Khodi *Language Testing in Asia* (2021) 11:30
https://doi.org/10.1186/s40468-021-00134-5

Language Testing in Asia

RESEARCH Open Access

# The affectability of writing assessment scores: a G-theory analysis of rater, task, and scoring method contribution

Ali Khodi

Correspondence: Ali.khodi@ut.ac.ir
University of Neyshabur, Neyshabur, Iran

## Abstract

The present study attempted to to investigate factors which affect EFL writing scores through using generalizability theory (G-theory). To this purpose, one hundred and twenty students participated in one independent and one integrated writing tasks. Proceeding, their performances were scored by six raters: one self-rating, three peers,-rating and two instructors-rating. The main purpose of the sudy was to determine the relative and absolute contributions of different facets such as student, rater, task, method of scoring, and background of education to the validity of writing assessment scores. The results indicated three major sources of variance: (a) the student by task by method of scoring (nested in background of education) interaction (STM:B) with 31.8% contribution to the total variance, (b) the student by rater by task by method of scoring (nested in background of education) interaction (SRTM:B) with 26.5% of contribution to the total variance, and (c) the student by rater by method of scoring (nested in background of education) interaction (SRM:B) with 17.6% of the contribution. With regard to the G-coefficients in G-study (relative G-coefficient ≥ 0.86), it was also found that the result of the assessment was highly valid and reliable. The sources of error variance were detected as the student by rater (nested in background of education) (SR:B) and rater by background of education with 99.2% and 0.8% contribution to the error variance, respectively. Additionally, ten separate G-studies were conducted to investigate the contribution of different facets across rater, task, and methods of scoring as differentiation facet. These studies suggested that peer rating, analytical scoring method, and integrated writing tasks were the most reliable and generalizable designs of the writing assessments. Finally, five decision-making studies (D-studies) in optimization level were conducted and it was indicated that at least four raters (with G-coefficient = 0.80) are necessary for a valid and reliable assessment. Based on these results, to achieve the greatest gain in generalizability, teachers should have their students take two writing assessments and their performance should be rated on at least two scoring methods by at least four raters.

**Keywords:** Classical test theory, Decision-making study, Generalizability theory, Writing assessment

## Introduction

Nowadays, testing and assessment are integrated with the contemporary life and students around the world are assessed continually for two purposes: first, to examine their educational progress and, second, to evaluate the quality of educational systems (Fulcher & Davidson, 2007). A plethora of studies has been done on the significance of language testing and assessment in the realm of education but still "A fundamental concern in the development and use of language tests is to identify potential sources of error in a given measure" (Bachman, 1990, p. 160). In order to increase the precision of measurement, we must explore the potential sources of variance and optimize the design of the assessment in a way that sources of error could have the least possible effect on the process. Hence, validity and reliability as proper indicators of the affectability of the scores were introduced. (Huang, 2012; Lynch & McNamara, 1998; Shavelson, 2004; Crawford et al., 2001).

Writing assessment scores are associated with error for EFL learners more than native speakers (Huang, 2012) Generally, factors affecting students' writing scores can be divided into two types: rater-related and task-related (Huang, 2011). Prior to the discussion of factors affecting writing assessment scores, it should be noted that different cultural backgrounds and linguistic abilities of students make the writing assessment a problematic issue. In fact, variation in measurement may stem from both systematic and unsystematic errors; in this case, if these systematic errors tend to show a positive effect on scores, then the performance would be overestimated, and if they tend to show a negative effect on scores, then the performance would be underestimated. Many research studies have investigated these factors as the major sources of systematic error in ESL writings: raters' linguistic and academic background, raters' tolerance for errors, rater training, and types of writing tasks (Ferris, 1994; Song & Caruso, 1996; Weigle, 1998). In line with what was said about rater-related factors, task-related factors could be pointed out; these include task type, task difficulty, and their interactions (Brown, 1991; Cumming, 1990; Sakyi, 2000). In spite of the fact that the research in this regard (Mehrani & Khodi, 2014) explored factors that influence the writing assessment scores, the relative impacts of these factors are neglected.

Different methodologies are used to assess learners' writing performance: reflective assessment, corrective feedback, formative assessment, and task-based assessments; what is common among them is that they try to look at writing ability as a unidimensional ability . d

In normal circumstances, the main source of variance should be the ability we intend to measure but how is it possible to make sure that the variance does not stem from construct-irrelevant factors? Calculating reliability will be a neat solution for the issue. In fact, classical test theory (CTT) suggests three approaches (e.g., parallel tests or internal consistency) to estimate reliability based on the sources of error in numerous testing situations. Additionally, it should be stated that the ignorance of the interaction between different sources of error brings about the inadequacy of CTT.

Shortcomings of CTT led to a broader model which is called "Generalizability theory" (hereafter G-theory), that is capable of estimating relative effects of different sources of variance. The brilliant feature of the generalizability theory is that it assesses the relative impacts of multiple sources of variation on test reliability simultaneously (Cardinet, Johnson, &Pini, 2010) and rests on ANOVA. In the implementation of G-theory, any

performance of test takers is considered a sample of infinite possible observations under the same condition (Bolus et al., 1982).

A test cannot guarantee that the result will be the same in other similar situations, and since it is impractical to observe students' performance in all conditions, we should search for a practical solution like generalizability theory which provides us with such enabling mechanism. The necessity of such an issue was stated by Fletcher (2006) who explained that a one-dimensional attempt in assessment is imperfect and the reason for this is nothing except considering one source of error and supposing an equal amount of ability for persons and equal level of difficulty for items.

In process of language assessment, different factors are involved such as task types, raters, and scoring methods affecting the accuracy which is a prerequisite for the validity of the assessment.

Defining different facets in generalizability theory increases both reliability and dependability of the results; thus in the present study it is tried to examine a wide range of contribution facets to writing assessment (Khodi, 2015). Moreover, the relative effects of the rater, task, and scoring methods simultaneously were analyzed.

### Research questions

1. What are the relative contributions of persons, raters, tasks, scoring method, and educational background in writing assessment?
2. What are the interactions between different facets in the writing ability assessment?
3. To what extent do changes in the writing test design contribute to its score dependability and optimization of the test design?

## Literature review

The fact that any test performance is affected by factors other than the abilities we want to measure leads to a fundamental concern of identifying potential sources of error in a given measure in the development and use of language tests (Bachman, 1990). By eliminating or minimizing the effect of the construct-irrelevant factors, a necessary condition for reliability could be satisfied. Conceptually, reliability is the number of individuals' real ability in a given measure (Bachman, 1990; Zabihi, Mehrani-Rad, & Khodi, 2019). Previously, it was assumed that the error of measurement was equal to mistake in the measurement procedure (Brennan, 2011) while what constitutes error is a matter of definition by the investigator of the construction to be studied.

Kyburg (1968) stated, "Error is a delicate concept; for if we can call on it at will, or willfully, then it no longer explains anything or accounts for anything. And if we can't call on it when we need it, none of our theories ... will stand up" (p.140). Measurement error does not exist by itself, and it makes sense when two measures of the same variable for the same person do not match; then, for resolving this dilemma, we assume error (Kane, 2010). Generally, the obtained score of measurement is based on a performance (on a test) or personal attribute and when the variability in the observed scores for a person is inconsistent with our expectations about the attribute of interest, errors of measurement are introduced to eliminate this inconsistency (Kane, 2010).

The demand of justification of inconsistencies in measurement and theory of errors was raised because of the stipulation which says the value of construct is invariant across several conditions and observations (Kane, 2010). What is consistent in different conditions is the construct in measurement rendered as the true score, while the observed score fluctuates around the fixed value of the construct because of random errors of measurement.

Since neither true score nor errors could be observed directly, the estimation of a true score will be tricky. One method to overcome the problems associated with the estimation of the true score is to consider a confidence interval using the concept of reliability. In a general sense, "reliability involves quantifying the consistencies and inconsistencies in observed scores" (Brennan, 2011, p.2). When we intend to explore reliability, it is necessary to regard the distinction between observable and unobservable abilities; in fact, all the language abilities which are subject of testing are unobservable, and in order to estimate the true score, a model which suggests a relationship between the observed score and the true score is required. Thus, the following formula was proposed for the estimation of the true score:

$$X = T + E \tag{1}$$

where $X$ stands for the observed score and $T$ stands for the true score while $E$ reports the error score of measurement. As Bachman (1990) stated, despite the fact that this linear equation shows a clear and simple relationship between scores, it ignores other potential sources of error and their interactions; it also treats error variance as homogeneous and random in its origin. These incompetencies in the elaboration of different sources of variation in measurements made the practitioners to introduce a new extensive conceptual framework to address numerous issues in measurement (Brennan, 2011), namely generalizability theory.

For the first time, G-theory was introduced by Cronbach et al. (1963, b) to the field of education and psychology. Briesch et al. (2014) stated that G-theory estimates the proportion of multiple sources of variances within the framework of ANOVA, and in line with this definition, Bachman (1990) asserted that "It constitutes a theory and a set of procedures for specifying and estimating the relative effect of different factors on observed test score." Thus, the means for relating the interpretation of scores to different factors and abilities is available. Many of the demerits, which exist in CTT, have been settled in G-theory and it is capable of measuring a single construct of interest under many different conditions.

