of single task writing assessments and on the effects of b) task topics and c) task-taking occasions on assessment reliability. We investigated analytic quantitative scores obtained from five CAF measures through a 1-day dataset and a 21-day dataset, consisting of 90 essays from 18 Chinese learners of English who did not follow any formal language instruction during the investigation. The results show that although some CAF scores (e.g., fluency) of single task assessments have distinctly higher reliability than other scores, the general conclusion is that single task assessments are not reliable from a GT perspective. Task topic introduces some score variance to the assessment result, yet this amount of variance differs profoundly between the CAF measures due to the functional variability, which corresponds with Complex Dynamic Systems Theory assumptions suggesting subsystems of an L2 do not develop synchronously. Finally, occasion, i.e., whether two samples were written on the same day or within 21 days, barely introduces score variance.

Writing assessments consisting of one task have been a prevalent instrument for measuring L2 writing and are frequently used in longitudinal studies tracking writing development over time using multiple measurements. Nevertheless, the reliability of such assessments is questioned by Generalizability Theory (GT) research, a statistical method studying the reliability of behavioral assessments (Brennan, 2001; Huang, 2008; Schoonen, 2012), which finds that writing assessments need multiple tasks and raters to be reliable because of the rich variation in assessment characteristics (Schoonen, 2005; Lee et al., 2002; Lee & Kantor, 2005; Lee & Kantor, 2007; Graham et al., 2016). Before transferring the GT findings to the assessments used in longitudinal writing studies, however, it should be noted that GT and longitudinal writing research do not share the same theoretical foundations, especially the way they view changes in assessment performance.

GT, on the one hand, holds that variation in writing performance is a deviation from a person's *universe score*, an expected value that can (theoretically) be obtained as the mean of all admissible (i.e., acceptable to the exam designer) measurements of an assessment (Brennan, 2001). Yet, in writing assessments, various external conditions (e.g., task topic, raters, etc.) and internal causes (e.g., changes in motivation, proficiency, etc.) attract score variance interfering with obtaining the universe score. GT helps decision makers

* Corresponding author.
E-mail address: y.wu@rug.nl (M.Y. Wu).
[1] orcid.org/0000–0001-8163–3608
[2] /orcid.org/0000–0002-3643–8704
[3] orcid.org/0000–0002-2241–0276

https://doi.org/10.1016/j.jslw.2022.100950
Received 30 April 2021; Received in revised form 8 September 2022; Accepted 11 November 2022
Available online 5 December 2022

to locate the external origins (assessment characteristics) of score variance, and design assessments arriving at a more reliable score (i.e., closer to the *universe score*). In longitudinal L2 writing research, the assessment characteristics that threaten the reliability of assessment outcomes are normally confined; studies usually use a series of carefully-restricted single task assessments over time to reduce variability caused by external factors (e.g., text genre) to capture changes demonstrating language development. Among those studies, the ones from a Complex Dynamic System Theory (CDST) perspective highlight *intra-individual* variability in L2 development, which is viewed as noise or random variability in traditional statistics, including GT. In CDST, intra-individual variability is perceived as a hallmark of learners' language development (Lowie, 2017; van Geert & van Dijk, 2002) and as *functional*, signifying the developmental status of an L2 system (van Dijk et al., 2011). More specifically, a high degree of intra-individual variability may signal a period of greater language development, while a low level of variability can indicate that the system is semi-stable, i.e., not developing (Verspoor et al., 2021).

These distinct stances towards variability in performance lead to different assessment regulations in GT and in (dense) longitudinal writing research. To investigate the impact of assessment characteristics on score variation, GT writing samples in one assessment may be from different genres (e.g., argumentative, narrative, etc.), or from the same genre but having different topics (Bouwer et al., 2015), etc. Through these assessment designs, GT has tested that characteristics such as task genre and topic may attract substantial score variance (Benton et al., 1995; Gebril, 2010; Graham et al., 2011; Schoonen, 2005). In contrast, those characteristics are normally restricted in longitudinal writing studies in order to observe intra-individual variability over time. An exception is task topic, which, to avoid repetition effects, is often less restricted in longitudinal studies than genre, timing, and other characteristics. Also, task topics are associated with learners' prior knowledge (both linguistic and content-related) that crucially affects writing performance (Benton et al., 1995), which is not fully controllable by the assessment conditions. Thus, regulations in longitudinal studies often include rater training and rubrics, while setting one task category (e.g., argumentative essays in academic writing) and ensuring task conditions are similar apart from task topic. Some CDST-inspired studies take one step further to enhance the reliability of samples by replacing subjective ratings on the quality of writing with quantitative measures (frequencies, ratios, formulas) on several language traits, typically from the domain of complexity, accuracy and fluency (CAF) measures (e.g., Hokamura, 2018; Hou et al., 2020; Penris & Verspoor, 2017). Such quantitative CAF measures can be consistently quantified and provide "more precise and objective accounts of an L2 learner's level within each (sub-)dimension of proficiency" (Housen & Kuiken, 2009, p. 464). These assessment regulations contribute to better observing developmental variability, rather than excluding changes caused by assessment characteristics entirely as CDST holds that every individual interacts with assessment conditions differently and that the resulting variability is considered meaningful. However, the effectiveness of such single task assessments scored on CAF measures has, to the authors' knowledge, not yet been examined using GT. In other words, it is not clear how reliable the well-regulated single task assessments scored on quantitative CAF would be under the impact of varying task topics.

In addition, the different views on performance change of GT and longitudinal studies also lead to different attitudes towards writing samples collected in a short period, e.g., within three months. While GT research equally accepts samples collected in a time window of several days up to three months as inter-replaceable samples of one assessment (e.g., Gebril, 2009; Graham et al., 2016, Schoonen, 2012), i.e., disregards potential variation associated with time, longitudinal studies regard changes in samples on a weekly or even daily basis as language development, based on the individuals' learning conditions and the stages of language acquisition (Lowie, 2017). These contrasting attitudes towards the performance change in writing samples collected in a small period (e.g., one day, one week, etc.) call out for a discussion on how to view variability between samples collected in a short period, where the tasks are carefully controlled and rated on quantitative CAF measures.

In short, the distinct stances of GT and of longitudinal writing studies raise two needs: to assess the effectiveness of carefully-restricted single task assessments rated on quantitative CAF measures such as used in longitudinal writing research, and to discuss the performance changes shown by same assessments in a short time. This has inspired the current study to compare the theoretical foundations of GT and longitudinal studies, and then investigate the reliability of such assessments. The goals of our study are to (1) study the reliability of single samples scored on different CAF measures and inspect the causes of score variance, (2) investigate the impact of task topic on the reliability of well-regulated single tasks scored on CAF measures, and (3) study the variability brought about by the length of the period within which multiple writing assessments are taken. Finally, this study will report estimations of the number of regulated tasks rated on quantitative CAF measures needed for obtaining a reliable writing score.

