Summer 2022

Analysis of the Correlation Between the Lexical Profile and Coh-Metrix 3.0 Text Easability and Readability Indices of the Korean CSAT From 1994–2022

Andrew Howie
Central Washington University, andrewshowie@icloud.com

ANALYSIS OF THE CORRELATION BETWEEN THE LEXICAL PROFILE AND COH-METRIX 3.0 TEXT EASABILITY AND READABILITY INDICES OF THE KOREAN CSAT FROM 1994–2022

A Thesis

Presented to

The Graduate Faculty

Central Washington University

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

English (TESOL)

by

Andrew Howie

July 2022

CENTRAL WASHINGTON UNIVERSITY

Graduate Studies

We hereby approve the thesis of

Andrew Howie

Candidate for the degree of Master of Arts

APPROVED FOR THE GRADUATE FACULTY

______________ ____
Dr. Charles Li, Committee Chair

______________ ____
Dr. Loretta Gray

______________ ____
Dr. Penglin Wang

______________ ____
Dean of Graduate Studies

ABSTRACT

ANALYSIS OF THE CORRELATION BETWEEN THE LEXICAL PROFILE AND COH-METRIX 3.0 TEXT EASABILITY AND READABILITY INDICES OF THE KOREAN CSAT FROM 1994–2022

by

Andrew Howie

July 2022

The Korean College Scholastic Ability Test (CSAT) is a highly competitive standardized assessment that graduating high-school seniors complete in the hope of getting a good score which will improve their chances of admission to a university of choice. The CSAT contains an English Section that has been described by scholars and educators alike as being far too difficult for the official English language curriculum to serve as sufficient preparation. The test's lack of construct validity has been the basis for calls to revise the test to be better reflective of the school curriculum so that it can serve the evaluative purpose for which it is intended. Use of automated text evaluation methods with the software Coh-Metrix 3.0 in recent years has allowed scholars to quantify different dimensions of the text of the CSAT English Section, such as cohesion and syntactic complexity, that contribute to its reading difficulty. Older research conducted before the introduction of this software into the field used word frequency counts in large corpora such as the British National Corpus (BNC) as a measure of word familiarity or unfamiliarity, which was thought to directly contribute to difficulty because as the proportion of low-frequency words in a text increases against the proportion of high-frequency words, the word knowledge burden of the text increases in proportion. Since the introduction of automated software-based tools like Coh-Metrix 3.0 and Lexical Complexity Analyzer (LCA), these corpus-based research methods have largely fallen by the wayside. In this paper, I maintain that despite its lower sophistication, corpus-based lexical analysis can still produce uniquely meaningful findings because of the degree of manual control the researcher is afforded in calibrating the parameters of the text base and, most importantly, in selecting the ranges of word family frequency that are best tailored to a text rather than having the ranges or functions of frequency assigned automatically by software. This study reports correlations between the outputs of these two methodologies that both inform us about the validity of Coh-Metrix 3.0's use in CSAT studies and quantify the strength of the role of word frequency in causing the excessive difficulty of the CSAT English Section.

ACKNOWLEDGEMENTS

This work began with the suggestion of my advisor, Dr. Charles X. Li, that I consider learning more about Coh-Metrix as a methodological tool when I came to him to confess that I was having difficulty deciding upon a research direction. Second language pedagogy has always been at the core of my academic interests, and my recent reading at the time had been in the area of second language assessment. When Dr. Li introduced me to this powerful toolset, I felt a world of possibilities opening before my eyes, and the very first text I wanted to try evaluating with the software was the CSAT English Reading Section, which I did as soon as I got home in the evening after that meeting. Now, it seems there are years, and even decades worth of research in this specific area that I could dedicate myself to and still find no end of new questions to be answered. It is all because Dr. Li sparked my curiosity with the right knowledge, at the right time, that I have practically resolved to make a career out of this one subfield.

I have said nothing about the other support, meaningful advice, and timely guidance Dr. Li has provided as my advisor during the course of my time spent at CWU. There is no way I could name everything without writing another whole thesis at the beginning of my actual thesis, so for now I must be brief. Dr. Loretta Gray has been such a wonderful instructor and mentor and kindly extended deadlines and made significant accommodations for my work and writing schedules, without which I would not have been able to succeed either in my classes or in writing this thesis. While studying for her courses, I always felt that I was learning something significant, much more than most other courses I have taken in my life, without feeling like I was being rushed or pressured to conform to any goals or standards other than the ones I set for myself. This is truly the mark of an excellent instructor, that learning from them could feel impactful and exciting but not tedious. I hope to learn from her example when I become an instructor.

My deep thanks go to Dr. Penglin Wang, who agreed to serve as committee member and went to great lengths to accommodate the evolving schedule of this project's composition alongside an already busy summer schedule. Dr. Wang's flexibility in accepting the task of reading my thesis while on vacation in order to allow for a change in defense date is deeply appreciated. I am endlessly grateful for the kind support I have received from the greater CWU academic community, and Dr. Wang's actions are exemplary of this.

Finally, the greatest thanks must go to my partner in life, Sol, who has empowered me to set no limitations on myself and has inspired me to pursue my dream of becoming a professor. Without her, I would still be in an unrelated career that was making me unhappy, and based on her trust, encouragement, and unconditional love, I have found a career goal that makes me excited to get up every day and just take one more small step forward. Without her help and support, it would have been difficult to make this paper come together in such a short time, and there's no way I can fully convey to her how much that means to me and my future. The journey ahead may still be full of difficulties, but I never doubt that as long as she is by my side, I will be successful in achieving my dreams, and that is why I dedicate this work to Sol, the love of my life and the one person in the world who is always there for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

English education is a significant national priority in South Korea because of its association with increased collective opportunity to engage with international markets and thus expand the reach of Korean industry and culture. As a matter of national economic advancement, the goal of improving the English proficiency of the populace has guided several targeted reforms in second-language pedagogy. In 1994, the new Korean College Scholastic Ability Test (CSAT) English Reading Section was introduced, alongside a pilot English Listening Section, to serve a direct evaluative purpose for the National Curriculum of English (NCE), a standardized second-language curriculum implemented throughout South Korea's public schools. Aside from English, the CSAT introduced new assessments in other content areas of the National Curriculum such as mathematics, social studies, science, and Korean language. The new English assessment of the CSAT broke with previous tradition by focusing exclusively on comprehension and eschewing recitation of explicit linguistic knowledge.

The CSAT English Reading and Listening sections represented a dramatic change from the previous College Entrance Academic Proficiency Test (J. Lee, 2009; O. Kwon, 2015). A written test devoted entirely to assessment of explicit grammatical and lexical knowledge, the College Entrance Academic Proficiency Test had been increasingly understood to be a poor measure of fluency in light of emerging second-language teaching theory, which emphasized communicative competence. Two separate sections—English Reading, which wholly consisted of reading comprehension questions, and English Listening, which assessed listening comprehension—took the place of the one outmoded English section when the CSAT was introduced in 1994. These represented the first in a series of English educational reforms intended to shift the pedagogical focus away from rote memorization of grammatical, lexical, and phonological information towards comprehension of whole texts. The fluency-first approach of the new English Section was intended to foster a greater communicative focus in secondary-school English classes through positive washback.

Historical Background

The role occupied by the CSAT in the education system is comparable to that of similar standardized tests in other countries in that it is a highly competitive annual comprehensive evaluation that serves the purpose of determining the college readiness of graduating high-school seniors. Among factors that affect an applicant's likelihood of admission to university, scores on the test are, by far, the most impactful, and are understood to influence the course of an individual's entire life (H. Lee, 2018; J. Lee, 2020; Lee & Lee, 2018). Notoriously difficult, the English Reading Section of the CSAT has made news headlines for such features as its emphasis on low-frequency vocabulary and high incidence of atypical collocations. The abnormally difficult language in the CSAT has become a pop-cultural touchstone in South Korea; in a video with over 10 million views, YouTuber Korean Englishman challenges British high-school students to answer the questions on the test, and very few of them are able to get many correct. Prompted for comments, the students all agree that the language was either unnatural or "not [real] English," with one student remarking that being a native speaker does not seem to provide the expected advantage in completing the test (Korean Englishman, 2021, 4'18"). An English-language Korea Times article from 2018 prompts readers to “check [their] English ability” with some of the passages to demonstrate that the test is difficult not only for second-language (L2) English learners, but also for first-language (L1) learners. In a provided example that is typical of many CSAT question types, a passage begins with the sentence “Although not the explicit goal, the best science can really be seen as refining ignorance,” where the underlined phrase becomes the subject of a set of multiple-choice questions (S. Park, 2018). The passage proceeds to very circuitously explain that scientists are always looking forward to penetrating the next frontier of the unknown, i.e., the areas about which they are ignorant, and that this is the default orientation of the scientific method rather than focusing on expanding the body of what is known. The unnatural terseness with which this and similar passages convey their concepts belies an astonishing commitment to lexical density. The passages presented by this article seem to reinforce the notion that the language in the test is artificial—not an appropriate representation of real-world target usage.

It is confirmed in recent scholarship that not only is the language of the test extremely difficult, but it is becoming more difficult each year (S. Kim, 2021; Koh & Shin, 2017; J. Lee, 2020; Moon & Kim, 2017). Scholars have increasingly observed that the test’s content seems to far exceed the targets established by the NCE, resulting in heavy reliance among the country’s youth on private cram schools. The construct validity of the test as an evaluation of aptitude in the NCE has been repeatedly called into question in these studies where the test content is found to be unreflective of the curriculum’s guidelines (Shin, 2019). The test has been steadily increasing in difficulty since its introduction in 1994, which further heightens concerns that there is a mismatch between the CSAT’s stated goal of measuring aptitude in the NCE and the actual content of the test, that increased need for reliance on the private education industry exacerbates the burden on low-income and provincial families, and that interest in acquisition of "test-wiseness skills" suffocates efforts to introduce communicative language teaching (CLT) into English as a foreign language (EFL) classrooms (Chon, 2014, p. 366; H. Lee, 2018).

