Title of Project:
Validating the Writing Part B of GSEEE:
Investigating the Topic Effect

Researcher:
Lumeng Fang
Beijing Foreign Studies University
fanglumeng@bfsu.edu.cn

Research Supervisor:
Prof. Xiaoying Wang
Beijing Foreign Studies University


Lumeng Fang

TIRF Research Topic Investigated:
Language Assessment

Motivation for the Research

In the past decades, the topic of writing tasks has been widely investigated as one of the fundamental prompt characteristics in both independent writing tasks and integrated writing tasks (Homayounzadeh, Saadat, & Ahmadi, 2019; Lee & Anderson, 2007; Lim, 2009; Weigle & Friginal, 2015). A significant concern for topic-based approach of testing is that test takers might be biased in terms of test performance when presented with a specific topic.

Technically, topic effect refers to the potential threat to the validity of a test that may result from topical factors (Jennings, Fox, Graves, & Shohamy, 1999), including test takers' prior knowledge, perceived relevance, interest, and opinions concerning the topic. The issue of topic effect matters because if topical factors are extraneous to the construct of language assessment, they are regarded as construct irrelevant variables, which would inevitably influence construct validity. Therefore, it is imperative for topic-based tests to investigate the possibility of topic effect as part of the ongoing process of test validation.

The current research contributes to the field of language testing by providing an innovative perspective on topic effect by investigating the possible presence of topic effect in writing tasks from the Graduate School Entrance English Examination (GSEEE), a highly influential language test in China. Therefore, this study is directly relevant to TIRF's current research priorities of language assessment in two aspects. First, the research addresses the issue of a potential construct-irrelevant factor, which echoes TIRF's call for validation for regional or local language assessment procedures and TIRF's commitment to ensure that English as a second or foreign language is tested in a manner that is demonstrably effective, expedient, and economical. Second, this research adopts MFRM and Coh-Metrix to investigate the scoring and written textual features of essays, which enriches evidence for claims of topic effect.

Research Questions

In this study, three tasks were selected as the target topics in educational, social, and personal domains from the original writing tasks of GSEEE, each writing prompt was given a description in several Chinese words: they are *Party at a phone age* (2015) from social domain, *Reading*

books (2017) from educational domain, and *Persistence* (2019) from personal domain, respectively. Three research questions given below will be investigated:

1. Are the three writing tasks from educational, social and personal domains in GSEEE comparable in difficulty?
2. To what extent do different topics affect the scoring of the essays generated in response to the three GSEEES writing tasks?
3. To what extent do the topics affect the textual features of the writing responses?

Research Methodology

Participants

Participants in this study were 45 college students preparing for the entrance test to Graduate Schools at universities in China aging between 20-24. They were 25 females and 20 males enrolled at 11 universities from a range of academic disciplines, including engineering, mathematics, medical science and humanities (all of them were non-English majors). They participated in this research after signing a consent form.

Four raters (three females, one male; L1 Chinese) were involved in the study. Two of the raters (Rater 2 and Rater 3) were experienced university EFL teachers who taught English majors in a prestigious university in China for more than 20 years, one of whom was a specialist in language testing. Rater 4 was a doctoral candidate majoring in language testing. Rater 1 was an MA student who was specializing in language testing.

Procedure

Two rounds of rater training were carried out to guarantee the formal tests and rating. Before the training, materials including the task prompts, rating scales, benchmark samples, and practice samples were handed out to all the raters. The formal test was conducted in a classroom in paper and pencil form with 35 minutes for each task. The GSEEE writing tasks were assigned once a week, and the whole data collection procedure lasted for 4 weeks, with one week for pre-test and three weeks for the formal test. After each test, data were collected and backed up in digital forms by the researcher. Upon receiving all the written essays, the researcher numbered each file and printed them for rating. After kicking out the invalid data (i.e., essays that failed to match the targeted topic), the final set of the data was composed of 39 essays on Topic 1 (social topic), 36 on Topic 2 (personal topic) and 34 on Topic 3 (educational topic), with a total of 109 written essays in response to the GSEEE writing task.

