RESEARCH

Test-Section Relative Strength Analysis: An Exploratory Study of Second-Year Students Core English Course Results

Kasin Janjaroongpak
CGEL, Thai-Nichi Institute of Technology
1771/1 Pattanakarn Rd., Suan Luang, Bangkok, Thailand 10250
Email: kasin@tni.ac.th

Received: 2020-04-06 Revised: 2020-07-13 Accepted: 2020-09-08

Abstract

With relative variations, formal paper-and-pencil tests in Thailand, especially nation-wide entrance examinations–O-NET, CU–TEP, TU–GET, NIDA TEAP, NT– were generally concerned with 4 test types: vocabulary, grammar, indirect conversation, and reading. While a large proportion of both students and teachers has been grappling with test specifications–how to beat the test–, to the best of my knowledge, little attention has been paid to examine the extent of relations amongst the four pillars. (Whittome, 2019) This paper attempted to shed light on the relations among test sections in a formal examination whether there was an indicative section to the total score. This exploratory study was conducted with the test results of more than 700 second-year students in a course that every student needed to pass. The results were exported and processed with simple regression statistics tools of Microsoft Excel by testing the relative strength between two test sections with the total score. The preliminary analysis suggested that the section designed with TOEIC as a model was a bellwether, statistically significant at the 0.05 level, for the overall achievement compared with the other section. A brief expository account was provided as well as possible pedagogical implications for developing an achievement examination.

Keywords: TOEIC, simple regression, test, quantitative research, ESP

Introduction

Traditionally, formal tests in Thai higher education were broadly organized in the two-phased system, a midterm exam which was widely used as a reliable yardstick for academic progress and a final exam as an overall assessment of the course. Compared with the midterm, students were concerned with the test result of the final exam as it was the key predictor of their

academic achievements. In this vein, with students' career prospects in mind, the test should be revisited and verified after the result has been produced to ensure that the test was both accurate and reliable.

Amongst a plethora of English test papers in Thailand, multiple choice format was one of the most popular forms used while the written open-ended questions or essay-like questions appearing in Cambridge IGCSE (International General Certificate of Secondary Education) were mainly used in English-major subjects (Whittome, 2019). Learners and instructors have been focusing on discovering the best practice in scoring points, in the fastest and correct manner. On the other hand, test makers developed their tests in a way that they maximally and accurately reflected natural linguistic abilities of the test takers, in other words, a test that could not be tricked by tutoring. In this sense, as far as I was concerned, little research has been studied on the relations within the test components. It was important to establish the correlations amongst the 4 categories as it would help both learners and teachers to direct their attentions and resources to an appropriate component of the test that has a strong influence on the total scores as well as specific parts of the test.

The test results in this study came from the final paper of second-year student for an English subject of Thai-Nichi Institute of Technology. There were 5 sections in the test: dialogue, vocabulary, grammar, reading, and TOEIC.

In this piece of exploratory study, I would like to shed light on what section was strongly predicative of the overall performance.

Literature Review

Background

According to (Grabe & Stoller, 2019), vocabulary was the key predictor of reading comprehension while other factors such as grammar or schematic knowledge played less important roles. In listening, (Rost, 2015) argued that lexical segmentation or mental lexicon was a defining factor in listening ability. The final test paper started with dialogue section, followed by vocabulary, reading, grammar, and TOEIC sections respectively. Dialogue test or indirect-speaking test came first because it was considered the easiest part of the test. Vocabulary and reading sections are bundled together as they had a clear correlation in that comprehension grew as the vocabulary size increased (Coxhead, 2017; Hinkel, 2016). Grammar and TOEIC sections sat together as the part taken from TOEIC (Part 5) was primarily concerned with syntax. The key difference between these two components was the fact that the former dealt with grammar test based on the textbook, *Business Plus 3*, whereas TOEIC section was a mock-test of authentic TOEIC test, which appears to be more difficult than its counterpart.

TOEIC section was added to this particular course because of two primary reasons. First, this course was considered a business English course and TOEIC was a standardized test

designed to be reflecting receptive communicative ability in business environment. Secondly, TOEIC was expected and intended to be used as an exit-exam for the institute; therefore, it is important to know in advance to what extent the students were ready for taking the actual test in their fourth year.