Overview of Positive and Negative Washback

The test exerts tremendous pressure on students and their families and strongly determines the orientation of English curricula and pedagogical practices. This pressure relates to the institutional aim in having an all-important test in the first place, which is to exert control and guidance over the teaching practices of the nation and ensure their compliance through the powerful influence of washback (J. Lee, 2020). The exceptional emphasis on the test in Korean society is not accidental, but by design. In 2000, six years after the introduction of the new test, O. Kwon (2015) found that a positive washback effect was discernible in Korean secondary-school English education. There was an increased focus on reading and listening comprehension and a much-reduced emphasis on analysis of grammatical structures. Indeed, the initial CSAT represented the first time that a listening section was included in an English national exam in the Korean education system (B. Chang, 2009). In terms of language teaching theory, it could be observed that English education in Korea was moving away from the theory-less approach of the Grammar-Translation Method and towards a soft version of the Communicative Approach, which eschews accumulation of linguistic knowledge alone in favor of equipping students with the means to negotiate meaning through active cognition in language contexts (Larsen-Freeman & Anderson, 2011). Moreover, English-language education had begun to be introduced as early as elementary schools starting in 1997, a change enacted via a national mandate that has since been identified as contributing to Korean students' overall greater fluency in English later in life compared to that of previous generations (Yoo, 2016). Around the early 2000s, it seemed like the instrumentality of the CSAT English Section in achieving the institutional aim of curriculum reform was upheld through the demonstrated effect of positive washback. The power of washback from the test was a double-edged sword, however. Unfortunately, manifestations of negative washback began to surface in the late aughts (J. Lee, 2020).

The types of negative washback from the CSAT English Section fall within two categories: those which affect the practice of English education itself, both in terms of curriculum and teaching, and those which are negative externalities of extreme competition. As an all-encompassing evaluation, the CSAT fostered intense competition from the very beginning (J. Lee, 2020). Reliance on private education across the society increased as a consequence of an ever-intensifying race to the top. The scoring of the test was, until 2018, norm-referenced rather than criterion-based. This meant that students' scores were judged not according to an absolute standard, but according to a percentile rank; this type of scoring system pits students directly against one another (H. Lee, 2018; J. Lee, 2020). Wealthy families were able to afford the best private English education for their sons and daughters, creating pressure on less-privileged families to ratchet up the percentage of their total income devoted to private-education solutions. A proposal to switch to a criterion-referenced assessment starting with the 2018 test came from a motivation to alleviate the burden on students to "spend much money on private education to get higher scores than others" and to restructure the test in a way that "does not focus on discrimination" (J. Lee, 2020, p. 23), but in practice, little has changed since this shift because the problems with the test stem more from the inherent nature of its textual composition than the methods used to score it. The test remains just as difficult as it had been before and just as out of step with the NCE, both in readability and easability measures and in terms of score outcomes and continued social problems (J. Chang, 2019a; S. Kim, 2021; J. Lee, 2020; Shin, 2019).

Having established the negative externalities engendered by the CSAT English Section in terms of social issues, I now turn towards the effect it has had on the practice of English education in Korea. I observed that after its introduction, there were indications of the test succeeding in its aim to shift the English curriculum of the country away from an approach resembling the Grammar-Translation Method and towards a more communicative approach. This goal was intended to be accomplished through a change in test composition that replaced the previous test's exclusive focus on explicit grammatical and lexical knowledge with an exclusive focus on the two areas of reading comprehension and listening comprehension. It was thought that the powerful effect of washback would cause teachers to develop communicative approaches in their classrooms that would stress overall fluency and the development of implicit linguistic knowledge rather than rote memorization of explicit linguistic knowledge. Although there was an initial improvement where teachers were now instructing students in the comprehension of whole texts through strategies that *are* also applicable to communication, such as prediction, the fiercely competitive nature of the test, which has only increased over time, meant that teachers were so pressured to "tailor their teaching according to the demands of the CSAT" that they "refuse or are unable to introduce CLT activities in class" (H. Lee, 2018, p. 281). And even though it is possible that students who learn English through CLT methods would likely score better on the CSAT than those who do not, H. Lee (2018) also notes that the "contents of the CSAT do not evaluate communicative competence"—i.e., there are no sections that evaluate productive skills such as speaking or writing (p. 280). The whole principle of washback from the CSAT being a force that can foster communicative English teaching is undermined on two fronts: first, the test lacks validity to what it is intended to encourage, i.e., communicative competency, because it does not evaluate this; and second, the nature of washback itself is at odds with language approaches that want to teach anything other than test-taking strategies—the test becomes its own objective separate from general English aptitude (Chon, 2014). Not only is the test not valid from the standpoint of executing the aim of indirectly making language classrooms more communicative, but it also lacks validity from the standpoint of accurately assessing competence in the NCE, which is a problem that could not be solved by changing scoring methodologies.

Justification for the Present Inquiry into the Validity of the Test

Studies of the text of the CSAT English Section have produced fruitful results in defining how, exactly, the test diverges from the instruction students receive in secondary school. For most of the history of the field, scholarship was mainly devoted to lexical examination of the vocabulary of the CSAT English Section, using corpus-based methods in an attempt to demonstrate the mismatch between its parameters and the guidelines set forth by the NCE. Chon's 2014 diachronic lexical analysis of the CSAT from 1994 to the present demonstrated that the CSAT contains not only a much greater lexical diversity than the NCE advises, but also an increasingly disproportionate quantity of low-frequency versus high-frequency words. In this work, Chon shows that the high-frequency band of 1000 most common word families was increasingly underrepresented in the test over time, and that the lower-frequency bands of word families—sixth and up—were noticeably increasing in proportion to the higher-frequency word families. Chon (2014) also found that each new iteration of the CSAT introduced a greater total of new words than the last. What resulted from calculating the mean word coverage of the CSAT were data showing that, as of the time of the author's writing, the lexical demands of the test were so great as to require knowledge of 13,000 word families for 98% word coverage of the test, compared to a lexical standard of 3,500 provided by the NCE (Chon, 2014).

A recent development over the last few years has been in finding ways to leverage the power of automated text evaluation software to diagnose the test's problems in the areas of difficulty and validity and point the way towards pedagogical solutions. Looking at the areas of reading easability and difficulty alongside textual attributes like word information, syntactic complexity, and cohesion—concepts which automated software algorithms can measure by calculating percentile scores from formulae based on the characteristics of a text that researchers have demonstrated to co-occur with the incidence of these concepts—scholars in the past couple of years have been able to find trends in these various attributes of the CSAT over time that tell us, across many more measures than just the lexical information provided by earlier studies, what exactly is wrong with the test and how it is evolving. The vast majority of recent scholarship on the CSAT English Section has engaged in automated text evaluation with software such Coh-Metrix 3.0 (McNamara et al., 2014) and Lexical Complexity Analyzer (Lu, 2022), universally finding, through the programs' built-in measures of textual attributes related to difficulty and easability, that not only is the test extremely difficult, but it is getting more difficult each year (S. Kim, 2021; Koh & Shin, 2017; J. Lee, 2020; Moon & Kim, 2017). The construct validity of the test as an evaluation of aptitude in the NCE has been repeatedly called into question in these studies (Shin, 2019).

The earlier work of Chon (2014) on the CSAT English Section's lexical profile and word-frequency thresholds remains of high importance, however, despite the recent shift to automated software-based evaluation in the scholarly community. It is one of the few diachronic studies of the CSAT overall and the only diachronic study to employ the methodology of corpus-based lexical analysis. Chon provides data that are able to compare directly with learning thresholds provided by the NCE, and, therefore, have the power to convincingly relate the findings of scholarship in this field to policymakers. The figures produced by automated text evaluation software may provide information in a wider range of conceptual categories and thus enrich our picture of the inner workings of a text, but what Chon (2014) finds in her more narrowly focused investigation, limited to the simple metric of word counts, is more easily quantifiable and communicable to a wider audience of stakeholders and influencers in education. Because of her wide, diachronic view, Chon (2014) is able to connect the massive shortfall in word knowledge between the official curriculum and the increasingly verbose CSAT English Section with social problems like the extreme over-reliance on private, after-school English education among the country's youth, a perspective which aligns with the conclusions new Coh-Metrix and Lexical Complexity Analyzer (LCA) studies have since brought forward in establishing the pedagogical ramifications of the test's textual attributes (J. Chang, 2018a; C. Kim, 2020; Kim & Kim, 2021; J. Lee, 2020; Om, 2021; Shin, 2019).

What is missing from the current body of research, however, is a synthesis of the findings from lexical analysis and automated text evaluation. The older studies that relied on corpus-based lexical evaluation still offer valuable insight, even though this type of research has not been significantly iterated since 2014. Chon (2014) was able to find, for example, that the test had been increasing its proportion of low-frequency vocabulary relative to high-frequency vocabulary over time across 14 bands of word frequency, which is not information that automated evaluation software is currently equipped to provide. Although the automated software-based approaches are meaningful when supported by reliable data collection methods, there is in fact a lot to be gained from inclusion of the lexical analysis data in this discussion, which puts the researcher in more direct contact with the source material and the corpora used for comparison. One powerful reason to synthesize multiple approaches is that it prevents the research from relying entirely on one piece of automated software as the basis for analysis and, therefore, the study stands a greater chance of producing meaningful results. Additionally, it is possible to comment upon the validity of methodologies by searching for correlations between their findings. This study was designed to show whether there is a correlation between Coh-Metrix 3.0 text readability scores in the CSAT English Section and changes in the lexical profile of the test over time. The methodologies of Chon (2014) and the various diachronic Coh-Metrix studies have provided a springboard for the use of a longitudinal focus that represents a summation of the most current data on the CSAT. The predicted finding is a negative correlation between text readability and lexical attributes that are thought to co-occur with high text difficulty. This study also provides a correlation analysis between Coh-Metrix indices and lexical profile of the CSAT over time, which is expected to result in strong correlations if Coh-Metrix is indeed a reliable tool for CSAT research.