### CTT vs. G-theory

The concept of observed score and error of measurement is closely associated with G-theory and its predecessor classical test theory (CTT). With an analytical approach to the CTT, it is understood that an observed score is the decomposition of "true" and a random "error" score (Fedelt & Brennan, 1989), while in reality errors are not the same and they have different contributions and interaction with themselves and with the target ability we want to measure, but manifest simply as one score called "error." CTT is based on correlation and tries to specify the hypothesized relationship between the

observed score and the true score on the test; the more correlation exists, the more reliable the test is (Bachman, 1990).

The canonical equation and mathematical representation of reliability, represented below, is the squared correlation between the true score and the observed score (Brennan, 2011).

$$\rho^2(X,T) = \rho\left(X,X^{'}\right) = \frac{\sigma^2(T)}{\sigma^2(X)} = \frac{\sigma^2(T)}{\sigma^2(T)+\sigma^2(E)} \tag{2}$$

Based on the above-mentioned formula (2), the correlation between the observed score ($X$) and true score ($T$) could be treated as the correlation of scores obtained from two parallel tests ($X$ and$X^{'}$). With more elaboration on Eq. 2, other types of estimate for reliability could be reached with explicit reference to true score variance. Actually, these estimates found the base of their applicability on the fact that covariance between the scores of two parallel tests would be the same as true score variance.

$$\sigma(X,X') = \sigma^2(T) \tag{3}$$

In spite of these statistical postulates, CTT is incapable of finding a solution for its incompetencies. There are a number of shortcomings for CTT: (1) sample dependency, (2) limitation to a specific situation, (3) dependence on the parallel forms, and (4) treating error as homogenous in origin.

In spite of the fact that CTT has the potential to estimate the reliability of a given measure employing test-retest, parallel tests, internal consistency, and rater consistency procedures (Bachman, 1990) which tends to materialize a powerful index of reliability, it was overlooked because of two major shortcomings; first, CTT calculates reliability with one of the above-mentioned procedures (Khodi, Alavi, & Karami, 2021) and finds one source of measurement error which leads to treating other uninvestigated potential sources of error either as an integral part of that source or mistakenly as true score. Second, it considers all errors of measurement as random and is incapable of investigating systematic error as well (Bachman, 1990).

The aforementioned shortcomings of CTT raised the need for a more powerful approach which could detect multiple sources of variability (Huang, 2008). An in-depth explanation of G-theory firstly was composed by Cronbach et al. (1972). G-theory, contrary to CTT which rests on correlation, rests on ANOVA. As Shavelson et al. (1989) stated, "G theory attempts to identify and estimate the magnitude of the potentially important sources of error in a measurement" (p. 923).

In simple representation of the above-mentioned quotation, it could be asserted that G-theory plays a role in measurement quite similar to ANOVA's role in research (Alavi, Karami, & Khodi, 2021). The connection between G-theory and CTT is parallel to the connection which exists between factorial and simple ANOVA. By applying simple ANOVA to a research study, researchers define two main sources of variance, namely, within group and between group variance contributing to error and systematic variances, respectively (Shavelson & Webb, 1991). In contrast, "the researcher acknowledges multiple factors contributing to the total variance in the observations" (Alkharusi, 2012); this scenario happens for both CTT and G-theory in considering the simple source of variance and multiple sources of variance.

With the unique conceptual framework in G-theory, it liberalized CTT by providing the capability of special models and methods which pave the way for investigators to disentangle multiple sources of error (formula 3); those simply contribute to $E$ in classical test theory (Brennan, 2000).

$$X = \mu_p + E_1 + E_2 + .... + E_H \tag{3}$$

where $\mu_p$ stands for universe score or in other words true score and $E$ stands for various sources of error summed up with the true score in measurement.

In sum, advantaged G-theory has several merits over CTT. It empowers researchers to manipulate the data in terms of the number of raters, tasks, items, etc. It provides an estimate of reliability that is comparable across different tests and testing contexts. It investigates the relative effects of sources of variance in a study, which is applicable to the process of increasing reliability by determining the viable scenario of selecting the best option. Finally, it enables the practitioners to have safe and reliable actual test conditions (Bachman, 1990).

### Previous applications of G-theory

The application of G-theory in the field of education and psychology serves to identify practical challenges of real-life assessment situations (Cardinet et al., 2010), and a comprehensive description is provided in Bain and Pini (1996). The first application of G-theory in the field of language testing dates back to 1980s. The first non-technical introduction to G-theory by Bolus et al. (1982) raised many research studies with the same theme. Afterwards, G-theory along with many-facet Rasch analysis was used by Lynch and McNamara (1998); it was used along with factor and cluster analyses by Kunnan (1992) and also by many other figures like Bachman et al. (1995), Brown (1999), Brennan (2001), and Zhang (2006) concerning the investigation of the relative effect of sources of variation. In spite of its applicability and usefulness, the underutilization of G-theory is a great concern and "may be due to an incomplete understanding of the conceptual underpinnings of GT" (Briesch et al., 2014).

### The contribution of G-theory to validity

G-theory is capable of providing information about the construct validity (Shavelson & Webb, 1991) by examining the relative size of variance components (Kraiger & Teachout, 1990; Karami & Khodi, 2021). Construct validity as the most important type of validity (Cronbach & Thorndike, 1971) necessitates both convergent and discriminant types of validity for its establishment. The required condition for convergent validity is met when the obtained scores of the same construct are invariant across different measurement procedures (Marcoulides, 1989); this quality could be assessed by checking the variance component (i.e., person). The large amount of variance shows convergent validity (Kraiger & Teachout, 1990). On the other hand, discriminant validity means dedicating different ranks for persons on measures evaluating different attributes (Kraiger & Teachout, 1990) and it means "the degree of consistency with which raters rate" (Huang, 2012, p.127) students' performance. Besides, achieving fair and valid test results is always a concern in the valuation process.

Since the term "Fairness" has a broad meaning in educational assessment, as a general guide to this concept, the American Educational Research Association, and

National Council on Measurement in Education, Joint Committee on Standards for Educational, and Psychological Testing (US). (1999) suggest the following qualities for a test to be fair: "interpreted as the absence of bias" (Huang, 2012; Karami & Khodi, 2021), and all examinees receive fair behavior during the testing process; the concepts like ethnicity, gender, disability, or other characteristics leave no effect on the outcome of the test; and previous participants had an equal chance in their learning process.

Operationally, if the variability and reliability of scores obtained from different groups tend to be the same, a condition for fairness has been met. One of the approaches to find variability and reliability of testing scores is generalizability theory. The results of generalizability analysis indicate the reliability (G-coefficient) and the relative and absolute variability of different facets in the study. To the extent that patterns match with different groups with no regard to their personal and social characteristics, the test is fair.

The ability to write is always considered as an absolutely essential element of language learning that is not only uneasy to learn but also difficult in its assessment and evaluation (Huang, 2012; Mehrani &Peterson, 2015; Mehrani, 2017). Writing in a new language rather than the mother tongue is a complex and complicated process which has some similarities with writing in L1 (Myles, 2002). Many research studies in the area of English as a second language and English as a foreign language have been suggesting that the assessment like learning is both challenging and complex (Connor-Linton, 1995; Hamp-Lyons, 1991; Huang, 2008; Sakyi, 2000). Typically, writing samples could be analyzed from different perspectives including organizational elements, content, and structure (Aryadoust, 2010). The proposed conceptual and theoretical frameworks in L2 writing suggest a host of factors affecting wiring prompts (Friedrich, 2008).

In deliberation of factors affecting writing prompts, both external and internal factors are important devices of contribution (Ballard & Clancy, 1991; Khodi, 2015; Khodi & Abbasi Sardari, 2015). The underlying factors have not exceeded a handful like cohesion, coherence, and grammar (Ferris, 2002). Writing assessment could be manifested in two different forms including analytical and holistic (sometimes impressionistic); as it was asserted by Weigle (2002) "In analytic writing, scripts are rated on several aspects of writing or criteria rather than given a single score. Therefore, writing samples may be rated on such features as content, organization, cohesion, register, vocabulary, grammar, or mechanics" (p. 114) while on the other hand in a holistic approach one single mark is assigned to writing prompts.

### From nature of writing to its assessment frameworks

There are two different approaches toward writing and its assessment which have been researched to a certain extent, namely holistic and analytical; research on issues of writing from a holistic perspective suggests a high level of variability due to grammatical elements including subordination, sentential connectors, errors, and length (Evola et al., 1980; Homburg, 1984) which leads to the dependency of this rating method to the writing enterprises.

Regardless of probable shortcomings investigated in the holistic perspective, it has found its place and has been advocated by various researchers (Mickan & Slater, 2003; Nakamura, 2004; Wiseman, 2012). One major advantage, asserted by Nakamura (2004),

is that the writing sample can be assessed by multiple raters quickly at the same cost and quality of analytical method.

Wiseman (2012) in his study conducted a many-faceted Rasch measurement and asserted that although different raters in the holistic approach may utilize different criteria, which is a threat to validity and reliability of measurement, a single latent trait of writing ability is evaluated totally similar to the analytical approach. Consensually, the analytical approach of rating has provided much of the fuel for research in writing assessment.