1. Background

1.1. Performance change in GT and in dense longitudinal writing research

GT is a statistical method targeting to assess "to what degree the observed measurement, made under a specific set of conditions, will generalize to all other sets of similar conditions" (Bolus et al., 1982, p. 248). It addresses how assessment variables, or characteristics, influence the generalizability of a measurement (Gebril, 2009). GT pinpoints the sources of score variance among multiple observations of an assessment to calculate the reliability of a measurement (e.g., L2 writing scores), and to study how assessment reliability would change when changing the measurement design (Lee & Kantor, 2007). Through a *G(eneralizability) study* on an L2 writing assessment, the variation is distributed to the object of measurement (*person*), *facet*(s), and *residuals*. The object of measurement is normally the person who takes the assessment; facets are assessment characteristics when a researcher investigates their impact on variability; residuals consist of the interaction between the universe score and the facet(s), some (potential) interaction between facets, and effects from unmeasured factors (e.g., task characteristics that are not investigated) or random events (*errors*) (Shavelson & Webb, 1991, p. 9). A G study aims at answering how much variability is introduced by the facet(s). Researchers can also

conduct a *D(ecision) study* based on the results of a G study, to calculate the *generalizability coefficient* ($E\rho^2$) and/or the *dependability coefficient* (Φ) of an assessment, which indicate the reliability of scores obtained from the assessment (cf. Brennan, 2001; O'Brien, 1995; Shavelson & Webb, 1991), and of similar assessments whose designs deviate from the original G study (Brennan, 2001; Bruckner et al., 2006). D studies answer how reliable an assessment is under the impact of facet(s), and how to limit score variance by manipulating the facet(s).

Facets attract the most spotlight in GT research. Empirical G studies on writing assessment have revealed characteristics including task genre, rater training and scoring rubrics to crucially affect score variance, while D studies show the poor reliability of single task assessments due to such score variance (e.g., Gebril, 2009; Graham et al., 2011; Schoonen, 2005; Schoonen, 2012). This variation caused by external factors differs from intra-individual variability over time (van Dijk et al., 2011; de Bot & Verspoor, 2021), which is not discussed in GT. The significance of this functional variability over time, however, is recognized by CDST research, which assumes language is a complex system comprising many continuously varying and interacting subsystems and factors (de Bot et al., 2007). Consequently, the learning trajectories of individuals are assumed to be non-linear and individually owned, as the changing interaction of subsystems over time is not likely to be the same for different persons (Lowie & Verspoor, 2019). One goal of longitudinal CDST research is to show that variability is potentially meaningful for language development, often by a continuous set of highly similar tasks (one at one moment in time) to portray the fluctuation of individuals' writing performances (e.g., Hokamura, 2018; Penris & Verspoor, 2017; Wang & Tao, 2020). Such research assumes that similar tasks avoid the potential (substantial) variation resulting from assessment characteristics and therefore mainly demonstrate functional variability (language development). If similar tasks were able to only leave functional variability as the major cause of score variance and to minimize the effect of external causes, the low reliability of single tasks as attested from a GT perspective would not be an issue in studies researching language development. However, whether this is the case is an open question, and the current study therefore aims at investigating it further. In the next section, this study first looks at the origins of variation in writing assessments, and then discusses if similar single task assessments used in longitudinal writing research may be able to control score variance and achieve high assessment reliability.

1.2. Factors affecting the reliability of single task writing assessments

The investigation of performance-based assessments, not limited to GT studies, shows that raters and tasks comprise two major sources of variation. While score variance related to raters could be eliminated by carefully-constructed scoring rubrics and systematic rater training, the interaction between person and task is very challenging, if not impossible, to control (Brennan, 1996; Brennan, 2000; Miller & Linn, 2000). In a review of earlier performance-based assessments, Dunbar et al. (1991) report that task is a substantial source of score variance in writing assessments, even when tasks are well-constructed. Two suggestions were provided to improve task reliability: one is to include multiple tasks in the assessment, the other is to restrict the task to a narrow domain (Brennan, 2000). The effectiveness of multi-task assessments has been tested by D studies on L1 and L2 writing, which agree that assessments with one task cannot obtain reliable scores (Gebril, 2009; Graham et al., 2016; Huang, 2008; Lee & Kantor, 2007; Schoonen, 2005; Schoonen, 2012).

CDST research assumes that a single assessment may not adequately reflect an overall performance, as many factors (such as task, topic, and especially functional variability) may be involved in each result of a single writing task (de Bot, K., & Verspoor, M. 2021). However, increasing the number of writing tasks is impractical as it means longer assessment time (Miller & Linn, 2000). Not surprisingly, many dense longitudinal writing studies therefore opt for a series of single, well-constructed tasks within a restricted domain to improve reliability. A representative set of restrictions for reducing variation introduced by assessment characteristics as common in longitudinal (CDST) studies, has three features: (1) to avoid inter-rater variability on the quality of the texts, task responses are usually measured through the quantity of some linguistic features, often by using CAF measures; (2) task characteristics, such as text genre, the communicative situation, or writing conditions are contrived to resemble one another in an attempt to minimize score variance; (3) tasks are completed in a period varying between a couple of months (e.g., Hokamura, 2018) to years (e.g., Penris & Verspoor, 2017). The role of the three features for score variance is inspected in turn below.

1.2.1. CAF measurements

Writing studies using quantitative CAF measures exclude the potential influence of raters. In practice, lexical complexity could be represented by type-token ratio, word frequencies, etc. (e.g, Penris & Verspoor, 2017); syntactic complexity is often measured by the ratio of subordinate clauses, the mean length of T-units, etc. (e.g., Kormos, 2011; Bulté & Housen, 2014); accuracy is commonly scored on the number of errors or error-free elements (e.g., Hou et al., 2020; Bulté & Housen, 2014); fluency can be studied via text length or words produced in a time frame (Kim et al., 2021). Such analytical scores on individual traits are suggested to be much less reliable than holistic scores rated on essay scales (Schoonen, 2005, p.15). But this suggestion has not been proven yet. Additionally, the choice of language traits (CAF measures) to be scored also affects assessment reliability. Schoonen (2005) demonstrated the influence of scoring methods and scored language traits. One of the findings is that scores on Language Use (errors) are more generalizable than Content and Organization scores (use of propositions). Admitting that these scoring methods involving raters are not in line with CAF studies using quantitative scores, the outcomes do suggest that different linguistic traits might show a different reliability. If so, the choice of traits to be scored would become a key factor in the reliability of writing scores. These two uncertainties beg a GT analysis on the reliability of independent CAF measures and on the causes of the variation in each measure.

1.2.2. Task characteristics

Apart from the measures used for scoring, some varying characteristics of writing tasks also contribute to score variance, among which task topic is difficult to control. Because prior knowledge of topics is hard to measure or regulate, it consequently begets task

topic as an unavoidable source of score variance. Research in L2 writing assessment has shown the significant impact of task topic on various dimensions of syntactic complexity (Yang et al., 2015). A repeated-task design (as used in, e.g., Larsen-Freeman, 2006) should prevent variation in task characteristics including topic, but learners' performance on the same topic might still vary as their prior knowledge and experience on the topic are developing through time. Some studies, therefore, opt for tasks belonging to a particular category (e.g., argumentative essays in academic writing) to constrain task genre and writing conditions without controlling task topic as, for example, Penris and Verspoor (2017). Considering the impact of task topic, it is not clear to what extent this strategy is able to lower unwanted score variance and restrict the observed variance to functional variability. If ineffective, there would not only be reliability issues for the single task assessments from a GT perspective, but also the question about the causes of CAF score variance from one data point to another in (dense) longitudinal studies: how much variation is caused by a task facet and how much by L2 development?

To answer this question, the present study examines the influence of different aspects of a writing task on assessment variability, taking the division of writing task aspects into task complexity, task conditions, and task difficulty by Robinson (2001) as a starting point—see Fig. 1.

Task complexity mainly regards the cognitive demands (use of cognitive resources) placed on the writer. *Task conditions* concern interactional factors consisting of participation variables, or testing conditions, and participant variables such as gender, while *task difficulty* is personalized and depends on the writer's affective factors and their ability. Among those, affective factors and participant variables, which are also individual differences, are varying over time (cf. Skehan, 1991; Lowie & Verspoor, 2019). Such individual differences always affect assessing L2 writing ability, and it is questionable whether they can or should even be sought to be controlled. Therefore, the present study regards the combination of individual factors (affective, ability, and participant variables) as an integrated object of L2 writing assessment, and these three types of variables will not be restricted. The remaining factors listed can almost all be manipulated to be highly similar, apart from prior knowledge (which is deeply associated with task topic). This brings us back to the concern on the effectiveness of restricting tasks in one category for improving assessment reliability. Even though studies could further mitigate the effect of topics (to some degree) by providing task-takers with the same amount of information on topics, as often done in schools and many standardized language proficiency tests (e.g., IELTS), no study has tested the impact of this method on the reliability of CAF scores and on the functional variability.