In rating session, this study adopted a fully-crossed design, which means all raters scored all of the 109 scripts on all the five rating criteria. The scores were manually assigned and then entered into Excel spreadsheets for later use. Missing data were found in this study, altogether there were 6450 data points. Multi-facet Rasch measurement (Linacre, 1989) was used for scoring data analysis and Coh-Metrix (Version 3.0) for textual data analysis in the current study.

Summary of Findings

For RQ1, the inferential statistics indicated that no significant group mean was found in the three topic domains. In Rasch analysis, separation index of 0.65, reliability of the index (0.31), chi-square test (p=.10) and fair measure average suggested that the difficulty level of the three tasks failed to separated, they were equally difficult.

With respect to RQ2, the inter-rater reliability was at an acceptable level with a range of .836 to 0.889. The bias/interaction analysis found only one exception of t score greater than 2

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

in which Examinee 5 presented bias in Essay 2 for the personal topic by writing an extensively long essay. Overall, there was no other significant bias or interaction between essay facets and other facets.

To answer RQ3, correlation and regression analysis were conducted. Textual indices of word count, lexical diversity, noun phrase density, and adjective incidence were significant predicators for social topic. Word count, verb cohesion, and content word overlap were significant predictors for personal topic. Word count, text easability passage coherence deep cohesion, and all connective indices were significant score predicators for educational topic. In addition, a high level of similarity and overlap of textual features among the three topics was found, it confirmed the frequent situations in GSEEE writing that a large number of students copy the writing template or widely apply the universal sentence structures taught by coaching programs.

To conclude, writing tasks for the three topics were comparable in task difficulty and no significant topic effect was found in the scoring results. Different significant textual features were found under each topic, but word count was a significant score predicators in all of the three tasks.

Implications

The implications of this current study are discussed with regards to GSEEE policymakers, second language writing instruction practitioners and test candidates respectively. First, it is recommended that test board could enrich the writing task bank to eliminate the possible topic effect and influence from writing templates. To guarantee a fair and efficient rating, it was suggested that policymakers provide guidelines to tackle the writing template phenomenon. Second, English teaching practitioners should pay more attention to essay coherence, repertoire building and originality in writing when giving lectures to students. Finally, findings of the study carry implications for test candidates as well. Writing requires knowledge beyond some salient textual features in templates, a large variety of topic knowledge input and a conscientious learning attitude might benefit each candidate much more in test preparation.

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

References

AERE, APA, & NCME (1996). *Standards for educational and psychological testing.* Washington, DC: AERA.

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Author.

Aryadoust, V. & Liu, S. (2015). Predicting EFL writing ability from levels of mental representation measured by Coh-metrix: A structured equation modelling study. *Assessing Writing, 24*(2), 35-58.

Ay, S., & Bartan, O. Z. (2012). The effect of topic interest and gender on reading test types in a second language. *The Reading Matrix, 12*(1), 62-79.

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

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2* (1), 1-34.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessment and justifying their use in the real world.* Oxford, UK: Oxford University Press.

Bai, Y. (2010). The impact of familiarity on group oral proficiency testing. *CELEA Journal, 32*(2), 114-125.

Bray, B. G., & Barron, S. (2004). Assessing reading comprehension: The effects of text-based interest, gender, and ability. *Assessing Writing, 9*, 107-128.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Chapelle, C. A., Enright, M. K. & Jamieson. M. J. (2008). Test Score Interpretation and Use. In C. A. Chapelle, M. K. Enright, & M. J., Jamieson, (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 1-25). New York, NY: Routledge.

Chen, M. (2018). The effect of topic familiarity and linguistic difficulty on EFL listening comprehension. *Overseas English*, 239-253.

Cho, Y., Rijmen, F., & Novak, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT integrated writing tasks. *Language Testing, 30*(4), 513-534.

Chuang, H. K., Joshi, R. M., & Dixon, L. Q. (2012). Cross-language transfer of reading ability: Evidence from Taiwanese ninth-grade adolescents. *Journal of Literacy Research, 44*(1), 97-119.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66-79.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in l2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1-16.

Davies, A., Brown, C., Elder, K., Lumley, T., & MaNamara, T. (1999). *Dictionary of language testing.* Beijing, China: FLTRP.

