Title: The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus

Authors: Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, Ulrich Boser

Vanderbilt University / Georgia State University / The Learning Agency / The Learning Agency / Georgia State University / Vanderbilt University/ / The Learning Agency / The Learning Agency / The Learning Agency

Citation:

Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, B., Picou, A., & Boser, U. (in press). Measuring second language proficiency using the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus. *International Journal of Learner Corpus Research.*

Abstract: This paper introduces the open-source English Language Learning Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus, which comprises ~6,500 essays written by English Language Learners (ELLs). All essays were written during state-wide standardized annual testing in the United States. The essays were written on 29 different independent prompts. Individual difference information is made available for each essay including economic status, gender, grade level (8-12), and race/ethnicity. Each essay was scored by two trained human raters for English language proficiency including an overall score of English proficiency and analytic scores for cohesion, syntax, vocabulary, phraseology, grammar, and conventions. The paper provides reliability on the human judgments of proficiency reported for the corpus. The

ELLIPSE corpus addresses many of the concerns found in existing learner corpora including unique holistic and analytic scores for each ELL essay. The corpus also includes limited demographic and individual difference data for each ELL.

Keywords: Corpus linguistics, English language learners, language proficiency

1. Introduction

Measuring language proficiency in English language learners (ELL) is an important element of assessment that can provide teachers, administrators and learning systems with information that can inform pedagogical interventions, track development, and recommend learning materials to help ELLs develop language skills. There have been various approaches used to measure language proficiency in the past including the use of standardized assessments such as the reading sections in the Test of English as a Foreign Language (TOEFL; Chapelle, Enright, & Jamieson, 2008) or the International English Language Testing System (IELTS; O'Sullivan, 2018), learner surveys of proficiency (Bailey & Kelly, 2010), and human ratings of proficiency (Cheng & Warren, 2005). More recently, researchers have begun to use corpora comprising user-produced texts and natural language processing (NLP) approaches to automatically assess language proficiency (Lagakis & Demetriadis, 2021). Such approaches use NLP techniques to extract language features from texts that have scales or individual scores for language proficiency for the learners. Statistical or machine learning models are built to predict learner proficiency based on the language features within the texts. NLP approaches hold promise because the models developed can be scaled to larger populations, generalized across populations, and automatized so they can be incorporated into testing platforms that provide rapid assessments of language proficiency to ELLs.

One concern with current approaches to automatically assessing language proficiency using corpus and NLP approaches is that the large English corpora that are available for the tasks often lack unique language proficiency scores for individual students.[1] Most of the large learning

[1] There are learner corpora available in languages other than English that contain unique proficiency scores for individual learners including MERLIN, which covers Czech, German, and Italian (Boyd et al., 2014).

corpora that are available provide proficiency information at a scalar level. As an example, texts within the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen, Alexopoulou, & Korhonen, 2013), which is a large corpus of ELL writings collected by English First (EF), are scaled from 1-16 based on Englishtown Skill Levels and are generally further rescaled to the six levels that comprise the Common European Framework of Reference (CEFR, Council of Europe, 2001, see also the Companion Volume to CEFR). Similarly, the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2013) contains ~10,000 speech and writing samples from ELLs in Asia. The corpus is scaled to four levels of the six levels found in the CEFR based on student-reported scores for high-stakes English tests and/or the results from a vocabulary size test (Laufer & Nation, 1999). Another example is the TOEFL11 corpus, which contains ~12,000 essays written by TOEFL participants scaled for three levels of writing proficiency (beginning, intermediate and advanced; Blanchard et al., 2013).

A second concern with the available large learner corpora that include proficiency information is that the metrics of proficiency included are holistic and not analytic, meaning they provide an overall score for language proficiency but do not provide indicators of language proficiency on individual metrics like lexical, grammatical, or syntactic proficiency. As noted above, many of the large-scale ELL corpora available scale proficiency to the six CEFR levels. These levels are aggregated scores that combine individual language skills including understanding, summarization, coherence, fluency, and meaning into a holistic proficiency score (Figueras, 2012). Lastly, the TOEFL writing and speaking rubrics include aggregated proficiency scores that focus on the production of linguistic features including topic development, coherence, syntax, word choice, and grammar errors.

A third concern with large learner corpora that include proficiency information is that they often do not provide demographic or individual difference measures. Researchers may want to co-vary demographic variables such as race/ethnicity and/or gender and individual difference such as working memory, motivation, time learning a language, grade level and socio-economic status in some analyses to examine sub-populations and control for potential variance in language proficiency scores. There are exceptions, of course. For instance, the TOEFL 11 database includes gender and first language of writers while ICNALE provides information on the nationality of students along with their age, sex, and grade level. EFCAMDAT includes information about students' nationality.

The purpose of this paper is to introduce the open-source English Language Learning Insight, Proficiency and Skills Evaluation (ELLIPSE) corpus as well as assess the reliability of the scores provided for each essay. The corpus comprises ~6,500 essays written by ELLs in an English as a Second Language (ESL setting). All essays were written during state-wide standardized annual testing in the United States. The essays were written on 29 different independent prompts that required no background knowledge on the part of the writer. Individual difference information is made available for each essay including economic status, gender, grade level (8-12), and race/ethnicity. Each essay was scored by two trained human raters for English language proficiency including an overall score of English proficiency and analytic scores for cohesion, syntax, vocabulary, phraseology, grammar, and conventions. The ELLIPSE corpus addresses many of the concerns found in existing learner corpora including unique holistic and analytic scores for each ELL essay. Additionally, the corpus includes limited demographic and individual difference data for each ELL.

1.2 Measuring proficiency

The development of reliable measures of language proficiency for language learners has been an important and productive research area in language learning for decades. Early research defined language proficiency narrowly as competency in grammar and lexis (Chomsky, 1972). Other approaches included a student's understanding of sociolinguistic knowledge and pragmatics (Widdowson, 1983) and communicative competence (Bachman & Palmer, 1996; Hymes, 1972) in attempts to include concepts of strategic competence along with linguistic knowledge.

Previous research has also focused on measuring language proficiency using linguistic features produced by learners related to complexity, accuracy, and fluency (CAF) measures (Larsen-Freeman, 1978). CAF measures generally ignore pragmatic and communicative contexts. For instance, Larsen-Freeman suggested the construction of a 'yardstick of development', which would be an independent measure from which the language development of English as a Second Language (ESL) students could be tracked. Her work indicated that students with more advanced proficiency tended to produce more words per T-unit (i.e., the smallest unit of language which can be bounded with punctuation). Since her work, CAF measures have been investigated in detail in a number of theoretical and data-driven studies. From a theoretical perspective, CAF dimensions incorporate important elements of L2 knowledge and proficiency that underlie production (Granena, 2019; Housen & Kuiken, 2009; Housen, Kuiken, & Vedder, 2012) and their use in investigating language learner performance has become common (Ortega, 2012). Models of second language (L2) development often include CAF as principal dimensions of proficiency (Skehan, 1989) with complexity defined as the use of elaborate and varied language (Ellis, 2003), accuracy as error free language (Housen & Kuiken, 2009), and fluency as a measure of language

produced without hesitation, pausing, or reformulation (Ellis, 2003). Complexity and accuracy are generally considered to represent productive L2 knowledge from a linguistic perspective, which is the focus of the ELLIPSE corpus, while fluency represents control of that knowledge (Housen & Kuiken, 2009).

