SCIENTIFIC PROOF AND CASE STUDIES: THE EFFECT OF FREQUENT WRITING ON ONE'S WRITING SKILLS

Dr. Sangjun Han

Vantage Technologies Knowledge Assessment
New Hope / United States

Abstract

"Frequent writing improves one's writing skill" is a hypothesis that has long been accepted as truth without any solid evidence. But is it true? Has any teacher or student kept track of all their writing works with grades? What affects one's writing skills? What kinds of role do factors such as gender, language fluency, ethnicity, and other demographic data play in writing improvement? No one has ever been able to answer these questions due to various reasons. The most compelling reason was the difficulty of collecting writings from people of diverse background and demographic information and yet ensuring the objectivity of the scoring standards as if all of them had been graded by one expert scorer.

This paper illustrates the effect of frequent writing practices on the overall quality of writing using the unique tool called MyAccess. MyAccess is an instructional writing tool that can grade any student's essays in a couple of seconds and return the holistic and five different domain scores as well as specific feedbacks on how a student can improve his/her essay. The massive amount of data from students in multiple grades and English proficiency levels with different genders and ethnic background has been collected and analyzed to test the long-time unchallenged hypothesis.

Keywords - Innovation, technology, research projects.

1 INTRODUCTION

With the evolving technology, we face the challenges of producing better writing more frequently than ever. Compared to 20 years ago when the sole method of written communication was the handwritten (or typed) letter, we write e-mails, send text messages, chat online almost everyday. The importance of writing continues to grow. We tend to write more these days and speak less at schools, at workplaces, and even at home. This new paradigm has raised the question of "how should we prepare for an era where the importance of writing will become higher and higher every day?" The most classic answer is "write as much and frequently as possible." But is it true? If we practice writing often, will it really improve our writing skills? Unfortunately, due to many unavoidable constraints, no one has been able to give a confident "Yes" to this question. To test the hypothesis, we need a pioneer who can invest time, money, and effort to force people to write, grade all the writings, save the data, and to continue the entire cycle until a sufficient number of writings is collected. But there are more challenges. The pioneer also must secure the consistency and reliability of scores. The writings must be graded as if one expert human scorer would handle them in his very best mind. He can neither get tired nor be influenced by any endogenous or exogenous factors; he simply has to be a grading machine. In addition, the sample writings have to be collected from people with diverse background to be a true representative of the population and to prevent the potential biases rising from a specific group of people. It is not difficult to understand why very few, if any, tried to pursue this interesting topic for research purposes.

Thanks to the technology and our writing software, we were able to overcome all the challenges. We collected over 2 million scored writings from over 700 thousand users that vary in gender, grades, ethnicity, and English proficiency. The methodology, tool, design of experiment, and comprehensive results are described in this paper.

2 AUTOMATED ESSAY SCORING

The core enabler of this research is Automated Essay Scoring (AES) technology. AES uses a machine to score written essays. Compared to traditional human scoring, AES provides immediate delivery of scores and feedback; customized tasks and scoring systems; consistent and bias-free scoring; and

enormous reduction in time and costs of scoring by teachers or professionals. Studies testing agreement rates on a "true" score between humans and computers consistently show that AES delivers higher congruence than human scoring [1]. Vantage's patented IntelliMetric™ is world's most advanced AES technology that is endorsed by myriad of prestigious institutions including Graduate Management Admission Council (GMAC) and Association of American Medical Colleges (AAMC).

2.1 IntelliMetric™

IntelliMetric™ is an automated essay scoring tool developed by Vantage Learning that uses Artificial Intelligence, Natural Language Processing, and Statistics in its scoring of essays. Since 1998, it has been used successfully to score open-ended essay-type assessments to become the first commercially thriving tool able to administer open-ended questions and provide feedback to students in a matter of seconds. Hundreds of studies have been conducted to evaluate the quality of IntelliMetric™ scoring. Agreement rates (exact, adjacent, and discrepant) with expert human scorers and correlations between IntelliMetric™ and human scores are the most common methods of evaluating the quality of IntelliMetric™ and other automated essay scoring engines. In essence, the expert human scoring is a baseline for the quality of automated essay scoring engines. IntelliMetric™ has been shown to be as accurate as or more accurate than expert scorers. In other words, IntelliMetric™ is able to agree with expert human scorers more often than experts agree with each other [2]. IntelliMetric™ emulates the process carried out by human scorers. IntelliMetric is theoretically grounded in a cognitive model often referred to as a "brain-based" or "mind-based" model of information processing and understanding. IntelliMetric draws upon the traditions of Cognitive Processing, Artificial Intelligence, Natural Language Understanding and Computational Linguistics in the process of evaluating written text. Among the key tools employed in this process are Natural Language Processing, Statistics and Machine Learning. The system must be "trained" with a set of previously scored responses with known scores as determined by experts. These papers are used as a basis for the system to "learn" the rubric and infer the pooled judgments of the human scorers. The IntelliMetric™ system internalizes the characteristics of the responses associated with each score point and applies this intelligence to score essays with unknown scores.

