Shih, JL. et al. (Eds.) (2023). Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Using Multidimensional Analysis to Investigate the Extrapolation Inference of a High-Stakes Test

Shengshu Lin
School of Foreign Languages, Nanjing University of Posts and Telecommunications, China
shengshu2016@hotmail.com

Abstract: The degree of correspondence of situational characteristics between test tasks and target domains is believed to affect how similar language use on tests is to the target domains (i.e., extrapolation). Multidimensional analysis (MDA) has gained in popularity in extrapolation research because it not only interprets language use functionally associated with situational characteristics but also provides a quantitative method to investigate co-occurring linguistic features. Using the automatic tool Multidimensional Analysis Tagger (MAT), the present paper applies MDA to explore the linguistic features between writing of the National Entrance Test of English for MA/MS (NETEM) and disciplinary writing on four dimensions in Biber's (1988) seminal work. Overall, NETEM writing differed significantly from all four disciplinary domains on Dimension 1 and Dimension 4, and showed similarities with Humanities and Arts papers but differed from the other three domains on Dimension 2 and Dimension 3. NETEM writing was weak in its extrapolation to real academic contexts.

Keywords: Multidimensional analysis, writing, linguistic features

1. Introduction

Whether test tasks elicit linguistic features that students will use in real academic contexts is an important extrapolation concern in argument-based validity (Kane, 2012). Take TOEFL iBT for example. The inclusion of source-based integrated writing task in 2005 makes TOEFL Writing Section more representative of language use in real learning environments (Llosa, Grapin, & Friginal, 2019). The underlying assumption of extrapolation inferences is that similar situational characteristics of the target domain elicit the linguistic features associated with it (LaFlair & Staple, 2017; Kyle et al., 2021). The emphasis on communicative functions associated with situational characteristics in tests justifies the application of Multidimensional analysis (MDA) (Biber, 1988) to the investigation of extrapolation of test performances to those in the target domain. MDA takes the register perspective, believing that some linguistic features are more frequently used in a register because "they are functionally adapted to the communicative purposes and situational contexts of texts from that register" (Biber & Conrad, 2019, p. 2). More importantly, this method provides a corpus-based quantitative method to compare a large number of linguistic features with shared communicative functions (LaFlair & Staple, 2017; Yan & Staples, 2019). The present study is the first empirical research that employs this method to examine the linguistic features elicited in the writing task of a high-stakes test in China. The purpose of this study is twofold: 1) to demonstrate how MDA can be used to investigate the extrapolation inference of this test; 2) to discuss the implications for test developers and learners.

2. Research Background

Language users make systematic lexical and grammatical choices appropriate to the situational characteristics of the register (Biber, 2019). Utilizing computational techniques, MDA takes a bottom-up approach to study a large number of fine-grained lexico-grammatical

features distributed in a register and identify co-occurrence patterns with the same underlying functional underpinnings (Biber, 2019). The systematic co-occurring sets of linguistic features are factors or dimensions, and should be interpreted functionally with respect to situational characteristics. It is based on co-occurrence linguistic features that important register differences can be revealed and explained.

In recent years, MDA has attracted mounting attention in validation research, particularly in the investigation of extrapolation inferences because MDA not only extends TLU domain but serves as a quantitative framework for examining the linguistic relevance between test performances and the domains to which they are extrapolated (LaFlair & Staples, 2017; Staples et al., 2018). Weigle and Friginal (2015) was the first empirical research using additive MDA to examine the comparability of TOEFL Independent writing with disciplinary papers along four dimensions established by Hardy and Römer (2013). Results indicated that TOEFL Independent writing was characterized by more narrative expressions of opinion and stance, and expressions of possibility, whereas the latter relied more on linguistic features associated with procedural information. They concluded that Independent writing task was limited in its generalization to real academic contexts. These findings were supported by Llosa et al. (2019) who used the same four dimensions and extended their research by comparing TOEFL Integrated writing with disciplinary papers. Llosa et al. found that this writing task complemented Independent task by eliciting linguistic features associated with information and argumentation. Based on these findings, researchers claimed that TOEFL Writing Section approximated successful academic writing than Independent task alone. Utilizing the four dimensions by Gardner et al. (2019), Staples et al. (2018) compared TOEFL writing with the narrower registers such as case studies and argumentative essays in disciplinary writing. Again, the previous findings about Independent writing were supported. However, Staples et al. pointed out that although TOEFL Integrated writing was better representative of disciplinary writing, its overall patterns were more similar to Independent writing than they were to most disciplinary registers. As a result, the extrapolation of TOEFL writing was still a problem.