A number of approaches have been used to measure language proficiency in terms of complexity and accuracy. These range from standardized test items that measure knowledge at the lexical, syntactic, and grammatical levels, the use of scoring rubrics to harness human judgments of proficiency, and corpus analyses using NLP techniques. Historically, the most common approach to assessing language proficiency has been through standardized tests based on close-ended assessment like multiple-choice questions. There are hundreds of examples of such tests from which to choose including the TOEFL (Blanchard et al., 2013), the vocabulary size test (Laufer & Nation, 1999), grammaticality judgment tasks (Ellis, 1991), and the American Council on the Teaching of Foreign Languages (ACTFL) reading proficiency test (Clifford & Cox, 2013; Lisken-Gasparro, 1984). One problem with close-ended assessment is that they are limited to the knowledge offered within the question and alternatives available in the answers. They do not let test-takers demonstrate knowledge beyond the questions (Foddy, 1993).

Open-ended responses do allow test-takers to produce knowledge in response to prompts that can go beyond the prompt. However, the assessment of open-ended responses generally requires human raters because they report the high internal consistency needed to ensure reliability in open-ended response scoring (Lumley, 2005). Expert raters have been used to assess the quality of language production in terms of writing proficiency (Crossley & McNamara, 2010), speaking proficiency (Lumley, 1998), lexical proficiency (Crossley, Salsbury, & McNamara , 2013), and overall language proficiency (Lim, 2011) with expert scores providing evidence to support

inferences about language ability (Kim, 2015). Rubrics that measure language proficiency are generally of two types – holistic rubrics which measure the quality of the language produced as a whole and analytic rubrics which divide language knowledge into a number of sub-domains (Weigle, 2004) like organization, vocabulary use, or content. Rubric-based approaches are most common in writing analytics in which both analytic and holistic measures have been found to be predictive of outcomes for students of all levels of competency (Wood & Schatschneider, 2021).

Open-ended assessments that rely on the production of language data by test-takers can lead to the development of large text corpora that can be computationally mined for patterns related to language learning. When these large corpora include human ratings of proficiency, then models of language proficiency based on linguistic information extracted from the texts can be derived. Such approaches rely on NLP techniques to calculate features of texts produced by test-takers at the lexical, grammatical, syntactic, cohesive, semantic, and discourse levels, among others. These features can then be used in statistical or machine learning models to predict the expert ratings assigned to each text so that proficiency level can be predicted and explained based on features that inform the models (e.g., higher proficiency ELLs produce more infrequent words). Such approaches have become commonplace with the release of large learner corpora such as EFCAMDAT, ICNALE, and the TOEFL11 and the NLP analyses of these corpora have been aided by the release of freely available and user-friendly NLP tools including Coh-Metrix (Graesser et al., 2004) and the Suite of Automatic Linguistic Analysis Tools (SALAT; Crossley et al., 2016, 2017; Kyle et al., 2018, 2021; Kyle & Crossley, 2018).

2. The ELLIPSE Corpus

This corpus report introduces the open-source *English Language Learning Insight, Proficiency and Skills Evaluation* (ELLIPSE) corpus, and its associated meta-data. This corpus report also provides reliability on the human judgments of proficiency reported for the corpus.

2.1 Initial Corpus

The initial essays selected to populate the ELLIPSE corpus came from a subsection of a larger corpus of essays collected from state-wide and national standardized testing in the United States (~600,000) used to assess writing skills at grade level. All essays were independent essays for which no background knowledge of the topic was required, and students were provided with no source texts. Writers were given between 25 and 30 minutes to write an essay on a computer. From this larger corpus, we selected essays that were labeled as written by ELLs as candidates for the ELLIPSE corpus. These labels were based on binary assignments provided by the states. We also selected essays that had corresponding demographic and individual difference measures for the ELLs and contained at least 75% correctly spelled words in English. The initial corpus comprised 8,890 essays written on 44 different prompts.

2.2 Proficiency scoring

A language proficiency rubric was developed to assign scores to each of the 8,890 essays in the initial ELLIPSE corpus. The rubric was based on a literature review of the components that comprise language proficiency and available language proficiency rubrics. In total, 18 published

articles were reviewed along with 38 rubrics. Feedback on the initial rubric developed was first given by a teacher advisory board that consisted of ten English teachers who taught ELLs. The rubric was next reviewed by a research advisory board consisting of experts in second language acquisition, English language learners, and composition. The rubric was modified to account for the feedback provided by the teacher and research advisory boards.

The final rubric comprises a single holistic score of overall language proficiency and six analytic scores related to specific features of language. These are cohesion, syntax, vocabulary, phraseology, grammar, and orthographic and punctuation conventions. The holistic score and the analytic scores are based on a 5-point Likert scale. A score of 5 relates to a "native-like facility" in English language proficiency while a score of 1 relates to limited ability in English language proficiency (see Figure 1 for the rubric).