The next chapter reviews the history of literature on the CSAT in the two categories of lexical analysis and automated text evaluation and provide supporting justification for this study's contribution to the field. The third chapter provides an overview of the methods used in collecting, interpreting, and analyzing data. Subsequently, the fourth chapter displays the results of each area of data analysis operation conducted and the fifth chapter engages in discussion of those results. Finally, the sixth chapter summarizes the findings with concluding remarks, pedagogical implications, and possible areas for future work in the field.

CHAPTER II

LITERATURE REVIEW

This study has been developed primarily out of concern for the absence of scholarship synthesizing recent findings with the long span of corpus-based lexical analysis of the CSAT English Reading Section that predates the current trend towards using Coh-Metrix to automatically evaluate the text. Valuable findings attained through corpus-based lexical analysis have been left underdeveloped as new, more refined methodologies for the application of automated software to study of the CSAT emerge year after year. Following the lead of H. Lee (2020) and M. Choi (2018), who in their reviews of the state of the literature on CSAT English reading items referred to this history of lexical analysis preceding the current focus on methodologies designed around use of Coh-Metrix, this chapter proceeds with a structure informed by this dichotomy.

The first section below provides an introduction to the research methods and terminology which are essential for proper understanding of both past scholarly contributions and the orientation of the present work. The second section below surveys past literature structured around corpus-based lexical analysis of the CSAT English Section. Finally, this chapter concludes with an overview of the current state of research in the field, which is centered around use of automated text evaluation with Coh-Metrix.

Introduction to Research Methods and Terminology

The first step in any research is proper definition of terms. In this field, there is an abundance of terminology that reflects the conventions of automated text evaluation software. When this study refers to automated text evaluation, this includes both Coh-Metrix 3.0 and LCA, both of which are web-based software that run computational algorithms on inputted text and calculate numerical scores in the categories of various indices, which are simply indicators or ways of measuring concepts, which then provide the researcher data about the possible operation of those concepts in the text. This paper does not make use of LCA because of its relatively low utilization among scholars in the field, although some reference to it is made throughout this chapter. The standard terminology of Coh-Metrix relies in large part on what McNamara et al. (2014), established in their manual on the use of the software in research. Coh-Metrix 3.0 is a web-based computational tool that evaluates language and discourse over 106 indices that can tell researchers, for example, how long the average word or sentence is in a given sample of text, how frequently arguments or syntactic constructions repeat, or the proportion of connective words. Applied to the written word, many of these indices can tell us, to greater or lesser degrees, about attributes of a given text that scholars have come to associate with reading difficulty.

The first, and perhaps most important, of these attributes is cohesion. McNamara et al. (2014), begin their introduction to the functions of Coh-Metrix with a whole chapter dedicated to cohesion, which they claim, based on an abundance of scholarly agreement, "critically determines both how challenging a text is and how well the reader will understand it" (p. 18). Cohesion refers to how well a text structures connections and logical relationships between content and to what extent it is able to, and how readily and without extensively relying upon the reader's knowledge of content outside of the text, foster coherent understanding when it is read. One of the most crucial assumptions of Coh-Metrix as a research tool is that cohesion is something that can be computationally measured (McNamara et al., 2014).

Coherence, on the other hand, is a phenomenon that can only be measured indirectly because it occurs in the mind of the reader, who makes inferences and maps the logical structure of the text onto their own understanding (McNamara et al., 2014). High-cohesion texts are likely to produce coherent understanding in the reader, and low-cohesion texts are correspondingly less likely to do so. Texts that are low in cohesion are much more difficult for readers who either lack sufficient knowledge of the conceptual world of the text, such as children, or whose skill in the target language is underdeveloped, as with beginning and intermediate L2 learners. Therefore, cohesion is an important attribute of text for the research and development of pedagogical materials such as textbooks and standardized tests. There are many indices programmed into Coh-Metrix that do not directly relate to cohesion, but many of them do relate to it in one way or another. The indices that are used in this research are enumerated in more detail, and the reasons for their inclusion are justified, during the chapter on methods.

This study makes careful use of the terms “evaluation” and “analysis” to differentiate the work performed by automated software, on the one hand, and the procedures of corpus-based lexical studies on the other. There is no underlying semantic reason for this differentiation other than perhaps the notion that software does not perform “analysis” in the same manner that a human researcher does when they are presented with the data output of the software or when selecting, performing, and reporting analytical statistics. The true reason is to guide the reader’s understanding of an important distinction in scholarly practice and to allow clear communication of the methodological orientation of this study. It is important to keep the two concepts of automated evaluation and corpus-based lexical analysis separate in this discussion because one of the critical concepts underlying this work is the lack of integration of two subfields which are largely kept separate on the grounds that the limitations of corpus-based analysis render it unable to move the field forward when automated evaluation offers more complete data (H. Lee, 2020). The reasons for this trend are critically examined later in this study during discussion of the results. Needless to say, corpus-based lexical analysis is also powered by software, so it is not correct to simply say that it is manual rather than automatic, but it is a more human-guided process in that interpretation of the output of data extracted from the corpus depends entirely upon human interpretation; there are no indices that fit multiple atomic measurements into conceptual packages or any built-in algorithms to execute formulas upon these packages. This can indeed be a limitation, but it can also be a strength. These distinctions are explored further in the chapter on methods.

When I talk about corpus-based lexical analysis, I am referring to the act of pulling data from an external corpus and using my understanding of what I find there to perform analysis on the text under examination. There are many corpora to choose from, but this study follows from the essential work of Chon (2014), which is perhaps the greatest single inspiration for the present research design, in choosing the British National Corpus (BNC) as a foundation. There is a software tool designed by Nation called RANGE that pulls words out of the text assigned to it and counts the word families present there (Nation, 2022). It then categorizes the word families by frequency of occurrence in the corpus with reference to 14 bands consisting of 1,000 words each that are ranked by frequency. Words in the first band are very common in the BNC, but words in the 14th band are very rare. This is, of course, simply a much faster and more efficient alternative to manually counting the word families in the corpus and the text and grouping them together based on how often they occur. There are other corpora and corpus tools available that are used by researchers in the field, but since this work is concerned with carrying forward the research design of Chon (2014), her justifications for choosing to use the BNC, "one of the largest corpora of present-day English usage," and RANGE, which, quite helpful for attaining convincing results, makes it so that "analysis of the corpus is possible at fourteen different levels," are carried forward in this study as well (p. 350).

The term "lexical profile" refers to the product of a lexical analysis of a text that is sufficient and complete from the standpoint of the study's objectives. In Chon (2014), because she is manipulating multiple different dimensions of the text in her lexical analysis, develops three different categorizations of results, one of which deals with type-token ratio (TTR), another which looks at word frequency, and another which formulates lexical thresholds. In this tripartite division, she uses the term "lexical profile" to refer to the first of the three headings (p. 348). I here depart from this usage, first, because "lexical profile" is a convenient designation, and second, this study produces only one type of result under the rubric of corpus-based lexical analysis, which is word frequency, the reasons for which are enumerated in the chapter on methods.

Corpus-Based Lexical Analysis Studies

Prior to 2017, studies of the CSAT English Reading Section were very strongly characterized by methodologies founded on corpus-based lexical analysis (G. Choi, 2015; Goh & Back, 2010a; Goh & Back, 2010b; Joo, 2008; J-W. Kim, 2016; N. Kim, 2008; J. Lee, 2011; C. Lee, 2010; Song, 2012; Yoon, 2006). The focus on vocabulary was very keen during the first decade of research in the field because it was a self-evident index of text difficulty (M. Choi, 2018). The greater the proportion of difficult or unfamiliar words in a text, the more likely it is to give L2 English students problems (Goh & Back, 2010b). Consequently, a high percentage of simple and familiar words such as often-repeated connectives and prepositions are understood to lend themselves to easier comprehension. Moreover, the ratio of types or “the number of different words within the number of ‘tokens’” to the total number of tokens, which is simply tantamount to the word count of the text, is expected to clue us in to how burdensome a text is to the reader’s memory, because of the number of new words that appear (Chon, 2014, p. 349). The hypothetical frequency of occurrence of a word or word family in actual, real-world language contexts, of which a corpus only hopes to be a representative sample, allows researchers to draw conclusions about the extent to which a text demands knowledge of a certain total number of word families on the part of the reader in order for them to be able to grasp the meaning (Chon, 2014; Joo, 2008; N. Kim, 2008). The proportion of a text occupied by low-frequency word families, and conversely, the proportion of a text occupied by high-frequency word families, is information that can stand as a proxy for grade level, based on the assumption that students at a certain grade level will be equipped with a certain number of word families that is, by necessity, lower than the number expected to be known by students of higher grade levels. This information also allows direct comparison with the word family quantity thresholds of the NCE (Chon, 2014).

