ARTÍCULOS

Complutense Journal of **English Studies**
ISSN-e: 2386-3935

EDICIONES COMPLUTENSE

https://dx.doi.org/10.5209/cjes.85798

Does topic choice affect high-stakes L2 writing scores?

Marian Amengual-Pizarro[1]

Abstract. This study sets out to investigate the potential effects of topic choice on test-takers' L2 writing scores in a high-stakes context. Data were collected from a total of 150 essays that were assessed by three qualified raters who participated as judges in the administration of the high-stakes English Test (ET), included in the Spanish University Admission Examination (SUAE), in July 2020. Although test-takers showed a clear preference for one writing topic choice over the other, results did not reveal statistically significant differences between the average scores awarded to both essay options. Therefore, the data clearly indicate that topic choice does not affect L2 writing quality. Findings also show that choice of topic had little impact on test-takers' overall performance on the ET. Additionally, no differences in choice patterns were either observed across test-takers' proficiency levels, which suggests that topic choice may be closely related to test-takers' characteristics (motivation, interest, relevance, etc.) rather than to the writing prompt itself. Lastly, the data show potential interactions between raters' characteristics and essay topics which may affect final writing scores.
Keywords: English proficiency tests; high-stakes testing; topic writing choice; students' performance; L2 writing assessment

[es] ¿Influye la elección del tema en las puntuaciones de los exámenes estandarizados de producción escrita en la L2?

Resumen. Este estudio se propone investigar los posibles efectos de la elección del tema de las redacciones en las puntuaciones de los candidatos en un examen en L2 estandarizado. Un total de 150 redacciones fueron evaluadas por tres correctores que participaron en la administración de la Prueba de Inglés (PI), incluida en las Pruebas de Acceso a la Universidad, en julio de 2020. Aunque los candidatos mostraron una clara preferencia por una opción temática frente a la otra, no se revelaron diferencias significativas entre las puntuaciones medias otorgadas a ambos temas de la redacción. Estos datos indican que la elección del tema no afecta la calidad de la producción escrita. Los resultados también señalan que la elección temática tuvo escaso impacto en las notas finales de los estudiantes en la PI. Asimismo, no se observaron diferencias significativas en la opción temática elegida en función del nivel de competencia lingüística de los candidatos, lo que sugiere que dicha elección estaría más estrechamente relacionada con las características de los estudiantes (motivación, interés, relevancia, etc.) que con la pregunta de la redacción en sí. Finalmente, se observan posibles interacciones entre las características de los evaluadores y los temas de la redacción.
Palabras clave: pruebas de nivel de inglés; pruebas estandarizadas; elección de tema de la redacción; rendimiento de los alumnos; evaluación de la escritura en L2

Cómo citar: Amengual-Pizarro, M. (2023). Does topic choice affect high-stakes L2 writing scores?, en *Complutense Journal of English Studies* 31, 1-13.

1. Introduction

Today, many high-stakes English proficiency tests include direct tests of writing in order to measure candidates' academic writing ability in English [e.g. TOEFL (i.e., Test of English as a Foreign Language, IELTS (i.e., Interna-

[1] Department of of Spanish, Modern and Classical languages (English philology), University of the Balearic Islands (UIB)
ORCID: 0000-0002-1645-3923
E-mail: marian.amengual@uib.es

tional English Language Testing System), FCE (i.e., Cambridge First Certificate in English, etc.]. In this assessment context, English second language (L2) students might be offered a choice between two different writing prompts, and then be asked to write an essay on the chosen topic under particular time constraints. The main reason for offering candidates to choose a particular topic or subject matter is to enable them to display their best writing skills and increase test validity, which is considered a key aspect in language testing (Bachman and Palmer 1996; Polio and Glew 1996; Chalhoub-Deville and Turner 2000; Weigle 2002, 2012; Cumming et al. 2005; Schoonen 2005, 2012; Chapelle, Enright and Jamieson 2008; Huang 2012; Eckes, Müller-Karabil and Zimmermann 2016; Slomp 2016; Zhao and Huang 2020; Eckes and Jin 2021; among others). In fact, some research suggests that variations in writing topic might have a clear impact on L2 writing quality and writing scores (Tedick 1990; Weigle 2002; Shaw and Weir 2007; Leblanc and Fujieda 2012; Bonyadi 2014; Yang, Lu and Weigle 2015; Calkins 2020). However, numerous scholars claim that, in high-stakes assessment settings, a choice of topic can affect equivalence of writing outcomes across topics, which may threaten the reliability of writing scores (Wainer and Thissen 1994; Jennings et al. 1999; Lee and Anderson 2007; O'Sullivan and Green 2011).

In addition to aspects within the prompt themselves, further sources of variance in writing scores have been attributed to raters' factors (e.g. gender, professional and linguistic background, teaching experience, training, etc.) (Weigle 1994, 1998; Song and Caruso 1996; Cumming, Kantor and Powers 2002; Barkaoui 2007; Huang 2011, 2012; Kim et al. 2017; Zhao and Huang 2020). Indeed, research has shown that differences in scores may be linked to raters' interaction with writing prompts rather than to actual test-takers' writing proficiency (Hamp-Lyons and Mathias 1994; Polio and Glew 1996; Weigle 1999, 2013; Weigle, Boldt and Valsecchi 2003; Schoonen 2005; Huang 2012; Kyle 2020). Raters may be influenced by linguistic and content features associated with different writing topics, leading them to score test-takers' writing performance differently. Thus, it seems that both task- and rater-related factors may affect the validity and reliability of L2 writing assessment in large scale testing situations (Weigle 2002, 2013; Schoonen 2005, 2012; Lee and Kantor 2007; Brown 2011; Huang 2012; In'nami and Koizumi 2016; Kim et al. 2017).

Of the many variables that may have an impact on examinees' writing outcomes, this study focuses on the effects of topic choice on L2 writing performance in high-stakes testing settings. High-stakes tests are used to make important decisions related to students' academic careers and their future lives. Given the serious consequences that such tests might have for test-takers, it becomes critical to ensure that writing prompts are equivalent, so that students can perform equally well on any of the choices provided (Jennings et al. 1999; Weigle 1999, 2012; Chalhoub-Deville and Turner 2000; Schoonen 2005, 2012). Indeed, in these assessment contexts, test developers are particularly compelled to ensure the comparability of topics in terms of difficulty in order to avoid scores being affected by particular prompts (Tedick 1990; Jennings et al. 1999; Weigle 1999; Lee and Anderson 2007; Cho, Rijmen and Novák 2013). This is a fundamental part of the test validation process. However, thus far, the extent to which writing topics may influence test-takers' writing outcomes has yielded inconclusive or contradictory results. Therefore, the question of whether or not students should be allowed a choice of prompts remains unclear. In addition, most empirical studies on topic effects have focused on score differences across genres, with less attention being paid to topic effects on highly similar tasks or genre-specific writing (see Yoon 2017). Thus, this study intends to contribute to the limited research that addresses the extent to which topic choice affects the writing scores of two argumentative/opinion essays in the context of a high-stakes English Test (ET) included in the Spanish University Entrance Examination (SUEE).

The rest of the study is divided as follows: Section 2 provides the background of the study, focusing on a review of related literature regarding writing prompts and test-taker- and rater- related factors. Section 3 presents the study methodology based on quantitative data. Section 4 deals with the analysis and discussion of results. Finally, section 5 provides the main concluding remarks.