The current study contributes to this area of study by investigating Writing Task 2 of the National Entrance Test of English for MA/MS (referred to as NETEM writing throughout the paper) in China, a national high-stakes test that aims to select candidates with required English proficiency for future intellectual pursuits. NETEM with millions of test takers is regarded as the most challenging English test for non-English majors. NETEM writing is an essay of 160-200 words based on one cartoon or a set of cartoons. Test takers are required to describe the cartoon(s) briefly, interpret the meaning, and give comments. The differences of NETEM writing from disciplinary papers in situational characteristics including the topic, the purpose, and the production circumstances may result in different choices of linguistic patterns. Our research into the linguistic features in these two situations will inform language testing and teaching. The following research question will be addressed:

How similar are the linguistic features of NETEM writing to those of graduate level disciplinary papers?

3. Methodology

3.1 Corpus

The corpora for the current study consist of NETEM writing corpus and a reference corpus. The former contains samples of NETEM writing from the past ten years' (2013-2022) real papers. These papers were compiled by ten different educational training and services providers and published by some well-known publishing companies in China. Although NETEM directions suggest 160-200 words, test takers are encouraged to produce over 180 words in order to score high. In the present research, the type-token ratio in MAT was set at 180 and four essays less than 180 words were excluded. 96 essays were used with a total of 20,891 words. The reference corpus is the Michigan Corpus of Upper-Level Student Papers (MICUSP). It is a corpus of A-graded papers by native and non-native English speakers at four levels of studies, and can be divided into four academic categories (Weigle & Friginal,

2015). See Table 1. The present study uses papers by graduate students in that they are in line with what NETEM candidates are expected to produce in future studies.

Table 1. *NETEM Writing and Disciplinary Writing*

Five domains	Number of papers	Tokens (words)
NETEM writing	96	20,891
Biological and Health Sciences	92	254,335
Humanities and Arts	78	374,569
Physical Sciences	79	248,383
Social Sciences	150	494,020

3.2 Additive MDA and Dimensions of Linguistic Features

Biber's (1988) dimensions have become stable references for additive MDAs across various domains (Nini, 2019; Berber Sardinha et al., 2019) in that this study described the overall patterns of register variation based on a wide range of language use situations. Additive MDAs allow researchers to focus on one dimension or several dimensions of interest in their studies. Another strength is that it makes research possible based on relatively small restricted corpus. This is the case in our study. The current study focuses on the first four dimensions because they have more explanatory power (Biber, 2019). The four dimensions are: Involved vs. Informational Production; Narrative vs. Non-Narrative Concerns; Explicit vs. Situation-Dependent Reference; Overt Expressions of Persuasion. In this study, 48 linguistic features were used in calculating mean scores of each dimension. Refer to Biber (1988, 2014) for details of the individual linguistic features. Dimension 1 and Dimension 3 include positive and negative features, which suggest they are distributed in a text in a complementary manner.

3.3 MAT and Data Analysis

All the fine-grained features on the four dimensions should be tagged and computed. Multidimensional Analysis Tagger (MAT) is an automatic tool developed by Nini (2019) that replicates the original Biber Tagger (not publicly available). It uses the Stanford Parser to parse lexical, syntactic and semantic information of the linguistic features in Biber (1988) and effectively replicates Biber's algorithms and results. The reliability of MAT is an important driving factor for the application of Biber's dimensions in additive MDAs (Berber Sardinha et al., 2019). This study uses MAT version 1.3.2 to map the five domains onto Biber's 4 dimensions. We used MAT to complete 3 of the steps in Berber Sardinha et al. (2019): 1) tagging linguistic features in each text, 2) computing and standardizing their normed frequency counts, 3) computing the dimension scores for the five domains. SPSS was used to compare the differences in frequency counts for the linguistic features, and means on each dimension of the five domains. Despite the normal distribution of data in this study, Levene's test of equality of variance was significant on all four dimensions. Therefore, Tamhane's T2(M) post-hoc tests, which don't assume equality of variance, were run to determine where significant differences were found.