This information, however, is critical for longitudinal (CDST) studies, many of which use data that are not specifically collected for the writing study, and therefore the task restrictions are somewhat lax. The Penris and Verspoor (2017) study is a case in point; the researchers used the past university writings of the subject and labeled them as academic writing. Studies without clear criteria on topics (e.g., Spoelman & Verspoor, 2010; Penris & Verspoor, 2017) acknowledge that the rich variability is introduced by various factors, and argue that some of the variability is developmental (de Bot & Verspoor, 2021). A GT analysis, which studies the reliability of single task writing assessments using distinct topics and the impact of topics on CAF measures, would contribute to a better understanding on language development trajectories portrayed by such measures and to the validation of the method used.

1.2.3. The duration of one datapoint

The last aspect that our study targets is the performance change brought by the period of data collection. For longitudinal research, particularly the studies that are from CDST perspectives, assessments containing a single sample each are distributed over a period of investigation lasting from several weeks (e.g., Larsen-Freeman, 2006) to several years (e.g., Penris & Verspoor, 2017). The time between two adjacent assessments, then, ranges from a week (e.g., Wang & Tao, 2020) to months (e.g., Hou et al., 2020; Larsen-Freeman,

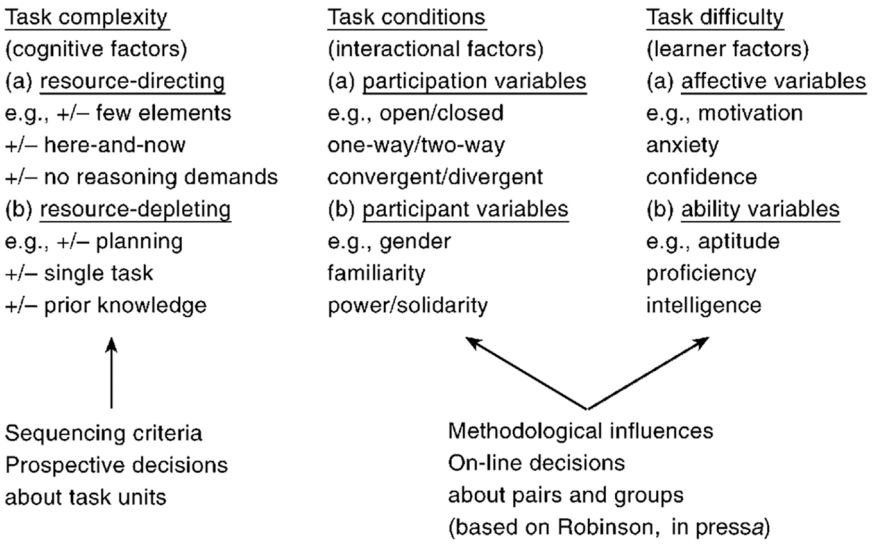

Fig. 1. The complexity, condition and difficulty aspects of a writing task (Robinson, 2001, p. 30).

2006). Considering individuals' learning stages and conditions, many of the CDST writing studies use changes on a weekly basis to build their argumentation, and interpret the weekly changes as (mainly) language development (e.g., Verspoor et al., 2021; Wang & Tao, 2020). This kind of variability, however, is perceived differently in GT research. A GT analysis needs several (group) samples to study the impact of particular facet(s) on assessment reliability. These samples may be collected in only a few days to several months, and the time between two adjacent samples varies from a couple of days (e.g., Gebril, 2009; Graham et al., 2016) to weeks (e.g., Schoonen, 2012). Unlike in CDST studies, the writing samples of a GT analysis are equally accepted as admissible observations of a person's writing performance, assuming learners' writing proficiency (i.e. their universe score) would not essentially change during investigation. In Schoonen (2012), for instance, although the participants were continuously learning English at school throughout the 2-year investigation, samples produced within each data wave of 2–3 months were regarded as interchangeable, since "we have little reason to believe that people's language proficiency changes rapidly from one occasion to the other within a measurement wave" (Schoonen, 2012, p.18).

This contrast between CDST and GT studies raises the question of how one should view the score variation between samples collected in a short period such as several weeks. Do the subsequent single samples reflect some functional variability, such as is assumed in longitudinal (CDST) studies, or do they only reflect score variance caused by assessment characteristics? The latter standpoint seems questionable. Some learners' language ability, or an aspect of it, could vary significantly within two or three months—the period used in Schoonen's (2012) study—due to factors such as intensive language exposure, a pause in L2 learning, affective variables, etc. In addition, compared to assessments taken at one moment in time, a duration of several weeks also heightens the possibility of random events affecting writing performance. It would therefore be enlightening to research the effects of differences in the length of the task-taking occasion on the reliability of writing assessment scores.

1.3. The present study

Through inspecting L2 writing assessments in GT and (dense) longitudinal studies, to date, we can identify three research gaps on the reliability of well-restricted single task assessments scored on CAF measures: (1) the reliability of different quantitative CAF measures, (2) the effect of task topic on the reliability of writing scores, and (3) the impact of differences in the length of the task-taking occasion on the reliability of writing scores. Therefore, we will use writing samples taken from tasks varying dominantly in topic with strict constraints on resource-directing factors, resource-depleting factors, and participation variables (Robinson, 2001) while regarding individual differences (writing competency, affective factors and participant variables in Fig. 1) as part of the universe score, to conduct two sets of GT studies. The first is a pair of G and D studies focusing on task topic, which defines the task facet, to investigate two questions:

RQ 1: What is the reliability of single task writing assessments rated on different CAF measures?

RQ 2: To what extent would task topic affect the reliability of CAF scores?

The second set of GT studies focuses on the effect of the time difference between samples in one assessment, i.e. the occasion facet, to answer the following question:

RQ 3: To what extent will differences in the length of the task-taking occasion affect the reliability of CAF scores?

Since empirical GT studies suggest that writing assessments containing one task are not reliable (Gebril, 2009; Graham et al., 2016; Huang, 2008; Lee & Kantor, 2007; Schoonen, 2005; Schoonen, 2012), and since Schoonen (2005) found analytic scores on single traits to be much less reliable than holistic scores and to show differing generalizability depending on the language trait being scored, for RQ 1 we predict that:

1. Single task writing assessments will yield low generalizability coefficients, and these will differ between different CAF measures.

Secondly, from our discussion on the effect of task topic built on Benton et al. (1995) and Robinson (2001), for RQ2 we predict that:

2. Task topic will result in different amounts of score variance on CAF scores.

For RQ3, no specific hypothesis was set up, considering the contrast between the GT assumption that samples collected in 6–9 weeks will not essentially differ from one another (Schoonen, 2012), and the CDST argument that individual language systems can develop in a time frame of one week or longer depending on the learning conditions and stages (Lowie, 2017).