Deluca, C., Cheng, L., Fox, J., Doe, C., & Li, M. (2013). Putting testing researchers to the test: An exploratory study on the TOEFL iBT. *System, 41*(3), 663-676.

Ebrahimi, S., & Javanbakht, Z. O. (2015). The effect of topic interest on Iranian EFL learners' reading comprehension ability. *Journal of Applied Linguistics and Language Research, 2*(6), 80-86.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234.

Hamp-Lyons, L. (1986). *Testing second language writing in academic settings* (unpublished doctoral dissertation).. Edinburgh, UK: University of Edinburgh.

Hamp-Lyons, L. (1991). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation.

He, L. (2010). The Graduate School Entrance Examination. In L. Cheng & A. Curtis (Eds.), *English language assessment and the Chinese learner* (pp. 145-157). London, UK: Taylor & Francis.

He, L., & Shi, L. (2012). Topical knowledge and ESL writing. *Language Testing, 29*(3), 443-464.

Hidi, S. (2001). Interest, reading, and learning: Theoretical and practical considerations. *Educational Psychology Review, 13*(3), 191-209.

Hinkel, E. (2002). *Second language writers' test: Linguistic and rhetorical features.* London, UK: Lawrence Erlbaum Associates.

Huang, F. (2002). The effects of topic familiarity and question type on EFL listening comprehension. *Journal of MIFLI, 83*, 11-21.

Huang, H. T., & Hung, S. T. A. (2013). Comparing the effects of test anxiety on independent and integrated speaking test performance. *TESOL Quarterly, 47*(2), 244-269.

Huang, H. T. (2010). *Modeling the relationships among topical knowledge, anxiety, and integrated speaking test performance: A structural equation modeling approach* (unpublished doctoral dissertation). Austin, TX: University of Texas.

Janosik, S. M., & Frank, T. E. (2013). Using e-portfolios to measure student learning in a graduate preparation program in higher education. *International Journal of Eportfolio, 3*(1), 13-20.

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing, 16*(4), 426-456.

James, C. L. (2008). Electronic scoring of essays: Does topic matter? *Assessing Writing, 13*, 80–92.

Ji, X. (2011). Topic effects on writing performance: What do students and their writings tell us? *The Journal of Asia TEFL, 8*(1), 23-38.

Kane, M. (1992). An argument-based approach to validation. *Psychological bulletin, 112*(3): 527-535.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*(3),135-170.

Kane, M. (2006). Validation. In R. Brennan(ed.). *Educational measurement* (4th edition) (pp. 17-64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing, 34*(1), 1-26.

Kobrin, J. L., Deng, H., & Shaw, E. J. (2011). The association between SAT prompt characteristics, response features, and essay scores. *Assessing Writing, 16*(3), 154-169.

Lavallee, M., & Mcdonough, K. (2015). Comparing the lexical features of EAP students' essays by prompt and rating. *Tesl Canada Journal*, *32*(2), 30-44.

Lee, H. K., & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing*, *24*(3), 307-330.

Leeser, J. M. (2004). The effects of topic familiarity, mode, and pausing on second language learners' comprehension and focus on form. *SSLA, 26*, 587-615.

Li, J. (2014). Examining genre effects on test takers' summary writing performance. *Assessing Writing, 22*, 75-90.

Li, J. (2018). Establishing comparability across writing tasks with picture prompts of three alternate tests. *Language Assessment Quarterly, 15*(4), 368-386.

Lim, G. S. (2009). P*rompt and rater effects in second language writing performance assessment.* Doctoral Dissertation. University of Michigan, Ann Arbor, MI.

Lim, G.S. (2010). Investigating prompt effects in writing performance assessment. In J.S. Johnson, E. Lagergren, & I. Plough. (eds). *Spaan fellow working papers in second or foreign language assessment* (Volume 8) (pp. 95-116). Ann Arbor, Michigan: University of Michigan.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: a longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543-560.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2007). *Facets Rasch measurement computer program*. Chicago, IL: Winsteps.com.