Chon’s (2014) longitudinal study is unique among the literature in its wide reach and the applicability of its findings. Chon noted in 2014 that the selections of the CSAT that researchers had been using for lexical analysis were too small, in terms of number or year range of tests analyzed, given how much time had now passed since the CSAT had been introduced. This was partly due to the fact that when those earlier studies had been conducted, the CSAT had not been around for as long (Chon 2014). There were also limitations in prior studies in terms of how narrow their objectives were; these included, for example, finding inconsistencies between CSATs (Joo, 2008), comparing the CSAT with certain English textbooks (Cho, 2013; N. Kim, 2008; Yoon, 2006), comparing it with standardized tests from other countries (Goh & Back, 2010b), or investigating the impact of a specific education policy change (Kwon & Shin, 2014). Chon's work was the first and last attempt to analyze the totality of the CSAT vocabulary from its beginning up until the latest iteration with general pedagogical implications that strike at the core assumptions of the NCE with respect to how much English vocabulary Korean high-school graduates should know (Chon, 2014).

These areas of inquiry into the lexical profile of the CSAT are the main contribution of Chon's work, which to this day has not been followed up on. Choi (2015) and Kim (2016) again relied upon lexical analysis to draw specific comparisons between textbook passages and the properties of vocabulary frequency and difficulty but did not engage with the thread of comprehensive analysis of the CSAT as a body of text that is evolving over time, which had been introduced by Chon (2014). There appears to be an absence of lexical analysis studies of the CSAT after 2016, which is coincidentally the same year the first major work employing Coh-Metrix to automatically evaluate the CSAT text emerged (M. Kim, 2016). One outlier to this generalization is the work of Yang and Lee, who in 2019 used lexical analysis and the BNC to compare word lists between the CSAT and textbooks with an updated view to recent tests, with the finding that textbook materials and the CSAT were dramatically out of step with one another on the level of vocabulary.

What Chon (2014) found by taking a comprehensive, longitudinal view of the CSAT must be broken down into her own three categories to be well understood. First, there is the type-token ratio of the test. Chon (2014) found that from 1994 to 2014, the test's type-token ratio had been relatively stable but that the total number of tokens had been increasing. A greater number of total tokens but a stable percentage of types naturally means more total types, and thus more total new vocabulary being introduced within the test, leading to a reading experience that is more taxing to the student's memory. Second, there is the proportion of low-frequency and high-frequency words in the test, represented as percentages of word families occupying the various 14 word family frequency bands of the BNC as tabulated with RANGE. Chon (2014) found that the test was increasing its overall proportion of low-frequency words at the expense of high-frequency words, which leads to greater demands upon a student's word knowledge. Finally, Chon (2014) established what the lexical thresholds were for a student to have a chance at success on the CSAT, which were knowledge of 6,000 word families for 95% coverage of the test's vocabulary and knowledge of 13,000 word families for 98% coverage. Both of these figures are well in excess of the 3,500 word family standard of the NCE (Chon, 2014).

Just as Chon wanted to update the existing body of scholarship in light of how many years had passed, each one bringing a whole new CSAT to add to the discussion, the present study extends her analysis up to the most recent test in 2022, work that has not been attempted since Chon (2014) initially introduced the methodology. The present time also affords the opportunity to integrate the now long-established methodology of Coh-Metrix into the research design so that the two subfields can be finally brought together. It is also of note which parts of Chon's analysis I am choosing to carry forward and which parts I am not. For the purposes of my research, an updated lexical profile rendered as a formulation of word family frequency in the various frequency bands of the BNC is the specific outcome I hoped to reach from lexical analysis. As has been established, there is a tripartite division of topics in Chon's work: type-token ratio, word frequency, and lexical threshold. Because type-token ratio was demonstrated to be very stable from 1994 through 2014, which according to Chon (2014), "is advisable for high-stakes tests such as the CSAT," and thus not a particularly surprising conclusion in itself (p. 354), it is assumed that if text length continues to increase, a variable that is easily captured by Coh-Metrix in the automated text evaluation portion of my research, there is no reason not to assume that the number of types will also increase based on the clearly established stability of the TTR. This assumption is additionally supported by several recent Coh-Metrix studies that have pointed to constant lexical density during the time periods 2016–2017 (J. Chang, 2019a; Shin, 2019) and 2015–2020 (S. Kim, 2021), a complete overview of which follows in the next section of this chapter. The final portion of Chon's work formulated lexical thresholds based on the specific vocabulary demands as displayed by the word frequency lexical profile of the test, but those thresholds are entirely predictable based on the groundwork laid in Chon's (2014) work. Out of respect for space and focus of objectives, only the non-predictable elements of lexical analysis are included here, and that leaves only the word frequency data furnished by the BNC, an external corpus, on the one hand, and, on the other, the ever-iterating CSAT English Section, which is apparently ramping up the rarity of its vocabulary year by year with no apparent direction or underlying motive that is known to researchers and with dire pedagogical consequences of increased reliance on private education and increasingly decontextualized language teaching practices (Chon, 2014). It is not possible to know what the word frequency profile of a new CSAT is until it is analyzed; therefore, the work must continue, and Chon's (2014) methodology employing the BNC tool RANGE remains a means to do so that is as excellent as it is under-utilized.

Automated Text Evaluation Studies

The way in which Coh-Metrix is most commonly used in research on the CSAT English Reading Section is with reference to individual indices such as referential cohesion, syntactic complexity, lexical diversity, or vocabulary frequency, which are chosen from more than a hundred indices available in Coh-Metrix because of their perceived relevance to the objectives of the research and the nature of the subject. Because the CSAT is a standardized assessment full of short, unrelated passages and not a natural text written for a wide audience, to be read from start to finish, special attention must be given to how the concepts underlying textual attributes are operationalized. Indices most readily applicable to a text full of paragraphs or passages that are constructed by an author with one consciousness, or in concert with other authors in the context of a social world and a discourse community, with awareness of the content of other passages and even an intent to make everything in the text work together, as in examples such as novels, plays, or academic articles, may not be as easily adapted to a standardized test, which has very different properties. In addition to the indices above, easability measures Flesch-Kincaid Grade Level, Flesch Reading Ease, and Coh-Metrix L2 Readability are commonly used because of their very direct applicability to teaching and assessment of L2 English reading skills (J. Lee, 2020).

The emergence of Coh-Metrix was a revolution in CSAT English studies. Researchers were now enabled to pursue any number of targeted investigations using the automated protocol and adapt the indices as they saw fit for their particular research designs. The earliest Coh-Metrix-based study of the CSAT is M. Kim's rarely referenced unpublished master's thesis from 2016 (M. Choi, 2018). That pioneering work investigated the differences in text difficulty measures of the CSAT and middle- and high-school English textbooks (M. Kim, 2016). This was the first application of Coh-Metrix to the CSAT to emerge in the research, although Coh-Metrix had been used previously to evaluate English textbooks in Korea (Jeon & Lim, 2009; Jeon, 2011; J. Kim & Yang, 2012). The first longitudinal approach employing Coh-Metrix to study change in the text of the CSAT text over time was developed in Moon and Kim (2017). The intention of Moon and Kim's (2017) research was, based on the simple assumption that a high-stakes assessment like the CSAT English Section ought to have relatively consistent difficulty and stable properties year over year, to apply Coh-Metrix indices to the CSAT as a way to operationalize its reading difficulty and observe both change between years of the test and change over the entire span of 1994–2016, such that general conclusions can be drawn. Based on the simple observation that this source is cited in almost every literature review of Coh-Metrix studies of the CSAT, Moon and Kim's (2017) work can be understood as having made a significant impact on the field of CSAT English Studies.

By far the most frequently cited Coh-Metrix study and the beginning point of most literature reviews of work in the field, Moon and Kim's (2017) publication found that item difficulty, operationalized as scores in the Flesch-Kincaid grade level and Flesch Reading Ease measures and specific changes in indices related to word information associated with ease and difficulty—including indices such as number of particles, number of negatives, incidence of personal pronouns and causative verbs—was increasing steadily in the CSAT over time. Although the design was and remains, among Coh-Metrix studies, novel for its temporal breadth, generalized focus, and broad prescriptions, its results are limited to only word and sentence information and readability measures, which are limited in scope to word and sentence length. There is no analysis of data on referential cohesion. J-R. Kim continues this work in his 2017 diachronic study of the CSAT English Section but divides the time period into three segments based on significant government interventions in the structuring of the test. In this work, J-R. Kim (2017) concludes, in alignment with Moon and Kim (2017), that the CSAT continues to get more difficult over time, and additionally that government interventions attempting to correct this have been unsuccessful. Alongside this publication, J-R. Kim, with M. Choi, published an analysis of cohesion and word information among CSAT English, which observed differences between question types across relevant Coh-Metrix indices (2017). A few months later, Koh and Shin's (2017) publication focused on one turning point in the history of CSAT policy when the government attempted to incorporate real-world language passages from the Educational Broadcasting System (EBS) into the CSAT instead of relying only on artificially constructed ones, with the conclusion that the reform did not work and did nothing to make the text any less difficult to read, and that all of the troubling trends observed in Moon and Kim’s (2017) work earlier that year had progressed steadily despite government interventions. Most importantly, Koh and Shin’s (2017) work continued to prove the efficacy of Coh-Metrix in allowing unprecedentedly broad conclusions and prescriptions to be made about the validity and consistency of the test.