Dialogue test was arguably an indirect test of interactive communication, speaking and listening. The institute has championed this pair of skills by providing a dedicated session, one hour, each week for international staff to give lectures following the core course book. In this sense, students were familiar with the topics and expressions concerned which were to appear in the final test paper. In terms of dialogue, the test items were in a form of long conversation between two or more people discussing about business topics appearing in the course book.

In addition, dialogue was long argued to be one of the easiest components of multiple choice test in that it generally reflected the natural language production in daily life conversation and it also included various aspects of language, indirectly signaling the extent of test takers' linguistics competencies. For another thing, national entrance exam (General Aptitude Test) sets dialogue test with approximately 15 items of its overall assessment compared with the other receptive-skill-oriented 45 items, suggesting that its difficulty threshold might be lower than the other.

On the contrary, although TOEIC was front and center, owing to syllabus constraints, it could not be directly incorporated into the course book, resulting in it appearing only in the final paper but disappearing from the classroom session. To avoid test-what-you-teach situation, supplementary materials were provided for students to study on their own at the beginning of the semester. Thus, it was possible to assume that the achievement in TOEIC section was virtually from students' autonomous learning.

TOEIC

In terms of vocabulary, TOEIC was business-oriented (Browne, Culligan, & Phillips, 2016). Students need to study a specialized group of vocabulary to perform in the test. The acceptable scores for TOEIC was 550 (Tannenbaum & Wylie, 2019) (CEFR: B1). On this front, TOEIC was reportedly more difficult compared with dialogue.

Furthermore, TOEIC expects test takers to not only perform linguistically but also possess well-organized time management especially reading section which includes incomplete sentences, cloze test, and reading passages. Candidates were expected to efficiently allocate time for each section within 75 minutes, causing knowledge-intense students with poor-management and good-management students with limited knowledge failed to do well on the test.

The test items in TOEIC section were similar to Incomplete Sentences section of TOEIC. Students are required to fill in the gap by choosing from multiple choice questions. The test specification ranged from vocabulary to collocations.

In this vein, I would like to anticipate that TOEIC should be a bellwether of the total

score because it requires both greater depth of knowledge and strategic decisions.

Methodology and Result

The 150 test items were divided into 5 sections and each section had 30 items. Having summed up the test results in each section in MS Excel, basic descriptive statistics was processed. The results suggested that there were two significantly low mean scores in both dialogue section and TOEIC section, approximately 40% each. Therefore, these two were being examined in greater detail.

The population of this study was 727 pre-intermediate sophomore students, aged between 18 to 20, from a higher education institute in Thailand enrolling in a required course of the institute. Students were from the three faculties in the institute, Faculty of Engineering, Faculty of Business Administration, and Faculty of Information Technology. The students passed their first and second English compulsory courses as a pre-requisite before studying this course. Purposive sampling was used as these second-year students would study this course as their last compulsory English course at the institute. This would ensure that results were the end-stage of their learning at the institute. The test results of both dialogue and TOEIC sections of 727 students were processed using Data Analysis function available in MS Excel. Simple regression laid out in (Brown, 2012) (Larson-Hall, 2012)was applied by pairing dialogue section with the total score and TOEIC section with the total score respectively. To get further insight into the difference among test takers, variance was confirmed by using VAR excel function. Also, to shed light on the distribution pattern, TOEIC-total score correlation and Dialogue-total score correlation were calculated.

Results

Table 1

Dialogue-Total Score Statistical Analysis

Dialogue	
Simple Regression R square	≈ 0.50
Simple Regression p-value	0.00
Correlation	0.70
Variance	≈ 9.9

Table 2

TOEIC-Total Score Statistical Analysis

TOEIC	
Simple Regression R square	≈ 0.70
Simple Regression p-value	0.00
Correlation	0.82
Variance	≈ 22.9

Table 1 has shown that dialogue section has had statistical significance in relation to total score with p-value at >0.05, the correlation has suggested a relatively strong directional movement between the two factors at 0.70. Variance was at approximately 10.

Table 2 has reported that TOEIC section has indicated a relatively high statistical significance when paired with total score with p-value at > 0.05, the correlation has been observed at 0.82 and the variance has been indicated at almost 23.