McNamara, T. F. (2003). Book review: Fundamental considerations in language testing. Oxford: Oxford University Press, Language testing in practice: Designing and developing useful language tests. *Language Testing*, *20*(4), 466-473.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd edition) (pp. 13-103). New York, NY: MacMillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Researcher, 12*(2), 9-15.

Miller, R. T., Mitchell, T. D., & Pessoa, S. (2016). Impact of source texts and prompts on students' genre uptake. *Journal of Second Language Writing, 31*, 11-24.

Mirshekaran, R., Namaziandost, E., & Nazari, M (2018). The effects of topic interest and L2 proficiency on writing skill among Iranian EFL learners. *Journal of Language Teaching and Research, 9*(6), 1270-1276.

Mo, Y. (2014). *Exploring task and genre demands in the prompts and rubrics of state writing assessments and the national assessment of educational progress (NAEP)*. Doctoratal Dissertation. East Lansing, MI: Michigan State University.

National Education Examinations Authority (2019). *GSEEE syllabus*. Beijing, China: Higher Education Press.

Ong, J., & Zhang, J (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing, 19*(4), 218-233.

Petersen, J. (2009). "This test makes no freaking sense": Criticism, confusion, and frustration in timed writing. *Assessing Writing, 14*, 178-193.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing, 26*(40), 561-587.

Riazi, M. (2016). Comparing writing performance in TOEFL-iBT and academic assignments: An exploration of textual features. *Assessing Writing, 28*, 15-27.

Schmidt-Rinehart, B. C. (2011). The effects of topic familiarity on second language listening comprehension. *Modern Language Journal, 78*(2), 179-189.

Spaan, M. (1993). The effect of prompt in essay examination. In D.Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 98-122). Alexandria, VA: TESOL.

Stapa, M (2001). *Assessing ESL writing performance: the influence of Background Knowledge on writing performance.* Paper presented in Annual Meeting of Midwest Association of Language Takers (MWALT). Ann Arbor, MI.

Tedick, D. J. (1990). ESL writing assessment: subject-matter knowledge and its impact on performance. *English for Specific Purposes, 9*(2), 123-143.

Tavakoli, P. (2009). Investigating task difficulty: Learners' and teachers' perceptions. *International Journal of Applied Linguistics, 19*(1), 1-25.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument* (updated edition). Cambridge, UK: Cambridge University Press.

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

Weigle, S. C., & Friginal, E. (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.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* New York, NY: Palgrave MacMillan.

Wiseman, C. S. (2009). *Rater decision-making behaviors in measuring second language writing ability using holistic and analytic scoring methods.* Paper presented at the annual meeting of the American Association for Applied Linguistics, Denver, Colorado.

Wolfe, E. W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27,* 1-10.

Yang, H. Z., & Weir, C. (1998). *Validation study of the National College English Test.* Shanghai, China: Shanghai Foreign Language Education Press.

Yang, W., Lu, X., & Weigle, S.C. (2015) Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing, 28,* 53-67.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236–259.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

何莲珍、孙悠夏 (2015) 提示特征对中国学生综合写作任务的影响研究，《外语教学与研究》(2)：237-250。

刘婷婷 (2015) 话题熟悉度和生词率对高中英语阅读成绩的影响，博士学位论文。南昌：江西师范大学。

罗凯洲 (2019) 整体效度观下语言测试四种效度验证模式: 解读、评价与启示.《外语教学》40 (6): 76-81。

孙悠夏 (2016) 综合写作测试的效度验证：提示特征的影响研究. 博士学位论文。杭州：浙江大学。

熊丹 (2013) 2005-2012 年考研英语阅读理解 B 部分内容效度分析，博士学位论文。长沙：湘潭大学。

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

张才丽 (2011) 考研英语传统阅读理解的测试内容效度分析——以 2010 年考研英语(一)传统阅读理解为例，《考试周刊》 (55)： 11-13。

张新玲、周燕 (2014) 任务类型对中国英语学习者写作表现的影响，《现代外语》 (4)：548-558。

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org

177 Webster St., # 220, Monterey, CA 93940 USA
Web: www.tirfonline.org / **Email:** info@tirfonline.org