The most recent phase of the CSAT started in 2018 with the switch from norm-referenced evaluation, which bases test scores on percentile achievement and thus compares students to other students taking the test in the same year, to criterion-referenced evaluation, which is scoring based on absolute standards that are not subject to change based on the student body’s achievement from year to year. J. Chang (2018a) used Coh-Metrix to evaluate the 2017 and 2018 tests, finding that they were very similar in text easability-readability measures and individual indices such as syntactic simplicity, word concreteness, and referential cohesion, meaning that the actual content of the test had not changed as part of the reform. Shin (2019) further found that during the time period of 2016–2019, overlapping with the year the evaluation approach of the CSAT was changed from norm-referenced to criterion-referenced in 2018, despite the fact that the change was intended to solve the problem of negative washback generated by extreme competitiveness, Coh-Metrix indices of syntactic complexity, lexical diversity, cohesion, and word information remained similar, and even slightly higher year over year in keeping with established trends. Chang further expanded her analysis in 2019 to include evaluation in 26 measures for lexical complexity, producing conclusions similar to the smaller-scale work from 2018 (2019a) and later replicated this design with similar outcomes for measures related to syntactic complexity (2019b). In 2021, S. Kim carried this research forward with updates from the 2020 test, reinforcing the understanding that reform of the evaluation method had done nothing to solve the problems of the CSAT, bolstering the underlying assertion that a simple change in grading method would do little to improve the function of the CSAT as an assessment, since even under criterion-referenced grading, all students in Korea are still competing against one another and the fierceness of that competition has test difficulty at its root.

A number of recent scholars have taken up the task of using Coh-Metrix to quantify the extent to which CSAT English Reading Section material differs from official national curriculum material. Previously, the lexical analysis study of Yang and Lee (2019) had found that the vocabulary level and topic distribution of the CSAT differed significantly from school textbooks and EBS materials and that the vocabulary difficulty levels of both the EBS and CSAT were much higher than those of school textbooks. In 2020, H. Lee used Coh-Metrix to evaluate both third-year high-school English reading and writing textbooks and passages from the two latest CSATs, the 2020 and 2019 tests, and found that in almost all of the Coh-Metrix indices and measures employed, the CSAT was significantly lower in easability and readability compared to the measures of the curriculum materials. J. Park completed a similar analysis in 2021 for the 2020 and 2021 tests, producing comparable findings (J. Park & D. Lee, 2021), which formed the basis of his master's thesis submitted two months later, which, in firm alignment with the scholarly consensus, called upon test writers to "establish objective standards…by using scientific language analysis tools such as Coh-Metrix" (J. Park, 2021, p. 92). Even if this call for change is not heeded, teachers may find the results of automated text evaluation of the CSAT useful in structuring pedagogical objectives that take into account the significant lack of sufficient preparatory material in the official curriculum textbooks.

Although it is known that the results of the various Coh-Metrix evaluations of the CSAT text, and of related material, produce conclusions that strongly agree with and correlate to one another, there is reason to investigate the extent to which Coh-Metrix itself provides a valid measure of item difficulty in this novel application to the CSAT. The question is whether or not it is possible to know that Coh-Metrix indices and measures actually evaluate linguistic difficulty in this type of text. In an attempt to answer this question, Hwang and J. Lee introduced three correlational studies that brought forward a new operationalization of item difficulty in the CSAT English Reading Section as percentage of wrong answers on mock tests (2020a; 2020b; 2020c). This allowed the authors to examine the validity of Coh-Metrix indices commonly referenced in the field by analyzing their correlation with percentage of wrong answers given on a per-item basis. In correlation analyses of item difficulty and indices related to syntactic complexity (Hwang & J. Lee, 2020a), readability/word information (Hwang & J. Lee, 2020b), and cohesion (Hwang & J. Lee, 2020c), strong correlations were found across the board. In agreement with previous scholarship, Hwang and J. Lee (2020c) conclude their third article with the broad prescription that maintaining consistent difficulty in the CSAT is of critical importance, and that since it is now known that demonstrably valid tools exist for measuring item difficulty, there is no reason why the CSAT could not be constructed in concert with the knowledge provided by these scholarly analyses. By strongly supporting the notion that commonly used Coh-Metrix indices are predictive of what they purport to evaluate, Hwang and J. Lee's three publications further justify the role of Coh-Metrix in evaluating the difficulty of passages in the CSAT.

The importance of these three works lies in their introduction of a methodology for external evaluation of Coh-Metrix results using a separate methodology, providing more than one operationalization of test item difficulty that can be correlated with the output of Coh-Metrix measures. These correlations inform the community on the validity of Coh-Metrix measures as applied to the field of CSAT studies and provide retroactive justification for the four years of Coh-Metrix studies published prior to these findings. Indeed, Hwang and J. Lee observe in their introduction that such correlational work is very hard to find outside of their own previously published work (2020c). The present study furthers this work by bringing the techniques of corpus-based lexical analysis to the forefront in correlation with Coh-Metrix indices and measures of readability and easability.

Other automated text evaluation programs exist, and one that has occasionally been used for CSAT English Section research is Lexical Complexity Analyzer (LCA). Om (2021) found that LCA's indices related to lexical sophistication and syntactic complexity were correlated with item difficulty, which was operationalized, similar to Hwang and J. Lee (2020a; 2020b; 2020c), as correct answers on items on the test. C. Kim (2020) used LCA to evaluate the lexical properties of the CSAT from 2015 to 2020 in light of the switch to criterion-referenced assessment, finding, in alignment with the Coh-Metrix studies, that the trend of high reading difficulty in the CSAT was unabated, and ought to be addressed. The LCA studies are far outnumbered by the Coh-Metrix studies, and the former platform contains far fewer indices than the latter. There is some small overlap in the area of readability measures because LCA also calculates Flesch-Kincaid Grade Level and Flesch Reading Ease as does Coh-Metrix, but because of its larger presence in the field and richer toolset, Coh-Metrix is the exclusive focus of this study.

Because Coh-Metrix provides its own measures of word frequency, one might question why I would not rely on these figures instead of resurrecting the methodology of corpus-based lexical analysis. Aside from the exigency of providing further external validation of Coh-Metrix measures, there is, perhaps, a need to close this small gap in justification for bringing the older methodology back into the present study's methods when, at least superficially, a newer alternative seems to exist. First of all, the BNC, the corpus chosen for this study, has over 100 million words (2009), whereas the corpus used by Coh-Metrix to calculate word frequency contains only 17.9 million (McNamara, et al., 2014). If a word is not included in that corpus, it is simply not calculated in the Coh-Metrix indices at all. On the other hand, the BNC tool RANGE provides an output of the number of words present in the text that are not included in the BNC, so if there were any significant confounding variable due to lack of data in the corpus, I would immediately know about it. Coh-Metrix offers no such feedback. Next, since the concern of this study is with the change over time in relative frequencies across numerous frequency bands, the BNC tool offers much richer data than the relatively narrow outputs of Coh-Metrix, which include only figures on frequencies of content words and frequencies of all words—providing only two levels of analysis rather than fourteen.

The current state of the research on the CSAT English Reading Section is such that there are many opportunities to develop the extensive findings produced over the years, through a variety of methodologies, into specific and definitive prescriptions for test writers and policymakers to enact meaningful changes that will solve the problem of negative washback from the test and begin to truly address its associated negative externalities and social ills. Coh-Metrix does not, on its own, speak directly to the guidelines of the NCE, which has no curriculum guidelines for concepts such as referential cohesion or lexical complexity. As Chon (2014) demonstrated, however, lexical thresholds calculated from the methodology of corpus-based lexical analysis allow the possibility of making direct connections with educational policy prescribing, for example, a certain number of word families that differs significantly from what the CSAT assesses (2014, p. 364). This demonstrates the necessity of adopting multiple methodologies and building connections between them with an interdisciplinary focus and an orientation towards making concepts more readily communicable. A unified front will bring together all subfields into the same conversation, synthesizing the relevant findings from these separate methodologies to both reinforce their individual validity and enable presentation of a cohesive scholarly statement to the greater public. The present correlational study between Coh-Metrix measures and indices and corpus-based lexical analysis of the CSAT begins with the step of expanding upon Chon's (2014) diachronic methodology for word frequency analysis and correlating it with an updated diachronic Coh-Metrix study inspired by those conducted by Moon and Kim (2017), J-R. Kim (2017), Choi (2018), and others.

CHAPTER III

METHODS

The preceding literature review pointed to a few absences in the scholarly output on the CSAT English Reading Section. First of all, the necessity of providing a synthesis of the findings of lexical analysis studies and automated text evaluation studies is upheld on two fronts. On the one hand, it provides a needed examination of the validity of automated text evaluation methodology as applied to the CSAT, a standardized test with unique properties that are not specifically accounted for either in the corpus upon which Coh-Metrix relies or in the indices themselves, which are intended for use with natural texts. On the other hand, it is a way to expand our insight into the properties of the CSAT English section, their evolution over time, and their influence on the negative externalities of washback.

In this chapter, I devote my attention to exhaustively outlining the methods I used so that they can provide a benchmark for further expansions upon the premise of bringing together the various types of inquiry into the CSAT English Section text under one critical umbrella. Many scholars have devoted their attention to one area or another, be it the pedagogical orientation of the test, the lexical thresholds, the content validity, or the automated indices of grade level or reading difficulty, but there is now enough history and a sufficient number of methodologies in the analytical toolset that it is time to start finding the correlations between the different linguistic, pedagogical, and institutional forces at work. In this view, the CSAT represents a sub-discipline in and of itself.

I begin with the exact procedures I followed for preparing the text of the CSAT English Reading Section for analysis, which involved extensive cleaning and adaptation to make it work optimally with the software. I explain the way in which I used Coh-Metrix 3.0 and what data I chose to pull from it, and why. I proceed to outline my lexical analysis protocols and further expand upon the reasons behind my inclusion of this type of data. Finally, I explain the analytical statistics used and what benefits they provide to the discussion.

Automated Evaluation of CSAT Text with Coh-Metrix 3.0

The first task in any automated text operation is selection of the text to be used for analysis. After having selected the texts to constitute the corpus, the researcher must format the corpus in a manner that accommodates the processing capabilities and constraints of the chosen automated text evaluation software while ensuring that the results have high validity within the context of the research project. Text that is processed in Coh-Metrix must be "cleaned" before it is run through the software (McNamara et al., 2014, p. 156). Based on the principle of "garbage goes in, garbage comes out," ensuring valuable outputs from a computer algorithm depends heavily on the accuracy and consistency of the inputs. One of the overarching prescriptions made by McNamara, et al. (2014) is that the text must appear as it would to the target reader, as if the author of the text had just typed it down and handed it over to be read.