Discussion and Implications

Discussion

With simple regression p-value both at 0.00, it is possible to argue that dialogue and TOEIC sections are contributing factors of the total score, which was suggested by the literature. Interestingly, TOEIC section R square was set at ≈0.70 compared with ≈0.5 of dialogue section R square, which pointed out that TOEIC closely correlated with the total score though TOEIC was not directly included in the core teaching materials. This implied that test takers striving to achieve high overall score needed to pay particular attention and perform well in TOEIC. A closer look into variance seemed to align with this argument in that TOEIC had the higher variance, roughly 23, compared with almost 10 for dialogue section, showing that there was a clear distinction among participants in TOEIC but less so in dialogue. In other words, in dialogue section, no matter if they were high or low achievers, the scores were relatively the same but in TOEIC section only high achievers could score points.

In terms of differences between dialogue section and TOEIC section, although both dialogue and TOEIC were considered a test in context in that they were not an elicitation or accuracy test like in minimal pair test in phonology, the variety of sources is different. The sources of dialogue were an adaptation from the dialogue appearing in the course book, resulting in closer variance proximity. On the other hand, TOEIC was adapted from various authoritative commercial sources, rendering it challenging for low achievers; therefore, acting as a defining factor for the overall achievement.

For another thing, the level of recognition between dialogue and TOEIC tests appeared to be significantly different. Dialogue test was not internationally recognized as a valid language test. Only certain countries, Thailand in particular, used it in an important test such as national standard examinations but TOEIC has been acknowledged as an accepted standard especially in business environment. One notable example was a dedicated TOEIC-score filtering search in a Thai recruitment agency JobsDB (jobsDB, 2019), which was one of the largest agencies in Thailand. Not only was TOEIC accepted in business but also in certain academic institute. One case in point would be an admission requirement at renowned universities in Japan (Ritsumeikan University, 2020) (University of Tsukuba, 2016) stating that students pursuing a particular discipline may submit TOEIC score for screening process, arguably on a par with TOEFL and IELTS as proof of English proficiency for studying.

Apart from the validity of the test, the objective of the test might be another source. According to Bloom's taxonomy (Bloom, 1956), most of dialogue test items were testing an ability to remember. For example, students were expected to choose from multiple choices among which were *"I just want to talk", "he just wants to say", "we have to say", and "I just want to tell"*. Students who were able to answer correctly needed to memorize the dialogue. On the other hand, a large number of TOEIC test items involved various linguistics components to choose the right choice. For instance, *"_____ Mr. Lee works…, he still hasn't been…"* was testing on the use of conjunction. Students not only had to understand the meaning of both subordinate and main clauses but also understand the relations between the two immediate constituencies as well as selecting the right connector, among those with similar meaning, to express such proposition.

Moreover, the proficiency level required for dialogue was not comparable with TOEIC. Dialogue test was mainly concerned with turn-taking. For example, after a dialogue line, "Thank you very much for your help", a test taker was expected to simply browse the most suitable reply such as "Never mind", "Talk to you soon", and so forth. At the same time, TOEIC was designed with accumulative learning in its design in that a test taker needed to have an extensive collection of linguistics knowledge to complete the task. Incomplete sentence test, word choice test in particular, could not be correctly answered without solid knowledge of word selection. For instance, the sentence, "Nash Library ____ announces…" requires one of the following words, "extremely, proudly, distantly, previously". In terms of part of speech, every choice was acceptable but only the one that is both grammatical and collocational could be chosen.

On readability front, according to Flesch-Kincaid Grade Level (Lindhout, Teunissen, & Lindhout, 2012) processed via MS Word, the level was reported at 3.3 for dialogue while it was 11.1 for TOEIC. It was possible to argue that the dialogue might be so assessable that the test cannot discriminate the level of participants, resulting in low dominance in total score whereas TOEIC was approximately at the range of 11-grade U.S. students. In CEFR terms, the former was A1 but the latter was B1-B2. It could possibly be proclaimed that the level of the TOEIC text was not appropriate for the learner's level in the first place, considering the overall Flesch-Kincaid Grade Level at approximately 6.3 of the whole test paper. On a close examination, this subject was designed to help develop, ideally, learners' proficiency to B1 by the end of the course. Although the level of TOEIC text was above the overall readability score, it corresponded with the course objective.