2. Background

2.1. Writing prompts and test-taker variables

Providing students with topic choice has been found to be beneficial for their writing performance (Flowerday and Schraw 2000; Graham 2006; Calkins 2020). Thus, different research studies have highlighted the positive influence of choice on students' writing quality in terms of syntactic complexity, lexical richness, coherence and cohesion (He and Shi 2012; Yang, Lu and Weigle 2015; Kim and Kim 2016; Yoon 2017, Shi, Huang and Lu 2020). Allowing learners to choose their writing topics has also been linked to enhanced motivation (Deci and Ryan 2000; Patall, Cooper and Robinson 2008; Troia, Shankland and Wolbers 2012; Basten et al. 2014; Schneider et al. 2018, Schneider 2021) and creativity (Chua and Iyengar 2008). Furthermore, differences in mean test scores have been related to prompt or topic effects in various testing contexts (Golub-Smith, Reese and Steinhaus 1993; Spaan 1993; Hamp-Lyons and Mathias 1994; Lee and Anderson 2007; Skehan 2009; He and Shi 2012; Bonyadi, 2014; Weigle and Friginal 2015). Thus, Tedick (1990: 138) argues that differences in topics have a clear impact on candidates' writing performance and advocates for the inclusion of topics familiar to the L2 students in order to encourage them to take risks, improve their writing performance and obtain "more accurate judgments about students' underlying writing ability".

With similar views, Kroll and Reid (1994: 235) explain that students perform more successfully when the prompt is familiar and within their experience. The effects of topic familiarity have been pointed out by other scholars who highlight the need to include general and everyday subject-matter topics in order to let examinees produce language and have their L2 writing skills adequately assessed (Bachman and Palmer 1996; Polio and Glew 1996; Skehan 1998, Yu 2010; He and Shi 2012; Shi, Huang and Lu 2020; Yang and Kim 2020; Green, 2021; Kessler, Ma and Solheim 2021). According to Kessler, Ma and Solheim (2021: 3-4), some writing topics might be considered unfair in certain testing situations due to test-takers' too little or lack of content knowledge, which may impede the fair assessment of their written production. In fact, several studies suggest that candidates could be unfairly penalised by being forced to write on only one single topic (Polio and Glew 1996; Shi, Huang and Lu 2020; Kessler, Ma and Solheim 2021). As Polio and Glew (1996: 45) point out: "Essays written on prompts students are forced to use may not be a good indicator of what a writing exam claims to be testing, even though such a test has high interrater reliability". Under this assumption, allowing topic choice in writing would be a better indicator of test-takers' general language proficiency, yielding a more valid and fairer score. Therefore, provided that the main goal of the test is not to assess specific content knowledge, denying choice of writing topics to examinees would seem to be both unfair and unreasonable (see Bridgeman, Morgan and Wang 1997: 273; Breland, Bridgeman and Fowles 1999; Bonzo 2008; Leblanc and Fujieda 2012; Bonyadi 2014; Hamel 2017; Calkins 2020).

Nevertheless, numerous testing researchers contend that a choice of prompts, especially in high-stakes assessment settings, may introduce an additional source of measurement error (Messick 1989), decreasing, in this way, test validity and reliability (Hughes 1989; Kroll 1990, 1991; Wainer and Thissen 1994; Bridgeman, Morgan and Wang 1997; Lim 2010; O'Sullivan and Green 2011). Wainer and Thissen (1994: 160), for instance, wonder whether allowing candidates choice is a 'sensible strategy'. According to these authors, most choice items are roughly equivalent and cannot be equated across topics: "If test forms are built that cannot be equated (made comparable), scores comparing individuals on incomparable forms have their validity compromised" (Wainer and Thissen 1994: 191). Jennings et al. (1999: 431) also note that "offering a choice among items in testing settings may threaten the principles of fair and ethical testing in which each test taker is faced with an equal challenge", and point to the effect of a 'topic effect' (i.e. interest, relevance or prior knowledge of the topic) as a potential threat to the establishment of test validity. Advocates of this latter point of view generally claim that there is no clear evidence that different exam prompts will affect candidates' final scores (see Jennings et al. 1999; Lee and Anderson 2007; Yang and Kim 2020; Aitken, Graham and McNeish 2022). Additionally, some researchers argue that, when given a choice of topics, students waste valuable time and do not always make the right choice, choosing the topic on which they cannot get their highest scores (See Kroll 1991; Wainer and Thissen 1994; Fitzpatrick and Yen 1995; Bridgeman, Morgan and Wang 1997; Powers and Fowles 1998). In fact, several studies point to the lack of relationship between candidates' perception of task difficulty and the quality of their responses (Elder, Iwashita and McNamara 2002; Kuiken and Vedder 2008; Cho, Rijmen and Novák 2013; Kim and Kim 2016; Yang and Kim 2020).

Relatedly, research studies have further noted that candidates' perceptions of writing topics must necessarily consider learners' backgrounds and their unique cultural and linguistic experiences in order to better understand the relationship between topic and written language complexity (Fox, Pychyl and Zumbo 1999; Hinkel 2002, 2003; Lo and Hyland 2007; Yoon 2017; Shi, Huang and Lu 2020). Several scholars also emphasise the need to take into account the effects of individual differences (i.e. cognitive abilities, L2 writing self-efficacy, motivation, interest, gender, etc.) on writing performance so as to gain a deeper insight into test-takers' motivation for their writing choices (see Patall, Cooper and Robinson 2008; Kormos 2012; Patall 2012; Harris et al. 2013; Preiss et al. 2013; Troia et al. 2013; Flowerday and Shell 2015; Schneider et al. 2018, Canziani, Esmizadeh and Nemati 2021; Golparvar and Khafi 2021; Schneider 2021).

2.2. Writing prompts and rater variables

Aside from test-takers' characteristics, different empirical studies attribute differences in writing scores to various assessment context variables, such as rater inconsistency. Raters have been found to show discrepancies in their assessment of writing quality due to numerous factors (e.g. age, gender, professional and linguistic background, teaching and language experience, training, rating criteria used, etc.) (see Hamp-Lyons and Mathias 1994; Weigle 1999, 2013; Weigle, Boldt and Valsecchi 2003; Lee and Kantor 2007; Huang 2011, 2012; Li and He 2015; Zhao and Huang 2020). Thus, Hamp-Lyons (1990) argues that the interaction of raters with essay topics can affect essay final scores. Indeed, Hamp-Lyons and Mathias (1994) found that raters were rewarding candidates for choosing topics judged to be more difficult, giving them higher scores as opposed to topics judged to be easier which, conversely, received lower scores (see also Polio and Glew 1996). Similarly, Purves (1992) admits that differences in writing scores might be attributable to rater variables rather than to actual student ability.

Therefore, rater reliability and task reliability have been identified as two major areas of concern in the writing assessment literature (Kroll 1998; Schoonen 2005, 2012; Lee and Kantor 2007; Brown 2011; Huang 2012; Kim et al. 2017; Zhao and Huang 2020; Eckes, Müller-Karabil and Zimmermann 2016; Eckes and Jin 2021). Thus far, evidence suggests that writing task effects (e.g. topic, genre, time limits imposed, etc.) seem to have a greater impact on writing scores than rater effects (Schoonen 2005; Brown 2011; In'nami and Koizumi 2016; Kim et al. 2017).