2. Methodology

To study the reliability of well-restricted single task writing assessments, and how task topic and time difference will affect the reliability of CAF scores,18 Chinese learners of English completed five IELTS Academic essay tasks online within 21 days. Three of the tasks were administered on Day 1, one on Day 11, and the last on Day 21, to be able to study the effect of time difference on tasks. The outcomes of the writing tasks served for two sets of GT analyses: the first set addressed the reliability of single task writing assessments (RQ1) and the role of task topic on the assessment reliability (RQ2); the second investigated if assessments written at one moment in time would essentially differ from the ones written at different moments in time (RQ3). In addition, two C-tests (completed on Day 1

and 21) served to monitor if any changes in learners' general English proficiency would occur (cf. Schoonen, 2012), which helped to determine the steps in the second GT analysis.

2.1. Participants

A total of 18 Chinese learners of English (eight males and ten females) who participated in the study were between 18 and 28 years old (Mean = 21.67). They were students studying or planning to study overseas who signed up to this longitudinal experiment in exchange for detailed feedback on their academic English writing ability. All participants had knowledge of IELTS and TOEFL, and 17 had taken one of the two exams before. Their self-reported English proficiency ranged from CEFR B1 to C1 levels. As the investigation took place during the 2020 COVID-19 lockdown in China, overlapping with the winter break of Chinese academic year, all students except one (who received four hours of instruction during the period of study) were on a break from English courses. Apart from this, no stimuli for potential distinctive changes in overall and general language proficiency were reported.

2.2. Materials

2.2.1. IELTS academic essay task

To constrain task characteristics belonging to Task Complexity and Task Difficulty as categorized by Robinson (2001), five published IELTS Academic essay tasks (Cambridge ESOL, 2013, p. 102; Cambridge University Press, 2016, p. 31; Cambridge University Press, 2017, p. 93; Cambridge University Press, 2018, p. 30, p. 52) differing in topics were completed by the participants on the online platform *Shimo.im* (similar to Google Docs) under supervision, which constrained the tasks to *closed* testing conditions (no access to *external sources*) with *one-way information flow* (only the task prompt was given to the participants and no questions could be asked), rendering them as close to IELTS test conditions as feasible. Participants were requested to plan and perform tasks within 40 minutes, as in IELTS exams, but they could hand in earlier or continue writing after the 40-minutes notification. The actual time in minutes spent on (planning and writing) a task was submitted. All tasks were completed separately at home due to the COVID-19 lock-down. This online testing mode was different from IELTS but consistent for all tasks.

Other characteristics were regulated by the tasks as such. While the IELTS Academic essay task limits the *discourse type* (task genre) to argumentative writing (Abdollahzadeh et al., 2017), the explicit requirement of using *prior knowledge* (Moore & Morton, 2007) in the task stem controlled the type of source used in writing. These *divergent* essay tasks share a deliberately unified task format, consisting of

(1) a controversial statement/opinion
+
(2) the prompt "To what extent do you agree or disagree with this opinion/statement"
+
(3) asking for reasons

regulated the aspects of +/- *few elements* and +/- *single task* in Robinson's (2001) framework, as the assessments utilized a single task each time with the same elements except task topic. No specific restrictions were made for "*here and now*", which meant participants could switch between tenses and use prior knowledge from different fields freely.

2.2.2. C-tests

To scrutinize the general L2 proficiency, the study applied two C-tests selected from a pilot test. In the pilot, C-test 1 and 2 were taken from Reuvers (2019) and Rouhani (2008) respectively, while C-test 3 was constructed with the author of C-test 1. Three female and one male Chinese volunteers performed every C-test online within 30 min independently; all tests were done in one day. The results of C-test 1 and 2 were close, as all individuals had score differences ($M = 4$) under 9 (out of 100). C-test 3 was excluded because of the larger score differences (ranging from 8 to 17).

2.3. Procedure

A digital poster was distributed to WeChat group chats consisting of (Chinese) students intending to study abroad. After interviewing every participant who responded to the poster, data collection started when they indicated they fully understood the experiment procedure and authorized the authors to use the data. Interviews showed that three tasks were the maximum that all participants were willing to perform in 24 hours. Sequentially, a *Shimo.im* document was created for each participant when they chose a starting date of the investigation individually. Since every learner started the investigation at different times, the tasks were assigned in the same order for everyone to prevent any sharing of tasks/task information among the participant group.

On each Day 1 of an investigation period, Task 1, 2, 3 and C-test 1 were completed. To match the number of tasks of Day 1 (the one-moment occasion), we planned for Task 4, 5 and 6 to be completed on Day 11, Day 21 and Day 31 respectively for the second GT analysis. However, five participants indicated that they could not perform Task 6 after the investigation had started, because of unexpected tasks from school programs. Therefore, C-test 2 was administered on Day 21 after Task 5 for all participants, closing the writing experiment. The time differences created by the five valid tasks ranged from 3 to 5 hours (one moment in time) to 21 days.

A paired sample t-test was conducted in parallel with the scoring process (see Data Coding), which found no significant difference

between C-test 1 and 2 (t (16) = 0.892, p = 0.386, 95% CI [− 2.998, 7.351]. The effect size was small, d = 0.126). Combining the insignificant result with the lack of stimuli for a significant proficiency change, the study assumed the participants' overall English proficiency to be stable (cf. Schoonen, 2012). Thus, tasks that could have been completed at any moment of the 21-day period were regarded as acceptable alternatives for the samples we had obtained. As there was no limitation on the number of admissible tasks and occasions, the task and the occasion facets can be regarded as 'unfixed' facets in GT designs (Shavelson & Webb, 1991, p. 65) in the Data Analyses below.

2.4. Data coding

The goal of this study is to examine the reliability of writing assessments rated on independent quantitative CAF measures, and study the impact of task topic and task-taking occasion on them. Ideally, the 90 (18 participants × 5 tasks) texts would be assessed on all existing quantitative measures in each CAF area, which is not realistic due to the scale of this study. Besides, since our focus does not lie on assessing language proficiency comprehensively but on comparing the reliability of individual CAF measures, the study did not need multiple measures for each aspect of CAF. Therefore, we firstly chose four global measures, as listed in Table 1, to give a comprehensive picture of writing performance in each CAF area (Housen et al., 2012), with the area of complexity being assessed on both lexical and syntactic levels as they are commonly studied as separate constructs (Hou et al., 2016).

Lexical complexity was measured by the log frequency of content words (FCW). This complexity measure reflects lexical sophistication and is more reliable than raw frequency of content words (Kormos, 2011). After the first author corrected the misspelt words using the Grammarly (2020) software, FCW was calculated using TAALES version 2.2 (Kyle & Crossley, 2015) based on the BNC written corpus (2007).

Syntactic complexity was computed as the mean length of T-units (MLTU). MLTU is not only the most employed complexity measure in L2 writing, but also an ideal measure for writing samples that are at intermediate level or above (matching to our data), because such data feature clauses and sentences (Norris & Ortega, 2009). The first author counted the number of T-units in each text manually, and subsequently divided the total number of words of a text, provided by Coh-Metrix 3.0 (McNamara et al., 2014), by this number.

The ratio of error-free T-units (EFT) to all T-units, was computed for accuracy. EFT is one of the best indices for measuring L2 development (Larsen-Freeman, 2006) and is commonly applied in longitudinal L2 writing research (e.g., Bulté & Housen, 2014; Casal & Lee, 2019). To compute EFT, the first author made correctness judgments for all texts, accepting a T-unit as error-free when it was without misuses of modifiers, wrong prepositions or pronouns according to the context, wrong word choices impeding understanding, and mistakes in number congruence and in verb conjugation (both tense and third-person forms); also, incomplete T-units were not accepted as correct. Misspelt words were not counted as errors.

Fluency was measured by the rate of production, a valid measure of writing fluency (Kim et al. 2021). It was operationalized as the average word production per minute (AWM), i.e., dividing the total number of words by the minutes spent on a task.