4. Results

MAT classified NETEM writing as general narrative expositions while graduates' papers as scientific expositions. This indicated NETEM writing used different language features from disciplinary papers. ANOVA results showed significant differences on all four dimensions: Dimension 1, $F(4, 220.24) = 43.223$, $p = .000$; Dimension 2, $F(4, 223.356) = 26.616$, $p = .000$; Dimension 3, $F(4, 219.012) = 18.350$, $p = .000$; Dimension 4, $F(4, 218.466) = 25.215$, $p = .000$. The means on four dimensions are displayed in Table 2 and visually shown in Figure 1, where the shared letters on each dimension represent mean scores that are not significantly different from each other. Post-hoc results with illustrative excerpts will be reported as follows.

Table 2. *Descriptive Statistics (Means and SDs) across Dimensions and Domains*

	Dimension 1 (S.D.)	Dimension 2 (S.D.)	Dimension 3 (S.D.)	Dimension 4 (S.D.)
Biological and Health Sciences	-14.44 (5.54)	-3.23(1.54)	7.58 (2.86)	0.56 (3.81)
Humanities and Arts	-11.16 (6.64)	-1.86 (1.61)	5.42(2.52)	-1.06 (2.83)
Physical Sciences	-15.19 (4.57)	-3.9 (1.36)	6.56 (2.83)	-0.46 (2.90)
Social Sciences	-13.06 (4.93)	-2.51(1.57)	7.64 (2.39)	-0.42 (2.66)
NETEM Writing	-4.04 (7.23)	-1.39 (3.36)	4.69 (4.18)	4.71 (5.07)

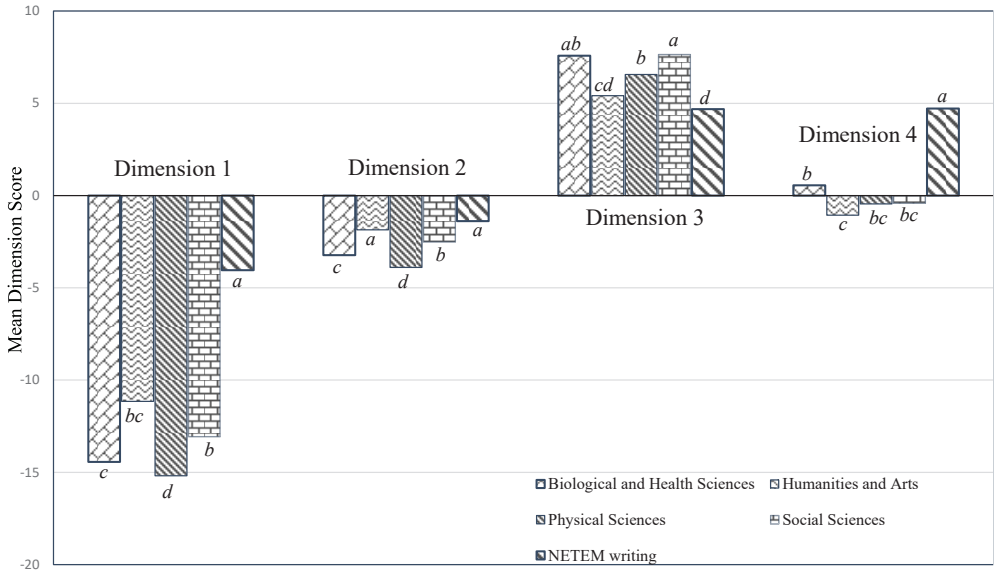

Figure 1. Mean Scores on Four Dimensions across Domains

On Dimension 1, Table 2 shows the mean score of NETEM writing is -4.04, significantly higher than four disciplinary papers, indicating the involved and non-informational focus. This textual style can be illustrated through the reliance on present verbs, 1st person pronoun and possibility modals. Expert 1 illustrates the use of some of these features.