Figure 1. Scoring rubric used to rate essays in the ELLIPSE corpus

English Proficiency Scoring Rubric for English Language Learners

	HOLISTIC	ANALYTIC					
	Overall	**Cohesion**	**Syntax**	**Vocabulary**	**Phraseology**	**Grammar**	**Conventions**
5	Native-like facility in the use of language with syntactic variety, Appropriate word choice and phrases; well-controlled text organization; precise use of grammar and conventions; rare language inaccuracies that do not impede communication.	Text organization consistently well controlled using a variety of effective linguistic features such as reference and transitional words and phrases to connect ideas across sentences and paragraphs; appropriate overlap of ideas.	Flexible and effective use of a full range of syntactic structures including simple, compound, and complex sentences; There may be rare minor and negligible errors in sentence formation.	Wide range of vocabulary flexibly and effectively used to convey precise meanings; skillful use of topic-related terms and less common words; rare negligible inaccuracies in word use.	Flexible and effective use of a variety of phrases, such as idioms, collocations, and lexical bundles, to convey precise and subtle meanings; rare minor inaccuracies that are negligible.	Command of grammar and usage with few or no errors.	Consistent use of appropriate conventions to convey meaning; spelling, capitalization, and punctuation errors nonexistent or negligible.
4	Facility in the use of language with syntactic variety and range of words and phrases; controlled organization; accuracy in grammar and conventions; occasional language inaccuracies that rarely impede communication.	Organization generally well controlled; a range of cohesive devices used appropriately such as reference and transitional words and phrases to connect ideas; generally appropriate overlap of ideas	Appropriate use of a variety of syntactic structures, such as simple, compound, and complex sentences; occasional errors or inappropriatenes s in sentence formation.	Sufficient range of vocabulary to allow flexibility and precision; appropriate use of topic-related terms and less common lexical items	Appropriate use of a variety of phrases, such as idioms, collocations, and lexical bundles; occasional inaccuracies and colloquialisms.	Minimal errors in grammar and usage.	Generally consistent use of appropriate conventions to convey meaning; spelling, capitalization, and punctuation errors few and not distracting.
3	Facility limited to the use of common structures and generic vocabulary; organization generally controlled although connection sometimes absent or unsuccessful; errors in grammar and syntax and usage. Communication is impeded by language inaccuracies in some cases.	Organization generally controlled; cohesive devices used but limited in type; Some repetitive, mechanical, or faulty use of cohesion use within and/or between sentences and paragraphs.	Simple, compound, and complex syntactic structures present although the range may be limited; some apparent errors in sentence formation, especially in more complex sentences.	Minimally adequate range of vocabulary for the topic; no precise use of subtle word meanings; topic related terms only used occasionally; attempts to use less common vocabulary but with some inaccuracy	Evident use of phrases such as idioms, collocations, and lexical bundles but without much variety; some noticeable repetitions and misuses.	Some errors in grammar and usage.	Developing use of conventions to convey meaning; errors in spelling, capitalization, and punctuation that are sometimes distracting.
2	Inconsistent facility in sentence formation, word choice, and mechanics; organization partially developed but may be missing or unsuccessful. Communication impeded in many instances by language inaccuracies.	Organization only partially developed with a lack of logical sequencing of ideas; some basic cohesive devices used but with inaccuracy or repetition.	Some sentence variation used; many sentence structure problems.	Narrow range of vocabulary to convey basic and elementary meaning; topic related terms used inappropri ately; errors in word formation and word choice that may distort meanings	Narrow range of phrases, such as collocations and lexical bundles, used to convey basic and elementary meaning; many repetitions and /or misuses of phrases.	Many errors in grammar and usage.	Variable use of conventions; spelling, capitalization, and punctuation errors frequent and distracting.
1	A limited range of familiar words or phrases loosely strung together; frequent errors in grammar (including syntax) and usage. Communication impeded in most cases by language inaccuracies.	No clear control of organization; cohesive devices not present or unsuccessfully used; presentation of ideas unclear.	Pervasive and basic errors in sentence structure and word order that cause confusion; basic sentences errors common.	Limited vocabulary often inappropriately used; limited control of word choice and word forms; little attempt to use topic-related terms	Memorized chunks of language, or simple phrasal patterns predominate; many repetitions and misuses of phrases.	Errors in grammar and usage throughout.	Minimal use of conventions; spelling, capitalizatio n, and punctuation errors throughout.

Each essay in the initial ELLIPSE corpus of 8,890 essays was scored by two trained raters at minimum. Twenty-six raters were recruited from a large research university in the Southeast of the United States. Of the 26 raters, 21 identified as female, 3 identified as male, and 2 identified as other. Seven of the raters were undergraduate students (seniors), 12 were masters' students, two had completed a master's degree, and five were PhD students. The majority of the raters were in an applied linguistics department (n = 24) and the remaining two were in an English department. All raters had experience teaching English as a second language. Most of the raters were between

the ages of 20-30 (n = 16). Half of the raters were white. The remaining raters were Asian (n = 4), black (n = 2), Hispanic (n = 5), or represented multiple ethnicities (n = 2).

Prior to rating, all raters took the Implicit Bias Module Series developed by the Kirwan Institute at The Ohio State University (https://kirwaninstitute.osu.edu/implicit-bias-training) to mitigate potentially harmful unconscious biases held by raters. The series covers a wide range of topics including the formation of implicit bias and feasible ways to prevent and intervene against the bias. The raters spent around 50 minutes on the online bias training along with additional time discussing and addressing how bias may appear during scoring. After the bias training, all raters were trained on similar writing samples not included in the initial ELLIPSE corpus. This training involved familiarity with the rubric scales, the wording within the rubric, group scoring of essays, and independent practice scoring. Once an acceptable Cohen's Kappa was reached between raters (k = .60; Cohen, 1992), raters scored essays independently. Essays were assigned randomly to raters without any context (i.e., no background information was available to raters). All scoring was conducted using TagTog (https://www.tagtog.com), an online annotation and scoring system.

Kappa values for pre-adjudicated ratings showed agreement that was not reliable at k = .60 for the 8,890 essays (see Table 1). After rating, essays were adjudicated by raters if the reported difference between the two scores was greater than 1. Raters were asked to discuss any differences in ratings and make changes to their scores if needed.

Table 1. Inter-rater reliability for ELL proficiency rubric

Feature	Cohen's Kappa
Overall	0.599
Cohesion	0.562
Syntax	0.559
Vocabulary	0.518
Phraseology	0.561
Grammar	0.593
Conventions	0.580

After scoring, a Many-Facet Rasch Measurement (MFRM) analysis for the raters and texts was conducted to check additional aspects of reliability. We used Facets Version 3.83 (Linacre, 2021) to compute the probability of receiving a particular score on a rating scale as a function of the abilities of the candidate (i.e., the language proficiency scores for each ELL writer), the severity of the rater, and the difficulty of the rated item (e.g., the rubric items). MFRM reports strata and reliability for each function. Strata can be interpreted as the number of distinct levels of ability (in the case of texts), severity (in the case of raters), and scale level (in the case of the holistic/analytic scores). The reliability estimates reported by MFRM are analogous to Cronbach's alpha and report reliability of texts in terms of ability, raters in terms of difficulty, and scales in terms of levels.

An MFRM analysis affords the removal of texts, raters, and scales that have extreme scores that do not allow for latent variables to be measured with precision. This is usually done with an infit measure that measures consistency among texts, raters, and scales. Infit statistics are information-weighted, inlier-pattern-sensitive, mean square fit statistic with expectation 1 and range 0 to infinity (Linacre, 2021). Scores higher than 1 may indicate unmodeled excessive variation (or noise) and scores lower than 1 may lack of independence in rating (or too little variation). In this study, acceptable infit scores were judged to be between .6 and 1.4 following guidance provided by McNamara, Knoch, & Fan (2019). This recommendation represents the Likert scales used in this study.