Different frameworks and components have been suggested which affect writing prompts like intelligibility, fluency, overall effectiveness and comprehension (McNamara, 1990, 1996), the structure of the text and organization of material (Mullen, 1977), cohesion and adequacy of vocabulary and punctuation (Weir, 1990), and organization of ideas and language use (Jacobs et al., 1981), and in various studies, the efficacy of the suggested frameworks (e.g., Astika, 1993; Connor, 1991; Harmer, 2004; Mullen, 1977) was the subject of analyses. They found the merits of implementation of analytical frameworks in writing assessments as a sound practice which helps to diagnose the problems associated with writing for English learners.

Research in related areas of writing assessment has shown that many factors affect writing assessment scores. Besides the real performance of students and their writing competence, factors affecting EFL writing assessment generally could be categorized into two different types: task-related and rater-related; task-related factors are defined as the level of difficulty and the type of the tasks. On the other hand, rater-related factors are considered to be different rating methods, rating criteria, raters' academic background, raters' linguistic background, professional experience, tolerance for errors, and rater training (Huang, 2012; Khalilzadeh & Khodi, 2021).

## Method

In the present study, to assure and determine the quality of the instruments used, both expert judgments and exploratory factor analysis (EFA) were used. Through fifteen sessions of essay writing instruction, the students got familiar with the main parts of an essay and its various elements based on their books. After essay writing instruction, the participants received information about how to rate a writing composition based on a TOFEL scoring rubric. To this purpose, they got familiar with different dimensions and components of writing samples and scoring rubric during three sessions of instruction. Finally, they took two writing tests and their performances (in two essays) were rated by several different raters six times .

### Participants

One hundred and twenty students, 90 females and 30 males, participated in the present study. All participants were undergraduate students who were studying either English Literature or Translation Studies at university Based on the ability of the participants from among the initially selected participants, only 100 students put into further analysis in the study. The participants were between 19 and 25 years of age.

## Instruments

### *Checklist*

One of the instruments used in this study was a 16-item checklist which was developed based on different writing assessment frameworks including Banerjee, Franceschina, and Smith (2007) and McNamara (1990) and investigates writing sub-skills. The checklist was used and developed by Aryadoust (2010) to examine the convergence and separability of writing sub-skills: cohesion and coherence, grammar and lexis, and arrangement of ideas.The checklistmainly attempted to assess the tone appropriateness, appropriate exemplification, well arrangement of ideas, full response to the writing tasks, and high level of relevance in the text.

### *Rating rubric*

The rubric introduced to the participants and used to rate independent and integrated tasks, was a modified version of the TOFEL Score Definitions. Based on the definitions in the TOFEL writing rubric, students (i.e., writers of the writing compositions) were rated on a Likert scale of 1 to 5. For instance, a writing sample which addresses the topic and task effectively providing a well-organized and well-developed body with clear and logical exemplification of details achieves 5 based on the rubric.

### *Data collection procedure*

The researcher followed the following steps to collect the data.

**Preliminary stage** Firstly, students who participated in regular essay writing classes and received instructions on how to write five-paragraph essays. During this process, the teacher of the class (who was not the researcher) used different instructional material to teach the essential ideas and techniques of writing a text in English. Each session of instruction was dedicated to one special aspect of writing including the body of the text, organization, content, etc. Following the above-mentioned process, students were suppoed to write essays on the predetermined topics at home.

These samples were evaluated by the instructor and students collaboratively on the cohesion, coherence, content, organization, and logical presentation of ideas with clear exemplification along with the proper use of language and grammar. Besides, for the examination of homework assignments, new comments of students concerning the writing were welcomed. Then, the teacher applied the true and proper suggestions to the writing prompts in the classroom. In order to increase the quality and validity of the work in this step, all participants were trained and got familiar with the rating criteria and rubric.

**Concluding stage** Since the ultimate purpose of the study was to investigate factors that affect EFL writing performance, in the concluding stage, the following steps were taken.

Firstly, all of the participants sat for two parallel writing tests, namely, independent and integrated. In the independent part of the study, students received a TOFEL writing topic and they were supposed to write a five-paragraph essay on that topic. In the writing procedure, students provide good content, organization, and structure for the

forthcoming writing prompts. During writing the essay, no dictionary usage and assistance were allowed to all participants. For the integrated part of the procedure, students read a short reading at first in 10 min as an introductory part of the process, then they were supposed to listen to the audio-track giving explanation and ideas against and for the topic and the students amalgamated their own ideas with those ideas mentioned in the reading and sound clip and work in the five-paragraph essays during 7 min. Then, their performances in the five-paragraph essays were the subjects of analyses and examination. In continuation, each essay was rated by three other students in the study called "peer" and two experienced and trained raters called "instructor." Besides, the students rated their own writing composition in a process called "self-rating."

All the raters received the proper instruction of rating procedure by participating in three sessions of essay writing rating. Moreover, they read and got familiar with a clear rating rubric in the TOFEL writing assessment.

### *Data analysis*

For the univariate analysis, a five facets mixed design generalizability analysis (S: B × R × T × M) with tasks (T) and raters (R) and methods of scoring (M) as random facets was used. To analyze the data, firstly, the distributional characteristics of the total data and each subset (rater, task, and method of scoring) were obtained. Secondly, because each student's writings were rated by six raters across both tasks and scoring methods, they are considered randomly parallel and fully crossed. Based on the suggestions by Brown (1999), ten generalizability analyses were conducted on the data to estimate the relative contribution of the person, task, rater, scoring method, and academic background: one analysis for the total data and nine separate studies for each subset. All the generalizability analyses were conducted using the EduG software (Cardinet et al., 2011). For the decision-making studies (D-studies) based on the results obtained from the G-studies, the researcher changed the design of the facets and applied an increase or decrease to the number of most effective facets, by this means he could promote the reliability and dependability of the results. Based on the capability of the software, some G-facet analyses were conducted to investigate the attributed variances to each level of the facets like rater or task.

## Results

### Observation and estimation designs

Since the generalizability theory rests on ANOVA, the basic assumptions of ANOVA should be checked previously. There are three assumptions in ANOVA, namely independence, normal distribution of raw data, and homogeneity of variances. The general assumptions of ANOVA were satisfied in the present study.

Table 1 shows the observation and estimation designs of the present generalizability study. Five facets feature in this application: student (S), rater (R), task (T), method of scoring (M), and background of education (B).

Based on Table 1, students were nested in their educational background (S:B); for this facet, 90 levels exist and they have been selected from an infinite number of other possible choices and an infinite universe of generalizations. As far as the estimation design

is concerned, we started by considering the facets method and task as fixed. As to students, raters and background of education were considered infinite random facets. Those six raters who participated in the study were labeled by R. In total, two tasks were proposed to the students and their performance was rated in two different methods, namely holistic and analytical. Besides, the educational field and background of the participants were limited to two fields of study. The aim was to see how well the real ability of students could be assessed in their writing compositions across performing different tasks, with different educational backgrounds, and against different raters. The measurement design is therefore S:B/RTM.

In following, the descriptive statistics are presented in Tables 2 and 3 for independent and integrated writing tasks, respectively. It is necessary to say that all of the scores indicated below are mean scores.

### Analysis of major sources of variance

One of the most important outputs of the generalizability study is the table of analysis of variance (ANOVA). Table 4 shows the contribution of the facets involved in the study to both true and error variance; we are interested in finding the contribution of the main facet (i.e.student) as the main source of variance because it is the differentiation facet and here the only facet which contributes to true variance is the student and any other amount of variance for rater, task, and method of scoring and their interaction is attributable to error variance.

The most interesting feature in Table 4 is the contribution of the student by task by method of scoring (nested in the background of education) interaction (STM:B) which was 31.8% of the total variance. The contribution of the interaction between student by rater by task and by method nested in the background of education (SRTM:B) is also noteworthy and more than 26.5% of the total variance because all facets that existed in this study interact with each other simultaneously.

The variance component estimates calculated and represented in Table 4 should have all positive values, while some of them have negative values, since theoretically it is not possible to consider the negative variance for facets, they were replaced by zero in the coming generalizability analyses. The results of these negative values were explained as sampling fluctuations around small component values in our examinations (Cardinet et al., 2011).

For each facet or facet interaction, Table 4 gives the sums of squares, degrees of freedom, and mean squares. More than 50% of the total variance is due to the interaction between student, task, method, and background of education. In addition, three sets of estimated variance components appear: one set relating to an entirely random effects model, one set relating to a mixed model, and a final set of components resulting from the application of Whidbey's correction factor. A random effect component estimates the global importance for the expected sources of variance in other situations in which the experiment will be repeated, while the levels of the observed will be different each time. When a facet is random, it has many levels, and in the present study, just a limited number of its levels are under observation. For this reason, the contributed variance due to these facets is more than other facets. The variance component for the facet S:B was equal to 9.01827 while if this facet was fixed this variance was equal $d$ to

4.59256. The variance for facets like rater, task, and background were −0.92167, −0.30613, and −0.08051, respectively. In an experimental sense, they had their own contribution to the total variance, but conventionally, because these amounts were negative, there were considered as zero. Although the variance for the method was not negative, it was very small and it was considered as zero, too. In addition to facets mentioned above, the contributions of facets RM, RB, TM, MB, RMB, TMB, and RTMB were estimated to be zero. The last column "SE" shows the standard error of measurement attributed to the measurement procedure; the biggest amount of SE belonged to STM:B, SM:B, and ST:B at 4.73811, 2.72418, and 2.38780, respectively. The smallest amounts of SE belonged to method and background at 0.09719 and 0.09678, respectively, then to those fully crossed designs including TM, TB, MB, RTB, and RTMB.