In sum, writing scores involve a complex set of interactions between multiple characteristics of the test-takers themselves, aspects of the prompts, and rater variables which may, either collectively or individually, account for variability in writing performance scores (Cho, Rijmen and Novák 2013)

2.3. The study context

The Spanish University Entrance examination (SUEE) is a high-stakes nationwide test taken by students at the end of upper secondary education or Bachillerato (Spanish Baccalaureate) in order to get admission to any Spanish University. The English Test (ET), which forms part of the SUEE, is a norm-referenced proficiency test whose main aim is to compare candidates' performance with each other. After completing the baccalaureate programme, students are supposed to have mastered a CEFR-level B1 (Common European Framework of Reference) in a second language (e.g. English) (see Díez Bedmar 2012; García-Laborda 2012). Despite some differences in test format, the task types of the ET are relatively similar across Spanish universities (Amengual-Pizarro 2010). The current ET at the University of the Balearic Islands (UIB) is a paper-based test which consists of two different exam options (A or B) of identical structure. Both options include a preceding reading passage, based on a wide range of general topics, and six text-related questions: A True / False section (item 1), an open comprehension question (item 2), a lexical comprehension section (item 3), a grammar or syntax section (item 4), a phonetics and phonology question (item 5) and, finally, a writing section (item 6). Spanish universities use a 10-point grading system, with 0 being the lowest mark (i.e. fail) and 10 being the highest mark (i.e. outstanding) that test-takers can earn. The minimum pass mark required for the ET is 5 points. Table 1 below summarises the number and nature of the items included in the ET, the scores assigned to each item and the testing techniques used to assess the different questions in the test.

Table 1. Components of the ET at the UIB for both options (A or B)

Item	Score	Type of item	Techniques
1	0-1	Objective	True / False
2	0-1	Subjective	Comprehension question (Open answer)
3	0-1	Objective	Matching synonyms
4	0-2	Objective	Grammar transformation
5	0-1	Objective	Phonetics and Phonology
6	0-4	Subjective	Non-directed essay

Test candidates are asked to choose one of the two ET options (A or B), and are given a maximum of 1:30h to complete the whole test. As can be seen, question 6 includes the writing of an essay (120-150 words) on a single general topic, and it is awarded a maximum of 4 points (with 0 being the lowest proficiency level and 4 the highest) on the ET. The essays are graded on an analytic rating scale which includes the following aspects: Task fulfilment (content and communicative goal, 1 point), Grammar (grammatical range and accuracy, 1 point), Organisation (cohesion and coherence, 1 point) and Vocabulary (lexical range and accuracy, 1 point). Although very little is known about the processes test-takers go through in choosing between both ET options (see Polio and Glew 1996), the high weight given to the essay (4 points out of a total of 10) might be a key aspect affecting their final decision. In fact, evidence indicates that those parts of the high-stakes test carrying most marks are the ones which are given greater emphasis in class (Lam 1993; Weigle 2002; Spratt 2005; Green 2007, 2021). Amengual-Pizarro (2010) also observed that secondary school teachers paid most attention to the essay part of the ET when preparing students for this examination in order to ensure students earned high grades on the test. Although students are not usually offered a choice of topic in the ET at the UIB, the current Covid-19 pandemic led Spanish universities to adopt more flexible assessment practices, allowing candidates to choose questions from any ET option (A or B), as long as they completed the 6 different types of questions on the test. Accordingly, during the past two years (2020-2022), students have been given the opportunity to choose the most convenient writing topic from the two ET options provided. This enables us to explore the extent to which topic choice may have affected writing test scores in a high-stakes assessment context.

2.4. The present study

As has been stated earlier, the main aim of this study is to explore the potential influence of topic choice on test-takers' writing quality and overall test-taker's performance in a high-stakes testing setting, adding to the limited amount of research examining this issue across genre-specific writing prompts. Additionally, possible differences between raters' judgements will be explored.

To attain these goals, the following research questions were posed:

1. What are the effects of topic choice on test-takers' final writing scores?
2. What is the relationship between topic choice and their final test scores on the ET?

3. Is there any relationship between choice in writing and test-takers' English proficiency?
4. Are there any statistically significant differences between raters' writing scores?

3. Method

3.1. Participants

A total of 150 English tests were evaluated by three experienced raters who took part in the administration of the ET in July 2020. The raters were recruited by the researcher after having obtained consent from the organizing committee for the university entrance examination at the UIB (OCUEE), which is responsible for the design and development of the test. All of the raters were female secondary education, qualified teachers who had already participated as examiners in previous SUEE administrations. Each rater evaluated individually a subset of 50 English tests randomly assigned to them by the OCUEE at the UIB (T = 150). Raters were asked to note down the individual and total scores they had awarded to each test, as well as to indicate the writing test option (A or B) chosen by each examinee. The tests were all blinded and hence all identification details from candidates (name, gender, age, etc.) were removed in order to minimise bias. Essays were rated on a 4-point analytic scale (with 0 being the lowest score and 4 being the highest score) based on the Common European Framework of Reference for Languages (CEFR, 2001), which included the following aspects: Task fulfilment, grammar, organisation, and vocabulary.

3.2. Prompt characteristics

Two different essay prompts were presented to students in the ET administration (June 2020) at the UIB. The prompts were based on a preceding reading text about two different topics which elicited the rhetorical mode of argumentative/opinion writing (i.e. writing where students hold a particular point of view on a debatable issue and try to illustrate and justify it). Therefore, the topics were controlled for genre (i.e., argumentative/opinion) in order to examine within-genre topic effects. The first test option (Option A) included a reading passage entitled: "Who is Greta Thunberg, the 'Fridays for Future' activist?". The essay prompt for this option asked students to discuss the following: "Do you think environmental activists help people become aware of the need to fight against climate change? What other social issues worry you the most? Explain". Option B was based on a reading passage on dating platforms entitled: "Love and dating after the 'Tinder' revolution". The essay prompt for this option asked students to comment on the following: "Do you think dating apps or dating websites are a good idea? What is the best or worst online dating experience you have ever had or heard of? Explain". The two writing topics, option A (hereafter, the environmental issue topic) and option B, (hereafter, the dating platform topic) tried to encourage candidates to discuss their personal perspectives and experiences so that students felt comfortable with the topic. Essays should be between 120 and 150 words. No dictionaries or digital tools were permitted to write them.

3.3. Data analysis procedures

A quantitative design was used to address the 4 research questions in this study. Data analysis was carried out by using the Statistical Package for the Social Sciences (SPSS), version 22.0. The Shapiro-Wilk Test was conducted to check the normality of the gathered data (i.e. p-values greater than .05). Descriptive statistics (i.e. means and standard deviation) were first computed. Independent-samples t-tests were run to identify possible differences between two categorical independent groups (i.e. topic choice) on the same continuous variables (i.e. writing scores and final ET scores). Pearson correlations were calculated to determine the strength and direction of association of categorical variables (i.e. writing scores and final ET scores). The chi-square test for independence was used to discover if there was a relationship between categorical variables (test-takers' level of English proficiency and topic choice). Finally, the one-way analysis of variance (ANOVA) was run to determine whether there were any statistically significant differences between the means of the three raters who participated in the study regarding the assessment of the test-takers' essays.

4. Results and discussion

4.1. What are the effects of topic choice on test-takers' final writing scores?

To address the first research question, descriptive statistics for each test writing option (A or B), including the mean and standard deviation were calculated (Table 2).

Table 2. Descriptive statistic for choice of topic

Essay option	N	Mean	Std. Deviation	Std. Error Mean
Essay A	48	2.66	1.07	.154
Essay B	102	2.60	.80	.080

Although the primary purpose of test developers is to develop and include two comparable topics, as can be seen, the vast majority of students (68%, N = 10) chose essay B (the dating platform topic) over essay A (the environmental issue topic) (32%, N = 48), contradicting previous research that suggests that students are likely to pay more attention and favour questions placed first (see Chiste and O'Shea 1988). Thus, the data show that examinees felt more confident in commenting on a more personal experience (the dating platform topic) than on a less personal topic (essay A, the environmental issue topic), which could be more challenging in terms of vocabulary or prior background knowledge of the topic. Topic relevance or candidates' interest in the topic could also be important factors in understanding examinees' motivation for their choice of topic (see Polio and Glew 1996; Patall, Cooper and Robinson 2008; He and Shi 2012; Kormos 2012; Patall 2012; Preiss et al. 2013; Troia et al. 2013; Flowerday and Shell 2015; Canziani, Esmizadeh and Nemati 2021; Golparvar and Khafi 2021).