The MFRM separated the essays into four levels of ability with a reliability of .94. Rater severity was separated into 13 levels with the most severe rater reporting severity of -1.21 and the most lenient rater reporting severity of 1.67. Reliability for rater severity was reported at .99. Infit for 25 out of 26 of the raters was between .6 and 1.4, which was acceptable. One rater reported an

infit of 1.45, which was outside of acceptability. The MFRM analysis for the scales was acceptable showing a 100% reliability and separated scales into 14 levels. The most difficult scale was for syntax (3.04) while vocabulary reported the easiest scale (3.24). The separation among scales is high, but it is a function of the many data points for each rating category.

In terms of essays, many essays showed high or low infit scores (i.e., below .6 or above 1.4), indicating that the distance from the mean scores had 40% more variation than predicted and were thus not reliably scored. After two iterations of pruning unreliable texts, 2,408 essays were removed leading to a corpus of 6,482 essays. A final MFRM analyses on the 6,482 remaining essays showed that the essays were still divided into 4 levels of difficulty with a reliability of .94. Rater severity was still separated into 13 levels with a reliability of .99, but all raters reported acceptable infit on the sub-corpus. The MFRM analysis for the scales was acceptable showing a 100% reliability and reported a separation statistic of 14. The most difficult scale was still syntax while vocabulary was still the easiest scale.

2.2 Final ELLIPSE corpus

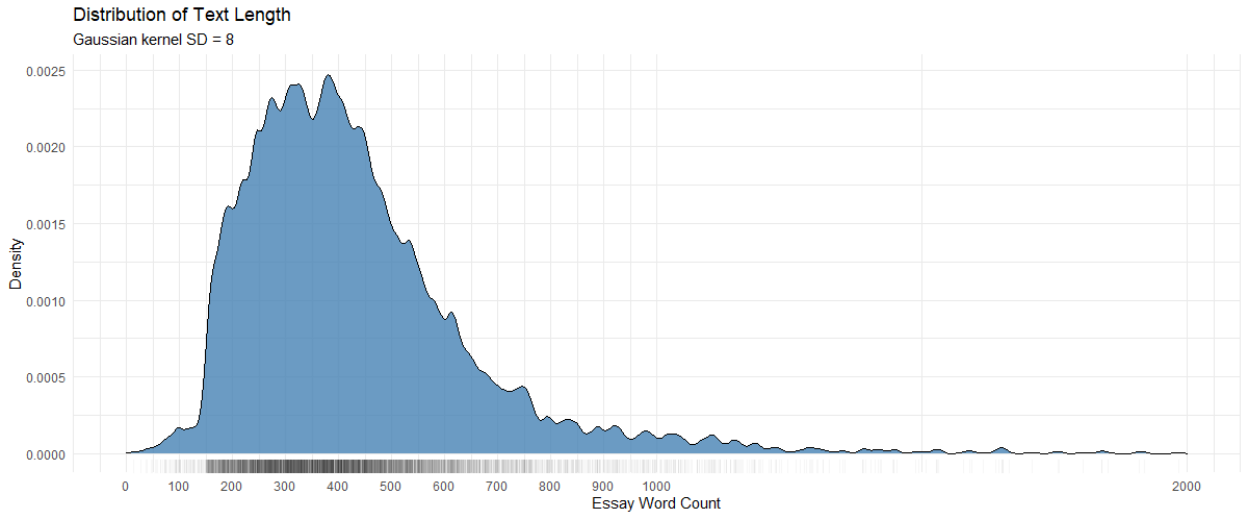
The final ELLIPSE corpus comprises the 6,482 texts that showed strong reliability in the MFRM analysis. Sixty percent of the corpus was released on Kaggle in 2022, which is an online community in which data scientists can enter competitions to solve data problems. The data was part of a competition (https://www.kaggle.com/competitions/feedback-prize-english-language-learning) to develop models to automatically assess language proficiency. The entire ELLIPSE dataset is available at https://github.com/scrosseye/ELLIPSE-Corpus. The dataset is stored in a dataframe that includes the ELL essays along with information about the essays including file

names, prompts, and simple descriptive data for each essay such as word count, sentence count, and paragraph count. The dataframe contains the average holistic and analytic scores from two raters for each essay along with demographic information about the writer including gender, race/ethnicity, grade level, and economic status. No information was available for learner background in terms of time in the United States and length of study. Descriptive statistics on the corpus are provided below.

2.2.1 Text statistics

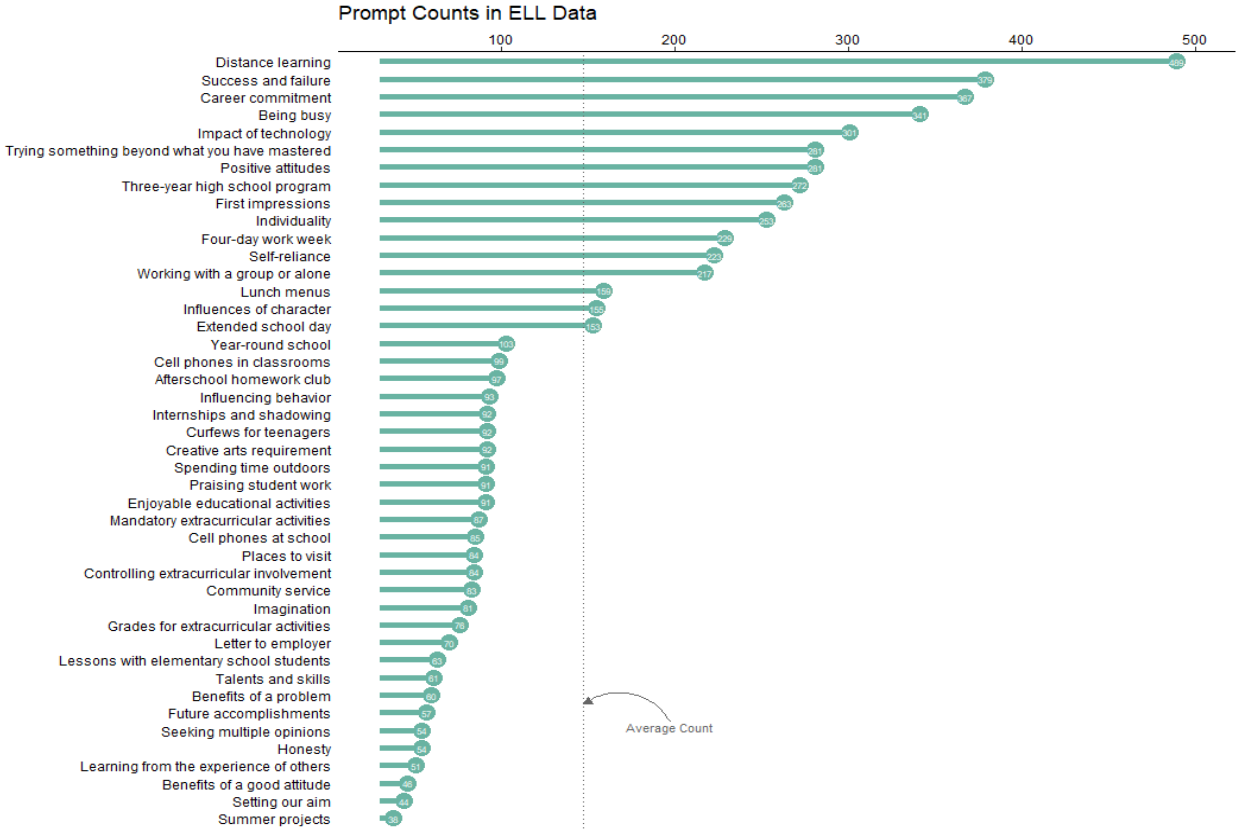
In general, the texts in the final ELLIPSE corpus are normally distributed in terms of word count although a number of essays are longer than average (see Figure 2). The shortest essay contains 17 words while the longest essay contains 1,274 words. On average, the essays contain 427.793 words (SD = 191.938).

