Direct Writing Prediction Models Identify At-Risk Writers

Charles James Harding Conrad II
james_c@ics.ac.th

International Community School Bang Na
1225 Parkland Road
Bang Na, Bangkok, Thailand

Received: 21/02/2020 Revised: 21/05/2020 Accepted: 22/05/2020

Abstract

Developing a sufficient level of writing proficiency takes time. It is also a complex skill difficult to measure. The history of writing assessments reveals changing views of construct validity, reliability and interpretation of results. This study used a binary logistic regression model with seven years of grades 3 to 12 annual direct writing assessments scored with the Oregon six traits rubric from 2012-2018. Three predictive models were developed to show how likely it would be for a participant to reach writing proficiency, and how long it may take to meet that expectation. The research question was, "To what extent can the Annual Writing Assessment scored with the six-traits writing rubric identify at-risk writers from Grades 3-12 at the International Community School Bang Na, Thailand?" Results indicated the bio-data did not prove significant in any of the three models. Increased direct writing data input improved the prediction accuracy. In the Year 1 model, only the average test scored proved significant. In the Year 2 model, the trait of conventions proved significant as one of the independent variables along with the first- and second-year averaged test scores, and the difference between those averages. In the Year 3 model, conventions and sentence fluency proved significant along with the first- and third-year averaged test scores. The process of developing the predictive models, and the results for identifying at-risk writers are presented in this quantitative longitudinal research.

Keywords: direct writing, writing assessment, longitudinal study, binary logistic regression, analytic scoring, six traits

Introduction

Before the invention of the sonogram, doctors and parents had less information on the baby's development in the mother's womb and had to wait for the birth before corrective intervention could take place. Direct writing assessments with analytic scoring are a teacher's ultrasound machine to check on student writing progress. Having access to snapshots of students' direct writing sample scores may alert teachers and other stakeholders to the potential need for early intervention. Just as a sonogram raises awareness of potential risk to a baby's development, an Annual Writing Assessment (AWA) may assist in raising stakeholder's awareness of intervention needs. The current study seeks to create three predictive models to assist stakeholders in identifying at-risk writers.

Students' writing development for those whose first language is not English (L2) differ in the time it takes to reach a sufficient level of writing proficiency when compared with students whose first language is English (L1). Ortiz (2018) notes for average L1 learners there

is no correlation between skill performances and increase of language demands in assessment testing. However, for L2 learners, there is a negative correlation. When the demand on L2 increases, writing performance decreases. It is important to note that the L2 production negative correlation is normal. If L2 development lags behind L1, then how long does it take L2s to reach a proficient writing level according to the six traits Oregon rubric? A key factor is educational exposure in L1 and L2. For example, L1 literacy development may have a positive effect on L2 development (Buckwalter & Lo, 2002). Learners starting in kindergarten and receiving quality dual-language (L1 and L2) schooling a minimum of 6 years take an average of 6 years to be academically successful in L2 while students in an L2 school take an average of 7–10 years to reach a similar sufficient level of proficiency (Collier & Thomas, 2017). Muñoz and Singleton (2011) also note the significance of the quality and amount of L2 input as well as the learner's attitude and orientation affecting language development.

Students in L2 immersion education who exhibit less than grade-level achievement are often referred to as long-term English learners (LTEL) (Collier & Thomas, 2017). Also in L2 immersion education, written L2 levels of proficiency developed slower than spoken fluency (Blanton, 2005). As new L2 students may struggle to communicate due to lack of exposure to L2, LTELs struggle academically despite years of L2 education immersion; not progressing in English language development when compared with other L2 learners. As LTELs get older and reach later grade levels, the gaps in performance widen and proficient levels of writing may no longer be attainable through standard instruction in mainstream classes, or English-as-a-second language programs (Kieffer, 2008; Mancilla-Martinez et al., 2011; Nakamoto et al., 2007).