### Main generalizability study

Table 5 shows the facets as well as their interactions as they are separated into differentiation facets and instrumentation facets (the first two columns).

The attributable variance to student as the differentiation facet is equal to 4.59 while a zero contribution for background of education was attributed. In this table, different error variances, based on the ANOVA table, are displayed. The most important and major error variance was 99.2%, in this measurement procedure, and it was due to the interaction between students and rater nested in background of education.

This implies that raters in some special fields of education show bias and severity in their performance and rating procedure of writing tests. In this regard, RB yielded about 8% of error variance. This information allows us to identify those sources of variance that have the greatest negative effect on the precision of the measurements, information that would be useful in a follow-up D-study to show how measurement precision might be increased. In the penultimate two rows of Table 5, the total differentiation variance and the total generalization variance are shown which means how much generalizable our results are, along with the standard deviations. The differentiation variance is about six times larger than the relative error variance (at 4.59256 and 0.76578, respectively). The two indices of generalizability—relative and absolute G-coefficients—are displayed at the bottom of the table; these two coefficients offer a global indication of the reliability of the measurement procedure and provide us information concerning the precision of measurement. The value for the relative G-coefficient was .86 and for the absolute G-coefficient was .85, suggesting that the measurement design and procedure with the current number of students, raters, and tasks were quite

**Table 1** Observation and estimation designs

| Facet | Label | Levels | Univ. |
|---|---|---|---|
| Student:background | S:B | 90 | INF |
| Rater | R | 6 | INF |
| Task | T | 2 | 2 |
| Method of scoring | M | 2 | 2 |
| Background of education | B | 2 | INF |

S:B means that each level of student is applicable to only one level of background of education, *Univ* universe of possible measures

**Table 2** Independent task mean scores

| Rater | Holistic method | Analytical method |
|---|---|---|
| Instructor | 16.25 | 15.64 |
| Peer | 18 | 17.20 |
| Self | 17 | 16.5 |

satisfactory. According to Cardinet et al., 2011 G-coefficients above 0.8 show a satisfactory reliability for the given measurement.

### *Independent generalizability study for rater*

Table 6 presents the results of three independent G-studies for different raters, namely self, peer, and instructor along with an overall G-study for this facet. Based on the results of the overall G-study, when three types of raters are involved in the measurement procedure of writing assessment, G-coefficient (in other words the reliability) is equal to 0.82. The greatest amount of variances stems from the differentiation facet (student) which was 0.463227 and 74%. In this regard, the main sources of variance belonged to the interaction between student-task and task-method with more than 6% contribution to the total variance.

In the second scenario, the only rater to assess writing composition was "self" which suggests a G-coefficient of 0.83 of the measurement procedure. Once more, the main source of variance was students' performance at 71% which was followed by the student by task and task by method interaction with almost 7% contribution to the total variance.

The highest reliability and generalizability belonged to peer assessment with a G-coefficient equal to 0.86. The main source of variance was students' performance. The major difference in this type of assessment was that the contribution of "method of scoring" was less than 1% suggesting that peers do not show bias in their assessments. Ultimately, the contribution which was calculated for students' performance in instructors' assessments was 73% with a G-coefficient of 0.81 suggesting an accurate and valid process of measurement. The contribution of other sources of variance including task and method of scoring was equal to 4%. The least and the most standard errors of measurement were for self-assessment and peer assessment at .039 and .043, respectively.

### Independent generalizability study for method of scoring

The scoring procedure was composed of two types, namely holistic and analytical methods.

**Table 3** Integrated task mean scores

| Rater | Holistic method | Analytical method |
|---|---|---|
| Instructor | 15.5 | 14.85 |
| Peer | 17 | 16.63 |
| Self | 16.25 | 15.66 |

**Table 4** Analysis of variance (ANOVA)

| | | | | Components | | | | |
|---|---|---|---|---|---|---|---|---|
| **Source** | **SS** | **df** | **MS** | **Random** | **Mixed** | **Corrected** | **%** | **SE** |
| S:B | 22864.374 | 178 | 128.45154 | 9.01827 | 4.59256 | 4.59256 | 3.7 | 1.48228 |
| R | 5.660 | 5 | 1.13204 | −0.92167 | −0.04230 | −0.04230 | 0.0 | 0.70181 |
| T | 4.260 | 1 | 4.26021 | −0.30613 | −0.47913 | −0.23956 | 0.0 | 0.25232 |
| M | 55.792 | 1 | 55.79297 | 0.20547 | 0.03248 | 0.01624 | 0.0 | 0.09719 |
| B | 19.244 | 1 | 19.24479 | −0.08051 | −0.05674 | −0.05674 | 0.0 | 0.09678 |
| SR:B | 16224.896 | 890 | 18.23022 | −1.03350 | 4.55756 | 4.55756 | 3.7 | 0.69505 |
| ST:B | 5979.231 | 178 | 33.59119 | −17.87501 | 1.82656 | 1.82656 | 1.5 | 2.38780 |
| SM:B | 27029.170 | 178 | 151.84927 | −10.67799 | 9.02358 | 9.02358 | 7.3 | 2.72418 |
| RT | 4556.687 | 5 | 911.33750 | 1.81791 | 2.43551 | 2.43551 | 2.0 | 1.40238 |
| RM | 11.652 | 5 | 2.33051 | −0.67677 | −0.05917 | −0.05917 | 0.0 | 0.36647 |
| RB | 157.949 | 5 | 31.58996 | −0.00040 | 0.03711 | 0.03711 | 0.0 | 0.08498 |
| TM | 37.012 | 1 | 37.01297 | −0.34599 | −0.34599 | −0.08650 | 0.0 | 0.18932 |
| TB | 162.394 | 1 | 162.39417 | 0.16326 | 0.09807 | 0.09807 | 0.1 | 0.19096 |
| MB | 6.939 | 1 | 6.93922 | −0.05054 | −0.11572 | −0.11572 | 0.0 | 0.14655 |
| SRT:B | 10388.464 | 890 | 11.67243 | −10.60103 | 5.83622 | 5.83622 | 4.7 | 0.82593 |
| SRM:B | 38773.985 | 890 | 43.56628 | 5.34589 | 21.78314 | 21.78314 | 17.6 | 1.29217 |
| STM:B | 47934.225 | 178 | 269.29340 | 39.40315 | 39.40315 | 39.40315 | 31.8 | 4.73811 |
| RTM | 1223.282 | 5 | 244.65651 | 1.23520 | 1.23520 | 1.23520 | 1.0 | 0.72954 |
| RTB | 172.776 | 5 | 34.55525 | 0.18576 | 0.12713 | 0.12713 | 0.1 | 0.12250 |
| RMB | 118.157 | 5 | 23.63155 | −0.05211 | −0.11075 | −0.11075 | 0.0 | 0.09759 |
| TMB | 188.343 | 1 | 188.34359 | −0.13036 | −0.13036 | −0.13036 | 0.0 | 0.29045 |
| SRTM:B | 29258.295 | 890 | 32.87449 | 32.87449 | 32.87449 | 32.87449 | 26.5 | 1.55665 |
| RTMB | 111.600 | 5 | 22.32018 | −0.11727 | −0.11727 | −0.11727 | 0.0 | 0.13369 |
| Total | 205284.39 | 4319 | | | | | 100 | |

Table 7 shows the results of G-studies conducted for investigating sources of variance across the implementation of different scoring procedures. The highest G-coefficient which shows the reliability of results belongs to mixed procedure (i.e., when both analytical and holistic methods are used in parallel form) at G-coefficient = 0.85. The pattern of variance distribution was as follows: students' real competence with 65% variability, rater with more than 21% variability, and task-rater, student-rater, and task with a total 10% variance.

In the holistic procedure, again the main source of variance stems from students' performance (about 65%) and then the major source was the rater with more 15% contribution in measurement. The third rank was the interaction between rater-task in assessment procedure with more than 6% variability. In continuance, the analysis revealed that about 69.5% of variance in the analytical method of scoring stems from students' competence. By taking into account the values of error variances and true variance, it could be understood that the highest G-coefficient was for the analytical procedure at 0.87 and also the least relative error variance was considered for the analytical method (Relative ErrV = 0.001763).