IntelliMetric™ has begun to have major impact on both classroom instruction and large-scale assessment. With virtually instantaneous electronic scoring, IntelliMetric™ dramatically reduces the cost and time required to evaluate student and professional writing. Moreover, IntelliMetric™ improves the instructional process by offering more frequent and immediate feedback to writers [3]. IntelliMetric™ shares much in common with the holistic scoring systems commonly employed to score large-scale writing assessments. Typically, a group of individuals asked to score essay papers are provided with examples of each score point determined by experts. After internalizing the characteristics associated with each score point and demonstrating calibration with the expert-assigned scores, the group is asked to score the remaining papers whose scores are unknown. Much like human scorers who are generally trained on each specific question or prompt, IntelliMetric™ creates a unique solution for each prompt. This process leads to high levels of agreement between the scores assigned by IntelliMetric™ and those assigned by human scorers. IntelliMetric™ learns the characteristics of the score scale through exposure to examples of essay responses previously scored by experts. In essence, IntelliMetric™ internalizes the pooled wisdom of many expert scorers. IntelliMetric™ benefits from the "expert judgments" reflected within the set of papers used to train the engine, not any single scorer's judgment. Since IntelliMetric™ scoring is a synthesis of many expert opinions it is more reliable (yet may not agree with any single opinion as reflected in a score for a particular paper).

IntelliMetric™ can be used for standardized assessments where a single essay submission is required as well as for various instructional applications where a student can provide multiple submissions of an essay response and receive frequent feedback. IntelliMetric™ Mentor, a complement to the IntelliMetric™ scoring engine, offers various editing and revision tools such as a spell checker, grammar checker, dictionary, and thesaurus. The IntelliMetric™ tool provides feedback on overall performance, diagnostic feedback on several rhetorical and analytical dimensions of writing (e.g., conventions, organization), and detailed diagnostic sentence-by sentence feedback on grammar, usage, spelling and conventions.

2.2 MyAccess!™

MyAccess! is a web-based instructional writing tool that fully leverages IntelliMetric's AES technology. In addition to score feedbacks (Fig. 1) on holistic and domain scores IntelliMetric provides, MyAccess! provides text feedbacks to allow the users to learn why they have received the scores in general and in

specific domains (Fig. 2). MyAccess! text feedback also encompasses the revision guides on how users can improve their writings. Specific feedback and revision suggestions are offered to users in essential five domains. These domains consist of Focus & Meaning, Content Development, Organization, language Use, Voice & Style, and Mechanics & Conventions. The content within each domain that MyAccess!TM analyzes and judges the scores upon are described:

Focus & Meaning (Focus): The extent to which the response establishes and maintains a controlling idea (or central idea), an understanding of purpose and audience, and completion of the task.

Content Development (Content): The extent to which the response develops ideas fully and artfully using extensive, specific, accurate, and relevant details. (facts, examples, anecdotes, details, opinions, statistics, reasons, and/or explanations)

Organization: The extent to which the response demonstrates a unified structure, direction, and unity, paragraphing and transitional devices.

Language Use, Voice & Style (Language): The extent the response demonstrates control of conventions, including paragraphing, grammar, punctuation, and spelling.

Mechanics & Conventions (Conventions): The extent to which response demonstrates an awareness of audience and purpose through effective sentence structure, sentence variety, and word choice that creates tone and voice.