Figure 2. Word distribution for the ELLIPSE corpus



For prompts, the average number of essays per prompt is 147.318 (SD = 110.672). *Distance Learning* is the most addressed prompt (n = 489) while *Summer Project* is the least addressed prompt (n = 38). Prompt distributions for the ELLIPSE corpus are presented in Figure 3.

Figure 3. Prompt distribution for the ELLIPSE corpus



2.2.2 Meta-data

Fifty-six percent of the ELL students in the ELLIPSE corpus identified as male (n = 3,636) with the remaining identify as female on binary scale. In terms of race/ethnicity, 72% of the students were Hispanic (n = 4,625) followed by Asians (n = 792 or 12%), blacks (n = 515 or 8%), whites (n = 471 or 7%), and those identified as two or more races or American Indian/Alaskan Native (both under 1%). Among these students, 70% were economically disadvantaged (n = 4,507). In addition, grade 11 students (average age ~ 16 years old) accounted for 35% of the population (n = 2, 280) followed by grade 12 (average age ~ 17 years old, n = 2,213 or 34%), grade 8 (average age ~ 13 years old, n = 1,627 or 25%), grade 10 (average age ~ 15 years old, n = 330 or 5%) and grade 9 (n = 32 or .5%). Thus, the ELLIPSE corpus represents a population rarely found second language studies because the participants are not students from North American

Universities (Plonsky, 2023). Rather, the participants are economically disadvantaged adolescents, many of whom likely come from central America (presuming the population sampled here follows the migration patterns found in the United States, U.S. Department of Education, 2017).

2.2.3. Score distribution

The holistic and analytic scores were normally distributed (see Figure 4). Scores by gender also demonstrated similar means across scales with few outliers (see Figure 5). Because of the sample size, *t*-test comparisons by gender across all scales were significant; however, all effect sizes, as measured by Cohen's *d*, showed no meaningful effects ($d < .20$). Scores by race and ethnicity also showed similar means. Because of sample sizes, ANOVAs demonstrated significant differences by groups. However, effect sizes, as measured by partial eta squared, showed no meaningful effects (partial eta squared < .01, see Figure 6). Scores by socio-economic status also demonstrated similar means across scales. Like the previous analyses, *t*-tests reported significant differences because of sample size, but these differences were not meaningful ($d < .20$, see Figure 7). Lastly, we examined score distributions by grade level. In general, similar mean scores were reported across grades for all scales. ANOVA results demonstrated significant differences for all scales, but this was a result of the sample size. A small effect size was reported for grammar (partial eta squared < .01) such that 8th graders reported higher grammar scores than 12th graders. No other meaningful effect sizes were reported (see Figure 8 for details).