Since the very object of the software is to evaluate the extent to which a text is readable for humans, the algorithms of Coh-Metrix are designed to process text that is in a human-readable format. That being said, Coh-Metrix can only process text, not pictures, tables, or figures. These elements cannot be included in the corpus provided to Coh-Metrix and therefore must be removed. The odd line breaks which often occur in a printed text may interfere with the analysis by artificially increasing the paragraph count and possibly confounding measures of cohesion, for example. These breaks should be corrected in order to preserve the logical grouping of a paragraph as understood by Coh-Metrix. Some degree of discretion on the part of the researcher is called for during this process because of the subjectivity with which, for example, a string of dialogue in a narrative text with multiple line breaks might be understood to belong to one or another adjacent paragraph. The researcher needs to make the decision, moreover, on whether or not to include intentional deviations in spelling, such as in representations of dialect, or errors not intended by the author, such as inconsistencies and lapses in spelling or grammar, and if not, how to address them. Another decision to be made is whether or not text which refers to the data in a figure or table which has been, by necessity, removed prior to analysis should also be removed in light of its dependence on absent information.

Aside from cleanliness, the other important aspect of text preparation is consistency. Perhaps more important than cleanliness, consistency demands that what is done to one text must be done to all texts. In comparative and correlative studies, as long as consistency is observed, some degree of uncleanliness can be tolerated—up to 5% in McNamara’s advice. If certain types of uncleanliness are constant across the corpus, they cannot represent a threat of confounding variables. The two overarching principles that govern corpus cleanliness and consistency in McNamara, et al. (2014) are as follows: “1. Unless there is a good reason to take it out, you should leave it in. 2. What you do to one, you do to all” (p. 156). The first principle advises to bias towards inclusiveness—when in doubt, do not take it out. The second principle unambiguously maintains that the same exact principles of text preparation should be applied to every text in the corpus regardless of the particular attributes of each individual sample. The researcher must avoid the pitfall of giving in to the temptation to modify their approach halfway through text preparation without retroactively applying the modifications to all other previously cleaned texts. Based on McNamara’s general guidelines, my text preparation approach was tailored to the unique attributes and constraints of the text being evaluated and to the objectives of this study.

Overview of the Structure of the Test

The CSAT English Reading Section spans a varying number of questions featuring passages of varying length from year to year. Although I observed throughout the process of text preparation that the distribution of question types, question length, and total number of questions showed a degree of continuity from year to year, such that the 1994 and 1995 tests look, superficially at least, very similar, the change in passage length and total number of passages over the course of the 29 years of the test is enormous. To be sure to capture these important changes in the final data, I chose to operationalize reading passages as paragraphs for the purpose of Coh-Metrix automated evaluation. That is, I redefined paragraphs as “entire individual reading passages,” so that Coh-Metrix can measure the change in length and number of the passages over time by making use of its built-in paragraph measuring protocol. Since this is a text where most passages are, in fact, single paragraphs, this practice seemed justifiable.

The passages of the CSAT text span a variety of genres and question types. In early tests, there were no visual elements such as charts and graphs, but in more recent tests there are several. Aside from the recent introduction of visual aids, the types of questions in the CSAT have remained relatively constant over the history of the test. In consultation with the terminology of J. Lee (2020), I observed an array of questions in the categories of "Understanding general ideas," "understanding details," and "understanding logical relations" (p. 26-27). Under these three umbrellas, questions may require students to provide the gist or a title for a given passage, identify the overall feeling or mood, use vocabulary in context, identify similarities and differences, fill in the blank, pick out the unrelated sentence, insert the correct sentence, sequence the sentences or paragraphs of a passage, or summarize. The structure of every question is multiple choice—none of them are open-ended.

Selection and Preparation of Texts for Evaluation

As mentioned before, the object of analysis of the present research is the entire body of CSAT English Reading Section text from the year the test was introduced in 1994 up until the most recent test year, which is 2022. Document files of each test were accessed from the official government website of the Korea Institute for Curriculum and Evaluation (KICE), the institution which develops and administers the CSAT (2022). These documents, which are provided in .pdf format, were run through the optical character recognition (OCR) algorithm in Adobe Acrobat. This provided a text which could be manipulated into conformity with the standards of readability of the automated text evaluation software using a word processor.

Each passage from the English Reading Section of each year's CSAT was pasted into a document labeled for the year. For example, extraction of the 1994 test's passages resulted in a single document consisting solely of paragraphs of text, each spaced one line apart from the other. Some passages of the test consisted of multiple paragraphs. For these passages, I observed the principle of grouping all paragraphs into one. As mentioned before, the reason for the adoption of this practice was that one of the important Coh-Metrix indices is paragraph length, which offered the possibility of measuring passage length over time. There is also concern that separating multi-paragraph passages into multiple units in a test that is heavily populated by short, unrelated texts, many if not most of which are single paragraphs, would confound the essential distinction between passages that is a hallmark of the conventions of standardized test composition and thus provide an unrepresentative evaluation of the text at hand.

After the conversion of the text of each test's passages into a word processor document for each of the 29 years that the test was administered, the corpus was ready to be cleaned for Coh-Metrix evaluation. In a text of this nature, there were a few considerations that required decisions to be made on what to exclude and what to keep intact. As mentioned before, the important rules are to exclude or alter text only when there is little doubt that the problematic text will interfere with the evaluation and to observe consistency across the corpus in the types of changes that are made. Misspellings and grammatical errors were left unchanged. This adheres to McNamara et al.'s (2014) injunction to "leave things the way you find them" unless there is a good reason to apply the principle underlying the correction, change, or omission across the entire corpus (p. 156). Furthermore, from the perspective of a Korean English language learner (ELL) taking the CSAT, a misspelling could very well signify an entirely different word from the correctly spelled item, and grammatical errors could be injurious to coherence when the student is trying to comprehend the passage. Therefore, in order to provide the greatest fidelity in measures of text easability and readability, both of which are determined in part by word overlap and syntactic simplicity, the soundest principle is to leave these unintended errors uncorrected.

Punctuation throughout the texts was corrected to standard North American usage. Instances of, for example, periods, exclamation points, or colons being braced by a space on each side were altered to reflect the standard North American practice of spacing only on the right side. Passages that were extracted through Adobe optical character recognition (OCR) superficially looked like paragraphs, but in fact were not because Adobe OCR does not conduct human language processing, only orthographic symbol recognition. The output of OCR was structured as individual, unconnected lines, which would have resulted in faulty measurements if left uncorrected. Additionally, OCR made many character recognition errors throughout the text which needed to be manually corrected through cross-referencing of the original document.

Depending on the question type, certain alterations needed to be made so that the content of the question fit within one intact paragraph. The passages of fill-in-the-blank question types were supplied with the correct answer according to the answer guide. The sentence represented as the correct answer was placed in the appropriate position within the paragraph. Incorrect answer choices were not included in the corpus in any fashion. Passages accompanying questions asking students to identify which sentence does not belong were simply left intact in order to preserve the whole original text as the students would read it. Here, there was no compelling reason to make changes since the text was intact as written and the object is to evaluate the test according to the demands it places upon the test taker. Arrangement question types featured passages with numbered, out-of-order sentences or paragraphs and a selection of options for arranging them. These passages were rearranged according to the answer guide based on the premise that students read such passages already expecting the numbered units to be out of order. Students are unlikely to accept the ordering provided in the passage by default based on this assumption, and indeed, I found no examples of the default ordering matching the correct order throughout the entire history of the test. The only justifiable decision, then, seems to be adoption of the canonical order.

I observed that figures, tables, and pictures started being introduced into CSAT English Reading Section questions in the mid-2000s. These questions seem to focus on comprehension of academic and scientific language used in the composition and analysis of visual representations of data. Because the text of these passages relies heavily upon information provided in the visuals, I could not find a way to include it in the evaluation. Therefore, these questions are omitted from the corpus outright. Passages accompanied by a picture did not share this impediment, however, and were able to be included. The pictures, however, cannot be read by Coh-Metrix and, therefore, were omitted. Numbered lists were generally excluded, but those that featured complete sentences and thus could be reformatted into paragraphs were included. This represented a very small percentage of question content across the body of text.

Completion of the corpus left only the question of whether to evaluate individual passages as texts or the content of one entire test as a text. That is, for the 1994 test, is it best to evaluate each paragraph of the extracted, reformatted, and cleaned document one at a time in the chosen Coh-Metrix indices and then average the results of each to come up with final numbers for that test? In contrast, is it better to simply treat the entire set of passages—formatted into paragraphs as described before—for one year as the unit to be evaluated? This would entail inputting the entire contents of one of the previously described word processor documents, e.g., the one produced for 1994, into Coh-Metrix, and using the output as the final numbers for that test year. Based on consideration of the pros and cons of both approaches, the latter seemed to best fit the type of text under analysis.

Treating the text of one iteration of the CSAT holistically adheres closer to the perspective of the student taking the test than does the approach of isolating individual passages. Despite the diversity of genres represented, the passages are presented to the student all together in one document; words and grammatical structures that repeat from one passage to the next can serve as comprehension clues. The isolating approach misses this important influence. Cohesion is one of the most significant, arguably *the* most significant, measures Coh-Metrix uses to evaluate easability and readability, because it directly relates to reading difficulty, and word overlap and syntactic overlap are critical factors influencing cohesion (McNamara et al., 2014). The isolation approach may provide greater understanding of the cohesion of individual passages depending on their length or genre, so it may be more useful for an analysis of how genre functions in the CSAT or investigation of the relationship between Coh-Metrix measures and indices that are typically assumed to be associated with certain genres. For example, narrative texts have greater word concreteness than science texts but lower cohesion (McNamara et al., 2014). Since the present study is focused on comparing the reading difficulty of entire tests to other entire tests, the question of genre is rendered irrelevant—all of the indices that are associated with text difficulty are covered by the broad readability and easability measures, so a discussion of genre can only tell us about genre itself, not the attributes of the texts relevant to this discussion.