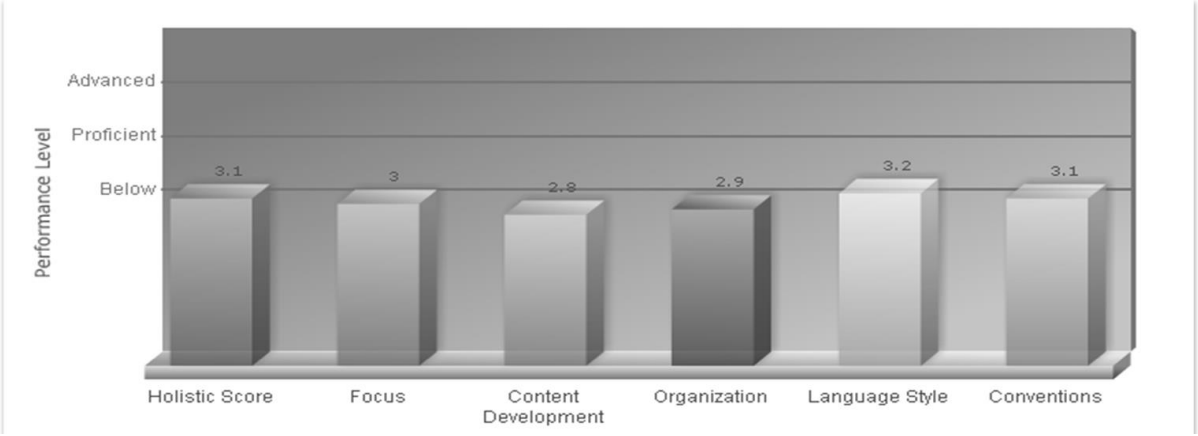

Figure 1: MyAccess! Score Feedbacks on Holistic and Domain Levels

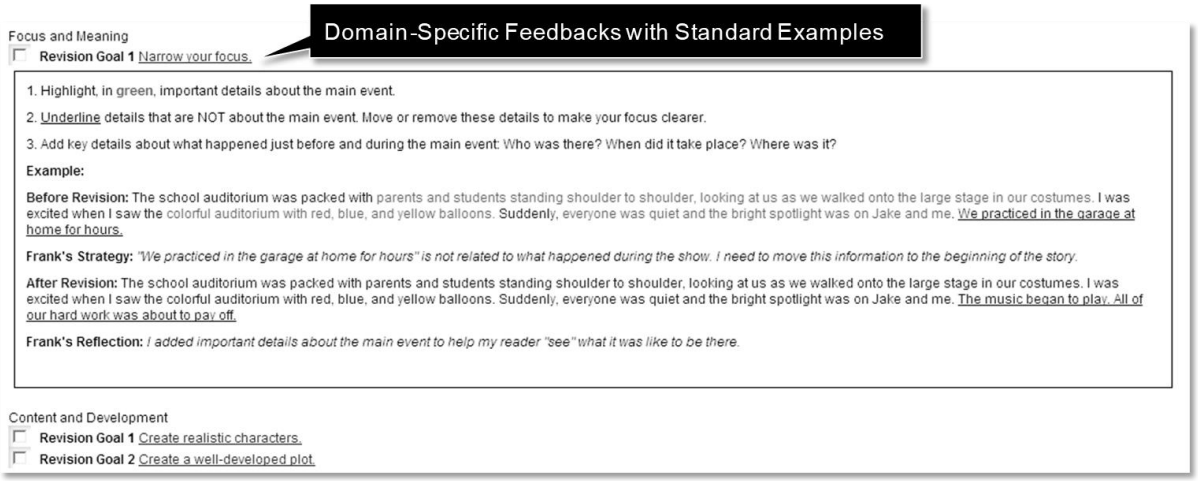

Figure 2: MyAccess! Text Feedback on Domains

3 RESEARCH METHODOLOGY

The purpose of this research was to determine if the frequent writing affects the improvement in writing skills. This study investigates the average score improvement per revision coupled with the total score improvement. MyAccess! provides web-based word processing interface for users to write and submit

their writings. As soon as the essays are submitted, IntelliMetric scores them within 2~3 seconds and returns the scores – both holistic and domain – to the users. In a testing environment, users are only allowed to submit their essays once. In instructional settings, users can receive the scores to their essay along with text feedback focused on how they can improve their essays. The users can revise essays either limited number of times or as many times as they want depending on the pre-settings of MyAccess!. The average score improvement per revision is defined as a discrepancy between the scores of newly and previously submitted essays for the identical prompt. The total improvement per prompt indicates the discrepancy between the scores of finally and initially submitted essays. For example, let's assume that a user's essay was scored 2.4 in his 1st attempt. He revises his essay and obtains 3.4 for his next submission. Not satisfied, he revises his essay again and obtains 4.4. In this case, the total improvement will be 2.0 (4.4 – 2.4: the score of the last submission subtracted by the score of the first submission) unless he decides to add more revisions to his essay. The average improvement per revision will be 1.0 (2.0 / 2 revisions: total improvement divided by total number of revisions).

3.1 Data Collection

The data used as a basis of this research was collected through MyAccess!. Within the database, MyAccess! possesses total of 2,024,518 writings during the period between 2006. 01. 01 and 2009. 02. 28. The number of unique users in the same period is 770,882. During this period, the average number of writings per user is 2.4. Users vary dynamically in their background information by gender, grade, ethnicity, and English fluency. The data fields used to analyze and categorize the data are shown in Table 1.

Table 1: Data Field

Data Field	Description
Vantage ID	Used for Protection of the user's private information
Grade	4~12
Gender	Male / Female
Ethnicity	American Indian or Alaska Native / Black, Not of Hispanic Origin / Hispanic / Asian or Pacific Islander / White, not of Hispanic Origin
English Fluency	First language / Second language / Limited English proficient / Non-English proficient
Country	Demographic Level 1
State	Demographic Level 2
Region	Demographic Level 3
District	Demographic Level 4
School	Demographic Level 5
Prompt Category	Narrative / Informative / Descriptive / Persuasive
Essay Raw Score	Holistic
Domain Score 1	Focus
Domain Score 2	Content
Domain Score 3	Organization
Domain Score 4	Language
Domain Score 5	Convention

MyAccess! has two score scales – 4 point and 6 point scales. All selected data only contains 6 point scale because it is easier to identify the distinctive improvement with the 6 point scale than with the 4 point scale. Each score within MyAccess! can be interpreted as English Proficiency level. The mapping of scores and their indications is given in Table 2.