Figure 4. Score distributions by scale

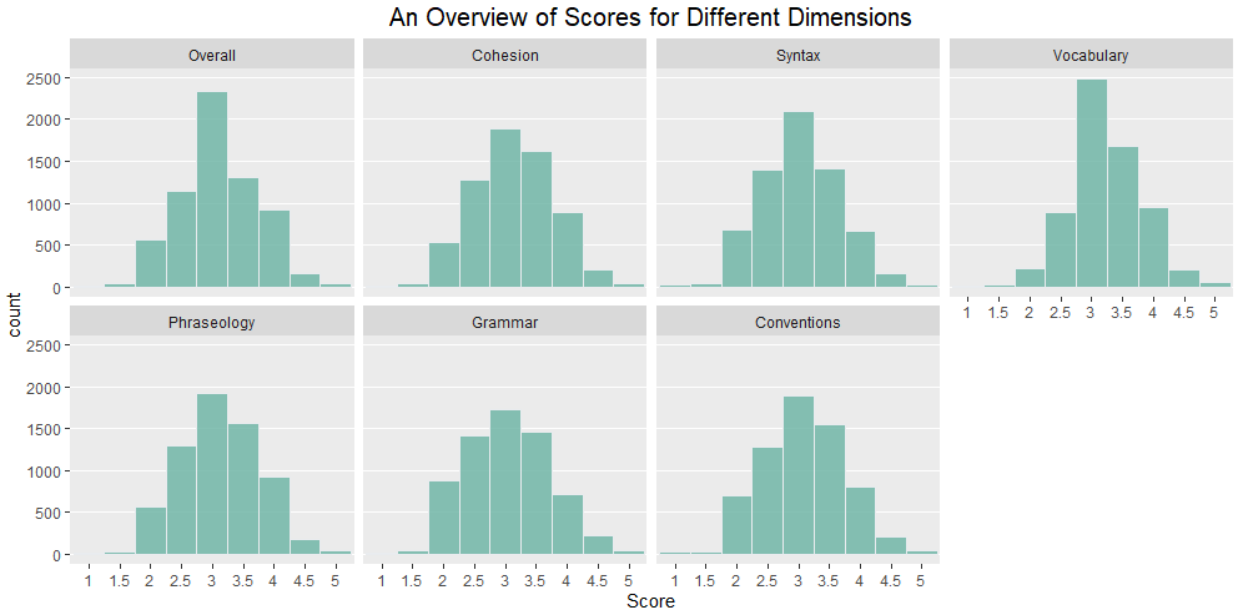


Figure 5. Score distribution by gender and scale

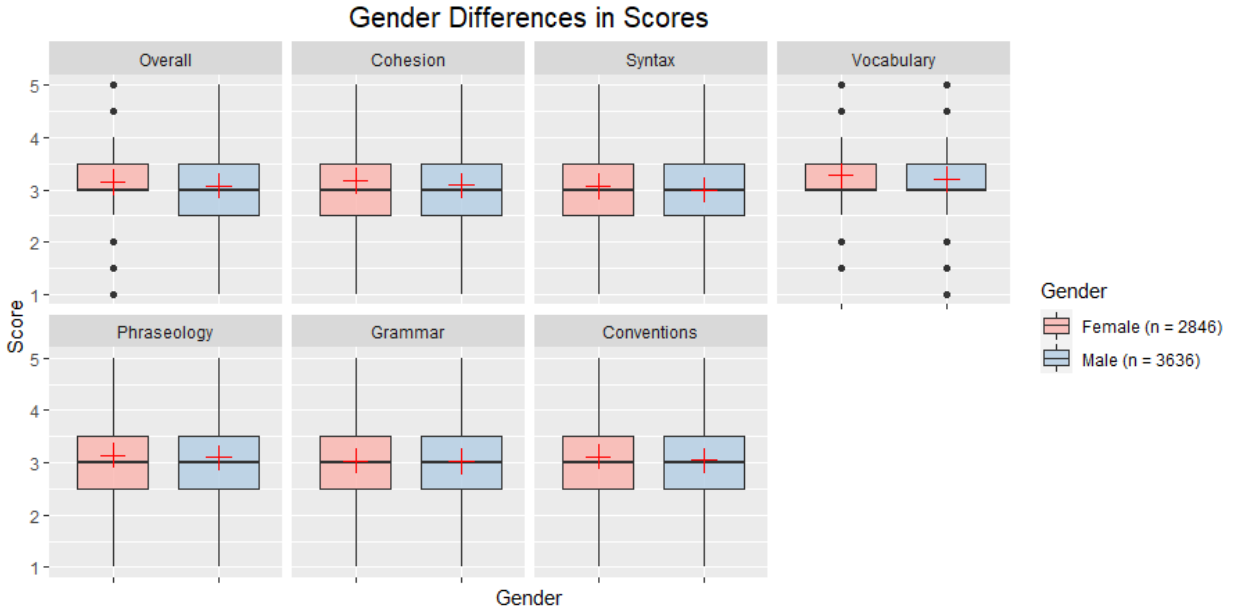


Figure 6. Score distribution by race/ethnicity and scale

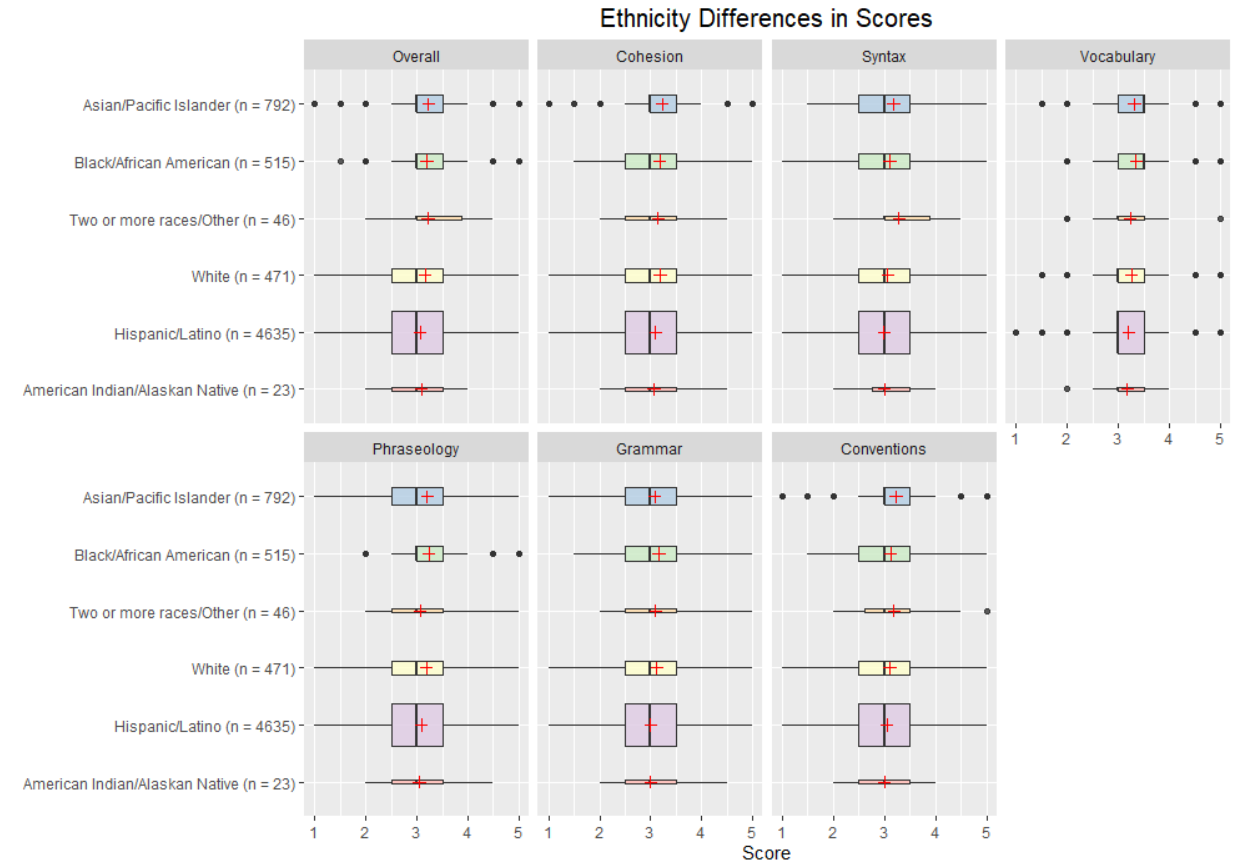


Figure 7. Scores by economic status and scale

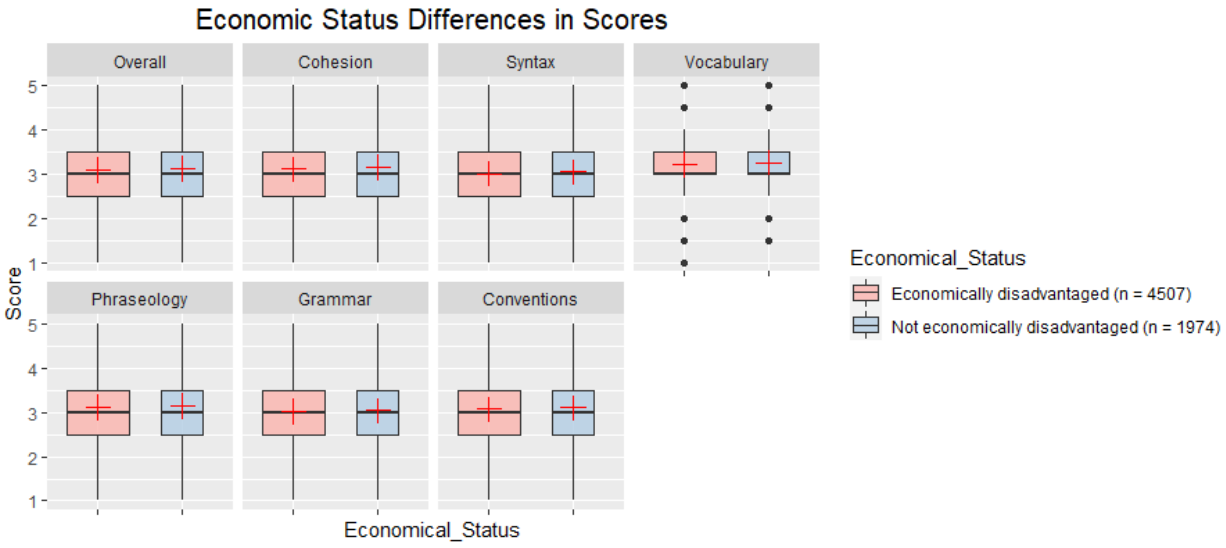
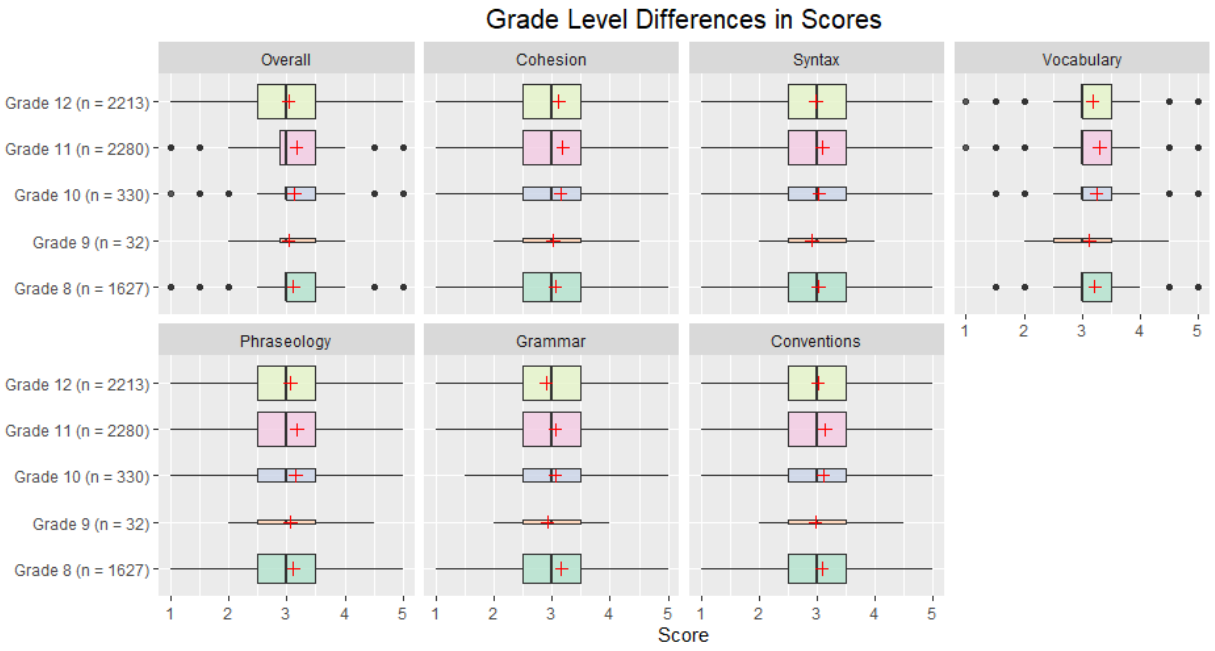


Figure 8. Scores by grade level and scale



3. Conclusion

This paper introduces the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus, a freely available corpus of ~6,500 ELL writing samples that have been scored for overall holistic language proficiency as well as analytic proficiency scores related to cohesion, syntax, vocabulary, phraseology, grammar, and conventions. In addition, the ELLIPSE corpus provides individual and demographic information for the ELL writers in the corpus including economic status, gender, grade level (8-12), and race/ethnicity. The corpus provides language proficiency scores for individual writers and was developed to advance research in corpus and NLP approaches to assess overall and more fine-grained features of proficiency.

Our assessments of reliability for the human scores of both the holistic and analytic ratings demonstrated lower than expected reliability for the initial corpus (N = ~8,000 essays) with all scales showing a Kappa value < .60. A Many-Facet Rasch Measurement (MFRM) analysis for the raters and texts was conducted to assess the reliability of specific raters, texts, and/or scales. The

initial MFRM analysis for raters indicated strong reliability with one rater reporting a level of severity that was not acceptable. The MFRM analysis indicated that all rating scales showed strong reliability. However, the MFRM analysis for the texts indicated that ~2,500 texts could not be reliably rated. These texts were removed creating a final ELLIPSE corpus of 6,482 essays. MFRM analysis of these essays indicated strong reliability in terms of raters, texts, and scales.