In addition, a specific, fine-grained measure of syntactic complexity was added as a complement to the global measures (see Table 1), contributing to understanding the reliability of CAF measures at different levels. Specifically, mean length of modifiers per noun phrase (MMN) was chosen in this study as learners would produce noun phrases with longer and more complex modifiers when developing their L2 ability (Crossley et al., 2011). MMN was obtained from Coh-Metrix.

2.5. Data analyses

Since tasks were nested into occasions unevenly, two sets of GT analyses using different designs were conducted. The first was a pair of *crossed one-facet* G and D studies that used scores obtained from all five tasks, targeting RQ 1 and 2. The second GT analysis was a *nested two-facet* G study that chose two tasks from Day 1 and two tasks from two different days.

2.5.1. The generalizability analysis of the crossed $p \times t$ design

RQ1 and RQ2 regarding the reliability of a single task writing assessment rated on CAF measures and the role of task facet were addressed by a pair of crossed one-facet *p(erson)* × *t(ask)* G and D studies on all samples since every participant wrote on every topic, using the `gtheory` package (Moore, 2016) in R (R Core Team, 2019) guided by Huebner and Lucht (2019). A generalizability coefficient of 0.8 was set as the desired level (cf. Bottema-Beutel et al., 2014; Schoonen, 2005; Schoonen, 2012). This design regarded samples written at any moment of the 21 days as providing matching indications of learners' writing ability. The $p \times t$ G study detailed the amount of score variance introduced by task topic. Based on the $p \times t$ G study results, the D study with the same design computed

Table 1
The four global and one specific CAF measures.

CAF Area	Measure Type	Measure
Lexical complexity	Global	Log. mean frequency of content word (FCW)
Syntactic complexity	Global	Mean Length of T-units (MLTU)
Syntactic complexity	Specific	Mean number of modifiers per noun-phrase (MMN)
Accuracy	Global	Ratio of error-free T-units (EFT)
Fluency	Global	Average word production per minute (AWM)

$E\rho^2$ and Φ estimates when the number of tasks varies from 1 to 25 in an assessment to estimate the generalizability coefficients when the assessment contained a single task.

2.5.2. The generalizability analysis of the nested $p \times (t: o)$ design

RQ3 regarding the impact of different lengths of task-taking occasions within 21-days was investigated through a nested two-facet $p \times (t : o)$ design. Based on the collected data, the study nested Task 2 and 3 (a 3–5 h interval) into the *one-moment occasion* while Task 1 and 5 (a 3-week interval) were nested into the *multi-moments occasion*, pursuing the largest time difference between occasions to enlarge the chance of detecting impacts of task-tasking occasions. The unfixed $p \times (t : o)$ G study allowed the researchers to analyze the writing scores in R following the same guide (Huebner & Lucht, 2019, p. 10).

3. Results

Table 2 and Fig. 2 present the descriptive statistics of scores obtained from the five CAF measures before zooming to the outcomes of G and D studies.

3.1. The crossed one-facet $p \times t$ design

Table 3 outlines the ratios of score variance introduced by three sources in each measure in the $p \times t$ design. The results suggested that for MLTU, EFT and AWM, *task* only brought 2.9% in MLTU while person contributed more than 55% score variance. For FCW and MMN, *task* introduced the least amount of variance (17.9% and 9.6%) while residuals were the major source of variance (45.3% and 76.6%, respectively).

The D study calculated the $E\rho^2$ and Φ estimates of the $p \times t$ design when the number of tasks ranged from one to 25. As shown in Table 4, when the assessment consisted of one task, the highest coefficient obtained was 0.74 (AWM), followed by 0.65 (EFT). To achieve the $E\rho^2$ of 0.8, the number of tasks needed ranged from two (AWM) to 23 (MMN). To achieve the Φ of 0.8, the measures required the same or more tasks as to achieve an $E\rho^2$ of 0.8.

3.2. The nested two-facet $p \times (t : o)$ design

Table 5 presents that, in the $p \times (t: o)$ design, score variance introduced by *occasion* was NULL in every measure except MLTU (6.8%). The majority of score variance was brought by *person* (in MLTU, EFT and AWM) and *residuals* (FCW and MMN). The contribution of *task nested in occasion* (*t, t: o*) and *person* × *occasion* was limited (0–18.7% and 0–12.6% respectively).

In short, the results distributed not much score variance to *task facet* in the crossed $p \times t$ design, no measure achieved an $E\rho^2$ or Φ coefficient of 0.8 when the assessment contained one task, and the occasion facet in the nested $p \times (t : o)$ design alone introduced no score variance other than the tiny amount in MLTU.

4. Discussion

4.1. The reliability of single task writing assessment rated on CAF measures

The initial concern on the reliability of single task assessments in (dense) longitudinal studies (RQ1) is answered by the crossed $p \times t$ D study. The generalizability coefficient estimates suggest that single task assessments scored on CAF measures are not reliable, and that the number of tasks needed for obtaining a reliable score differs from trait to trait. These results confirm one part of the first prediction, namely that CAF scores obtained from single task assessments differ in reliability, corresponding with Schoonen (2005). However, the part that the reliability of every CAF measure would be low is not confirmed. Some measures, particularly AWM (0.74), are relatively reliable albeit not meeting the threshold of 0.8. To achieve the high threshold set by this study, the number of tasks needed in most CAF measures is in agreement with those studies using holistic scores (Gebril, 2009; Graham et al., 2016; Huang, 2008; Lee & Kantor, 2007; Schoonen, 2005; Schoonen, 2012), with the exception of the specific syntactic complexity measure (MMN) that needs 23 tasks to be reliable. These outcomes correspond with empirical experiments discouraging the use of single task measures (Brennan, 2000; Graham et al., 2016; Huang, 2008; Lee & Kantor, 2007; Miller & Linn, 2000).

Interestingly, global analytic quantitative scores on individual language traits require a similar number of tasks as the holistic scores reported in the literature. This is likely because the well-constructed tasks restricted in IELTS academic writing (a very small

Table 2
Descriptive statistics on the five measures of the five tasks.

Measure	Mean	SD	Min	Max	Median
FCW	-0.66	0.12	-1.08	-0.41	-0.65
MLTU	15.66	3	9.96	24.75	15.6
MMN	0.74	0.22	0.09	1.96	0.72
EFT	39.09	23.24	0	94.44	35
AWM	5.89	1.35	3.34	9.06	5.9