Table 2: Score Scale and Indication

Score	Indication
6	Excellent

5	Advanced Proficient
4	Proficient
3	Marginal
1~2	Below Proficient

3.2 Sampling

From the population, the data from 2008. 01. 01 to 2008. 12. 31 were selected to conduct the analysis. The number of unique users was 350,401 and the total number of writings was 964,055, yielding the average number of writings per user to be 2.74 during this period. The sample was selected for three reasons. One, they are the latest updated one-year data within our database. Two, the period of one year was selected to eliminate any potential seasonality issue in writing practice. Due to different curriculum schedules in international environment, US students may write more intensively during certain months of year whereas Asian students may write more in different months. Choosing one year duration eliminates such seasonality and could provide more general outcomes. Three, the proper mix of existing and new users are taken into consideration. Since its launch in 2006, MyAccess! successfully accumulated students over time. An existing user can be more adapted to the system whereas new users may not be familiar with the system and this may affect the writing skills through MyAccess! By choosing the latest set of data, this potential bias could be prevented.

4 RESEARCH FINDINGS

To isolate as many unique influences that come from users' diverse background and that could bias the results as possible, the research was conducted and viewed from multiple angles. The average improvement per revision and the total improvement per prompt were investigated from multiple categories including gender, grade, English proficiency, and ethnicity.

4.1 General Statistics

The average improvement per revision in holistic, focus, content, organization, language, and convention were 0.25, 0.22, 0.19, 0.19, 0.20, and 0.20, respectively (Fig. 3), and the total improvement per prompt were 1.03, 0.90, 0.80, 0.78, 0.84, and 0.82, respectively (Fig. 4).

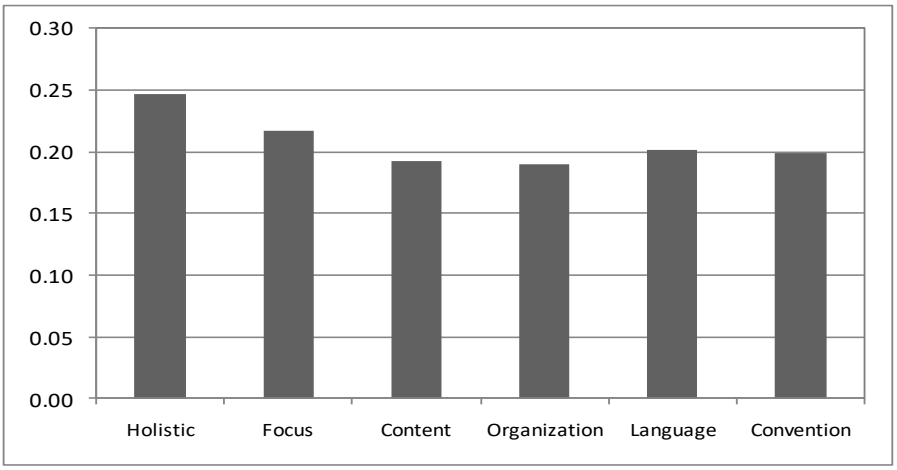

Figure 3: Average Improvement per Revision for All Sample Data

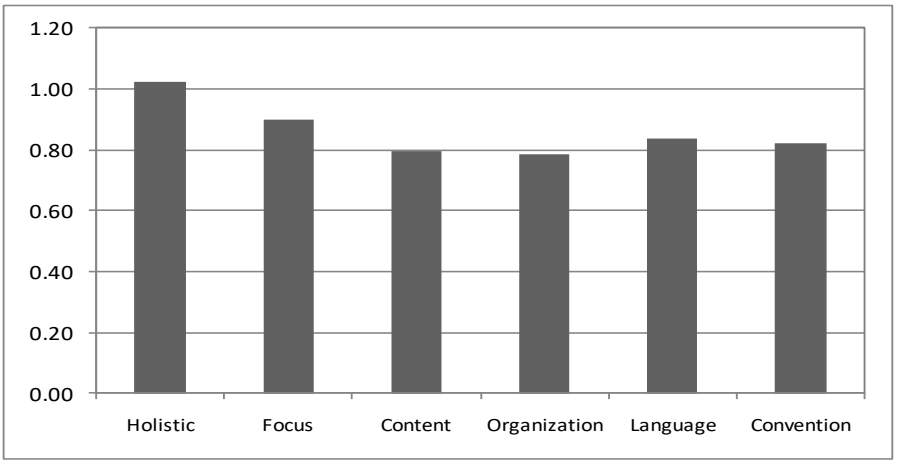

Figure 4: Total Improvement per Prompt for All Sample Data

4.2 Categorized Statistics

The following sections describe the further investigation into whether or not diverse background that 350,000+ users have effects on the hypothesis of "frequent writing improves one's writing skills." The results of analyzing average improvements per revision and total improvement per prompt within categories of gender, grade, English proficiency, and ethnicity are presented.

A. Results by Gender

The average improvements per revision in holistic were 0.25 and 0.24, respectively for female and male (Fig. 5). The total improvements per prompt were 1.01, and 1.03, respectively for female and male (Fig. 6). Although female scored slightly higher in average improvement per revision, male scored slightly higher in total improvement in holistic and all domain scores. The average difference in all scores of average improvement per revision and total improvement per prompt are 0.01 and 0.02, respectively. This is only 4% and 1.9% of the average scores.