Nevertheless, as can be observed in Table 2, the mean scores of both essays (M= 2.66, essay A vs. M = 2.60, essay B) indicate that the two writing topics were found to be equivalent and did not involve great difficulty, since in both cases the value obtained was above the midpoint (2 points out of 4). Data also show that the least preferred option (essay A, the environmental issue topic) produced only slightly higher scores than the most preferred one (essay B, the dating platform topic). This finding contradicts that of Hamp-Lyons and Mathias's (1994) which suggested that examinees are likely to score higher on public topics than on more personal ones. In fact, the independent-samples t-test used to compare the means between the two topic choices, showed no statistically significant differences between the average scores awarded to both writing options [$t(148)=.404$, $p=.715$]. This clearly indicates that topic choice did not elicit different writing test scores. That is, test-takers did not perform significantly better when they had a choice. However, without further research, we cannot rule out the possibility that test-takers would have been disadvantaged if they had been forced to choose a different writing topic option (See Polio and Glew 1996: 45; Jennings et al. 1999; Bonzo 2008; Leblanc and Fujieda 2012; Yang, Lu and Weigle 2015; Calkins 2020). Likewise, the lower values of the standard deviation (i.e. how much individual scores differ from the mean) obtained indicate that most of the writing test scores in both options clustered around the mean.

In sum, the findings of this study reveal that the clear preference that most test-takers showed towards writing topic B (the dating platform topic) did not have any significantly major effect on the quality of their performance. These findings align with those of previous studies that observed little or no relationship between test candidates' preference for a particular topic and the scores obtained in their writing tests (Powers et al. 1992; Gabrielson, Gordon and Engelhard 1995; Powers and Fowles 1998; Jennings et al. 1999; Lee and Anderson 2007; Yang and Kim 2020; Aitken, Graham and McNeish 2022). Admittedly, the limited amount of choice offered (i.e. only two writing topics) could have reduced the potential influence of topic choice on essay scores (see Gabrielson, Gordon and Engelhard 1995; Jennings et al. 1999). In any event, both topic options yielded comparable scores, which indicates they were also sufficiently general to be used in the high-stakes ET (see Polio and Glew 1996; Skehan 1998, Lee and Anderson 2007, Yu 2010; He and Shi 2012; Shi, Huang and Lu 2020; Yang and Kim 2020; Green, 2021; Kessler, Ma and Solheim 2021). This is also a positive finding since it means that test-takers were neither rewarded nor penalised by having chosen one topic over the other in terms of their test results (see Jennings et al. 1999; Bonzo 2008; Leblanc and Fujieda 2012; Bonyadi, 2014; Yang, Lu and Weigle 2015). Therefore, writing test scores on the ET have validity (Chalhoub-Deville and Turner 2000; Weigle 2012, 2013; Cumming et al. 2005; Chapelle, Enright and Jamieson 2008; Lim 2010; Schoonen, 2012; Kim et al. 2017; Zhao and Huang 2020; Guapacha 2022).

4.2. What is the relationship between topic choice and examinees' final test scores on the English test (ET)?

In order to examine the effect of topic choice on overall test-takers' performance on the ET, Pearson correlations between both test-takers' writing scores on both options (essay A and B) and their final scores on the English test (ET) were first calculated. Correlations were found to be large, positive, and statistically significant at the 0.01 level (2-tailed) across both topic choices: option A ($r = .918$, $n = 48$, $p = .001$) and B ($r = .860$, $n = 102$, $p = .001$), which further suggests that topic choice had little impact on test-takers' overall performance on the ET. Therefore, allowing test-takers a choice does not constitute a potential source of construct-irrelevant variance (Messick 1989).

Differences between the two writing choices and examinees' final test scores on the ET (Mean = 6.10; SD = 2.02) were calculated using the independent-samples t-test. Findings showed no statistically significant differences between topic choice and candidates' final ET results [$t(148)= 1.336$, $p=.220$]. Thus, test scores seem to be insensitive to topic or prompt variation and hence do not represent a threat to the assessment of English language proficiency, which is the construct of interest (see Messick 1989; Jennings et al. 1999; Weigle 2002, 2013; Cumming et al. 2005; Chapelle, Enright and Jamieson 2008; Lim 2010; Schoonen 2012; Kim et al. 2017, Zhao and Huang 2020, among others). In other words, differences in examinees' final test scores are not attributed to choice of writing topics. From

this data, it can be concluded that choice on writing cannot be considered a predictor variable of enhanced performance on the ET.

4.3. Is there any relationship between choice in writing and test-takers' English proficiency?

Further analysis was carried out to examine the relationship between students' English language proficiency, based on their final test scores in the ET, and their choice of writing topic. Students were categorised into three main levels of language proficiency: low (scores less than 5), intermediate (scores between 5 and 7) and advanced (scores between 8 and 10). The chi-square test for independence (Tables 3 and 4) was used to discover if there was a relationship between test-takers' levels of English proficiency (i.e. low, intermediate and advance) and their choice of writing. Chi-square results show that there was not a statistically significant association between choice of writing topic (essay A or B) and test-takers' language ability [$\chi^2(2) = 4.185$, $p = .123$]. Therefore, test-takers' level of English proficiency cannot account for examinees' high preference for essay topic B (the dating platform topic) over essay topic A (the environmental issue topic). These findings contradict those of Jennings et al. (1999) who observed differences in the topic choices made by high-proficiency and low-proficiency test-takers. In sum, contrary to some research which points to the possibility of an interaction between these two latter variables, topic choice (essay A or B) does not seem to be related to participants' mastery of English (Spaan 1993; Hamp-Lyons and Mathias 1994; Lee, Breland and Muraki 2004; Yoon 2017).

Table 3. Type of essay A or B*Participants' proficiency Cross tabulation

		Participants' proficiency							
		Low		Intermediate		Advanced		Total	
		N	%	N	%	N	%	N	%
Type of essay A or B	Essay A	12	29.3%	21	27.3%	15	46.9%	48	32.0%
	Essay B	29	70.7%	56	72.7%	17	53.1%	102	68.0%
Total		41	100.0%	77	100.0%	32	100.0%	150	100.0%

Table 4. Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.185[a]	2	.123
Likelihood Ratio	4.016	2	.134
Linear-by-linear Association	2.178	1	.140
N of Valid Cases	150		
a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 10.24.			

4.4. Are there any statistically significant differences between raters' writing scores?

Finally, since research points to variation in raters' characteristics as a major factor that might affect test-takers' scores, a comparison between the mean scores awarded by each of the three raters to the two writing topics was also made. Summary descriptive statistics of the two writing prompts are shown in Table 5.