Another source that might account for the relations was an insight from a focus group describing test takers' direct experience. For dialogue, the conversation used in the test was a recurring one, not a brand new dialogue. Memorization was the key to answer the questions rather than linguistics knowledge. In addition, the dialogue was in general one page long which was tedious for students, given that the final paper contains 32 A4 pages. Potentially, students

might give answers at random rather than carefully study the questions before selecting the best answer, reflecting its distant correlation with the total score. For another thing, some students informed that the context embedded gave away too much information, especially turn-taking fixed idiomatic expressions such as “Good Morning”. Instead of reading and grasping the gist of the conversation, students selected a non-context sensitive answer.

For TOEIC, though the test was considered difficult for students, they said that it required considerable knowledge to answer each question, judging from not only the intensity of business vocabulary that appeared in the paper but also its novelty. For instance, some students said that a formal connector i.e., “regardless”, was new to them so prior preparation was needed. Also, they contended that incomplete sentences were less intimidating compared with dialogue because each test item was one or two lines. They could examine each in great detail before making a decision. This gave a more accurate picture of learners’ profile.

Pedagogical Implications

This preliminary research shed light on the telling factors of the total score by using simple regression. Statistical analysis suggested that TOEIC section was a bellwether for total score rather than dialogue. One of the key elements of TOEIC’s incomplete sentence was its analytic nature which gave rise to the ability to discern and accurately judge both sentential grammaticality and appropriate language use. To improve the test, each section should contain analytical component rather than memorization. For further study, other domains of the test should be put under consideration to establish the strength of influence each section has on the total score in order to better design and maintain an appropriate balance among sections in the test paper.

For dialogue section (Tim, 2000), there were two approaches to be considered, a radical and a moderate method.

Dialogue test was supposed to be a proficiency test, not an achievement test, since identical recurrence of a text did not facilitate learning or bolster assessment accuracy, making it irrelevant. The business of turn-taking, discursive competence, was far more complex than simply anticipating rhetorical moves on paper removed from conversational reality. To prove this hypothesis, a piece of thorough study to investigate the correlations between scores from dialogue section test and interactive conversational tests based on the same referential content should be conducted.

For a more modest measure, dialogue section could be developed by incorporating more business-oriented environment into the test items. Instead of daily conversation in the workplace which was indistinguishable from general English, such as greetings, test items should include idiomatic expressions in business (McCarthy & O'Dell, 2005) such as *“go part time”, “carve a niche for myself”, “moving up the ladder”, “fit the job description”, and “realize his potential”.* Another possible source to consult was commercial TOEIC books under listening sections especially short talk and monologue sections, given that this core

course was intended to be an introduction to business English, an ESP course, which involved specific set of language features (Anthony, 2018; Biber & Conrad, 2009). As a newcomer, this unfamiliar genre needed to be introduced and made as a prime to make sure that learners would experience a smooth transition into the business world where English was arguably the lingua franca. On the other hand, technical terms and jargons were moving towards the central ground of language use in daily life as internet of things (IOT) was increasingly ubiquitous, resulting in a swift from being a jargon to plain English lexicon. This development did not take place in business English as a large proportion of business jargons, especially corporate terms, remained highly technical. One case in point would be the fact that many students knew the term, *"export"*, meaning making a file available and accessible on another platform but only business administration students knew the term as an antonym to the term, *"import"*. On a broader scale, expressions belonging to technology domain were more familiar than that of business. For instance, students were comfortable to deal with an expression, *"freeze up",* in a sentence, *"My computer has frozen up"*, but they struggled with "*My bank account was frozen."*

Another major consideration that should be taken into account was the length of each conversation. There was a proposal stating that in short talk section of TOEIC, there should be no more than four turns (Craven, 2013). Therefore, it was quite unusual and out-of-range to use an excessively long 14-turn conversations in a given test paper. In addition, real-life conversation by and large varied in size and shape meaning some were quite long while others were brief and concise. Having only extended conversation was not only tiresome but also counterintuitive. One notable example to adapt was to redesign the dialogue in line with TOEIC by mixing among stimuli-response, short talk, and monologue. Apart from referring to a standardized test format, test format in dialogue should also reflect social development (Tim, 2000). For instance, people were using social media as well as surfing the internet almost all the time. In this setting, not only was sound presented via embedded videos but also visuals. For another thing, in workplace communication, especially in a multinational context, telecommunication was ubiquitous via telephone in particular which now and then required both parties to both look at a data set, diagram, table, map, or infographic while performing turn-taking. In this sense, pure verbal-simulation dialogue test should be revisited.