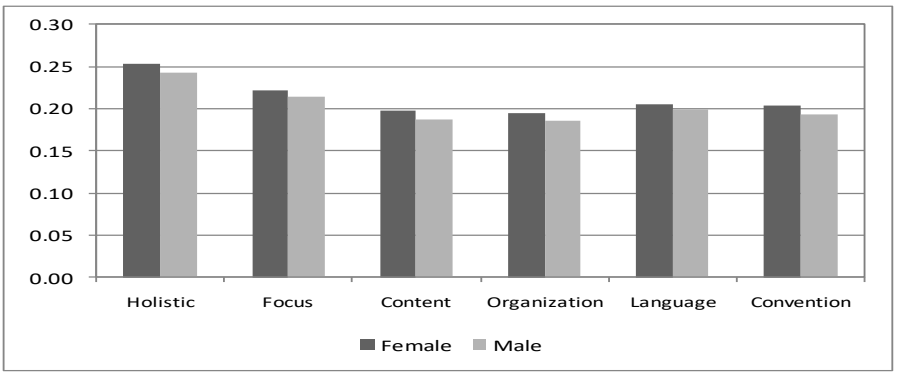

Figure 5: Average Improvement per Revision by Gender

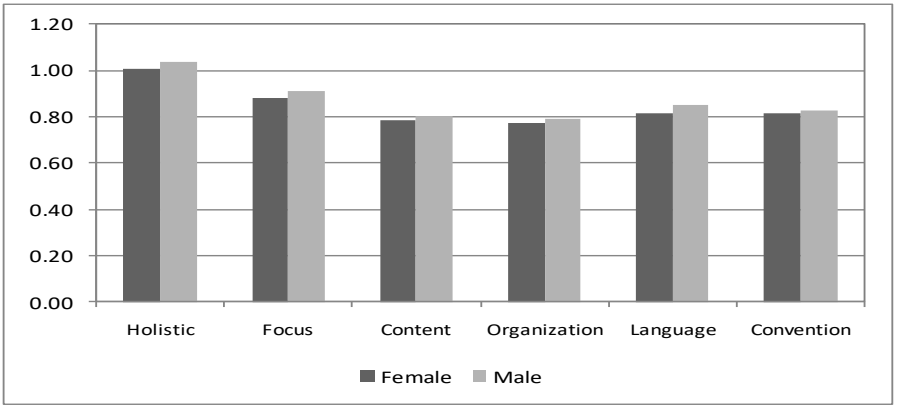

Figure 6: Total Improvement per Prompt by Gender

B. Results by Grade

The average improvements per revision in holistic were 0.22, 0.21, 0.22, 0.22, 0.24, 0.28, 0.27, 0.33, and 0.35 for grade 4 to 12 (Fig. 7). The total improvements per prompt were 1.15, 1.07, 1.10, 0.94, 1.01, 1.02, 0.94, 0.97, and 0.95 for grade 4 to 12 (Fig. 8). Grade 12 scored the highest in average improvement per revision. However, Grade 4 improved most in total scores. The difference is significant in this case since the maximum difference in average improvement per revision and total improvement per prompt are 0.14 and 0.21, respectively. Though a further study must be conducted, we can induce, solely from data, that lower graders are likely to benefit more from using MyAccess in total improvement and that higher graders will achieve faster improvement in writing.