Table 5. ANOVA descriptives

Type of essay A or B		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Essay A	Rater 1	23	2.9348	1.00628	.20982	2.4996	3.3699	.00	4.00
	Rater 2	17	2.4412	1.02519	.24864	1.9141	2.9683	.50	3.75
	Rater 3	8	2.3750	1.28869	.45562	1.2976	3.4524	.00	3.75
	Total	48	2.6667	1.07106	.15459	2.3557	2.9777	.00	4.00
Esay B	Rater 1	27	2.7222	.72169	.13889	2.4367	3.0077	1.25	4.00
	Rater 2	33	2.2879	.80069	.13938	2.0040	2.5718	.50	4.00
	Rater 3	42	2.7738	.81302	.12545	2.5205	3.0272	1.00	4.00
	Total	102	2.6029	.80864	.08007	2.4441	2.7618	.50	4.00

The one-way analysis of variance (ANOVA) was run to determine whether there were significant writing score differences between the three raters that participated in our study. Results revealed no statistically significant differences between the scores given to essay A across raters [$F(2,45) = 1.419$, $p = .253$]. However, results from ANOVA showed a statistically significant effect of raters on essay B, which was the preferred writing option by candidates [$F(2,99) = 3.955$, $p = .022$]. Bonferroni post hoc tests were used to identify where differences between the raters on essay B occurred. The data showed significant differences between rater 2 (M = 2.28) and rater 3 (M = 2.77), $p = .028$, which indicates that the writing scores of these two raters on the assessment of essay B differed significantly from each other. The value of the effect size shows that the strength of association was medium (.074), indicating that about 7.4% of the variation in test-takers' scores on essay B was accounted for by the different scores awarded to this essay by raters 2 and 3. These rating inconsistencies could be attributed to numerous factors such as test-takers' writing ability, raters' interaction with the chosen topic, or both (see Weigle 1999, 2002; Schoonen 2005; Lee and Kantor 2007; Brown 2011; Huang 2012; In'nami and Koizumi 2016; Kim et al. 2017; Zhao and Huang 2020). The use of a mixed methods approach, combining quantitative and qualitative data (e.g. verbal protocol analysis, interviews, etc.), could help to gain an in-depth understanding of these factors in future research. In any event, these results raise concerns about the appropriateness of using a single rater to assess students' writing proficiency in large-scale assessments (see Weigle, 1999, 2002, Schoonen 2005; Slomp 2016; Kim et al. 2017; Zhao and Huang 2020). As Schoonen (2005: 21) notes, the use of multiple raters should be almost a "conditio sine qua non" in high-stakes testing in order to ensure test reliability and fairness.

The results of this study also align with previous research findings that show that the interaction of raters with essay topics may affect final essay scores (Hamp-Lyons 1990; Purves 1992; Hamp-Lyons and Mathias 1994; Weigle 1999; Elder, McNamara and Congdon 2003; Barkaoui 2007; Huang 2011, 2012; Eckes, Müller-Karabil and Zimmermann 2016; Kim et al. 2017; Zhao and Huang 2020; Eckes and Jin 2021). Further research is needed to be able to understand the potential aspects of the writing topic that raters take into account in the assessment of L2 writing (see Hamp-Lyons and Mathias 1994; Cumming, Kantor and Powers 2002; Weigle, Boldt and Valsecchi 2003; Schoonen 2005; Li and He 2015; Huang 2012; Kyle 2020).

5. Conclusion

The main purpose of this study was to investigate the extent to which within-genre topics may have an influence on test-takers' writing scores in a high-stakes testing context. Despite clear differences in candidates' prompt choice or preferences, the findings of this study concur with previous research which suggests that choice does not have a major effect on writing outcomes (Powers and Fowles 1998; Jennings et al. 1999; Lee and Anderson 2007; Yang and Kim 2020; Aitken, Graham and McNeish 2022). In fact, the independent-samples t-test revealed no statistically significant differences between the average scores awarded to both writing essays, which clearly indicates that topic choice did not lead to enhanced writing performance. This might also be considered a reassuring result since topic preference for one option (essay B) over the other (essay A) cannot be linked to prompt difficulty or variation. The data show that both writing options were equivalent and yielded comparable scores. Therefore, writing scores on the ET have validity (see Bachman and Palmer 1996; Jennings et al. 1999; Weigle 1999, 2002, 2013; Cumming et al. 2005; Chapelle, Enright and Jamieson 2008; Lim 2010; Schoonen 2005, 2012; Kim et al. 2017; Zhao and Huang 2020; Eckes and Jin 2021; Guapacha 2022).

Furthermore, correlations between writing scores on both topic choices and overall test-takers' performance on the ET were found to be large and statistically significant, which also suggests that scoring across both writing topics was comparable and had little impact on test-takers' overall performance on the ET. Similarly, the independent-samples t-test showed no statistically significant differences between topic choice and examinees' final ET results (M= 6.10; SD = 2.09). In other words, difference in test-takers' final scores on the ET cannot be attributed to choice of writing topics. Therefore, contrary to prior research (Tedick 1990; Golub-Smith, Reese and Steinhaus 1993; Spaan 1993; Hamp-Lyons and Mathias 1994; Lee and Anderson 2007; Skehan 2009; Shaw and Weir 2007; Leblanc and Fujieda 2012; Bonyadi, 2014; He and Shi 2012; Yang, Lu and Weigle 2015), this study shows that different writing topics do not yield different test outcomes in a high-stakes context and, consequently, do not constitute a potential source of construct-irrelevant variance (Messick 1989).

With regard to the relationship between test-takers English proficiency (i.e. low, intermediate and advance) and their choice of writing topic, the chi-square test for independence showed no statistically significant association between both variables. That is, no differences in choice patterns were observed across candidates' proficiency levels (see also Spaan 1993; Hamp-Lyons and Mathias 1994; Lee, Breland and Muraki 2004). Thus, topic preference does not seem to be linked to test-takers' English language ability. The findings of this study appear to be consistent with prior research that suggests that topic choice may be more closely related to candidates' variables such as prior background knowledge of the topic, topic familiarity or test-takers' personal interests or concerns rather than to specific prompt characteristics (Kroll 1994; Powers and Fowles 1998; Hinkel 2002; Patall, Cooper and Robinson 2008; Kormos 2012; Patall 2012; Troia, Shankland and Wolbers 2012; Cho, Rijmen and Novák 2013; Harris et al. 2013; Preiss et al. 2013; Flowerday and Shell 2015; Canziani, Esmizadeh and Nemati 2021; Golparvar and Khafi 2021).

Finally, the data in this study reveal that differences in the writing average scores could be partially attributed to raters' characteristics. Thus, results from the one-way analysis of variance (ANOVA) showed a statistically significant effect of raters on the assessment of essay B, which points to the complex interaction between raters and essay topics and its influence on writing performance scores (Hamp-Lyons 1990; Purves 1992; Hamp-Lyons and Mathias 1994; Weigle 1999, 2012; Lee and Kantor 2007; Brown 2011; Huang 2012; In'nami and Koizumi 2016; Zhao and Huang 2020). This finding is in line with those of other studies which suggest that, in addition to writing prompt characteristics, raters' variables (i.e. expectations, background, training, etc.) appear to have a clear impact on essay scores (Weigle 1999, 2003; Schoonen 2005, 2012; Huang 2012; Li and He 2015; Kim et al. 2017).

In sum, from this data it cannot be concluded that topic choice will lead to improved writing performance and yield better writing outcomes. However, the findings of this study also show that offering a choice does not pose a threat to the validity of the ET. In fact, given that most test-takers admit they feel that they should be offered a choice, allowing candidates a choice on writing would be desirable for the sake of face validity (See Polio and Glew 1996; Jennings et al. 1999; Flowerday and Schraw 2000; Graham 2006; Hamel 2017; Schneider et al. 2018; Brown and Abeywickrama 2019; Calkins 2020; Schneider 2021; Aitken, Graham and McNeish 2022).