Direct Writing Assessments

In the 1980s, the six traits writing model was developed from Diederich's (1974) and Purves' (1988) descriptive and theoretical work of classroom-based analytical assessments. The six traits writing model (Culham, 2003) integrates process writing, planning, drafting, assessing, and revising building on the work of Emig (1971), Flower and Hayes (1981), and Applebee (1986). Coe and associates (2000) used the six traits writing rubric in a correlation study to predict writer performance on the Washington Assessment of Student Learning (WASL). Participants scoring less than 3.0 on any of the six traits were predicted to have a 28.6% chance to pass the simulated WASL, and were considered at high risk of not meeting Washington state standards for writing. Students with scores between 3.0 to 3.4 on all six traits had an 83.1% expectancy to pass the simulated WASL, and writers scoring 3.5 or above on all six traits had a 93.8% expectancy to pass the WASL. Formative assessments with six traits assessments identified writing strengths and challenges to help teachers adjust writing instruction and student practice (Coe, 2000). Follow-up six traits writing research by Coe et al. (2011) provided empirical evidence in which the experimental fifth grade students who were taught the six traits significantly outperformed the control group with an effect size of 0.109 (p = .023) using Glass' delta.

Analytic scoring of direct writing assessments has benefits and challenges. Regarding evaluation time per writing sample, analytic scoring may take 1 to 2 minutes per trait compared to 1 to 2 minutes per paper for holistic scoring (Spandel & Stiggins, 1980). However, the benefit is procedural reliability as analytic scores have been the most reliable of all direct writing assessment procedures (East, 2009; Scherer, 1985; Veal & Hudson, 1983; Weigle, 2002). Additionally, with regards to time, direct writing assessments are easier to administer, have a shorter turn-around time than portfolio assessments (Cho, 2003), and can

be used in a variety of settings (White, 1995). While challenges of rater reliability, writing prompts, background knowledge and time constraints have been addressed through empirical research (Cho, 2003), critics claim that timed direct writing essay tests are not grounded in theory or real-life contexts (Camp, 1993; Huot, 1996; Wiggins, 1994).

There is a lack of longitudinal research on direct writing assessment scores of L1 and L2 writers attending international schools outside English-speaking countries. The gap in research provides an opportunity to demonstrate how the Annual Writing Assessment (AWA) scored with the six traits writing rubric may be used to predict at-risk writers as early as third grade to provide intervention and avoid the LTEL lack of progress. Therefore, the research question in the current study proposes, "To what extent can the Annual Writing Assessment scored with the six-traits writing rubric identify at-risk writers from Grades 3-12 at the International Community School Bang Na, Thailand?"

Method

Participants

The overall sample size was 1,784 participants. All the participants had at least one year of writing sample data. However, not all of these participants had all seven years of writing samples. The participants for the AWA included students in grades three to twelve who attended a private international (K-12) English immersion school in Bangkok, Thailand between the years 2012-2018 (Conrad, 2020). The bio-data variables gathered each year, but not tested for in the binary regression model included the year of the AWA, student name, grade level, and homeroom teacher (elementary)/language arts teacher (secondary). Variables gathered each year and tested for in the binary regression model included new student status, sex, nationality, and ESL pullout program status.

The context of the samples gathered was in the AWA in which administration wanted a way to take a snapshot of students' writing abilities on a macro level. The English as a Second Language (ESL) Subject Area Curriculum Team (SACT) partnered with administration in implementing the AWA. The AWA initiation was part of the accreditation recommendation from the *Association of Christian Schools International (*ACSI) and Western Association of Schools and Colleges (WASC) to make efforts to improve language skills of the student population.

Regarding ethical issues, the use of the second hand data collection was approved by the school's Director of Curriculum and Instruction who was assured no participants would suffer physical or emotional harm, or be subjected to unwarranted stress from the research study. Each participant was assigned an ID number, so there would be no reference to student names in this report.

Instruments

The data collection instruments included the direct writing assessment samples gathered by classroom teachers from 2012-2018. The direct writing assessment samples were responses to the different expository essay prompts each year (see Appendix A). Two scorers from National Scoring Services in the United States used the Oregon six traits rubric to score each writing sample (see Appendix B). The resulting scores were statistically analyzed using Excel 2013 and the SPSS version 22.0. In the data gathering process, validity and reliability measures included using classroom teachers as proctors with scripted directions provided by the Director of Curriculum and Instruction to ensure participants completed their writing

independently as well as setting the time limit for the task and collecting the papers at the end of the allotted time.

With regards to the data preparation, the validity and reliability measures included using the Pearson correlation (parametric) then Spearman's Rank correlation coefficient (non-parametric) for normality, and the interclass correlation coefficient for interrater reliability on the seven years of writing samples (see Table 1). The two evaluators' scores from National Scoring Services intraclass correlation coefficient descriptive statistics, being close to 1.0, can be interpreted as a strong correlation in the scoring of the writing assessments between the two evaluators for each trait.