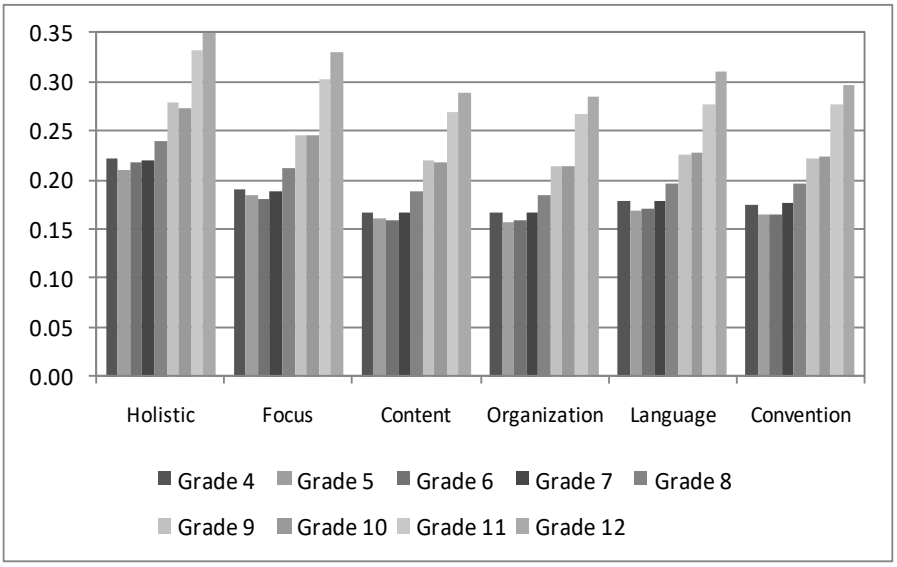

Figure 7: Average Improvement per Revision by Grade

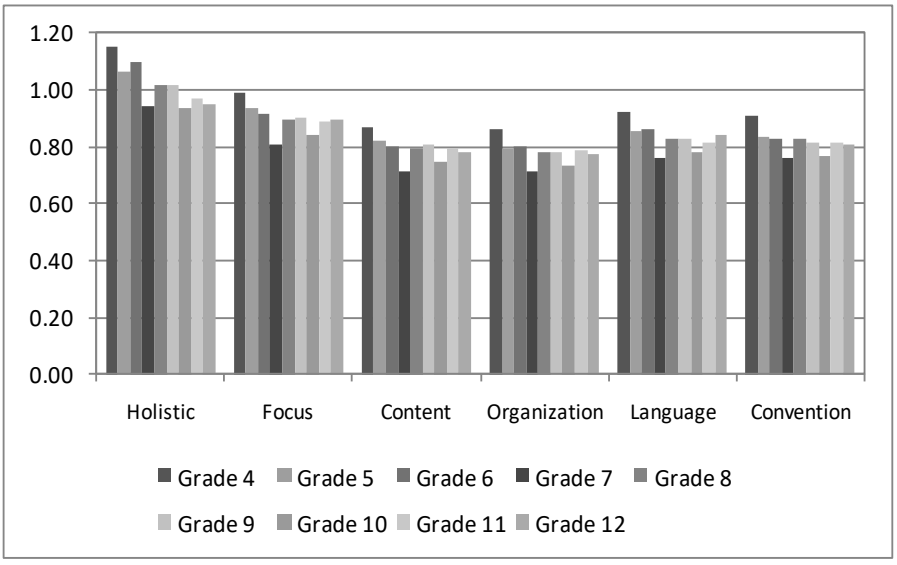

Figure 8: Total Improvement per Prompt by Grade

C. Results by English Proficiency

The average improvements per revision in holistic were 0.28, 0.29, 0.27, 0.24, respectively for students with English as their first language, as a second language, with limited proficiency, and with non-English proficiency (Fig. 9). The total improvements per prompt were 0.99, 1.16, 1.14, and 1.02, respectively in the same consequence (Fig. 6). Students who speak English as their second language achieved highest improvement in both average improvement per revision and total improvement per prompt in most of the categories.