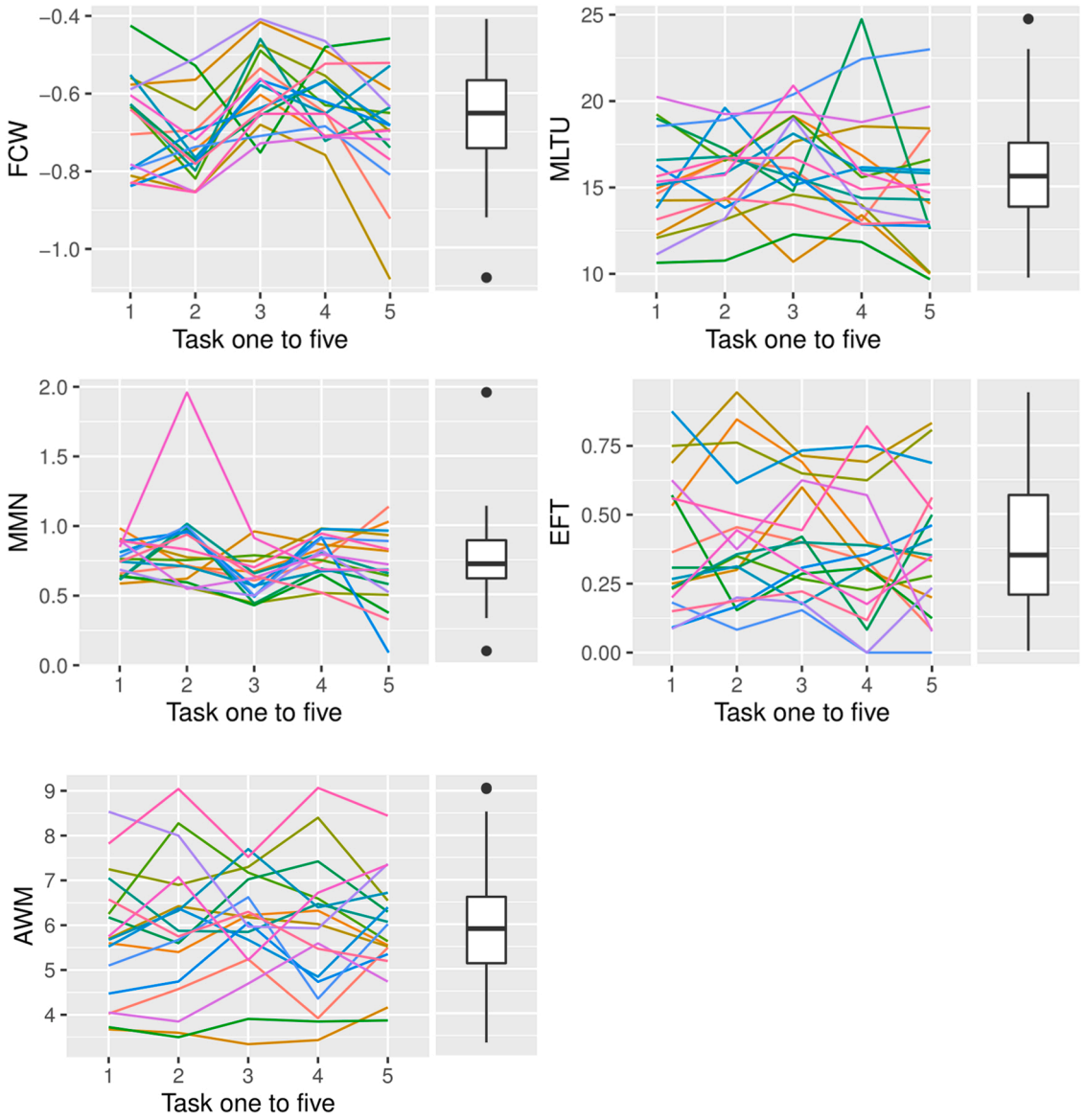

Fig. 2. Boxplots of each measure and the line graphs of participants' performance on task one to five for five CAF measures (90 texts).

Table 3
The percentage of variance introduced by different sources for the five measures in the $p \times t$ design.

Source (facet)	FCW	MLTU	MMN	EFT	AWM
Person (*p*)	36.8	55.2	13.8	65.3	74
Task (*t*)	17.9	2.9	9.6	0	0
$p \times t$ and errors (*residuals*)	45.2	42	76.6	34.7	26

domain) exclude most assessment variables, and the global measures are more representative of each CAF aspect. The contrast between the reliability of global and specific measures calls for a large-scale GT investigation on a wide range of CAF traits rated by computers to frame a systematic generalizability coefficient reference.

4.2. The role of task topic

The first pair of G and D studies uncover that task topic per se contributes little variance in writing scores while having relatively more impact on two of the three complexity measures, corresponding with the crucial effect of task topic found on syntactic complexity

Table 4
The $E\rho^2$ and Φ estimates of the five measures with the single task assessment design and the minimal number of tasks that they needed for reaching the $E\rho^2$ estimates of 0.8.

Measure	Number of tasks = 1		Number of tasks needed
	$E\rho 2$	Φ	$E\rho^2 > 0.8$
FCW	0.4485602	0.3680537	5
MLTU	0.5679115	0.5516696	4
MMN	0.1525434	0.1378254	23
EFT	0.6534084	0.6534084	3
AWM	0.7400124	0.7400124	2

Table 5
The percentage of variance introduced by different sources in the $p \times (t : o)$ design.

Source (facet)	FCW	MLTU	MMN	EFT	AWM
Person (p)	29.2	55.9	8.7	67.1	74.6
Occasion (o)	0	6.8	0	0	0
Task (t), t: o	18.7	0.3	11.2	0	0
$p \times o$	12.6	0	10.5	0.4	0
$p \times (t: o)$ and errors (*residuals*)	39.5	37	69.6	32.5	25.4

(Yang et al., 2015). This score variance attests that task topic is hard to be fully controlled even under such severe assessment restrictions, and the interaction between task and person in residuals is unavoidable (Brennan, 1996; Brennan, 2000; Miller & Linn, 2000). Overall, our restrictions maintain *person*, or intra-individual variability, as the major source of score variance in measures of accuracy, fluency, and global syntactic complexity; in the other two measures, the interaction between person and topic (*residuals*) overtakes the main contribution of score variance.

The unavoidable score variance caused by or related to *person* reflects functional variability in the participants, the amount of which differs in each trait (shown by GT group analyses). This confirms that the complex language system is continually varying due to changes in its constituent elements, subsystems (different language traits), their internal interactions and the interplay with external factors (including task topic), which make variability "an inherent property of any complex, self-organizing system" (Spoelman & Verspoor, 2010, p. 535). Based on these results, future GT studies would be advised to take functional reliability into account as a critical reason for the low reliability of single task assessments.

4.3. The role of occasion

The last interest of this study is the role that a (short) task-taking period plays for the reliability of CAF scores, and this was studied by the $p \times (t: o)$ design. The analysis suggests that the occasion facet alone introduces NULL score variance in the measures apart from MLTU; its interaction with tasks, jointly with the task facet, introduces some variance to FCW and MMN scores; the interaction among *person*, task-taking occasion and task topic attracts a varying amount of score variance in each measure. Overall, the close-to-zero main effect of occasion facet suggests that samples taken from a one-moment occasion were not profoundly different from samples written on a multi-moment occasion in the current dataset. Due to the nested G study design, i.e., participants writing the same tasks in the same order in each occasion (task nested into occasion), the study cannot detect the main effect of task facet, as it is intertwined with the nested effect of task and occasion. What we can suggest here is that the main effect of occasion is subtle, and the nested effect of task and occasion is rather modest.

This addresses the concern about some GT studies that samples collected across several weeks might not be admissible replacements for each other. However, the current results were obtained when there were no stimuli for changes in learners' language proficiency, which could explain the small role occasion plays in the current G-study design. This condition does not apply to all GT studies (e.g., Schoonen, 2012). We therefore suggest that future studies may use *crossed* designs to test the influence of a short task-taking period on language learners who are exposed to strong stimuli as in traditional L2 writing studies (e.g., Schoonen, 2012; Larsen-Freeman, 2006) with more occasions.

4.4. Performance change in single task assessments rated on CAF measures

The results suggest that single task assessments rated on independent CAF measures may not be reliable due to the rich dynamics of a language system. Nevertheless, well-regulated tasks restricted to a small domain can still be used to trace language development since very little score variance is caused by assessment characteristics. The global CAF measures (EFT, and AWM) that reflect mostly intra-individual variability at one moment in time, are suggested to be prioritized in longitudinal writing studies. Less reliable CAF measures (MMN and FCW) also demonstrate functional variability as a combination of variability in *person* and variation introduced by the interaction between persons and assessment characteristics (*residuals*) among assessments.