The output of Coh-Metrix 3.0 consists of a large table of scores in the form of percentiles, z-scores, and means of the various indices. McNamara et al. (2014) distill these more than 100 different measures of the properties of text, such as sentence length, word length, or argument overlap, for example, into eight principal "easability components" (p. 84). Among these, several such as referential cohesion and syntactic simplicity are frequently employed in the literature on Coh-Metrix evaluation of the CSAT English Section text. Out of the eight easability components, McNamara, et al. (2014) note that the first five account for the majority of the variance between texts in the corpora used by the program and are thus the most predictive of reading difficulty or grade level. The scores for each of these easability components are presented in percentile format. The way to understand this type of figure is that the given numerical percentile is equivalent to the percentage of texts in the corpora that exemplify the linguistic trait signified by the particular easability component to a greater degree than the text under evaluation. A percentile score of 10 in referential cohesion, for example, means that only ten percent of other texts in the corpora used as the bases for Coh-Metrix 3.0 are less referentially cohesive than the text currently under evaluation. For the purposes of the current research, I selected from the first five of the easability measures those which I felt were most relevant to the current research based on both established findings and ongoing conversations in the literature and the particular conceptual demands and constraints of the text of the CSAT English Section, being a document comprised of many short unrelated passages intended to challenge second-language

English students' reading comprehension skills. I am using the syntactic simplicity percentile score, the word concreteness percentile score, and the referential cohesion percentile score, numbers 2, 3, and 4, respectively, among the eight principal component scores of Coh-Metrix 3.0.

I chose to ignore the narrativity percentile score for this study because of my belief that the textual properties associated with genre, not genre itself, are the true predictors of text difficulty and that these properties are adequately captured by other scores. Additionally, there is, as of yet, no way to properly adjust the methodology to account for genre distribution in a text of this nature, which is a patchwork of different texts from all genres—some of which may not be adequately understood by Coh-Metrix at all, such as billboards, instructions, announcements, advertisements, or conversations. The best practice at this time is to forego the use of narrativity in the study of the CSAT English Reading Section. Genre does not form a significant part of this study because of the need for further research into its applicability to the CSAT and is only alluded to superficially as a way of accounting for possible confounding influences.

I avoided the use of the deep cohesion percentile score in constructing my analysis because, again due to the fragmented nature of a text like this, there is little opportunity for the establishment of "causal and logical relationships within the text" (McNamara et al., p. 85). It is questionable the extent to which the concept of deep cohesion could even apply to a collection of short, isolated texts that have no internal relationships with one another. It merits being mentioned that the same could be said, to a limited extent, of referential cohesion. I cannot expect to find much "overlap…across the entire text" in the CSAT English Section, but I do expect, agnostic of the content, genre, or intention of each individual passage, to see the influence of varying extents of accidental semantic overlap—perhaps in passages of different genres that cover similar subject matter or in passages of the same genre that cover different subject matter—reliance on connective words, and the inevitable repetition—or lack of repetition—of certain words and phrases throughout the body of the text, which can serve as aids to coherent understanding of the language of the test on the part of the test taker (McNamara et al., p. 85).

As is now standard practice in research on the CSAT English Section, I also used standard unidimensional measures of readability and easability alongside these three principal components. These include the Flesch-Kincaid Grade Level and Flesch Reading Ease, both of which include a formula that assigns a score based on sentence length and word length, and Coh-Metrix L2 Readability, which "considers content word overlap, sentence syntactic similarity, and word frequency" (McNamara et al., 2014, p. 80). Including these measures in my package of Coh-Metrix 3.0 automated evaluation data serves to increase the validity of my findings, both with regard to what they can say about the CSAT and what they can say about how the different parts of Coh-Metrix 3.0 work together with respect to this specific body of research. Moreover, most models of text comprehension align on the important point that the reading comprehension process is not unidimensional but multidimensional, so restricting myself to only one lens or only one or two dimensions of text difficulty would reduce the fidelity of the picture of the CSAT that I hope to capture. Finally, in the discussion of the results of these automated evaluations, reference is made to McNamara's (2014) data on Coh-Metrix Indices Norms for high-school grade levels as a baseline point of comparison for the changes in textual properties of the CSAT over time.

Lexical Analysis of CSAT Text

Chon's (2014) work on the lexical profile of the CSAT English Reading Section produced the finding that the test was increasing its proportion of low-frequency word families at the expense of high-frequency word families. This led to the conclusion that the current trend in the test was one of increased difficulty, operationalized as an increasingly high vocabulary burden on test takers that is far out of step with the NCE. My first assumption in conducting the present study is that this change in lexical profile would correlate to a decrease in measures of referential cohesion, a concept that strongly relies upon high-frequency words and content overlap (McNamara et al., 2014). The more words that repeat within a text, and the more frequent the use of function words, which are by nature high in frequency within any given corpus, the more coherent the reader's perception of the text will be. Additionally, I wanted to push the research forward by determining whether there is a correlation between the findings of the lexical analysis approach and those of the automated evaluation approach and establishing, in the form of corpus-based lexical analysis, an external point of comparison that will either strengthen or weaken the case for Coh-Metrix 3.0's position as the primary tool relied upon for study of the text of the CSAT English Section.

Corpus-based lexical analysis is fundamentally different from automated text evaluation because it is purely descriptive, not interpretive. It informs the researcher's interpretation, but it does not perform the interpretation itself. Coh-Metrix evaluation is heavily relied upon in studies of the CSAT text, and has, in recent years, become the dominant methodological tool. Although there is no salient reason to doubt the validity of automated evaluation with Coh-Metrix, which is based on decades of research into unidimensional and multidimensional measures of the properties of text and discourse and did originally derive from corpus studies (McNamara et al., 2014), critical awareness of the dominance of this one methodology is absent in the field of CSAT studies. Lexical analysis also provides information that Coh-Metrix cannot, specifically in the pedagogically significant area of lexical thresholds. As Chon has demonstrated, as of the 2014B test, the CSAT requires students to have knowledge of an excessive number of word families— 13,000—compared to the standard 3,500 of the NCE (2014, p. 364). By directly speaking to the vocabulary demands of the test in a way that is quantifiable in terms similar to those of the NCE guidelines, lexical analysis is a significant arrow in the quiver for scholars pushing for change in the CSAT English Section.

Lexical Analysis of CSAT Text Using RANGE

Following the procedures outlined in Chon (2014), I used Nation's RANGE program, which calculates word frequency across 14 1,000-word-family bands with data drawn from the BNC (Nation, 2022), to analyze the vocabulary of the text of the CSAT English Reading Section from each year, utilizing the cleaned text documents I created for use with Coh-Metrix 3.0. The 14 frequency bands used by RANGE collectively represent the 14,000 most frequently used word families in the BNC. There are additional words that fall outside of this range, but they are so exceptionally rare as to be insignificant for this type of work. The program itself is a simple executable that produces a text output of numbers of words in each frequency band for each document loaded into the queue. Using this information, I was able to replicate the study of Chon

(2014), whose findings up until the 2014B test, the harder of two versions of the test that were released that year—the only year that saw two versions of the CSAT—demonstrated a clear trend of decreasing proportions of high-frequency words and increasing proportions of low-frequency words. More importantly, I was able to extend this methodology up until the present year of 2022, providing data that will allow correlation of the output of various analyses of changes in the test's composition over its entire history. I did not choose to look at type-token ratio, or TTR, which Chon demonstrated to be relatively constant over the history of the test (2014), despite the changes in word frequency distribution and overall increase in number of word families. The information on the NCE lexical threshold in Chon (2014) is brought back into the discussion when examining the findings of the RANGE lexical analysis in order to provide an external point of reference for the significant change in the test over time.

Procedures for Statistical Analysis

Having meticulously completed the text extraction and preparation procedures and having finally retrieved the data for all thirty tests from Coh-Metrix 3.0 and RANGE, I arrayed all of the data in a spreadsheet for the graphical representation and statistical analysis phase of the research. I used Pearson's correlation coefficient to investigate the possible correlations between decreases in the RANGE proportion of word families in the high-frequency bands, the first through the fifth, and increases in proportion of word families in the low frequency bands, the sixth through the 10th, with changes in the Coh-Metrix principal component percentile scores of syntactic simplicity, word concreteness, and referential cohesion, and with changes in the standard unidimensional readability and easability measures over the course of 30 tests from 1994–2022. An analysis of deep cohesion was included to test my assumption that the standardized test format would be uniquely resistant to this component and to provide a sort of control group; that is, a set of data where I do not expect to find any effect or correlation. Pearson's *r* was calculated for each of the correlation analyses and a *p*-value was calculated for statistical significance. Graphs were produced for ease of visual comprehension of the data, which covers such a long time that the changes, although subtle from year to year, produce a much stronger impression when arrayed all together.

CHAPTER IV

RESULTS

This chapter includes the findings of the automated text evaluation in Coh-Metrix in the various components and measures stated in the previous chapter on methods, the results of the lexical analysis in the BNC word family frequency bands, and the results of the correlation analysis using Pearson's correlation coefficient. Finally, supporting analytical statistics assessing the strength of the relevant correlations are included at the end of this chapter. Numbered tables and figures are included in each section where appropriate.