Table 1 Intraclass Correlation Coefficient

		95% Confidence Interval	
Writing Trait	Intraclass Correlation	Lower Bound	Upper Bound
Ideas	0.938	0.934	0.941
Organization	0.942	0.939	0.945
Voice	0.935	0.931	0.938
Word Choice	0.934	0.930	0.937
Sentence Fluency	0.940	0.937	0.943
Conventions	0.969	0.968	0.971

Procedures

The methodological approach underpinning the longitudinal quantitative descriptive research was the development of three prediction models (Year 1, Year 2, Year 3) based on a binary logistic regression model. The three predictive models analyzed participants' scores in their first, second, and third AWA. While prediction accuracy increased with each additional year of participant data, after three years the model would be more descriptive than predictive.

The binary logistic regression model allows for continuous and categorical independent variables. However, there may only be one binary dependent variable in the model. Participants' nominal bio-data as well as both interval and binary coded test scores were used as independent variables. The binary dependent variable was coded as "0" for students who never scored 4.0 or more on a test average range of 0.0 - 6.0 on any of the tests, and "1" for students who scored a 4.0 or above average on at least one of their tests.

The four assumptions for binomial logistic regression in Statistical Package for the Social Sciences version 22 (SPSS) include the following: Assumption 1 - the dependent variable should be measured on a dichotomous scale, Assumption 2 - the presence of one or more independent variables, Assumption 3 - independence of observations in the independent variables and mutually exclusive and exhaustive categories for the dependent variable, and Assumption 4 – there is a linear relationship between any continuous independent variable and the logic transformation of the dependent variable (Laerd Statistics, 2019).

The independent variables of the participant bio data were coded and tested in the binary logistic regression model (see Table 2). Nationality was coded based on the passport country the participants used during the admissions process when enrolling at the school. In addition, L1/L2 status was based on the participants' passport country (see Table 3). Because the data used was secondary data, determining individual participant actual L1/L2 status went beyond the scope of this study.

Table 2 Bio data binary coding

0	1
female	male
not new	new
non-ESL	ESL
non-Thai	Thai
non-USA	USA
non-Indian	Indian
non-Korean	Korean
non-other	other
not English as L1	English as L1

Table 3 Total participants

	Percent of participants	Number of participants
Total	100%	1784
male	48%	863
female	52%	921
Thai	56%	993
non-Thai	44%	791
USA	14%	245
non-USA	86%	1539
India	5%	82
non-India	95%	1702
Korea	14%	256
non-KOR	86%	1528
Other	11%	205
non-Other	89%	1579
English as L1	15%	272
English not as L1	85%	1512
ESL	24%	434
non-ESL	76%	1350

Each year for Grades 3 -12, the data was collected early September, in the first period of the school day. There were 40 minutes for actual writing not including additional administration time. Homeroom teachers proctored the writing for the elementary students and secondary students sat with their English sections. The writing samples were mailed to National Scoring Services in the United States. Having the writing sample scoring outsourced allowed the school to save time and avoid putting undue pressure on new teachers; especially at the beginning of a school year. After each paper was analyzed by two examiners, the raw interval scores were recorded in Excel and sent to the school.

After receiving the data in Excel file format from National Scoring Services, the data preparation involved creating a merged file of the seven years with each row separated by test score (n = 5,964) was transferred to SPSS to check for a normality of distribution. First, the data was checked for normality to know if the data was parametric or not to reject using parametric Pearson correlation. From the SPSS results, with the dependent variable as the writing results, the independent variables were each year the test was given. For each year and each trait there was a significance of .000 on the Shapiro-Wilk test for normality meaning

acceptance of the null hypothesis, so the data was not normally distributed, or non-parametric. Therefore, the binary regression model could be implemented. Finally, using the binary logistic regression model was a justifiable approach to answering the research question because, one of the primary outputs of the model is the predictability that a dichotomous event would or would not occur based on the input variables. The three models were considered acceptable if the reliability percent of the model was higher than the reliability percent of the null hypothesis once all the independent variables with non-significance were removed from the equation.

