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

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 GSEE 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
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 predictors 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 predictors 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 predictor 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.
References


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。
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。