**Table 5** G-study table. Measurement design (SB/TMR)

| Source of variance | Differentiation variance | Source of variance | Relative error variance | % relative | Absolute error variance | % absolute |
|---|---|---|---|---|---|---|
| S:B | 4.59256 | | ..... | | ..... | |
| | ..... | R | ..... | | (0.00000) | 0.0 |
| | ..... | T | ..... | | (0.00000) | 0.0 |
| | ..... | M | ..... | | (0.00000) | 0.0 |
| B | (0.00000) | | ..... | | ..... | |
| | ..... | SR:B | 0.75959 | 99.2 | 0.75959 | 99.2 |
| | ..... | ST:B | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | SM:B | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | RT | ..... | | (0.00000) | 0.0 |
| | ..... | RM | ..... | | (0.00000) | 0.0 |
| | ..... | RB | 0.00619 | 0.8 | 0.00619 | 0.8 |
| | ..... | TM | ..... | | (0.00000) | 0.0 |
| | ..... | TB | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | MB | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | SRT:B | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | SRM:B | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | STM:B | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | RTM | ..... | | (0.00000) | 0.0 |
| | ..... | RTB | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | RMB | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | TMB | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | SRTM:B | (0.00000) | 0.0 | (0.00000) | 0.0 |
| | ..... | RTMB | (0.00000) | 0.0 | (0.00000) | 0.0 |
| Sum of variances | 4.59256 | | 0.76578 | 100% | 0.76578 | 100% |
| Standard deviation | 2.14302 | | Relative SE 0.87509 | | Absolute SE 0.87509 | |
| Coef_G relative | 0.86 | | | | | |
| Coef_G absolute | 0.85 | | | | | |

Variance error of the mean for levels used 0.03283

### *Independent generalizability study for task*

In Table 8, the results of G-studies for both independent and integrated tasks are presented.

In Table 8, three G-studies with regard to different writing tasks were performed. Students' performances were 53%, 56%, and 60% in mixed, integrated, and independent tasks, respectively. In mixed conditions, when both tasks were applied in measurement, a G-coefficient of 0.79 was obtained. This index of reliability was 0.78 in the independent task and 0.80 in the integrated task of writing. The main similarity in these three G-studies was the second-rank source variance which was the student-rater interaction; in the mixed and independent conditions, it was about 25% contribution while it was about 12% in the integrated task. In integrated and independent tasks, the third source

of variability was rater—with 6 to 8% contribution—and in mixed conditions, the student by method interaction was more than 9% contribution.

### Decision-making study (D-study)

Generalizability theory distinguishes a decision (D)-study from a G-study. The G-study is associated with the development of a measurement procedure and the D-study uses information from a G-study to design a measurement that minimizes error for a particular purpose (Shavelson & Webb, 1991). Due to contributions investigated in both G-studies related to task and scoring method, the rater or its related interactions were the main sources of variance.

In Table 9, five D-studies are presented. In the main G-study with 6 raters, the dependability was 0.857. In the analyses done on the data, it became clear that with increasing number of raters from 2 to 12, the generalizability and dependability of assessment scores increase from 0.66 to 0.8. It is recommended that for obtaining the least acceptable amount of reliability in writing assessment, there should be at least 4 raters require in normal circumstances (Coef_G rel. = .79993).

By increasing the number of raters, the relative error variance decreases and it ranges from 0.297 for two raters up to .0382 for 12 raters. Since for the fixed facets it is not possible to conduct an optimization study, the D-studies were limited to optimizing the number of raters because both it is possible and cost-effective.

## Discussion, conclusion, and implications

In order to answer the first research question and investigate the relative contribution of different facets, four separate G-studies were conducted: one main including all participants and three independent G-studies for rater, task, and scoring method.

Based on the results of the mixed design analysis presented in Tables 4 and 5, the students' performances on writing tests were affected by factors like the interactions existing among rater, task, and scoring method. The contribution of facets such as rater, task, and scoring method was zero showing that the instrumental facets by nature are not sources of error but when they interact with each other the dependability of test scores is affected. Based on Table 5, 99.2% of the total error variance was due to student by rater (nested in educational background) interaction, which was a proper and normal amount of contribution. Besides, the interaction of rater by background of education made .08% of the total error variance, which was a normal and expectable amount of variance. This result is in line with the findings of many research studies concerning rater effects (Brown, 1991; Hamp-Lyons, 1996; Kobayashi, 1992; Santos, 1988; Weigle et al., 2003) and concerning task effects (Brown et al., 1991; Hamp-Lyons & Mathias, 1994). Student facet only contributed 3.7% of the total variance in writing assessment scores. Importantly, facets such as rater, task, and method of scoring by themselves did not contribute to error variance and just the interactions between facets (SR:B, RB) generated around 96% of the total variance. Although in cases of interaction between student and only one other facet (e.g., rater, task, method of scoring), an amount of interaction exists (SR:B = 3.7%, ST:B = 1.5%, SM:B = 7.3%) but they are not significant and around 10% of the total variance is due to all these three interactions ( SR:B = 3.7%, ST:B = 1.5%, and SM:B = 7.3%) among facets.

**Table 6** G-study (rater)

| Mixed | | | | Self | | | Peer | | | Instructor | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Source** | **SS** | **VC** | **%** | **SS** | **VC** | **%** | **SS** | **VC** | **%** | **SS** | **VC** | **%** |
| **S** | 23634.03 | 0.4632 | 75.148 | 18934.3 | 0.3616 | 71.857 | 11478.32 | 0.219236 | 81.665 | 16678.713 | 0.318563 | 73.05 |
| **T** | 1565.416 | 0.0306 | 4.977 | 1465.416 | 0.0279 | 5.5613 | 409.416 | 0.00782 | 2.912879 | 943.408 | 0.018019 | 4.13 |
| **M** | 1117.49 | 0.0219 | 3.553 | 1117.49 | 0.0213 | 4.240 | 98.952 | 0.00189 | 0.704016 | 1001.34 | 0.019126 | 4.38 |
| **ST** | 2017.153 | 0.0395 | 6.413 | 2017.153 | 0.0385 | 7.655 | 861.326 | 0.016451 | 6.128091 | 1509.669 | 0.028835 | 6.61 |
| **SM** | 994.642 | 0.0194 | 3.162 | 994.642 | 0.0189 | 3.7747 | 161.358 | 0.003082 | 1.148017 | 1087.752 | 0.020776 | 4.76 |
| **TM** | 2120.93 | 0.0415 | 6.743 | 1820.93 | 0.0347 | 6.9105 | 1046 | 0.019979 | 7.441994 | 1608.36 | 0.03072 | 7.04 |
| **Total** | 31449.66 | | 100 | 26349.93 | | 100 | 14055.37 | | 100 | 22829.24 | | 100 |
| **Coef_G relative** | .82 | | | .83 | | | .86 | | | .81 | | |
| **Coef_G absolute** | .80 | | | .82 | | | .84 | | | .80 | | |
| **Relative ErrV** | .00365 | | | .00321 | | | .00345 | | | .00341 | | |
| **Absolute ErrV** | .00381 | | | .00369 | | | .00358 | | | .00350 | | |
| **Relative SE** | .0413 | | | .0391 | | | .04321 | | | .04190 | | |
| **Absolute SE** | .4329 | | | .0399 | | | .04563 | | | .04368 | | |

**Table 7** G-study (method of scoring)

| Mixed | | | | Holistic | | | Analytical | | |
|---|---|---|---|---|---|---|---|---|---|
| **Source** | **SS** | **VC** | **%** | **SS** | **VC** | **%** | **SS** | **VC** | **%** |
| **S** | 25956.32 | 0.508744 | 65.06321 | 18934.3 | 0.361645 | 64.69993 | 15478.32 | 0.295636 | 69.54869 |
| **T** | 1106.856 | 0.021694 | 2.774492 | 1465.416 | 0.027989 | 5.007437 | 1009.416 | 0.01928 | 4.535606 |
| **R** | 8659.85 | 0.169733 | 21.70715 | 4586.256 | 0.087597 | 15.67158 | 3098.952 | 0.05919 | 13.92451 |
| **ST** | 856.14 | 0.01678 | 2.146037 | 1463.25 | 0.027948 | 5.000035 | 861.326 | 0.016451 | 3.870194 |
| **SR** | 1256.369 | 0.024625 | 3.149268 | 994.642 | 0.018998 | 3.398766 | 161.358 | 0.003082 | 0.725029 |
| **TR** | 2058.47 | 0.040346 | 5.159848 | 1820.93 | 0.03478 | 6.222255 | 1646 | 0.031439 | 7.395967 |
| **Total** | 39894.01 | | 100 | 29264.79 | | 100 | 22255.37 | | 100 |
| **Coef_G relative** | .85 | | | .81 | | | .87 | | |
| **Coef_G absolute** | .84 | | | .80 | | | .86 | | |
| **Relative ErrV** | .00189 | | | .001810 | | | .001763 | | |
| **Absolute ErrV** | .00201 | | | .001873 | | | .001819 | | |
| **Relative SE** | .04690 | | | .05132 | | | .04189 | | |
| **Absolute SE** | .04731 | | | .05322 | | | .04236 | | |

*SS* sum of squares, *df* degree of freedom

In addition, no variance was investigated for the background of education (B) because it was fixed and fixed facets do not contribute to measurement error (Cardinet et al., 2010). As it was mentioned, rater, task, and method of scoring contribute neither to relative nor absolute error of variance while the interaction between student and rater was very high (SR:B = 99.2%).

Table 5 shows the results of overall and main G-study by design of S:B/TMR in which the students nested in background education (S:B) was the differentiation facet and rater, task, and method of scoring were considered as sources of error variance. In sum, the minimum acceptable value of G-coefficients is considered to be 0.80 (Cardinet et al., 2011) and all G-coefficients we have come up with here, 0.86 and 0.85 for relative and absolute G-coefficients, are higher than standards. It shows that the writing assessment procedure enjoys a standard level of reliability. Although none of the facets contributes significantly to the total variance, but the student by rater interaction was the major source of variance.