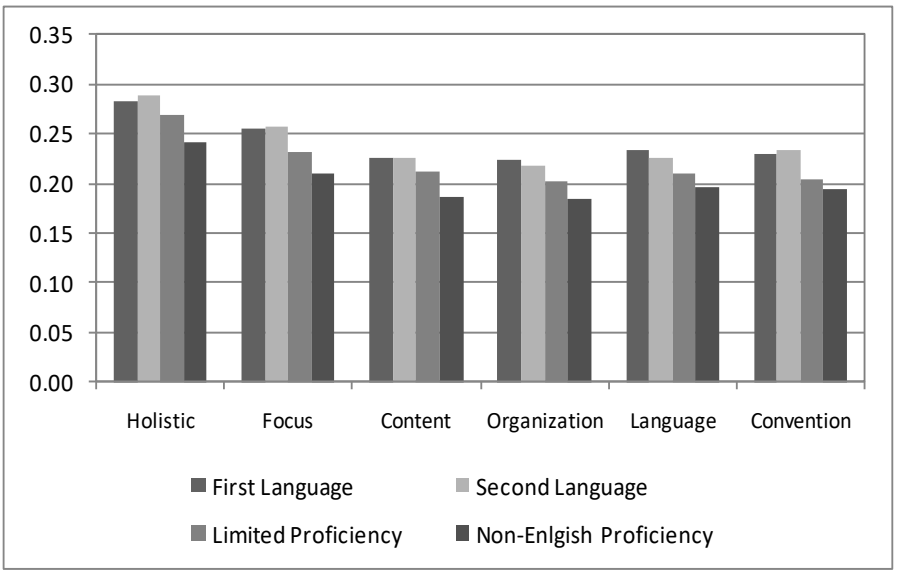

Figure 9: Average Improvement per Revision by English Proficiency

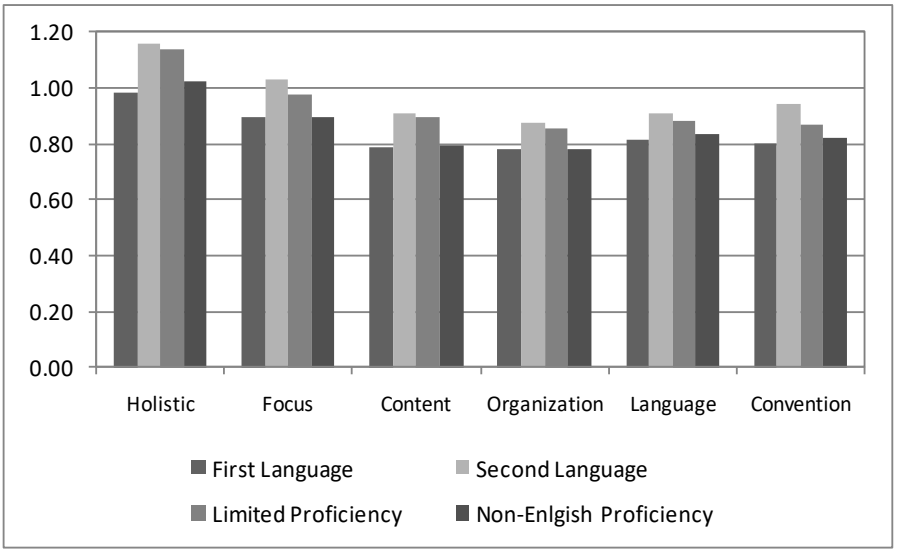

Figure 10: Total Improvement per Prompt by English Proficiency

D. Results by Ethnicity

The average improvements per revision in holistic were 0.28, 0.23, 0.28, 0.23, and 0.28, respectively for American Indian, Black, Hispanic, Asian, and White people (Fig. 11). The total improvements per prompt were 1.19, 1.18, 1.11, 0.93, and 1.09, respectively in the same order (Fig. 12). American Indians, Hispanic, and White achieved similar level of improvement in average improvement per revision and American Indian and Hispanic achieved the highest total improvement within a very close range.

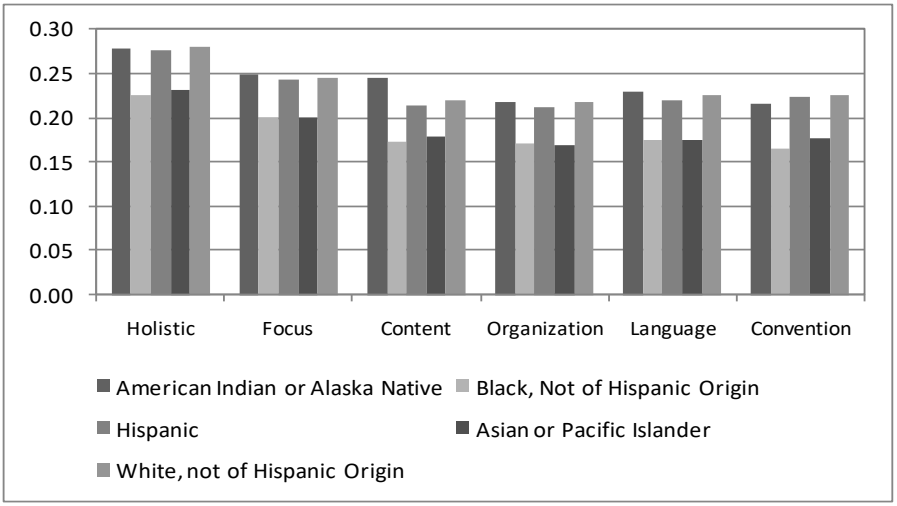

Fig. 11: Average Improvement per Revision by Ethnicity

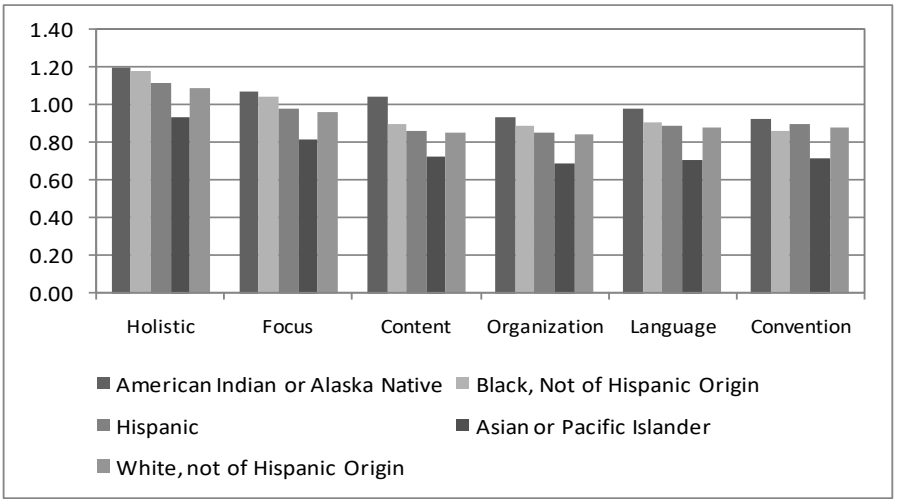

Fig. 12: Total Improvement per Prompt by Ethnicity

4.3 Summary

The findings of this research show a concrete trend that regardless of diverse background of each student, all of them accomplished, on average, a high growth in improving their English writing skills. All results are summarized in Fig. 13 and 14. Though there are discrepancies in what one category person can achieve (Fig. 13), all the students in different categories are closely gathered around improving the total score by 1.0 (Fig. 14). As can be seen from Table 2, this means that each student was able to "upgrade" their writing skills to the next level; those who were at below proficient level to marginal level, from marginal level to proficient level, from proficient to advance proficient, and from advanced proficient to excellent.