Analysis

One challenge was aligning the tests for comparison because not all participants attended the 2012 school year. Therefore, the school year labels were replaced with Test 1, Test 2, Test 3... In an attempt to show due diligence, all the bio data variables were introduced in the model. However, they showed little to no significance in comparison to the initial test score. Gathering any additional bio-data would have involved several hours of attempting to cross reference student records with the admissions office and went beyond the scope of this research.

It was still difficult to see a pattern of predictability from the averaged initial test score alone in the Year 1 model. If the initial test score was not robust enough to identify the participants who may be at-risk writers, the next question was to consider if the difference in growth rate, or decline of scores between the initial test and the second test was an indication of a pattern for at-risk writers. The differences between Test 1 and Test 2 were averaged and compared between the groups of participants. Using the interval raw data of individual six trait scores increased the reliability of the Year 2 prediction model.

Three prediction models: Year 1, Year 2, and Year 3 were developed with the focus of increasing the predictive accuracy of the initial model by including the raw scores for each of the six traits to identify if the additional independent variables were significant. For example, do participants who reach 4.0 have a different combination of strengths and challenges, than those who don't reach 4.0; even though their initial Test 1 average score was the same? I started approaching the six trait scores as independent variables and compared them to the averaged trait scores, and finally tried mixing the average trait scores with individual trait scores which proved most accurate for predictability for years 2 and 3.

For the Year 1 model, the significant independent variable was the Test 1, the initial test score average (0.0 – 6.0). For the Year 2 model, the significant independent variables were the Test 1 score average (0.0 – 6.0), the Test 2 test score average (0.0 – 6.0), the combined interval Test 2 score from the two evaluators for Conventions (0.0 – 12.0), and the difference between the Test 1 and Test 2 averages. For the Year 3 model, the significant independent variables were the Test 1 score average (0.0 – 6.0), the Test 3 test score average (0.0 – 6.0), the combined interval Test 3 score from the two evaluators for Conventions (0.0 – 12.0), and the combined interval Test 3 score from the two evaluators for Sentence Fluency (0.0 – 12.0) (see Table 4).

Table 4 Highest predicted accuracy of independent variables for Year 1, 2, 3 models

	Prediction accuracy percent	
	null	Independent variables: Test averages
Year 1	64.60%	72.4% Test 1 average
Year 2	74.30%	79.6% Test 1 and Test 2 averages, Test 2 (Conv), 0.55 cut off (between Test 1 & Test 2)
Year 3	80.30%	85.6% Test 1 and Test 3 averages and Test 3 (SF and Conv)

For Year 1, a prediction table could be made, but for Year 2 and 3, there were too many variables, so I put the prediction model formulas directly into the Excel rows so each participant's prediction will be based on their individual scores. I also calculated the average number of years it took for participants to reach 4.0 at each point on the scale from 1.0 to 3.9. Therefore, the Excel sheets for each year's model included the prediction in percent of how likely it would be a participant reach 4.0 based on the independent variables, and how much time it may take each participant to reach 4.0 based on the average scores of participants whose initial score was below 4.0 but did reach 4.0 or above.

To set up a hypothetical context for the models, at-risk writers were identified based on two criteria: criteria #1 the initial test score, unless they got 4.0 or above on any tests; criteria #2 4.0 should be reached within 14 months beyond the average time it took other participants with the same initial score to reach 4.0 or above.

Results

Descriptive statistics

The following are the frequency counts for the bio-data binary variables (see Table 5), and six traits binary variables (see Table 6):

Table 5 Descriptive statistics for the binary bio-data

	Frequency	Percent
Male	899	50.4
Female	885	49.6
Thai	791	44.3
Non-Thai	993	55.7
USA	1539	86.3
Non-USA	245	13.7
Indian	1702	95.4
Non-Indian	82	4.6
Korea	1528	85.7
Non-Korean	256	14.3
Other nationalities	1577	88.4
Not Other nationalities	207	11.6
L1	296	16.6
L2	1488	83.4
ESL	1416	79.4
Non-ESL	368	20.6
Total	1784	100.0

Table 6 Descriptive statistics for binary six traits variables

	Frequency	Percent
Scored 4.0 or higher	451	25.3
Scored less than 4.0	1333	74.7
Difference between Test 1 and Test 2 < 0.55	1477	82.8
Difference between Test 1 and Test 2 > 0.55	307	17.2

Next are the means and standard deviations for the interval variables for Tests 1-3 (see Table 7).