Admittedly, further research is needed to substantiate our interpretations and gain greater insight into the prompt characteristics that might affect candidates' choices. In fact, this study confirms the need to consider multiple factors in order to understand test-takers' choice of topic under time constraints. Future areas of research could examine test-takers perceptions of topic difficulty and the potential strategies that they use to make their final decisions regarding topic choice in evaluative contexts. It would also be informative to investigate the value of choice in a testing situation from the test-takers' perspective (see Polio and Glew 1996; Jennings et al. 1999; Lim 2010; Shaw and Weir, 2007; Troia et al. 2013; Slomp 2016; Brown and Abeywickrama 2019). Given the study findings, it would also be useful to explore raters' behaviour and their interaction with prompt characteristics while assessing different writing topics (see Hamp-Lyons and Mathias 1994; Weigle 1999, 2012; Cumming, Kantor and Powers 2002; Weigle, Boldt and Valsecchi 2003; Huang 2012; Cho, Rijmen and Novák 2013; Kyle 2020). Additionally, the current practice of using a single rater in large scale writing assessments should be reviewed in order to improve score reliability (see Weigle, 1999, 2002, Schoonen 2005; Slomp 2016; Kim et al. 2017; Zhao and Huang 2020). Since critical decisions are taken on the basis of the scores obtained in high-stakes testing settings, it is believed that the importance of investigating topic variables such as prompt effects cannot be underestimated in order to ensure the validity and fairness of L2 writing assessment decisions (Jennings et al. 1999; Tedick 1990; Schoonen 2005, 2012; Lee and Kantor 2007; Brown 2011; Huang 2012; In'nami and Koizumi 2016; Slomp 2016; Kim et al. 2017; Zhao and Huang 2020).

References

Aitken, A. Angelique, Graham, Steve and Daniel McNeish (2022). The effects of choice versus preference on writing and the mediating role of perceived competence. *Journal of Educational Psychology 114,* 8: 1844-1865. DOI: https://doi.org/10.1037/edu0000765

Amengual-Pizarro, Marian (2010). Exploring the washback effects of a high-stakes English test on the teaching of English in Spanish upper secondary schools. *Revista Alicantina de Estudios Ingleses* 23: 149-170. DOI: https://doi.org/10.14198/raei.2010.23.09

Bachman, Lyle. and Adrian Palmer (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Basten, Melanie, Meyer-Aherns, Inga, Fries, Stefan and Matthias Wilde (2014). The effects of autonomy-supportive vs. controlling guidance on learners' motivational and cognitive achievement in a structured field trip. *Science Education* 98, 6: 1033-1053. DOI: https://doi.org/10.1002/sce.21125

Barkaoui, Khaled (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing* 12: 86–107. DOI: https://doi.org/10.1016/j.asw.2007.07.001

Breland, Hunter M., Bridgeman, Brent and Mary E. Fowles (1999). *Writing assessment in admission to higher education: Review and framework* (ETS Research Report No. 99-3). Princeton, NJ: Educational Testing Service.

Bridgeman, Blent Morgan, Rick and Ming-mei Wang (1997). Choice among essay topics: Impact on performance and validity. *Journal of Educational Measurement* 34, 3: 273–286. DOI: https://doi.org/10.1111/j.1745-3984.1997.tb00519.x

Bonyadi, Alireza (2014). The effect of topic selection on EFL students' writing performance. *Sage Open* 4, 3: 1-9. DOI: https://doi.org/10.1177/2158244014547176

Bonzo, Joshua D. (2008). To assign a topic or not: Observing fluency and complexity in intermediate foreign language writing. *Foreign Language Annals*, 41: 722-735. DOI: https://doi.org/10.1111/j.1944-9720.2008.tb03327.x

Brown, H. Douglas and Priyanvada Abeywickrama (2019). *Language assessment: Principles and classroom practices* (3rd ed.). Pearson Longman.

Brown, James Dean (2011). What do the L2 generalizability studies tell us? International Journal of Assessment and Evaluation in Education, 1: 1–37.

Calkins, Lucy (2020). *Teaching writing*. Heinemann.

Canziani, Bonnie Farber, Esmizadeh, Yalda and Hamid R. Nemati (2021). Student engagement with global issues: the influence of gender, race/ethnicity, and major on topic choice. *Teaching in Higher Education* 1-22. DOI: https://doi.org/10.1080/13562517.2021.1955340

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching and Assessment*. Council of Europe, Language Policy Unit: Strasbourg. www.coe.int/lang-cefr

Cumming, Alister, Kantor, Robert, Baba, Kyoko, Erdosy, Usman, Eouanzoui, Keanre and Mark James (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing* 10, 1: 5-43. DOI: https://doi.org/10.1016/j.asw.2005.02.001

Cumming, Alister, Kantor, Robert and Donald E. Powers (2002). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal* 86: 67–96. DOI: https://doi.org/10.1111/1540-4781.00137

Chalhoub-Deville, Micheline and Carolyn E. Turner (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System* 28, 4: 523-539. DOI: https://doi.org/10.1016/S0346-251X(00)00036-1

Chapelle, Carol A., Mary K. Enright, and Joan M. Jamieson, eds. (2008). *Building a validity argument for the Test of English as a Foreign Language TM*. Routledge.

Chiste, Katherine Beaty and Judith O'Shea (1988). Patterns of question selection and writing performance of ESL students. *TESOL Quarterly* 22: 68l-684. DOI: https://doi.org/10.2307/3587275

Cho, Yeonsuk, Rijmen, Frank and JaKub Novák, (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT™ integrated writing tasks. *Language Testing* 30, 4: 513–534. DOI: https://doi.org/10.1177/0265532213478796

Chua, Roy Y. J. and Sheena S. Iyengar (2008). Creativity as a matter of choice: Prior experience and task instruction as boundary conditions for the positive effect of choice on creativity. *The Journal of Creative Behavior* 42, 3: 164-180. DOI: https://doi.org/10.1002/j.2162-6057.2008.tb01293.x

Deci, Edward L. and Richard M. Ryan (2000). The "what" and "why" of goal pursuit: "Human needs and the self-determination of behavior. *Psychological Inquiry* 11, 4: 227-268. DOI: https://doi.org/10.1207/S15327965PLI1104_01

Díez Bedmar, María Belén (2012). The use of the Common European Framework of Reference for Languages to evaluate compositions in the English exam section of the university admission examination. *Revista de Educación* 357: 55-80

Douglas, Dan and Carol A. Chapelle, eds. (1993). *A new decade of language testing research: Selected papers from the 1990 Language Testing*. Alexandria, VA: Teachers of English to Speakers of Other Research Colloquium Languages.

Eckes, Thomas and Kuan-Yu Jin (2021). Examining severity and centrality effects in TestDaF writing and speaking assessments: An extended Bayesian many-facet Rasch analysis. *International Journal of Testing* 21, 3-4: 131-153. DOI: https://doi.org/10.1080/15305058.2021.1963260

Eckes, Thomas, Müller-Karabil, Anita and Sonja Zimmermann (2016). Assessing writing. In Tsagari, Dina and Jayanti Banerjee, eds.,147-164. Berlin, Boston: De Gruyter.

Elder, Catherine, Iwashita, Noriko and Tim McNamara (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing* 19, 4: 347-368. DOI: https://doi.org/10.1191/0265532202lt23

Elder, Catherine, McNamara, Tim and Peter Congdon (2003). Rasch techniques for detecting biasing performance assessments: An example comparing the performance of native and non-native speakers on a test of academic English. *Journal of Applied measurement*, 4, 2: 181–197.

Fitzpatrick, Anne R. and Wendy M. Yen (1995). The psychometric characteristics of choice items. *Journal of Educational Measurement* 32: 243-259. http://www.jstor.org/stable/1435296.

Flowerday, Terri and Gregory Schraw (2000). Teacher beliefs about instructional choice: A phenomenological study. *Journal of Educational Psychology* 92, 4: 634-645. DOI: https://doi.org/10.1037/0022-0663.92.4.634

Flowerday, Terri and Duane F. Shell (2015). Disentangling the effects of interest and choice on learning, engagement, and attitude. *Learning and Individual Differences* 40: 134–140. DOI: https://doi.org/10.1016/j.lindif.2015.05.003

Fox, Janna, Pychyl, Timothy A. and Bruno Zumbo (1997). An investigation of background knowledge in the assessment of language proficiency. In Huhta, Ari, Viljo Kohonen, Liisa Kurki-Suonio and Sari Luoma eds., 367-383. Jyvaskyla, Finland: University of Jyvaskyla Press.