### Investigation of interactions between facets in writing assessment

Based on Table 4, interactions between different facets do exist and this interaction between student and other facts (e.g., rater, task, method of scoring) was as follows: SR:B = 3.7%, ST:B = 1.5%, and SM:B = 7.3%. This amount of interaction is not too high and around 10% of the total variance is due to all these three interactions because the complexity of writing assessment is accentuated when different factors are involved in the measurement procedure and one to one interaction may not find a great opportunity of effectiveness (Connor-Linton, 1995; Hamp-Lyons, 1991; Huang, 2008; Sakyi, 2000).

The greatest amount of contribution happens when two or three of these facets interact with student (student × rater × T: B = 4.7%, student × rater × method of scoring: B = 17.6%, and student × task × method of scoring: B = 31.8% of total variance) and

**Table 8** G-study (task)

| Mixed | | | | Independent | | | Integrated | | |
|---|---|---|---|---|---|---|---|---|---|
| **Source** | **SS** | **VC** | **%** | **SS** | **VC** | **%** | **SS** | **VC** | **%** |
| **S** | 16932.887 | 0.4578 | 53.1089736 | 17713.1774 | 0.28855254 | 56.0555035 | 17023.5723 | 0.01684049 | 60.68 |
| **M** | 1218.1297 | 0.0195 | 3.82058987 | 769.108311 | 0.02233211 | 2.43393676 | 1385.1477 | 0.00109826 | 4.93 |
| **R** | 1177.4291 | 0.1527 | 3.69293598 | 2060.3667 | 0.06989268 | 6.52028093 | 2308.49738 | 0.00337168 | 8.22 |
| **SM** | 3069.5684 | 0.0151 | 9.6275141 | 1882.70675 | 0.0222994 | 5.95805439 | 2047.14838 | 0.00093711 | 7.29 |
| **SR** | 8049.2835 | 0.0221 | 25.24609 | 7962.3315 | 0.01515829 | 25.1977661 | 3476.96691 | 0.00017556 | 12.39 |
| **MR** | 1435.9906 | 0.0363 | 4.50389641 | 1211.66409 | 0.02775058 | 3.83445834 | 1810.32567 | 0.00179088 | 6.45 |
| **Total** | 31883.288 | 0.45784 | 100 | 31599.355 | | 100 | 28051.658 | | 100 |
| **Coef_G Relative** | .79 | | | .78 | | | .80 | | |
| **Coef_G absolute** | .76 | | | .76 | | | .79 | | |
| **Relative ErrV** | .00561 | | | .00632 | | | .003473 | | |
| **Absolute ErrV** | .00582 | | | .00752 | | | .003625 | | |
| **Relative SE** | .0362 | | | .0259 | | | .03789 | | |
| **Absolute SE** | .04731 | | | .02822 | | | .03936 | | |

**Table 9** D-study (rater)

| | G-study (6 raters) | 2 raters | 4 raters | 8 raters | 10 raters | 12 raters |
|---|---|---|---|---|---|---|
| Observ. | 4320 | 720 | 1440 | 2880 | 3600 | 4320 |
| Coef_G rel. | 0.85709 | 0.66656 | 0.79993 | 0.88884 | 0.90905 | 0.92304 |
| Rounded | 0.86 | 0.67 | 0.80 | 0.89 | 0.91 | 0.92 |
| Coef_G abs. | 0.85709 | 0.66656 | 0.79993 | 0.88884 | 0.90905 | 0.92304 |
| Rounded | 0.86 | 0.67 | 0.80 | 0.89 | 0.91 | 0.92 |
| Rel. Err. Var. | 0.76578 | 2.29733 | 1.14867 | 0.57433 | 0.45947 | 0.38289 |
| Rel. Std. Err. Of M. | 0.87509 | 1.51570 | 1.07176 | 0.75785 | 0.67784 | 0.61878 |
| Abs. Err. Var. | 0.76578 | 2.29733 | 1.14867 | 0.57433 | 0.45947 | 0.38289 |
| Abs. Std. Err. of M. | 0.87509 | 1.51570 | 1.07176 | 0.75785 | 0.67784 | 0.61878 |

finally the interaction between all facets (SRTM:B) was equal to 26.5% of the total variance which is more than one-fourth of the total variance.

In order to investigate contributing facets of writing assessment, three independent G-studies were conducted for rater, task, and method of scoring; the one for rater explains whether raters are consistent in their rating or not. Moreover, when different raters are involved in the rating process, the results are applicable and generalizable to various contexts, a fact which would not happen when just one single rater participates in the rating procedure.

### *Rater G-study*

In Table 6, concerning the impacts of raters, it was investigated that the highest amount of reliability could be achieved when the rater is defined as peer and it had the biggest amount of relative and absolute G-coefficients (relative G-coefficients = 0.86 and absolute G-coefficient = 0.84); this was followed by self-rating and instructor rating, respectively. This fact is counterintuitively suggesting that peer rating enjoys a high level of reliability when the rubric and procedure are defined well as the TOFEL procedure. The mixed design of raters, that is having all self, peer, and instructors rating simultaneously in a study, had an acceptable amount of reliability (relative G-coefficients = 0.82 and absolute G-coefficient = 0.80).

Peer rating with its high reliability could be considered as a helpful method of writing assessment which requires much less effort, energy, and costs if they receive the proper training and clear rubric of assessment (Davidson, 1991; Reid & O'Brien, 1981). Since in the present study, peer raters received suitable training, a satisfying result is achieved. For the instructor ratings and their distance from the peer ratings, it is possible to find the answer in experience and knowledge. The severity, which exists for the instructors toward students, does not exist between peers; the existence of this severity among raters makes their ratings inconsistent and reduces the reliability. In addition, the effects of raters' academic discipline of instructors also affect their ratings, which may lead to a lower amount of reliability in comparison with peer rating (Santos, 1988). Moreover, because of repetition of errors in different writing compositions (due to the large number of writings they rate), the tolerance of instructors reduces and the expectations of instructors to find an error-free composition increases (Casanave & Hubbard, 1992; Janopoulos, 1995; Kobayashi, 1992); this fact may affect the professional

performance of instructors in the rating process. This was the fact suggested as the difference between instructors and other raters (Cumming, 1990; Vaughan, 1991). The point to mention here is that the most significant variance component (around 70% in all G-studies) is due to students' real performance indicating that raters were able to discriminate between students with different abilities properly (Briesch et al., 2014) and this variance component shows that how much the students are spread out by the raters, or in other words how much rating the test can differentiate between examinees. The next highest variance component is the TM, or task and method interaction. The percentage of this variance ranges from 6.74 for mixed design to 7.44 for the peer ratings showing that raters are biased in rating different tasks with different methods of scoring. Raters around 7% may vary in their ratings of different tasks with different methods of scoring. Maybe, it stems from the method of scoring they utilized in rating procedures (because analytical scoring is more exact than a holistic scoring method).

### Method of scoring G-study

In a similar analysis for methods of scoring, again the most significant variance component was attributable to the student facet showing that both methods of scoring were stable in assessing students' writing ability. In addition, the rater in different methods of scoring plays an important role (Casanave & Hubbard, 1992) and the rater may affect students' scores; its contribution ranges from 13.9 to 15.6% of the total variance. Finally, in Table 8, two sets of G-coefficients: relative and absolute, are reported. The analytical method of scoring had the highest amount of G-coefficients (relative G-coefficients = 0.87 and absolute G-coefficient = 0.86) while the relative and absolute G-coefficients for the holistic method of scoring were 0.81 and 0.80, respectively. The finding of this section is in line with other research studies (Evola et al., 1980; Homburg, 1984; Weigle, 2002). Weigle (2002) asserted, "In analytic writing, scripts are rated on several aspects of writing or criteria rather than given a single score. Therefore, writing samples may be rated on such features as content, organization, cohesion, register, vocabulary, grammar, or mechanics" (p. 114) and enjoys a higher amount of reliability. This reliability may stem from the rubric and the framework, which exists in the analytical rating procedure for the practitioners.

#### *Task G-study*

The synthesized results of Table 8 show that the most significant variance is attributable to the student facet, which accounts for about 45 to 60% of the total variance in different studies. Although in the integrated task 60% of the variance was due to student, the interaction between rater and student (SR) is noteworthy which was equal to 12%. Likewise, the same pattern happened to the other two G-studies (mixed and independent) with a level of 25% contribution for this interaction between rater and student. Xi and Mollaun (2006) and In'nami (2016) in separate studies asserted that student by task interaction (ST) would be larger when both independent and integrated tasks are used. Maybe this result is due to the broader construct definition, which happens with the involvement of more task types and freedom of action for students. Consequently, it increases the interaction between student and task (Xi & Mollaun, 2006).

Based on Table 8, students' performance reached about more than 50% of the total variance. The G-coefficient of this analysis showed that integrated tasks provided more reliable methods for examining the writing ability (relative G-coefficients = 0.80 and absolute G-coefficient = 0.79). Although not all the G-coefficients are above 0.80, they are very close to that standard amount (as it was suggested by Cronbach et al., 1972) and it could be stated that they were reliable. After student's variance, the interaction between student and rater (SR) was the main source of variance, which ranged from 12 to 25%.