The functional variability shown in tasks taken within several hours (as the first three tasks in this study) reveals the status of a

language system (e.g., stable or not), which might be called *synchronic* variability. It differs from the functional variability in tasks from time to time, mixing the synchronic variability and developmental variability, which might be called *diachronic* variability. Although the current study did not detect diachronic variability through inspecting the occasion facet (possibly) because of lacking stimuli of language development in the second GT analysis, it would be intriguing for future studies to investigate the two types of functional variability in a longer experimental period (e.g., one year) when L2 learners are under a common language learning or immersion condition.

4.5. Pedagogical implication

The experiment outcomes confirm that L2 writing cannot be reliably measured by one writing task, corresponding with empirical GT research on writing tests (Brennan, 2000; Gebril, 2009; Graham et al., 2016; Huang, 2008; Lee & Kantor, 2007; Schoonen, 2005; Miller & Linn, 2000). Consequently, educators are suggested to use multi-task assessments or multiple samples collected in a short time window to assess L2 learners' writing. When using analytic quantitative CAF measures to evaluate writing, educators are advised to consult global measures that are more reliable (e.g., EFT). This is not only because the three dimensions of CAF represent different aspects of L2 ability, but also because the writing performance in each CAF dimension is not equally stable due to the nature of each trait, the functional variability, and the impact from assessment characteristics such as task topic. It therefore seems necessary to look at multiple CAF dimensions to better capture learners' writing ability.

5. Conclusions and future research

This paper compares the contradicting ways of viewing performance changes in GT research and longitudinal L2 writing studies, leading to different attitudes on the reliability of single task assessments. The outcomes show that writing scores of quantitative CAF measures obtained from a single task are overall unreliable, but the reliability differs among measures. Task topic and task-taking occasion are not similarly influential on writing results, and some variability shown between assessments is functional. This variability might be divided into synchronic and diachronic variability.

These outcomes, however, are based on a rather small experimental group compared to GT studies in writing assessment that include tens (e.g., Bouwer et al., 2015, Schoonen, 2005) and hundreds of participants (e.g., Graham et al., 2016, Huang, 2008). Thus, it would be highly valuable if future studies use large samples to investigate the reliability of possibly many CAF measures and then formulate a generalizability coefficient reference, to investigate the score variance brought by task topic on different linguistic traits, and to address the current limitations. The subtle variance introduced by occasion suggests the existence of synchronic variability across assessments at one moment in time, differing from developmental variability. Yet this finding is based on only two occasions in 21 days, during which the participants were not in a process of active language learning. Therefore, the current results cannot tell us about the stability of writing performance in 2–3 months when learners are continuously learning an L2 as in common L2 writing studies (e.g., Schoonen, 2012), since what happens in between measurements may largely depend on the L2 development within that time frame (Lowie, 2017). Researchers could stretch the investigation period and include more occasions in a 3-month limitation (e.g., back-to-back, two days, one week, etc.) and investigate learners that are exposed to stimuli for potential language development (e.g., an L2 course), as in common L2 writing experiments. Such an investigation would provide a critical insight into the role of occasions. Last but not least, we believe the discovery of the distinct reliability of CAF measures and the inspection of functional variability in contrast to traditional views on performance changes not only add important insights to L2 writing development research, but also have great potential in future investigations.

Conflict of interest

We have no conflicts of interests to disclose.

Data availability

The authors do not have permission to share data.

References

Abdollahzadeh, E., Amini Farsani, M., & Beikmohammadi, M. (2017). Argumentative writing behavior of graduate EFL learners. *Argumentation, 31*(4), 641–661. https://doi.org/10.1007/s10503-016-9415-5

Benton, S. L., Sharp, J. M., Corkill, A. J., Downey, R. G., & Khramtsova, I. (1995). Knowledge, interest, and narrative writing. *Journal of Educational Psychology, 87*(1), 66–79. https://doi.org/10.1037/0022-0663.87.1.66

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning, 32*(2), 245–258. https://doi.org/10.1111/j.1467-1770.1982.tb00970.x

Bottema-Beutel, K., Lloyd, B., Carter, E. W., & Asmus, J. M. (2014). Generalizability and decision studies to inform observational and experimental research in classroom settings. *American Journal on Intellectual and Developmental Disabilities, 119*(6), 589–605. https://doi.org/10.1352/1944-7558-119.6.589

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32*(1), 83–100. https://doi.org/10.1177/0265532214542994