García-Laborda, Jesús. (2012). De la Selectividad a la Prueba de Acceso a la Universidad: pasado, presente y un futuro no muy lejano. *Revista de Educación* 357: 17-27.

Gabrielson, Stephen, Gordon, Belita and George Engelhard (1995). The effects of task choice on the quality of writing obtained in a statewide assessment. *Applied Measurement in Education* 8: 273-290. DOI: https://doi.org/10.1207/s15324818ame0804_1

Golparvar, Seyyed Ehsan and Afshin Khafi (2021). The role of L2 writing self-efficacy in integrated writing strategy use and performance. *Assessing Writing*, 47: 100504. DOI: https://doi.org/10.1016/j.asw.2020.100504

Golub-Smith, Marna, L., Reese, Clyde M. and Karin Steinhaus (1993). Topic and topic type comparability on the Test of Written English™ (TOEFL Research Report No. 42; ETS RR-93-10). Princeton, NJ: ETS.

Graham, Steve (2006). Writing. In Alexander, P. and P. Winne, eds., *Handbook of educational psychology* (pp. 457– 478). Erlbaum.

Green, Anthony (2021). *Exploring language assessment and testing: Language in action* (2nd ed.). Routledge.

Green, Anthony (2007). *IELTS washback in context: Preparation for academic writing in higher education* (Vol. 25). Cambridge University Press.

Guapacha Chamorro, María Eugenia (2022). Cognitive validity evidence of computer-and paper-based writing tests and differences in the impact on EFL test-takers in classroom assessment. *Assessing Writing* 51: 100594 DOI: https://doi.org/10.1016/j.asw.2021.100594

Hamel, Fred L. (2017). *Choice and agency in the writing workshop*. Columbia University. New York and London: Teachers College Press.

Hamp-Lyons, Liz (1990). Second language writing: assessment issues. In Kroll, Barbara, ed., 69-87. Cambridge: Cambridge University Press.

Hamp-Lyons, Liz and Sheila Prochnow Mathias (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing* 3: 49–68. DOI: https://doi.org/10.1016/1060-3743(94)90005-1

Harris, Karen R., Graham, Steve, Friedlander, Barbara, and Leslie Laud (2013). Bring powerful writing strategies into your classroom! Why and how. *The Reading Teacher* 66, 7: 538–542. DOI: https://doi.org/10.1002/trtr.2013.66.

He, Ling and Ling Shi (2012). Topical knowledge and ESL writing. *Language Testing* 29, 3: 443– 464. DOI: https://doi.org/10.1177/0265532212436659
Hinkel, Eli (2002). *Second language writers' text: Linguistic and rhetorical features*. Mahwah, NJ: Lawrence Erlbaum.
Hinkel, Eli. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly* 37: 275–301. DOI: https://doi.org/10.2307/3588505
Huang, Jinyan (2011). Generalizability theory as evidence of concerns about fairness in large Scale ESL writing assessments. *TESOL Journal* 2, 4: 423–443. DOI: https://doi.org/10.5054/tj.2011.269751
Huang, Jinyan (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing. *Assessing Writing* 17, 3: 123–139. DOI: https://doi.org/10.1016/j.asw.2011.12.003
Hughes, Arthur. 1989: *Testing for language teachers*. Cambridge: Cambridge University Press.
Huhta, Ari, Viljo Kohonen, Liisa Kurki-Suonio and Sari Luoma, eds., (1997). *Current developments and alternatives in language assessment: Proceedings of LTRC 1996*. Jyvaskyla, Finland: University of Jyvaskyla Press.
In'nami, Yo and Rie Koizumi (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language testing* 33, 3: 341-366. DOI: DOI: https://doi.org/10.1177/0265532215587390.
Jennings, Martha, Fox, Janna, Graves, Barbara and Elana Shohamy (1999). The test-takers' choice: an investigation of the effect of topic on language-test performance. *Language Testing* 16,4: 426–456. DOI: https://doi.org/10.1177/026553229901600402
Kim, Bomin and Haedong Kim (2016). Korean college EFL learners' task motivation in written language production. *International Education Studies* 9, 2: 42–50. DOI: https://doi.org/10.5539/ies.v9n2p42
Kim, Young-Suk Grace, Schatschneider, Christopher, Wanzek, Jeanne, Gatlin, Brandy and Stephanie Al Otaiba (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and writing* 30: 1287-1310. DOI: https://doi.org/10.1007/s11145-017-9724-6.
Khabbazbashi, Nahal (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing* 34, 1: 23–48. DOI: https://doi.org/10.1177/0265532215595666
Kessler, Matt, Ma, Wenyue and Ian Solheim (2021). The Effects of Topic Familiarity on Text Quality, Complexity, Accuracy, and Fluency: A Conceptual Replication. *TESOL Quarterly*: 1-28. DOI: https://doi.org/10.1002/tesq.3096.
Kormos, Judit (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing* 21, 4: 390-403. DOI: https://doi.org/10.1016/j.jslw.2012.09.003
Kroll, Barbara, ed. (1990). *Second Language Writing* (pp.69-87). Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139524551
Kroll, Barbara (1991). Understanding TOEFL's Test of Written English. *RELC Journal* 22, 1: 20–33. DOI: https://doi.org/10.1177/003368829102200102
Kroll, Barbara and Joy Reid (1994). Guidelines for designing writing prompts: clarifications, caveats and cautions. *Journal of Second Language Writing* 3: 231-255. DOI: https://doi.org/10.1016/1060-3743(94)90018-3
Kroll, Barbara (1998). Assessing writing abilities. *Annual Review of Applied Linguistics* 18: 219- 240.
Kuiken, Folkert and Ineke Vedder (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing* 17, 1: 48-60. DOI: https://doi.org/10.1016/j.jslw.2007.08.003
Kyle, Kristopher (2020). The relationship between features of source text use and integrated writing quality. *Assessing Writing* 45: 100467. DOI: https://doi.org/10.1016/j.asw.2020.100467
Leblanc, Catherine and Miho Fujieda (2012). Investigating effects of topic control on lexical variation in Japanese university students' in class timed-writing. *Humanities Review*: 17: 241-253.
Lee, Hee-Kyung and Carolyn Anderson (2007). Validity and topic generality of a writing performance test. *Language Testing* 24, 3: 307–330. DOI: https://doi.org/10.1177/026553220707720
Lee, Sunjung and Diana Pulido (2016). The impact of topic interest, L2 proficiency, and gender on EFL incidental vocabulary acquisition through reading. *Language Teaching Research* 22, 1: 118–135. DOI: https://doi.org/10.1177/1362168816637381
Lee, Yong-Won, Breland, Hunter, and Eiji Muraki (2004). Comparability of TOEFL CBT prompts for different native language groups. *TOEFL Research Reports*, RR-04-24. Princeton, NJ: Educational Testing Service. DOI: https://doi.org/10.1002/j.2333-8504.2004.tb01951.x
Lee, Yong-Won and Robert Kantor (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing* 7: 353–385. DOI: https://doi.org/10.1080/15305050701632247
Li, Hang and Lianzhen He (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly* 12, 2: 178-212. https://doi.org/10.1080/15434303.2015.1011738
Lim, Gad S. (2010). Investigating prompt effects in writing performance assessment. *Spann Fellow Working Papers in Second or Foreign Language Assessment* 8: 95–116.
Linn, Robert L. ed., (1989) (3rd ed.) *Educational measurement.* New York: American Council on Education and Macmillan.
Lo, Julia, and Fiona Hyland (2007). Enhancing students' engagement and motivation in writing: The case of primary students in Hong Kong. *Journal of Second Language Writing* 16: 219-237. DOI: https://doi.org/10.1016/j.jslw.2007.06.002
Messick, Samuel (1989). Validity. In Linn, Robert L. ed., (3rd ed.), 13-103. New York: American Council on Education and Macmillan.
O'Sullivan, Barry and Anthony Green (2011). Test taker characteristics. In Taylor, Lynda, ed., 36-64. Studies in Language Testing 30. Cambridge: UCLES/Cambridge University Press.
Patall, Erika A., Cooper, Harris and Jorgianne Civey Robinson (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. *Psychological Bulletin*, 134: 270–300. DOI: https://doi.org/10.1037/0033-2909.134.2.270
Patall, Erika A. (2012). The motivational complexity of choosing: A review of theory and research. In Ryan, Richard M., ed., Oxford handbook of human motivation (pp. 249–279). Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780195399820.013.0015