Excerpt 1: To be exact, mobile phones have occupied so much of our time that we even don't have time to talk with people around us! In this sense, we can say that mobile phones have separated us from each other…(NETEM writing, Dimension 1 score = -9.29)

Contrary to these features, Excerpt 2 from Physical Sciences is characterized by negative features such as nouns, attributive adjectives, and prepositions. Meanwhile, word length as a negative feature suggests the use of advanced vocabulary in disciplinary writing.

Excerpt 2: If a simple transport model is sought, it is recommended that a spatially varying reaction rate be employed with a constant diffusion coefficient. Future work could incorporate hydrologic models to assess the impact of stormwater surges on contaminant transport. (CEE.G3.03.1 Dimension 1 score = -21.89)

On Dimension 2, the mean score of NETEM writing is -1.39. Figure 1 indicates no difference between NETEM writing and Humanities and Arts papers. They both used more narrative features than the other three disciplinary groups. Narrative features like 3rd person pronoun, present participial clauses are very common in NETEM writing (see Excerpt 3).

Excerpt 3: … he is convinced that it is his interest to explore the deep space that significantly contributes to his present achievement, making it easier for him to survive in this competitive commercial world. (NETEM writing, Dimension 2 score = 3.1)

On Dimension 3, NETEM scored 4.69, close to Humanities and Arts papers but significantly lower than the other three groups. This low score indicated more use of context-dependent features. Explicit references such as "nominalization and pied piping construction"

characteristic of informational writing are significantly less used in NETEM writing. Instead, it used more place adverbials and other adverbials as illustrated in Excerpt 4, while nominalization was found to be a salient feature in all disciplinary papers.

Excerpt 4: Many students are facing various choices ahead …studying abroad… (NETEM writing, Dimension 3 score = -0.26)

On Dimension 4, the mean of NETEM writing is 4.71. Figure 1 indicates NETEM writing is clearly distinct from all disciplinary writing. The higher the score is, the more persuasive the textual style is. NETEM writing used many linguistic features associated with persuasion such as infinitives and obligation modals as a way of suggesting what should be done to solve the problem mirrored in the prompt. Excerpt 6 is an example. Also, split auxiliaries are common in NETEM writing, usually in the form of formulaic expressions such as "as is vividly shown/illustrated in the picture" and "It is universally held that", to name a few.

Excerpt 6: College students should not miss any opportunity to acquire knowledge to achieve their all-round development. (NETEM writing, Dimension 4 score = 5.83).

5. Discussion and Conclusion

Using corpus linguistic method MDA, the present study found that NETEM writing differed significantly from all four disciplinary domains on dimensions 1 and 4, and showed similarities with Humanities and Arts but differed from the other three domains on dimensions 2 and 3

Regarding Dimension 1 and Dimension 4, NETEM writing was characterized by interpersonal interactions and persuasion, while disciplinary papers used more linguistic features associated with information density. These differences can be primarily explained by the differing purposes in writing tasks. Disciplinary tasks as a type of source-based writing focus on the presentation of propositional information while NETEM writing requires test takers to describe the cartoons vividly and argue convincingly for the viewpoints generated. Given that Dimension 1 is the most basic variation among registers in nearly all MDA studies (Biber, 2014) and therefore, an important marker for varying degrees of information density, a weak link exists between NETEM writing and real academic contexts. That is, it is impossible to predict how learners will perform linguistically on these two dimensions based on test scores (LaFlair & Staples, 2017). It is worth noting that the major syntactic functions of the linguistic features with negative weights in Dimension 1 are phrasal modifiers in noun phrases. The dense use of phrasal modifiers is the unique style of academic prose (Biber et al., 2011). Yet these features are disproportionately underrepresented in NETEM writing. Just as Kyle et al. (2021) pointed out "…if key linguistic features of a particular domain are not required for the successful completion of a task, support for the extrapolation inference is also weakened", NETEM writing failed to elicit the language patterns required in graduate level studies.