Table 7 Descriptive statistics for Tests 1-3

	N	Mean	Std. Deviation
Test 1 average as one score (0-6)	1784	3.75	.789
Test 2 average as one score (0-6)	1297	3.91	.670
Test 3 average as one score (0-6)	1039	4.21	.640
Test 2 Conventions (0-12)	1297	7.32	2.298
Test 3 Conventions (0-12)	1039	7.94	2.020
Test 3 Sent. Fluency (0-12)	1039	8.44	1.446

In order to address the research question concerning the extent to which the AWA could identify at-risk writers, three models of the prediction were constructed by inputting the participants' test scores and bio-data into SPSS and analyzed with the binary logistic regression model. The dependent variable was coded the same for all three models (0 = no, the participant did not have any test year average score of 4.0 or above (0.0 – 0.6); 1= yes, the participant had one or more test year average scores of 4.0, or above (0.0 – 0.6)).

The Year 1 model independent interval variable was the Test 1 averaged score (sig. .000) from the two evaluators raising the null predicted percent correct from 64.60% to 72.4%. None of the nominal bio-data input variables proved significant and were therefore excluded from the model.

In the Year 2 model, the four independent interval-level variables were the Test 1 (sig. .000) averaged scores, Test 2 (sig. .000) averaged scores, Test 2 (sig. .000) the interval variable of the combined raw scores for the Conventions trait (sig. .001), and the binary coded variable of the difference in averaged scores between Test 1 and Test 2 with the cut- point of 0.55 (sig. .000) (0 = no; 1 = yes) raising the null predicted percent correct from 74.3% to 79.6%. None of the nominal bio-data input variables proved significant and were therefore excluded from the model.

The four independent interval variables in the Year 3 model were the interval Test 1 (sig. .000) averaged scores, Test 3 (sig. .000) averaged scores, the interval variable of the combined raw scores for the Conventions trait (sig. .001), and the interval variable of the combined raw scores for the Sentence Fluency trait (sig. .005) raising the null predicted percent correct from 80.3% to 85.6%. None of the nominal bio-data input variables proved significant and were therefore excluded from the model.

While the likelihood a participant will reach 4.0 or above is predicted in the three models, administration will still need to decide at what percentage point of likelihood

intervention may take place. To further assist identifying at-risk writers, two criteria were developed.

Criteria #1 was met if the initial test score was 3.0 or below and the participant never scored 4.0 or above on any tests. Criteria #2 was met if 4.0 was not reached within 14 months beyond the average time it took other participants who had the same initial score to reach 4.0.

Criteria #2 took into account the average number of years it took participants to reach 4.0. The 14 months was chosen as a cutoff point to be interpreted that the participant should reach 4.0 or above no more than one year more than the average time it took for other participants with the same initial score. For example, if the participant's initial test score was 3.0, the average time it took participants to reach 4.0 or above was 3 years. Therefore, if a participant had an initial score of 3.0, and had not reached 4.0 or above by the fifth test year, they met the requirement for criteria #2. However, for the participants (n = 23) with an initial score of 1.0, 1.6, 1.7, or 2.1, the average number of years to reach 4.0 was set to 12 years because none of the participants with these initial scores ever reached 4.0. At-risk writers were identified as participants who matched both criteria (see Table 8).

Table 8 At-risk writers

At-risk writers for…	n	Participant meets both Criteria #1 AND Criteria #2	% of total participants	Participant meet Criteria #1 only	Participant meets Criteria #2 only	Participant meets either Criteria #1 OR Criteria #2	% of total participants
Year 1 Model	1,785	23	1.29%	239	0	239	13.39%
Year 2 Model	1,297	34	2.62%	95	49	144	11.10%
Year 3 Model	1,039	25	2.41%	56	25	81	7.80%

Discussion

All participants (n =1,785) had at least taken one AWA test. With one year of data, it is very difficult to identify who an at-risk writer may be, but the lower the initial score the more likely a student may be in need of intervention. If the participant's initial score was between 2.0 – 2.5, then the actual percent of students that ever reached 4.0 in seven years was 3 to 23%. If the participant's initial score was less than 2.0, then the actual percent of students that ever reached 4.0 in seven years was 0 to 11%. A challenge with using only one score in the Year 1 model is the inability to identify those participants that score between 3.0 to 3.9 and never reach 4.0. This is important because 47% of the students with an initial score of 3.0 never reached 4.0, and even with an initial score of 3.9, around 11% of those participants never reached 4.0.