Polio, Charlena and Margo Glew (1996). ESL writing assessment prompts: How students choose. *Journal of Second Language Writing* 5: 35–49. DOI: https://doi.org/10.1016/S1060-3743(96)90014-4

Powers, Donald E., Fowles, Mary E., Farnum, Marisa and Kalle Gerritz (1992). Giving a choice of topics on a test of basic writing skills: does it make any difference. The praxis series: professional assessments for beginning teachers, *Educational Testing Service, Research Report*: 92–19.

Powers, Donald E. and Mary E. Fowles (1998). Test takers judgments about GRE writing test prompts (GRE No. 94-13). Princeton, NJ: Educational Testing Service.

Preiss, David D., Castillo, Juan Carlos, Flotts, Paulina and Ernesto San Martín (2013). Assessment of Argumentative Writing and Critical Thinking in Higher Education: Educational Correlates and Gender Differences. *Learning and Individual Differences* 28: 193–203. https://doi.org/10.1016/j.lindif.2013.06.004

Purves, Alan C. (1992). Reflections on research and assessment in written composition. *Research in the Teaching of English* 26: 108–122.

Schneider, Sascha (2021). Are there never too many choice options? The effect of increasing the number of choice options on learning with digital media. *Human Behavior and Emerging Technologies* 3, 5: 759-775. DOI: https://doi.org/10.1002/hbe2.295

Schneider, Sacha, Nebel, Steve, Beege, Maik and Günter Daniel Rey (2018). The autonomy-enhancing effects of choice on cognitive load, motivation and learning with digital media. *Learning and Instruction* 58: 161-172. DOI: https://doi.org/10.1016/j.learninstruc.2018.06.006

Schoonen, Rob (2005). Generalizability of writing scores: An application of structural equation modeling. *Language testing* 22, 1: 1-30.

Schoonen, Rob (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In Van Steendam, Elke, Tillema, Marion, Rijlaarsdam, Gert and Huub Van Den Bergh, eds., 1-22. Leiden: Koninklije Brill. https://doi.org/10.1163/9789004248489_002

Shaw, Stuart D. and Cyril J. Weir (2007). *Examining writing: Research and practice in assessing second language writing.* Studies in Language Testing 26, Cambridge: Cambridge University Press.

Shi, Bibing, Huang, Liyan and Xiaofei Lu (2020). Effect of prompt type on test-takers' writing performance and writing strategy use in the continuation task. *Language Testing* 37, 3: 361-388. DOI: https://doi.org/10.1177/0265532220911626

Skehan, Peter (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, Peter (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics* 30, 4: 510-532. DOI: https://doi.org/10.1093/applin/amp047

Slomp, David H. (2016). An integrated design and appraisal framework for ethical writing assessment. *The Journal of Writing Assessment*, 9, 1: 1–14.

Song, Bailin and Isabella Caruso (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. DOI: https://doi.org/10.1016/S1060-3743(96)90023-5

Spaan, Mary (1993). The effect of prompt in essay examinations. In Douglas, Dan and Carol A. Chapelle, eds., 98-122. Alexandria, VA: Teachers of English to Speakers of Other Research Colloquium Languages.

Spratt, Mary (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research* 9, 1: 5-29. DOI: https://doi.org/10.1191=1362168805lr152oa

Taylor, Lynda, ed., (2011) *Examining speaking: Research and practice in assessing second language speaking*- Studies in Language Testing 30. Cambridge: UCLES/Cambridge University Press.

Tedick, Diane J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes* 9: 123-143. DOI: https://doi.org/10.1016/0889-4906(90)90003-U

Troia, Gary Alan, Shankland, Rebecca K., and Kimberly A. Wolbers (2012). Motivation research in writing: Theoretical and empirical considerations. *Reading and Writing*, 28, 1: 5–28. DOI: https://doi.org/10.1080/10573569.2012.632729

Troia, Gary A., Harbaugh, Allen G., Shankland, Rebecca K., Wolbers, Kimberley A. and Ann M. Lawrence (2013). Relationships between writing motivation, writing activity, and writing performance: Effects of grade, sex, and ability. *Reading and Writing* 26, 1: 17–44. DOI: https://doi.org/10.1007/s11145-012-9379-2

Wainer, Howard and David Thissen (1994). On examinee choice in educational testing. *Review of Educational Research* 64: 159-195. DOI: https://doi.org/10.3102/00346543064001159

Weigle, Sara Cushing (1994). Effects of training on raters of ESL compositions. *Language Testing* 11: 197-223.

Weigle, Sara Cushing (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing* 6: 145-178. DOI: https://doi.org/10.1016/S1075-2935(00)00010-6

Weigle, Sara Cushing (1998). Using FACETS to model rater training effects. *Language Testing* 15: 263-287.

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

Weigle, Sara Cushing. (2012). Assessing writing. In Coombe, Christine A., Peter Davidson, Barry O'Sullivan and Stephen Stoynoff, eds., 218-224. Cambridge: Cambridge University Press.

Weigle, Sara Cushing (2013). Assessment of writing. In Chapelle, Carol A., ed., 1-7. Oxford: Blackwell. DOI: https://doi.org/10.1002/9781405198431.wbeal0056.

Weigle, Sara Cushing, Boldt, Heather and María Inés Valsecchi (2003). Effects of task and rater background on the evaluation of ESL writing: A pilot study. *TESOL Quarterly* 37: 345–354. https://doi.org/10.2307/3588510

Weigle, Sara Cushing and Eric Friginal (2015). Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: Effects of language background, topic, and L2 proficiency. *Journal of English for Academic Purposes* 18: 25-39. DOI: https://doi.org/10.1016/j.jeap.2015.03.006

Wolfe, Christopher, Britt, M. Anne and Jodie A. Butler (2009). Argumentation schema and the myside bias in written argumentation. *Written Communication* 26, 2: 183–209. DOI: https://doi.org/10.1177/0741088309333019

Yang, Weiwei and YouJin Kim (2020). The effect of topic familiarity on the complexity, accuracy, and fluency of second language writing. *Applied Linguistics Review* 11, 1: 79–108. DOI: https://doi.org/10.1515/applirev-2017-0017

Yang, Weiwei, Lu, Xiaofei and Sara Cushing Weigle (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. DOI: https://doi.org/10.1016/j.jslw.2015.02.002

Yoon, Hyung-Jo (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System* 66: 130-141. DOI: https://doi.org/10.1016/j.system.2017.03.007

Yu, Guoxing (2010). Lexical diversity in writing and speaking task performances. Applied Linguistics 31: 236-259. DOI: https://doi.org/10.1093/applin/amp024

Zhao, Changhan and Jinyan Huang (2020). The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: Implications for assessment policy makers. *Studies in Educational Evaluation* 67: https://doi.org/10.1016/j.stueduc.2020.100911