The ELL students in the final corpus represented a greater proportion of male versus female students (56% to 44%). The majority of these students were Hispanic and ~70% of the students were in the 11th or 12th grade with remainder from grades 8, 9, and 10. Around 70% of the students were economically disadvantaged. The scores for the essays were normally distributed and showed no meaningful effect sizes in terms of gender, race/ethnicity, and socio-economic status. There was a small effect size reported for grammar across grade level, but no other differences were reported for human scores by grade level.

There are a few limitations to the ELLIPSE corpus that are mostly a result of convenience and stratified sampling used to ensure a representative sample of students in terms of demographics and individual difference measures. The stratification process was difficult because ELL students in the United States are more likely Hispanic and from lower economic status (U.S. Department of Education, 2017), as seen in the population sampled. Another limitation in the ELLIPSE corpus is grade level representation, which is not evenly distributed across the corpus. In addition, the ELLIPSE corpus does not have an even prompt distribution with an upper-level count of 489 observations for the distance learning prompt and a lower-level count of 38 for the summer projects prompt. Additionally, the human scores indicated that students from lower grade levels received higher grammar scores. This may reflect random variance in the data, but this is unlikely given that similar scores from overall, syntactic, and phraseological proficiency (all of which likely

overlap with grammar scores) did not show differences. A more likely explanation is that ELL students in later grades may have immigrated to the United States at later ages (i.e., higher age of arrival) when grammar acquisition becomes more difficult (Birdsong, 2005).