1,297 participants had taken at least two AWA tests. With the additional information of the second-year scores in the Year 2 model, the predictability percent increased where even the null predicted percent was higher than the Year 1 model. The Year 2 model was the only model where the binary coded cut-off point of 0.55 was significant. 0.55 was quite a large change for most participants accounting for many of the participants who scored 4.0 on Test 2 after initially scoring below 4.0 on Test 1. The Year 2 model also introduced Conventions as a significant factor in the model.

1,039 participants had taken at least three AWA tests. Not surprisingly, the Year 3 model had the highest prediction accuracy percent by having more data to input into the independent variables thereby increasing accuracy. Test 3 Sentence Fluency became a significant independent variable joining the other variables of Test 3 Conventions, Test 1 and Test 3 average scores. It may be worth future research to better understand why Sentence Fluency and Conventions became independent variables while the other four skills did not. However, supporting any claims should involve discourse analysis research of the writing samples which lies beyond the scope of this study.

A low percentage of the participants were identified as at-risk writers based on the direct writing assessment data using three prediction models and two criteria which may indicate the international school English language immersion environment may be beneficial in developing English writing skills. However, Muñoz and Singleton (2011) noted the significance of the quality and amount of English language input as well as the learner's attitude and orientation on language development. Students' writings may vary because of the cultural diversity they bring to the classroom.

Longitudinal studies observed a plateau effect (Collier & Thomas, 2017; Kieffer, 2008; Mancilla-Martinez et al., 2011; Nakamoto et al., 2007) from secondary LTEL students similar to writers who were unable to develop to the 4.0 level of proficiency while consistently scoring in 3.0 to 3.9 range of the average test scores. This is important because 47% of the students with an initial score of 3.0 never reached 4.0, and even with an initial score of 3.9, 11% of those students never reached 4.0. The plateau effect may be attributed in part to writers feeling they write well enough to pass their classes, and in most cases this would be true. In addition, direct writing assessments are only a small percentage of a student's overall Language Arts grade, and even less so for other content subject classes.

In general, language development is a complex process, and the sub-skill of writing is no exception. What adds to the complexity is having an immersion international school in an L2 context which provides unique opportunity and challenges for both L1 and L2 learners. The English-speaking learners may be unique by living in a non-English speaking foreign country; specifically Thailand. The English-speaking learners may have gaps in their education if they have moved often, or simply were not immersed in an English-speaking environment outside the school community. Further research at other international schools may provide more evidence to the plateau effect found in LTEL students and participants in this study who were unable to reach the proficient level of 4.0 on the six traits writing rubric. If similar results occurred, the models may have wider implications than the local school used in this study. However, it may be the case that predictive models need to be developed for each unique school environment in which some aspects of the bio-data may prove to be significant contributing factors for identifying at-risk writers, or other traits besides Sentence Fluency and Conventions may be significant as independent variables. Nevertheless, a positive aspect of the predictive models from this study is that the independent variables of the six traits did not have bio-data bias.

Limitations

One limitation of this study is the fact that all the writing samples were collected from one school. While the seven years of data provided a robust sample, future research may provide an opportunity to explore if the three predictive models could identify at-risk writers

in another international school environment where English is used as a medium of instruction. In addition, because nationality was coded based on the passport country the participants used during the admissions process when enrolling at the school, the consequence of coding with such generalities means that some participants may have been miscoded regarding their L1/L2 background.

Collecting the direct writing assessment allows for a large number of samples to be gathered simultaneously. However, the data collection method reflects a limitation in the nature of direct writing assessments. Writing to a prompt under a time constraint is equivalent to writing a first draft and lacks the opportunity for participants to respond, or interact with articles, or other literature. In addition, there is no opportunity for participants to select their best writings as in portfolio assessments.

Another limitation is the trade-off due to the nature of longitudinal studies. In the general field of writing instruction and assessments, trends have changed over those years in diverse directions such as: the inclusion of literature in text-based writing assessments, the use of technology for scoring writing assessments, and the practice of collecting several writing samples over the period of the academic year for portfolio assessment. If the AWA had tried to incorporate these trends over the past seven years in gathering the participants' samples, it would have been very difficult, if not impossible, to identify the independent variables influencing the participants' scores over that time.