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 19–58). Washington DC: National Center for Education Statistics.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339–353. https://doi.org/10.1177/01466210022031796
Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer,.
Bruckner, C. T., Yoder, P. J., & McWilliam, R. A. (2006). Generalizability and Decision Studies: An Example Using Conversational Language Samples. *Journal of Early Intervention, 28*(2), 139–153. https://doi.org/10.1177/105381510602800205
Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005
Cambridge ESOL. (2013). *Cambridge IELTS 9 Student's Book with Answers: Authentic Examination Papers from Cambridge ESOL*. Cambridge: Cambridge University Press. https://www.worldcat.org/title/904727386.
Cambridge University Press. (2016). *Cambridge IELTS 11 Academic Student's Book with Answers: Authentic Examination Papers*. Cambridge: Cambridge University Press. https://www.worldcat.org/title/1031571249.
Cambridge University Press. (2017C). *Cambridge IELTS 12 Academic Student's Book with Answers: Authentic Examination Papers*. Cambridge: Cambridge Universtiy Press. https://www.worldcat.org/title/999606124.
Cambridge University Press. (2018b). *Cambridge IELTS 13 Academic Student's Book with Answers: Authentic Examination Papers*. Cambridge: Cambridge University Press. https://www.amazon.com/Cambridge-IELTS-Academic-Students-Answers/dp/1108450490/ref=pd_bxgy_img_sccl_1/132-1704683-1881849?pd_rd_w=QUo6z&content-id=amzn1.sym.7f0cf323-50c6-49e3-b3f9-63546bb79c92&pf_rd_p=7f0cf323-50c6-49e3-b3f9-63546bb79c92&pf_rd_r=SW9JEHAHJ3X4017KQ4WM&pd_rd_wg=B2RlY&pd_rd_r=7a7494a9-8d69-49f7-a236-e6828083c3e9&pd_rd_i=1108450490&psc=1.
Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing, 44*, 51–62. https://doi.org/10.1016/j.jslw.2019.03.005
Core Team, R. (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria, (URL) ⟨http://www.R-project.org/⟩.
Crossley, S. A., Weston, J. L., 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. https://doi.org/10.1177/0741088311410188
de Bot, K., & Verspoor, M. (2021). Measuring short-term effects in task repetition and variability. In S. Bátyi, & Z. Lengyel (Eds.), *Bilingualism: Hungarian and Non-Hungarian context, Studies in honor of Judit Navracsics* (pp. 37–49). University of Pannonia Press.
de Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism, 10*(1), 7–21. https://doi.org/10.1017/S1366728906002732
Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289–303. https://doi.org/10.1207/S15324818AME0404_3
Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507–531. https://doi.org/10.1177/0265532209340188
Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15*(2), 100–117. https://doi.org/10.1016/j.asw.2010.05.002
Graham, S., Harris, K., & Hebert, M. (2011). *Informing Writing: The Benefits of Formative Assessment*. Washington: DC: Alliance for Excellence in Education and New York: Carnegie Corporation of New York,.
Graham, S., Hebert, M., Sandbank, M. P., & Harris, K. R. (2016). Assessing the Writing Achievement of Young Struggling Writers: Application of Generalizability Theory. *Learning Disability Quarterly, 39*(2), 72–82. https://doi.org/10.1177/0731948714555019
, 2020Grammarly®, 2020, Grammarly Inc. Retrieved 15 October 2019, from https://www.grammarly.com/.
Hokamura, M. (2018). The Dynamics of Complexity, Accuracy, and Fluency: A Longitudinal Case Study of Japanese Learners' English Writing. *JALT Journal, 40*(1), 23–46. https://doi.org/10.37546/jaltjj40.1-2
Hou, J., Verspoor, M., & Loerts, H. (2016). An exploratory study into the dynamics of Chinese L2 writing development. *Dutch Journal of Applied Linguistics, 5*(1), 65–96. https://doi.org/10.1075/dujal.5.1.04loe
Hou, J., Loerts, H., & Verspoor, M. H. (2020). Coordination of linguistic subsystems as a sign of automatization? In G. G. Fogal, & M. H. Verspoor (Eds.), *Complex Dynamic Systems Theory and L2 Writing Development* (pp. 27–48). John Benjamins Publishing Company. https://doi.org/10.1075/LLLT.54.02HOU.
Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461–473. https://doi.org/10.1093/applin/amp048
Housen, A., Kuiken, F., & Vedder, I. (2012). *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins,.
Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? A generalizability theory approach. *Assessing Writing, 13*, 201–218. https://doi.org/10.1016/j.asw.2008.10.002
Huebner, A., & Lucht, M. (2019). Generalizability theory in R. *Practical Assessment Research, and Evaluation, 24*. https://doi.org/10.7275/5065-gc10
Kim, M., Tian, Y., & Crossley, S. A. (2021). Exploring the relationships among cognitive and linguistic resources, writing processes, and written products in second language writing. *Journal of Second Language Writing, 53*, Article 100824. https://doi.org/10.1016/j.jslw.2021.100824
Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20*(2), 148–161. https://doi.org/10.1016/J.JSLW.2011.02.001
Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly, 49*(4), 757–786. https://doi.org/10.1002/TESQ.194
Larsen-Freeman, D. (2006). The Emergence of Complexity, Fluency, and Accuracy in the Oral and Written Production of Five Chinese Learners of English. *Applied Linguistics, 27*(4), 590–619. https://doi.org/10.1093/applin/aml029
Lee, Y. W., & Kantor, R. (2005). Dependability of new ESL writing test scores: evaluating prototype tasks and alternative rating schemes. *ETS Research Report Series, 2005(1), i–*, 76. https://doi.org/10.1002/j.2333-8504.2005.tb01991.x
Lee, Y. W., & Kantor, R. (2007). Evaluating Prototype Tasks and Alternative Rating Schemes for a New ESL Writing Test through G-theory. *International Journal of Testing, 7*(4), 353–385. https://doi.org/10.1080/15305050701632247
Lee, Y. W., Kantor, R., & Mollaun, P. (2002). Score reliability as an essential prerequisite for validating new writing and speaking tasks for TOEFL. *the annual meeting of Teachers of English to the Speakers of Other Languages (TESOL)*. UT: Salt Lake City,.
Lowie, W. M. (2017). Lost in state space? Methodological considerations in Complex Dynamic Theory approaches to second language development research. In L. Ortega, & Z. Han (Eds.), *Complexity theory and language development: In celebration of Diane Larsen-Freeman* (pp. 123–141). John Benjamins Publishers. https://doi.org/10.1075/lllt.48.07low.
Lowie, W. M., & Verspoor, M. H. (2019). Individual Differences and the Ergodicity Problem. *Language Learning, 69*(S1), 184–206. https://doi.org/10.1111/lang.12324
McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press,.
Miller, D. M., & Linn, R. L. (2000). Validation of Performance-Based Assessments. *Applied Psychological Measurement, 24*(4), 367–378. https://doi.org/10.1177/01466210022031813
Moore, C.T. (2016). gtheory: apply generalizability theory with R. R package version 0.1.2. Retrieved from: https://CRAN.R-project.org/package=gtheory.
Moore, T., & Morton, J. (2007). Authenticity in the IELTS Academic Module Writing test: a comparative study of Task 2 items and university assignments. In L. A. Taylor (Ed.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 197–248). Cambridge: Cambridge University Press.
Norris, J. M., & Ortega, L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics, 30*(4), 555–578. https://doi.org/10.1093/applin/amp044
O'Brien, R. M. (1995). Generalizability coefficients are reliability coefficients. *Quality & Quantity, 29*(4), 421–428. https://doi.org/10.1007/BF01106066

Penris, W., & Verspoor, M. (2017). Academic writing development: a complex, dynamic process (Second language acquisition). In S. Pfenniger, & J. Navracsics (Eds.), *Future Research Directions for Applied Linguistics* (Vol, pp. 215–242). Bristol; Tonawanda, NY; North York: Ontario: Multilingual Matters Ltd.

Reuvers, M. (2019). .. *Teach me to do it on my own: Language learning in Montessori education*. University of Groningen: [Unpublished bachelor thesis],.

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27–57. https://doi.org/10.1093/applin/22.1.27

Rouhani, M. (2008). Another Look at the C-Test: A validation study with Iranian EFL learners. *The Asian EFL Journal, 10*(1), 154–180.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing, 22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Schoonen, R. (2012). The validity and generalizability of writing scores: the effect of rater, task and language. In E. van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring Writing: Recent Insights into Theory, Methodology and Practice* (pp. 1–22). Leiden: Brill. https://doi.org/10.1163/9789004248489_002.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Newbury / London / New Delhi: Sage Publications, Inc,.

Skehan, P. (1991). Individual Differences in Second Language Learning. *Studies in Second Language Acquisition, 13*, 275–298. https://doi.org/10.1017/S0272263100009979

Spoelman, M., & Verspoor, M. (2010). Dynamic Patterns in Development of Accuracy and Complexity: A Longitudinal Case Study in the Acquisition of Finnish. *Applied Linguistics, 31*(4), 2006–2008. https://doi.org/10.1093/applin/amq001

van Dijk, M., Verspoor, M., & Lowie, W. M. (2011). Variability and DST. In M. H. Verspoor, K. de Bot, & W. M. Lowie (Eds.), *A Dynamic Approach to Second Language Development: Methods and techniques* (pp. 55–84). John Benjamins Publishing Company. https://doi.org/10.1075/LLLT.29.04VAN.

van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development, 25*(4), 340–374. https://doi.org/10.1016/S0163-6383(02)00140-6

Verspoor, M., Lowie, W. M., & Wieling, M. (2021). L2 Developmental Measures from a Dynamic Perspective (Cambridge Applied Linguistics). In B. Le Bruyn, & M. Paquot (Eds.), *Learner Corpus Research Meets Second Language Acquisition* (pp. 172–190). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108674577.009 (Cambridge Applied Linguistics).

Verspoor, M., Lowie, W., & de Bot, K. (2021). Variability as normal as apple pie. *Linguistics Vanguard, 7*(s2), 20200034. https://doi.org/10.1515/lingvan-2020-0034

Wang, Y., & Tao, S. (2020). The dynamic co-development of linguistic and discourse-semantic complexity in advanced L2 writing. In G. G. Fogal, & M. H. Verspoor (Eds.), *Complex Dynamic Systems Theory and L2 Writing Development* (pp. 49–78). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.54.03wan.

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

May (Yue) Wu is a PhD candidate in the Department of Applied Linguistics, Faculty of Arts, University of Groningen, Groningen. Her current research is concerned with the variability in the development of L2 production.

Rasmus Steinkrauss is an assistant professor in the Faculty of Arts, University of Groningen. His expertize is in L1 and L2 development and Usage-based/Cognitive Linguistics.

Wander Lowie holds a PhD in Applied Linguistics from the University of Groningen and is chair of Applied Linguistics at this university. He is associate editor of The Modern Language Journal. His main research interest lies in the application of Dynamic Systems Theory to second language development (learning and teaching). He has published more than 50 articles and book chapters and (co-)authored five books in the field of Applied Linguistics.