About tasks in writing assessment, it is investigated that task difficulty could be a major source of variance (Brown et al., 1991). In this particular case, the interaction between student and rater (SR) means a type of biased exists in favor of some students in two writing tasks. As it was indicated by Hamp-Lyons and Mathias (1994) that consciously or unconsciously raters compensate for more difficult tasks by assigning higher scores, in total, using two tasks for writing assessment could be useful and suggestible (Weigle, 1999). In addition to what has been said, these findings could check the bias which may exist in favor or against students; in G-theory, bias is defined as interaction, that is, when the variance component for the interactions between existing facets is high, it indicates bias (Brennan, 2000); and in this particular case, the variance of SR:B was 3.7% and this shows that some raters were more lenient across students of particular field of study than others.

As it was mentioned, rater, task, and method of scoring neither contributed to neither relative nor absolute error of variance while the interaction between student and rater was very high and could be a source of bias (SR:B = 99.2% of relative error variance) implying that the rater could be considered the main source of error in measurement procedures. Even the interaction between rater and background of education (which is a fixed facet) caused 0.8% of error variance.

Based on the results, raters in writing assessment are the most effective factors and other facets are under the consequences of rater severity toward scoring method, task, or field of study (Santos, 1988; Weigle et al., 2003).

### D-study

In order to answer the research question about design optimization, the following analysis was conducted.

The synthesized results of G-studies helped to plan the subsequent D-studies. Following the identification of the main sources of variance, it would be possible to proactively change the design of the study to reach more reliable and generalizable results by changing the levels of any existing facet (Briesch et al., 2014). The goal of the D-study is to define the proper number of levels for any given facet, which results in smaller error variance and higher G-coefficients (Brennan, 2000).

Within a decision-making study, those facets that are considered to be fixed by the investigator do not contribute to error variance; in some cases, they reduce the error variance and increase the coefficients.

Three facets in this study, namely task, method of scoring, and background of education, were fixed and it was not possible to go beyond these levels in a D-study and also it was not possible to reduce the number of levels because at least two levels for each facet are required in the G-study (Briesch et al., 2014). In this regard, the only

manipulation of the data was applicable to levels of both student and rater. Since the student facet was the object of measurement, any increase in its number would lead to an increase in reliability coefficients, so the manipulation was applied to the levels of rater and five D-studies were conducted.

The results showed that when the number of raters was two, the relative G-coefficient was 0.66 and by increasing it from 4 to 12 the relative G-coefficient increases from 0.79 to 0.92. In order to suggest the proper number of levels for rater, it is clear that the minimum amount of reliability and generalizability could be achieved only when the minimum number of raters is six (the main G-study) with a relative G-coefficient equal to 0.85.

As it was expected by the researcher and stated by Wu and Tzou (2015), the error variance became smaller when the number of raters increases (for two raters, SE = 2.29; for 12 raters, SE = 0.38) because when the number of raters increases the accuracy of measurement increases too. Although the more raters used in an assessment procedure the more reliable it will be, using a large number of raters (e.g., 12) is neither logical nor possible for our educational system and the solution to this is choosing the minimum level, which provides the required reliability (i.e., 4 raters).

## Implications

The results of this study partly suggest that different facets affect EFL writing performance. The implication is that different sources of variance are involved in writing assessment; a fact that English teachers, test developers, and administrators should be informed of its undesirable consequences for test reliability. The fact that the students' performance had the biggest amount of contribution to total variance implies that the validity could be guaranteed through assessing writing proficiency with different methods of scoring and different rater and tasks (Marcoulides, 1989). The findings of this study also had implications in the following areas:

The investigation of the main sources of variance in writing assessment makes the practitioners capable of increasing the reliability and generalizability of measurements in this area. Designing studies through which the true abilities of students could be elicited with spending the minimum cost, energy, and time brings the fairness to measurements in an educational setting by providing an in-depth analysis of factors, which affect the observed scores.

Moreover, the convergent and discriminant validity of tests could be examined through the generalizability analysis (Kraiger & Teachout, 1990). Finally, the flexibility of G-studies and D-studies provides a situation for researchers to design their future standard setting agenda.

The present study explored the factors affecting EFL writing performance with regard to generalizability theory. Of further interest in this study was the interaction among the facets of the study (student, rater, task, scoring method, and background of education) which was categorized as measurement error (Tadayon & Khodi, 2016) by Bachman (1990). One hundred students completed the instruments of the study (independent and integrated tasks of writing test).

The generalizability analysis showed that the main source of variance across different raters, tasks, and methods of scoring is attributable to the student in measurement procedures while the interaction between student by rater nested (SR:B) in background is totally significant (99.2% of total variance) and it may be that raters' severity has been applied to rating across different students with different fields of study.

To examine factors affecting EFL writing performance, an initial model was used which included direct paths from rater, task, method of scoring, and background of education to students' test score. Finally, it became clear that the most effective relationship as the source of error variance was the interaction between rater and student. The interaction between background of education and rater also had a significant effect, in comparison with other interactions, on students' scores and test generalizability.

This study also investigated the relationship between different facets in three separate G-studies on writing tests. It was found that in all G-studies the main source of variance was the student, which explains almost more than 50% of the total variance. Additionally, it was investigated that the analytical method of scoring in comparison with the holistic method had a greater G-coefficient or reliability; maybe it is due to its exact examination of writing components by its detailed rubric. Concerning the effect of the rater, it was found that peer assessment with a relative G-coefficient of 0.86 enjoyed a higher amount of reliability rather than instructor and self-assessments. Finally, with regard to the analysis of tasks, it was found that integrated task of writing had a better ability to assess the writing performance of students and it had higher G-coefficients in comparison with mixed and independent tasks. Finally, five D-studies for optimization of the operational procedure of writing assessments were done and the result suggested that for a reliable measurement at least 4 raters were required.

**Acknowledgements**
I would like thank the editor and all the anonymous reviewers for their comments on the first draft of the paper.

**Authors' contributions**
The author read and approved the final manuscript.

**Funding**
We received no funding.

**Availability of data and materials**
The data will be available upon request.

## Declarations

**Competing interests**
The author declares that he has no competing interests.

Received: 3 May 2021 Accepted: 26 July 2021
Published online: 01 October 2021

**References**

Alavi, S. M., Karami, H., & Khodi, A. (2021). Examination of factorial structure of Iranian Englishlanguage proficiency test: An IRT analysis of Konkur examination. *Current Psychology*, 1–15.

Alkharusi, H. (2012). Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of Studies in Education*, *2*(2), 157–164 https://doi.org/10.5296/jse.v2i2.1495.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. American Educational Research Assn.

Aryadoust, V. (2010). Investigating writing sub-skills in testing English as a foreign language: A structural equation modeling study. *TESL-EJ*, *13*(4), 1–20.

Astika, G. G. (1993). Analytical assessment of foreign students' writing. *RELC Journal*, *24*(1), 371–389.