Another point of investigation was the qualitative analysis of the test items which learners found difficult; an error analysis. Rigorous scrutiny over low-correct-response items could unearth the zone of proximal development in learners as a whole as they were shared problems among test takers (Lloyd & Fernyhough, 1999). Another source that could be useful was a focus group on particular test specifications and it worked both ways in that a sizeable number of students who could not answer correctly could have a chance to explain their think-out-loud methods deriving the answers and those who could answer could provide insights to what steps were taken to come up with the right answers.

Moreover, the test results established that TOEIC could not be addressed solely by low

proficient learners. Proper scaffolding mechanism should be in place; therefore, a session dedicated to TOEIC should be incorporated. In terms of scope and sequences of TOEIC session, as time constraint could set in when it came to formal education, only those identified marked problems would be targeted to optimize the time spent on each session. Following the change, an in-house supplementary material should be developed to accommodate learners' needs since outright commercial books might not be able to cater the needs or target specific problems of a particular group of learners. Some might argue that the all in-house material could risk running into practical concerns such as, updating paces or consistency. Commercial books in line with learners' problems could be a valid alternative. For instance, learners at high-elementary level often faced vocabulary issues given that TOEIC was an ESP in nature. A commercial text focusing on building lexico-grammar should be prime rather than a grammar-oriented publication.

A quick survey over TOEIC test items revealed that a substantial number of test specifications were testing the command of conjunctions. To tackle this, a graded reader in business context should be put into consideration because it gave context of the proper use of conjunctions as well as embedded context. Instructors might source appropriate materials for students to study, such as English newspapers published in non-native countries.

References

Anthony, L. (2018). *Introducing English for specific purposes.* London: Routledge.

Biber, D., & Conrad, S. (2009). *Register, genre, and style.* Cambridge University Press.

Bloom, B. (1956). *Taxonomy of educational objectives, handbook 1: Cognitive domain.* New York, NY: Addison-Wesley Longman Ltd.

Brown, J. D. (2012). *Understanding research in second language learning : A teacher's guide to statistics and research design.* New York: Cambridge University Press.

Browne, C., Culligan, B., & Phillips, J. (2016). *TOEIC service list*. Retrieved from New General Service List: http://www.newgeneralservicelist.org/toeic-list

Coxhead, A. (2017). *Vocabulary and English for specific purposes research.* London: Routledge.

Craven, M. (2013). *Pass the TOEIC test advanced course.* Fulbourn: First Press ELT.

Grabe, W., & Stoller, F. L. (2019). *Teaching and researching reading.* New York: Routledge.

Hinkel, E. (2016). *Handbook of research in second language teaching and learning.* New York, NY: Routledge.

jobsDB. (2019). *JobsDB*. Retrieved from https://th.jobsdb.com/th

Larson-Hall, J. (2012). How to run statistical analyses. In A. Mackey, & S. M. Gass, *Research methods in second language acquisition : A practical guide* (pp. 245-274). Malden: Wiley-Blackwell.

Lindhout, P., Teunissen, T., & Lindhout, M. (2012). Quick effective CEFR level evaluation of

complete documents: Integrated readability appraisal of text and graphics with the L-scale algorithm. *International Journal of Language Studies*, 37-56.

Lloyd, P., & Fernyhough, C. (1999). *Lev Vygotsky : Critical assessments.* London: Routledge.

McCarthy, M., & O'Dell, F. (2005). *English collocations in use:How words work together for fluent and natural English : Self - study and classroom use.* New York: Cambridge University Press.

Ritsumeikan University. (2020). *How to apply.* Retrieved from International Admission: http://en.ritsumei.ac.jp/file.jsp?id=426743

Rost, M. (2015). *Teaching and researching listening.* New York: Routledge.

Tannenbaum, R. J., & Wylie, E. C. (2019). *Mapping the TOEIC® tests on the CEFR.* Princeton, NJ: Educational Testing Service.

Tim, M. (2000). *Language testing.* New York: Oxford University Press.

University of Tsukuba. (2016). *Undergraduate application information (Japanese programs)*. Retrieved from University of Tsukuba: http://www.tsukuba.ac.jp/en/study-tsukuba/under-graduate/undergrad_exam_schedule

Whittome, E. (2019). *Cambridge international AS & A level literature in English coursebook.* Cambridge: Cambridge University Press.