Limitations aside, the ELLIPSE corpus will advance a number of research threads in learner corpus applications. We envision the ELLIPSE corpus advancing studies that automatically model language proficiency and eventually leading to systems that can provide feedback to students, teachers, and administrators about language proficiency in the moment and about development over time, likely in an intelligent Computer Assisted Language Learning (CALL) system (Choi, 2016; Meurers et al., 2016). The demographic and individual differences data can provide opportunities to examine differences in language development that may be related to non-cognitive factors and/or societal bias. The corpus also affords rich qualitative analyses of data in terms of language proficiency assessment. Lastly, the ELLIPSE corpus can be further annotated for aspects of writing related to pragmatics, discourse features, and rhetorical structures (among others) to examine more nuanced aspects of language proficiency not captured in the current annotation scheme.

Acknowledgments

We would like to acknowledge Sara Cushing for her assistance in the Many-Facet Rasch Measurement (MFRM) analysis.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bailey, A. L., & Kelly, K. R. (2010). The use and validity of home language surveys in state English language proficiency assessment systems: A review and issues perspective. *Evaluating the Validity of English Language Proficiency Assessment*. https://doi.org/10.4324/9780429491689-5

Birdsong, D. (2005). Interpreting age effects in second language acquisition. *Handbook of bilingualism: Psycholinguistic approaches*, 109, 127.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). TOEFL11: A corpus of non-native English. *ETS Research Report Series*, *2013*(2), 1–15. https://doi.org/10.1002/j.2333-8504.2013.tb02331.x

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., & Vettori, C. (2014). The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp: 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

Cheng, W., & Warren, M. (2005). Peer assessment of language proficiency. *Language testing*, *22*(1), 93-121. https://doi.org/10.1191/0265532205lt298oa

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. London: Routledge. https://doi.org/10.4324/9780203937891

Choi, I. (2016). Efficacy of an ICALL tutoring system and process-oriented corrective feedback.

Computer Assisted Language Learning, *29*(2), 334-364. https://doi.org/10.1080/09588221.2014.960941

Chomsky, C. (1972). Stages in Language Development and Reading Exposure. *Harvard Educational Review*, *42*(1), 1–33. https://doi.org/10.17763/haer.42.1.h78l676h28331480

Clifford, R., & Cox, T. L. (2013). Empirical validation of reading proficiency guidelines. *Foreign Language Annals*, *46*(1), 45-61.

Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, *1*(3), 98-101.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press

Crossley, S., Salsbury, T., McNamara, D.S. (2013). Validating lexical measures using human scores of lexical proficiency. *Vocabulary knowledge: Human ratings and automated measures.* Amsterdam: John Benjamins, 105-134. https://doi.org/10.1075/sibil.47.06ch4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, *49*(3), 803–821. https://doi.org/10.3758/s13428-016-0743-z

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, Coherence, and Expert Evaluations of Writing Proficiency. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32).

Ellis, R. (1991). Grammatically judgments and second language acquisition. *Studies in second language acquisition*, *13*(2), 161-186. https://doi.org/10.1017/s0272263100009931

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford university press.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, *66*(4), 477–485. https://doi.org/10.1093/elt/ccs037

Foddy, W. (1993). *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge University Press. https://doi.org/10.1017/cbo9780511518201

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013, October). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project* (pp. 240-254).

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Granena, G. (2019). Cognitive aptitudes and L2 speaking proficiency: Links between LLAMA and Hi-LAB. *Studies in Second Language Acquisition*, *41*(2), 313-336. https://doi.org/10.1017/s0272263118000256

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics*, *30*(4), 461-473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32). John Benjamins Publishing. https://doi.org/10.1075/lllt.32.01hou

Hymes, D. (1972). Editorial Introduction to Language in Society. *Language in Society*, *1*(1), 1–14. https://doi.org/10.1017/S0047404500006515

Ishikawa, S. I. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*, *1*(1), 91-118.

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing*, *32*(2), 227-258. https://doi.org/10.1177/0265532214558457

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757-786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. *The Modern Language Journal*, *102*(2), 333–349. https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Lagakis, P., & Demetriadis, S. (2021). Automated essay scoring: A review of the field. *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*, 1–6. https://doi.org/10.1109/CITS52676.2021.9618476

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability.

Language Testing, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Larsen-Freeman, D. (1978). An ESL Index of Development. *TESOL Quarterly*, *12*(4), 439. https://doi.org/10.2307/3586142

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543–560. https://doi.org/10.1177/0265532211406422

Linacre, J. M. (2021). *A User's Guide to FACETS Rasch-Model Computer Programs Program Manual* 3.83.5.

Lisken-Gasparro, J.E. (1984). The ACTFL proficiency guidelines: Gateway to testing and curriculum. *Foreign Language Annals 17(5), 475-489.* https://doi.org/10.1111/j.1944-9720.1984.tb01736.x

Lumley, T. (1998). Perceptions of Language-trained Raters and Occupational Experts in a Test of Occupational English Language Profficiency. *English for Specific Purposes*, *17*(4), 347–367. https://doi.org/10.1016/S0889-4906(97)00016-1

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Lang.

McNamara, T., Knoch, U., Fan, J., & Rossner, R. (2019). *Fairness, justice & language assessment*. Oxford University Press.

Meurers, D., De Kuthy, K., Nuxoll, F., Rudzewitz, B., & Ziai, R. (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics, 39,* 161-188. https://doi.org/10.1017/S0267190519000126

Ortega, L. (2012). Epilogue: Exploring L2 writing–SLA interfaces. *Journal of Second Language Writing, 21*(4), 404-415. https://doi.org/10.1016/j.jslw.2012.09.002

O'Sullivan, B. (2018). IELTS (international English language testing system). *The TESOL*

Encyclopedia of English Language Teaching, 1-8. https://doi.org/10.1002/9781118784235.eelt0359

Plonsky L. (2023). Sampling and Generalizability in Lx Research: A Second-Order Synthesis. *Languages*. 8(1):75. https://doi.org/10.3390/languages8010075

Skehan, P. (1989). *Individual differences in second-language learning*. Edward Arnold.

U.S. Department of Education. (2017). *Our nation's English learners*. Washington, DC: US Department of Education. Retrieved from https://www2.ed.gov/datastory/el-characteristics/index.html.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*(1), 27–55. https://doi.org/10.1016/j.asw.2004.01.002

Widdowson, H. G. (1983). *Learning purpose and language use*. Oxford University Press.

Wood, C., & Schatschneider, C. (2021). Examining Writing Measures and Achievement for Students of Varied Language Abilities and Linguistic. *Reading and Writing Quarterly*, *37*(1), 65–81. https://doi.org/10.1080/10573569.2020.1716284