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238–257. https://doi.org/10.1177/026553229501200206.
Bain, D., & Pini, G. (1996). *Pour évaluer vos évaluations—La généralisabilité: Mode d'emploi*. Geneva: Centre for Psychoeducational Research of the Orientation Cycle.
Ballard, B., & Clancy, J. (1991). Assessment by misconception: Cultural influences and intellectual traditions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, (pp. 19–36). Norwood: Ablex Publication Corporation.
Banerjee, J., Franceschina, F., & Smith, A. M. (2007). Documenting features of written language production typical at different IELTS band score levels. *International English Language Testing System (IELTS) Research Reports 2007: Volume 7, 1*.
Bolus, R., Hinofotis, F., & Bailey, K. (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.
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-Verlag Press. https://doi.org/10.1007/978-1-4757-3456-0.
Brennan, R. L. (2011). Using generalizability theory to address reliability issues for PARCC assessments: A white paper. In *Center for Advanced Studies in Measurement and Assessment (CASMA). Iowa: University of*.
Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of school psychology*, *52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008.
Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *Tesol Quarterly*, *25*(4), 587–603. https://doi.org/10.2307/3587078.
Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing*, *16*(2), 217–238. https://doi.org/10.1177/026553229901600205.
Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics minimizing the effect of mean differences. *Written Communication*, *8*(4), 533–556. https://doi.org/10.1177/0741088391008004005.
Cardinet, J., Johnson, S., & Pini, G. (2010). *Applyinggeneralizability theory using EduG*. New York, NY:Routledge - Taylor & Francis Group
Cardinet, J., Johnson, S., & Pini, G. (2011). *Applying generalizability theory using EduG*. Taylor & Francis. https://doi.org/10.4324/9780203866948.
Casanave, C. P., & Hubbard, P. (1992). The writing assignments and writing problems of doctoral students: Faculty perceptions, pedagogical issues, and needed research. *English for Specific Purposes*, *11*(1), 33–49. https://doi.org/10.1016/0889-4906(92)90005-U.
Connor, U. (1991). Linguistic/rhetorical measures for evaluating ESL writing. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts*, (pp. 215–226). Norwood: Ablex Publication Corporation.
Connor-linton, J. E. F. F. (1995). Looking behind the curtain: what do L2 composition ratings really mean? *Tesol Quarterly*, *29*(4), 762–765. https://doi.org/10.2307/3588174.
Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment*, *7*(4), 303–323. https://doi.org/10.1207/S15326977EA0704_04.
Cronbach, L., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1963). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). Theory of generalizability for scores and profiles. The dependability of behavioral measurements.
Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x.
Cronbach, L. J., & Thorndike, R. L. (1971). Educational measurement. *Test Validation*, 443–507.
Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7*(1), 31–51. https://doi.org/10.1177/026553229000700104.
Davidson, F. (1991). Statistical support for training in ESL composition rating. *Assessing second language writing*. In L. Hamp-Lyons (Ed.), *Assessing second language writing*, (pp. 155–165). Norwood: Ablex.
Evola, J., Mamer, E., & Lentz, B. (1980). Discrete point versus global scoring for cohesive devices. *Research in language testing*, 177–181.
Fedelt, L.S, & Brennan, R L. (1989). Reliability. In R.L. Linn (Ed), *Educational measurement* (3rd ed., pp. 105-146). New York: American Council on Education and MacMillan.
Ferris, D. (2002). *Treatment of error in second language student writing*. Ann Arbor: University of Michigan Press.
Ferris, D. R. (1994). Rhetorical strategies in student persuasive writing: Differences between native and non-native English speakers. *Research in the Teaching of English*, 45–65.
Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading*, *10*(3), 323–330. https://doi.org/10.1207/s1532799xssr1003_7.
Friedrich, P. (2008). *Teaching academic writing*. NY: Continuum Press.
Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London and New York: Routledge.
Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. *Assessing second language writing in academic contexts*, 241–276.
Hamp-Lyons, L., (1996). The challenges of second language writing assessment. In E. White. Lutz and S. Kamusikiri (eds.), *Assessment of writing: Policies, politics, practice* (pp. 226-240). New York: Modern.
Hamp-Lyons, L., & Mathias, S. P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, *3*(1), 49–68. https://doi.org/10.1016/1060-3743(94)90005-1.
Harmer, J. (2004). *How to teach writing*. Essex: Longman Press.
Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL quarterly*, *18*(1), 87–107. https://doi.org/10.2307/3586337.
Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? A generalizability theory approach. *Assessing Writing*, *13*(3), 201–218. https://doi.org/10.1016/j.asw.2008.10.002.
Huang, J. (2011). Generalizability Theory as Evidence of Concerns About Fairness in Large-Scale ESL Writing Assessments. TESOL Journal, 2(4), 423-443.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, *17*(3), 123–139. https://doi.org/10.1016/j.asw.2011.12.003.
In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language testing*, 33(3), 341-366.
Jacobs, H. L., Zinkgarf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley: Newbery House.
Janopoulos, M. (1995). Writing across the curriculum, writing proficiency exams, and the NNS college student. *Journal of Second Language Writing*, *4*(1), 43–50. https://doi.org/10.1016/1060-3743(95)90022-5.
Kane, M. (2010). Errors of measurement, theory, and public policy. William H. Angoff Memorial Lecture Series. *Educational Testing Service*.
Karami, H., & Khodi, A. (2021). *Differential item functioning and test performance: A comparison between the Rasch model, logistic regression and Mantel-Haenszel*.
Khalilzadeh, S., Khodi, A. (2021). Teachers' personality traits and students' motivation: A structural equation modeling analysis. *Curr Psychol*, *40*, 1635–1650. https://doi.org/10.1007/s12144-018-0064-8.
Khodi, A. (2015). Revisiting Mobile Assisted Language Learning in EFL Writing Classes. *Enjoy Teaching Journal*, 3(2).
Khodi, A., & Abbasi Sardari, S. (2015). The effect of metalinguistic corrective feedback on students' writing performance. *International Journal of Educational Investigations*, 2(4), 102–8.
Khodi, A., Alavi, S. M., & Karami, H. (2021). Test review of Iranian university entrance exam: English Konkur examination. *Language Testing in Asia*, 11(1), 1–10.
Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly*, *26*(1), 81–112. https://doi.org/10.2307/3587370.
Kraiger, K., & Teachout, M. S. (1990). Generalizability theory as construct-related evidence of the validity of job performance ratings. *Human Performance*, *3*(1), 19–35. https://doi.org/10.1207/s15327043hup0301_2.
Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. *Language Testing*, *9*(1), 30–49. https://doi.org/10.1177/026553229200900104.
Kyburg, H. (1968). *Philosophy of science: A formal approach*. New York: Macmillan.
Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158–180. https://doi.org/10.1177/026553229801500202.
Marcoulides, G. A. (1989). Measuring computer anxiety: The computer anxiety scale. *Educational and Psychological Measurement*, *49*(3), 733–739. https://doi.org/10.1177/001316448904900328.
McNamara, T. F. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing*, *7*(1), *52–75*. https://doi.org/10.1177/026553229000700105.
McNamara, T. F. (1996). *Measuring second language performance*. NY: Longman.
Mehrani, M. B., & Khodi, A. (2014). An appraisal of the Iranian academic research on English language teaching. *International Journal of Language Learning and Applied Linguistics World*, *6*(3), 89–97.
Mehrani, M. B. (2017). A Narrative Study of Iranian EFL Teachers' Experiences of Doing Action Research. *Iranian Journal of Language Teaching Research*, 5(1), 93–112.
Mehrani, M. B., & Peterson, C. (2015). Recency tendency: Responses to forced-choice questions. *Applied Cognitive Psychology*, *29*(3), 418–424. https://doi.org/10.1002/acp.3119.
Mickan, P., & Slater, S. (2003). Text analysis and the assessment of academic writing. *IELTS Research Reports Volume*, *4*, 59–88.
Mullen, K. A. (1977). Using rater judgments in the evaluation of writing proficiency for non-native speakers of English. *On TESOL*, *77*, 309–320.
Myles, F. (2002). Second Language Acquisition (SLA) research: Its significance for learning and teaching. *The guide to good practice for learning and teaching in languages, linguistics and area studies*.
Nakamura, Y. (2004). A comparison of holistic and analytic scoring methods in the assessment of writing. In *3rd annual JALT Pan-SIG Conference*.
Reid, J. M., & O'Brien, M. (1981). *The application of holistic grading in an ESL writing program. Paper presented at the annual convention of Teachers of English to Speakers Other Languages*. MI: Detroit.
Sakyi, A. (2000). Validation of holistic writing for ESL writing assessments: How raters evaluate ESL compositions. In: A. Kunnan(Ed), *Fairness and validation in language assessment* (pp. 129-152). Cambridge University Press.
Samar, R. G., Mehrani, M. B., & Kiyani, G. (2012). An investigation into the generalizability of quantitative research studies in Iranian ELT context. *Comparative Language & Literature Research*, *3*(4), 193–213.
Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *Tesol Quarterly*, *22*(1), 69–90. https://doi.org/10.2307/3587062.
Shavelson, R. J. (2004). Editor's Preface to Lee J. Cronbach's "My Current Thoughts on Coefficient Alpha and Successor Procedures". *Educational and Psychological Measurement*, *64*(3), 389–390. https://doi.org/10.1177/0013164404264117.
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park: Sage Publications.
Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922–932. https://doi.org/10.1037/0003-066X.44.6.922.
Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing*, *5*(2), 163–182. https://doi.org/10.1016/S1060-3743(96)90023-5.
Speck, B. W., & Jones, T. R. (1998). Direction in the grading of writing? In F. Zak, & C. C. Weaver (Eds.), *The theory and practice of grading: Problems and possibilities*, (pp. 17–29). Albany: SUNY Press.
Tadayon, F., & Khodi, A. (2016). Empowerment of refugees by language: Can ESL learners affect the target culture? *TESL Canada Journal*, 129–137.
Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In L. Hamp-Lyons (Ed.), *Assessing second language writing*, (pp. 111–126). Norwood, NJ: Ablex.
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287. https://doi.org/10.1177/026553229801500205.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), 145–178. https://doi.org/10.1016/S1075-2935(00)00010-6.
Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997.
Weigle, S. C., Boldt, H., & Valesecchi, M. I. (2003). Effects of task and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly*, *37*(2), 345–354. https://doi.org/10.2307/3588510.
Weir, C. (1990). *Communicative language testing*. NJ: Prentice Hall Regents.
Wiseman, C. S. (2012). A comparison of the performance of analytic vs. holistic scoring rubrics to assess L2 writing. *Iranian Journal of Language Testing*, *2*(1).
Wu, Y. F., & Tzou, H. (2015). A multivariate generalizability theory approach to standard setting. *Applied Psychological Measurement*, *39*(7), 507–524. https://doi.org/10.1177/0146621615577972.
Xi, X., & Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST). *ETS Research Report Series*, 2006(1), i–71.
Zabihi, R., Mehrani-Rad, M., & Khodi, A. (2019). Assessment of authorial voice strength in L2 argumentative written task performances: contributions of voice components to text quality. *Journal of Writing Research*, *11*(2), 331–355. https://doi.org/10.17239/jowr-2019.11.02.04.
Zhang, S. (2006). Investigating the relative effects of persons, items, sections, and languages on TOEIC score dependability. *Language Testing*, *23*(3), 35369.

## Publisher's Note