The alignment of NETEM writing with Arts and Humanities papers on Dimension 2 and Dimension 3 is not surprising since this domain consists of many papers from philosophy, history and classical studies where narration and description are common. This finding is consistent with Weigle and Friginal (2015), Llosa et al. (2019), and Staples et al. (2018). However, similarities on these two dimensions do not support the extrapolation of test task to other disciplines because the sharp differences from disciplinary writing on other dimension(s) still make the authenticity of the task problematic (Staples et al., 2018).

In addition to the nature and requirements of NETEM writing, the stereotype of academic language as clausally complex is also at work, influencing the choices in NETEM writing of linguistic features different from disciplinary papers. For example, the bestseller among NETEM candidates published by a renowned press in China encourages the use of participial clauses and adverbial clauses as a way to score higher. Some education and training providers encourage test candidates to use more personal examples as an effective way to argue for viewpoints, thus making narrative features salient. All this is no different from placing an obstacle to learners' academic language development.

The present study also affirms that independent writing alone is not representative of linguistic features that learners are expected to learn in the university setting. Although NETEM did undergo changes in test tasks, cartoons as writing prompts has been used since

the 1980s. It is time to re-examine and reform the test task. Admittedly, it is hard to solve the extrapolation problem once and for all. Yet test developers should give more attention to source-based writing skills, an important academic literacy in today's higher education.

Acknowledgements

This study was supported by the 11th China Foreign Language Education Fund (ZGWYJYJJ11Z025) and Research on Philosophy and Social Sciences of Jiangsu Universities in 2022 (2022SJYB0101).

References

Berber Sardinha, T., Veirano Pinto, M., Mayer, C., Carolina Zuppardi, M., & Herique Kauffmann, C. (2019). Adding register to a previous Multi-Dimensional analysis. In T. Berber Sardinha, & M. Veirano Pinto (Eds.), *Multi-dimensional Analysis: Research Methods and Current Issues* (pp.165-186). Bloomsbury Academic.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, *14*(1), 7-34.

Biber, D. (2019). Multi-dimensional analysis: A historical synopsis. In T. Berber Sardinha, & M. Veirano Pinto (Eds.), *Multi-dimensional Analysis: Research Methods and Current Issues* (pp.11–26). Bloomsbury Academic.

Biber, D., & Conrad, S. (2019). *Register, Genre, and Style*. Cambridge University Press.

Biber, D., Gray, B., & Poonpon., K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5-35.

Gardner, S., Nesi, H., & Biber, D. (2019). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing *Applied Linguistics*, *40*(4), 646-676.

Hardy, J. A., & Römer, U. (2013). Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-Level Student Papers (MISCUP). *Corpora*, *8*(2),183-207.

Kane, M. T. (2012). Articulating a validity argument. In G. Fulcher, & F. Davidson (Eds.), *The Routledge Handbook of Language Testing* (pp. 34-47). Routledge.

Kyle, K. A. T., Choe, A. T., Eguch, M., LaFlair, G., & Ziegler, N. (2021). A comparison of spoken and written language use in traditional and technology-mediated learning environments. *ETS Research Report Series*, (1),1-29.

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, *34*(4), 451-474.

Llosa, L., Grapin, S. E., Friginal, E., Cushing, S. T., & Malone, M. E. (2019). Linguistic dimensions of TOEFL iBT essays compared with successful student disciplinary writing in the university. *TESOL Quarterly*, *54*(1), 251-265.

Nini, A. (2019). The multi-dimensional analysis tagger. In T. Berber Sardinha, & M. Veirano Pinto (Eds.), *Multi-dimensional Analysis: Research Methods and Current Issues* (pp. 67-94). Bloomsbury Academic.

Staples, S., Biber, D., & Reppen, R. (2018). Using corpus-based register analysis to explore the authenticity of high-stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. *The Modern Language Journal*, *102*(2), 310-332.

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.

Yan, X., & Staples, S. (2019). Fitting MD analysis in an argument-based validity framework for writing assessment: Explanation and generalization inferences for the ECPE. *Language Testing*, *37*(2), 189-214.