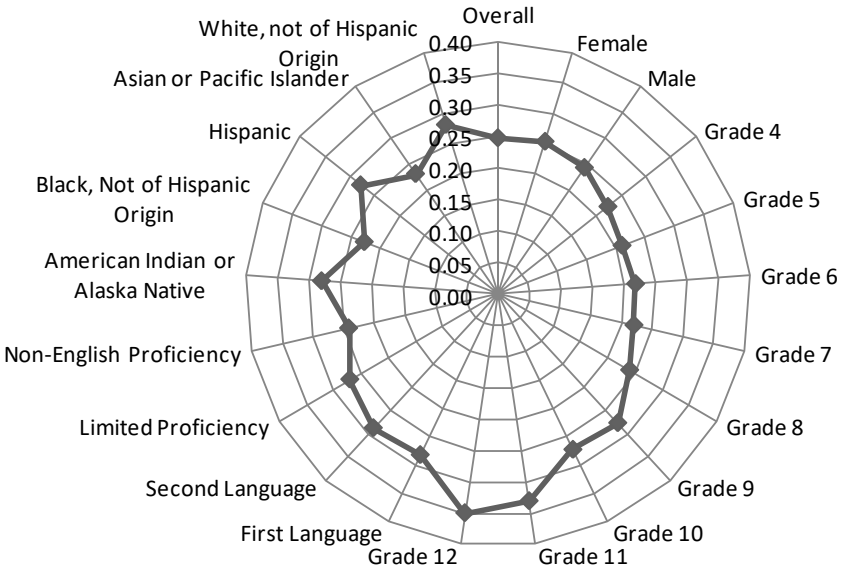

Figure 13: Summary of average improvement per revision

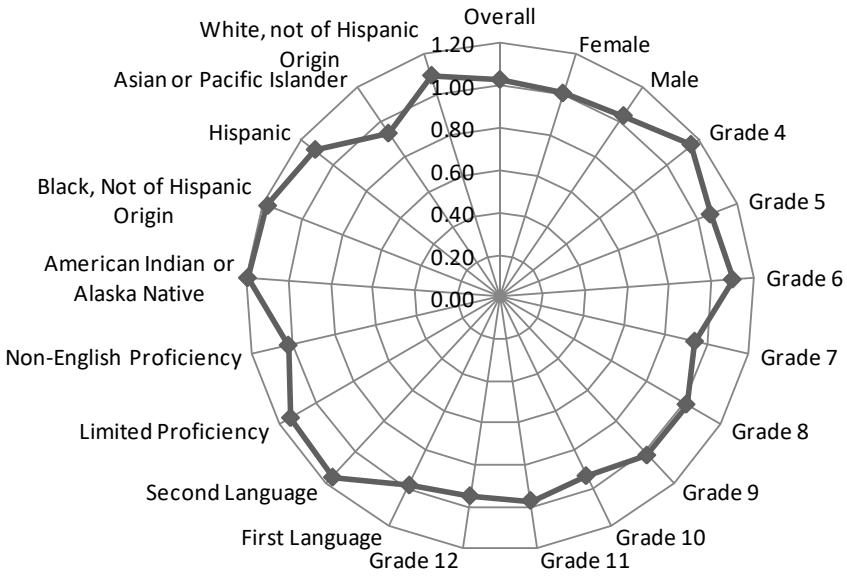

Figure 14: Summary of total improvement per prompt

5 CONCLUSIONS & FUTURE WORKS

Data strongly implies that the hypothesis is partially proven to be valid; frequent writing does improve one's writing skills. Diverse background and different ages affect the rate of improvement but not significantly in most cases. Regardless of different demographic, geographic information, and improvement per revision, all students tried to improve their total holistic score by at least 1.0, which represents that they were able to upgrade their writing skill to the next level.

Future research will be targeted at enhancing the current findings by further analysis of data. Further research may be conducted to determine the causality of significant discrepancy in average improvement per revision and the total improvement. In addition, there are many interesting topics to explore thanks to the abundant amount of existing data within MyAccess!. One of the potential topics is "determining the optimum number of writing revisions to maximize the efficiency in writing skills improvement." Educators will be in a better position to provide consulting to students when we attain meaningful results from this break-through future research.

References

[1] Bronwyn Davies and Tracey Gralton, Trial of Automated Essay Scoring: new directions for national assessment in Australia, Curriculum Corporation, 2009

[2] The Effects of Inclusion of Native Speakers' Writing Samples on the Domain Scoring Accuracy of Automated Essay Scoring of Writing Submitted by Taiwanese English Language Learners, Paul Edelblut, Cathy Mikulas, The 32nd Annual Conference of the International Association for Educational Assessment, Singapore, March 2006

[3] From Here to Validity. In *Automated Essay Scoring,* Elliot, S, ed. M. Shermis and J. Burstein. New Jersey: Lawrence Erlbaum Associates, 2002.