Lastly, while this study did measure the inter-rater reliability between the two independent scorers, the possible effects of varied writing prompts year to year were not analyzed. In other words, this study did not investigate equivalence or differences in the AWA prompts given. Some prompts might have been more difficult resulting in lower scores. Future studies may want to consider the influence of the writing prompts and the possible difficulties or misinterpretations the prompts may have contributed to the variation of scores; particularly relating to participants who never scored 4.0.

Conclusions

If the student has had three years of writing assessments, the Year 3 model would prove most reliable in terms of prediction accuracy. The Year 2 model also has the benefit of taking in account the student's previous year's performance. Even the Year 1 model, while the most limiting in assisting data-driven decisions, still may prove of value, for example, in assisting administrators' development of a watch list of potential at-risk writers.

The effectiveness of the prediction models on assisting at-risk writers is beyond the scope of this research as the purpose was to determine to what extent the Annual Writing Assessment scored with the six-traits writing rubric could identify at-risk writers from Grades 3-12 at the International Community School Bang Na, Thailand. Administrators can use the prediction models to identify at-risk writers. Future application of the current study may involve creating a database of the writing scores and designing a user-friendly interface with integrated prediction models to assist administrators in making data-driven decisions.

Just as a sonogram assists in raising awareness of potential risk to a baby's development, the AWA Year 1, Year 2, and Year 3 models provide predictions of direct writing assessment for individual participants as well as for cohorts or other sub-populations

that may alert administrators and other stakeholders to the potential need for early intervention before a gap widens between low levels of writing proficiency and grade level expectations.

References

Applebee, A. N. (1986). Problems in process approaches: Toward a reconceptualization of process instruction. *The Teaching of Writing*, *85*, 95-113.

Blanton, L. (2005). Student, interrupted: A tale of two would-be writers. *Journal of Second Language Writing, 14*, 105–121.

Buckwalter, J., & Lo, Y. (2002). Emergent literacy in Chinese and English. *Journal of Second Language Writing, 11*, 269–293.

Camp, R. (1993). Changing the model for the direct assessment of writing. *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*, 45-78.

Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing Writing*, *8*(3), 165-191.

Conrad, C. (2020), 2012-2018 *Direct Writing Assessment Scores for an international school (K-12) in Bangkok, Thailand*, Mendeley Data, v1 http://dx.doi.org/10.17632/hfkt24zxcw.1

Coe, M.T. (2000). Direct writing assessment in action: Correspondence of six-trait writing assessment scores and performance on an analog to the Washington Assessment of Student Learning writing test. Northwest Regional Educational Laboratory. Portland, OR.

Collier, V. P., & Thomas, W. P. (2017). Validating the power of bilingual schooling: Thirty-two years of large-scale, longitudinal research. *Annual Review of Applied Linguistics*, *37*, 203-217.

Culham, R. (2003). *6+1 Traits of Writing: The Complete Guide Grades 3 and Up*. New York: Scholastic Inc.

Diederich, P. B. (1974). *Measuring growth in English*. National Council of Teachers of English. Urbana, IL.

East, M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing writing*, *14*(2), 88-115.

Emig, J. (1971). The composing processes of twelfth graders. In Culham, R. (2003). *6+1 Traits of Writing: The Complete Guide Grades 3 and Up*. Scholastic Inc. New York.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365-387.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, *47*(4), 549-566.

Kieffer, M. J. (2008). Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. *Journal of Educational Psychology*, *100*(4), 851.

Laerd Statistics. (2019). *How to Perform a Binomial Logistic Regression in SPSS Statistics*. Retrieved from https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php

Mancilla-Martinez, J., Kieffer, M. J., Biancarosa, G., Christodoulou, J. A., & Snow, C. E. (2011). Investigating English reading comprehension growth in adolescent language minority learners: Some insights from the simple view. *Reading and Writing*, *24*(3), 339-354.

Muñoz, C., & Singleton, D. (2011). A critical review of age-related research on L2 ultimate attainment. *Language Teaching*, *44*(1), 1-35.

Nakamoto, J., Lindsey, K. A., & Manis, F. R. (2007). A longitudinal analysis of English

language learners' word decoding and reading comprehension. *Reading and Writing*, *20*(7), 691-719.

Ortiz, S. (2018). *Testing with English Learners and the C-LIM: Myths and Misconceptions.* [PowerPoint slides]. Retrieved from https://www.youtube.com/redirect?event=video_description&v=A0X5ljIyI1M&redir_token=o7lyzITFtDi9dFz4MYi1PmOXdh8MTU5MDU2MDMxNkAxNTkwNDczOTE2&q=https%3A%2F%2Fdrive.google.com%2Ffile%2Fd%2F1eSx4gUsYUHsIUHzllCWYSNdSAlV2A-pD%2Fview%3Fusp%3Dsharing

Purves, A. (Ed.). (1988). *Writing Across Languages and Cultures: Issues in Contrastive Rhetoric*. Newbury Park, CA.: Sage.

Spandel, V., & Stiggins, R. J. (1980). *Direct Measures of Writing Skill: Issues and Applications*. OR.: Northwest Regional Educational Laboratory.

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English*, *17*(3), 290-296.

Weigle, S. (2002). *Assessing Writing*. Cambridge University Press. Cambridge, UK.

White, E. M. (1995). An apologia for the timed impromptu essay test. *College Composition and Communication*, *46*(1), 30-45.

Wiggins, G. (1994). The constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing*, *1*(1), 129-39.

Appendix A

Annual Writing Assessment Prompts 2013

Elementary	Think about something special you have or would like to have. It might be a toy, a gift, or a common object you use every day. What makes this thing special? Share at least one real or imaginary experience about your special thing.
Secondary	We have many items that make our lives comfortable or better. Choose an object you own or would like to own. How does this object improve your life? Share at least one specific example that shows how this object is used and how it makes a difference in your life.

AWA prompts International Community School, Bang Na 2013

Appendix B

	Not Proficient			Proficient		
	Beginning 1	Emerging 2	Developing 3	Capable 4	Experienced 5	Exceptional 6
Ideas / Content	• ideas are limited • minimal development • too short	• insufficient details • irrelevant details • extensive repetition	• easily defined purpose • difficulties moving from general to specific	• supporting details are relevant • topic is explored and explained	• clarity, focus, and control • details well-suited to audience • use of relevant resources	• main ideas stand out • carefully selected details • in-depth explanations
Organization	• failure to provide identifiable beginning, body, or ending • main point obscured	• extremely undeveloped beginning, body, or ending • random details	• skeletal or rigid structure abrupt beginning, or ending such as "My topic is . . ." or "That is my reason."	• clear sequencing • developed beginning, body, and ending • basic transitions	• effective sequencing • inviting beginning and satisfying ending • smooth, effective transitions	• effective, creative sequencing • strong beginning • smooth transitions throughout
Word Choice	• extremely limited range • words that do not fit the context	• images that are fuzzy • monotonous, repetitions	• words work, but rarely capture interest • clichés and overused expressions	• attempts at colorful language • rare experiments with language	• accurate, specific words that energize • words that evoke clear images	• figurative language • fresh, original expression • striking and varied in choice of words
Fluency	• confusing word order, jarred • obscured meaning • disjointed, rambling	• portions difficult to read or follow • monotonous sentence patterns	• good control of simple sentences • sentences lacking energy	• natural sound • some lapses in stylistic control • rare fragments	• variation in sentence structure, length, beginnings • sentence glides along	• extensive variation in sentences adding interest to the text
Voice	• no sense of the reader • no sense of interaction b/n writer and reader	• overly informal and personal • writer is flat, stiff, lifeless	• limited ability to shift to a more objective voice • a sense of personality to text	• questionable or inconsistent level of closeness or distance from the audience • sincerity, humor	• strong sense of audience • sense the topic has come to life • may contain dialogue, exclamations	• writer shows originality, conviction, excitement
Conventions	• basic punctuation omitted, or incorrect • frequent spelling errors • major grammatical errors	• paragraphs run together • substantial need for editing • many spelling errors	• misspelling of common words • text may be too simple • internal punctuation lacking	• spelling is usually correct • moderate need for editing • sound paragraph breaks	• strong control of conventions • little need for editing • correct spelling of difficult words	• skill in using a variety of conventions in long and complex writing

Six Traits Oregon Rubric adapted by ICS, BangNa, Thailand

About the Author

Dr. Charles James Harding Conrad II has an education doctorate in TESOL from Anaheim University and has been teaching in Thailand since 1996. He taught 4 years with Sarasas Affiliated Schools and has been teaching ESL at the International Community School, Bang Na since 2000.