Results of Automated Evaluation of CSAT Text with Coh-Metrix

This subsection includes the results of the Coh-Metrix automated text evaluation. Coh-Metrix 3.0 provided percentile scores in the principal components of syntactic simplicity, word concreteness, referential cohesion, and deep cohesion. It provided formula results for the Flesch-Kincaid Grade Level, Flesch Reading Ease, and Coh-Metrix L2 Readability scores. These findings are arranged below according to each type of data.

Principal Component Scores

The first principal component percentile score is syntactic simplicity. The results of this part of the evaluation are presented in the form of Table 1 and Figure 1 below. As noted before, each year represents the entirety of the English Reading Section text cleaned and prepared for analysis according to the protocols described in the chapter on methods.

Table 1

Syntactic Simplicity in the CSAT English Reading Section over Time

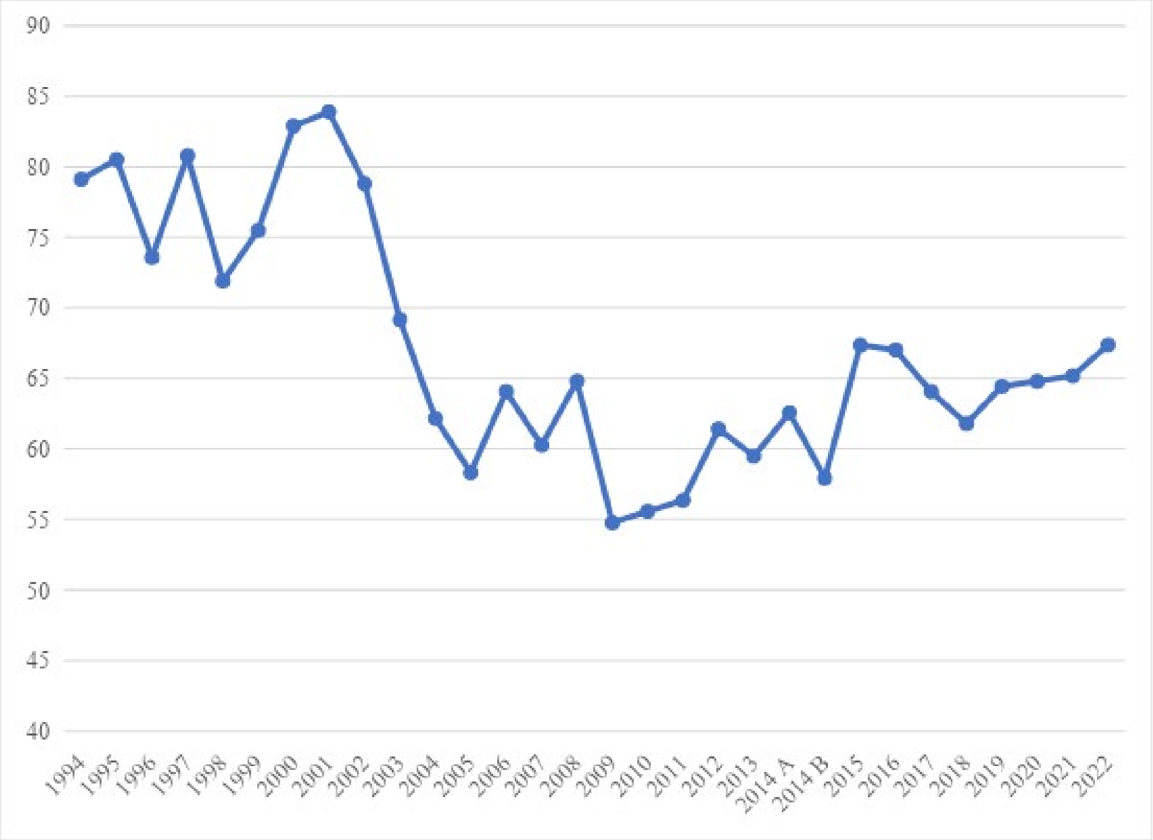
Year	Syntactic Simplicity Percentile Score
1994	79.1
1995	80.51
1996	73.57
1997	80.78
1998	71.9
1999	75.49
2000	82.89
2001	83.89
2002	78.81
2003	69.15
2004	62.17
2005	58.32
2006	64.06
2007	60.26
2008	64.8
2009	54.78
2010	55.57
2011	56.36
2012	61.41
2013	59.48
2014 A	62.55
2014 B	57.93
2015	67.36
2016	67
2017	64.06
2018	61.79
2019	64.43
2020	64.8
2021	65.17
2022	67.36

In Table 1, a remarkable decrease in the percentile score corresponding to syntactic simplicity can be observed over time. The most recent test rated a 67.36 whereas many of the early tests exceeded 80. No test after 2002 received a score over 70.

There seems to be a general downward trend leading to the present day. This means that in each successive year, I expect the test to feature longer sentences and less familiar syntactic structures than the year before. Sentence length is also directly captured by the formulas underlying the Flesch-Kincaid Grade Level and Flesch Reading Ease measures, so I expect to see some correlation with those findings. The results are represented visually in Figure 1.

Figure 1

Syntactic Simplicity in the CSAT English Reading Section over Time



A precipitous drop in the syntactic simplicity score is shown in the early 2000s, after which it recovered briefly and became relatively stabilized during the late 2010s and early 2020s.

Table 2 shows word concreteness percentile scores for each year in the same format as Table 1.

Table 2

Word Concreteness in the CSAT English Reading Section over Time

Year	Word Concreteness Percentile Score
1994	36.32
1995	57.53
1996	43.64
1997	26.11
1998	39.36
1999	54.38
2000	56.75
2001	55.17
2002	50
2003	54.38
2004	53.19
2005	52.39
2006	67
2007	61.41
2008	66.64
2009	69.15
2010	46.81
2011	35.57
2012	39.74
2013	44.04
2014 A	55.96
2014 B	44.04
2015	35.94
2016	41.68
2017	34.46
2018	33.72
2019	35.94
2020	37.83
2021	18.94
2022	31.56

There is not as clearly discernable a trend as was seen with syntactic simplicity, but there does still appear to be a gradual decline. The first test from the year 1994 scores quite low, and contrary to the results on syntactic simplicity, the peak is concentrated around the year 2009, not the early 2000s. As is the case with syntactic simplicity, word concreteness, which is seen in Figure 2, is an index of reading ease. It is thought that more concrete words will be easier to comprehend than more abstract words, so a higher concentration of them would be expected to positively correlate with text easability. Nevertheless, the 2009 test scored simultaneously very low in syntactic simplicity but very high in word concreteness.

Figure 2

Word Concreteness in the CSAT English Reading Section over Time

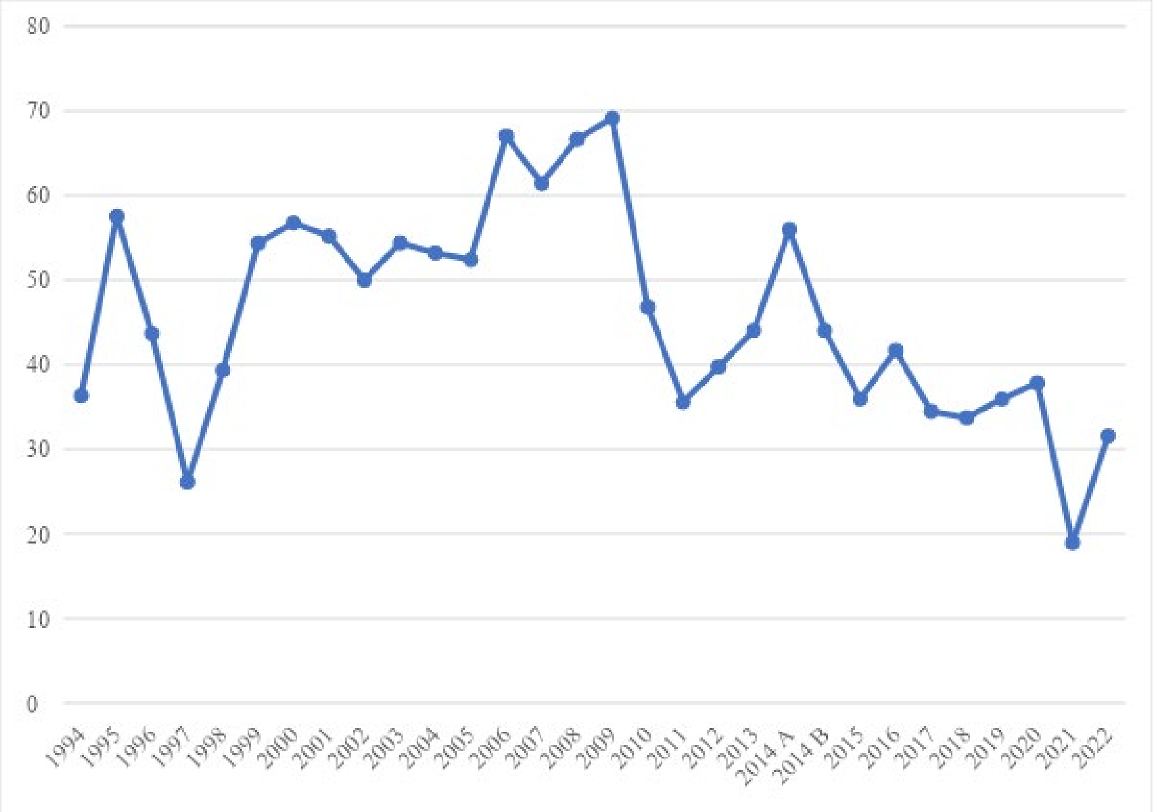


Table 3 shows the results in the index of referential cohesion, and as before, is followed by a graph in Figure 3 representing the data visually.

Table 3

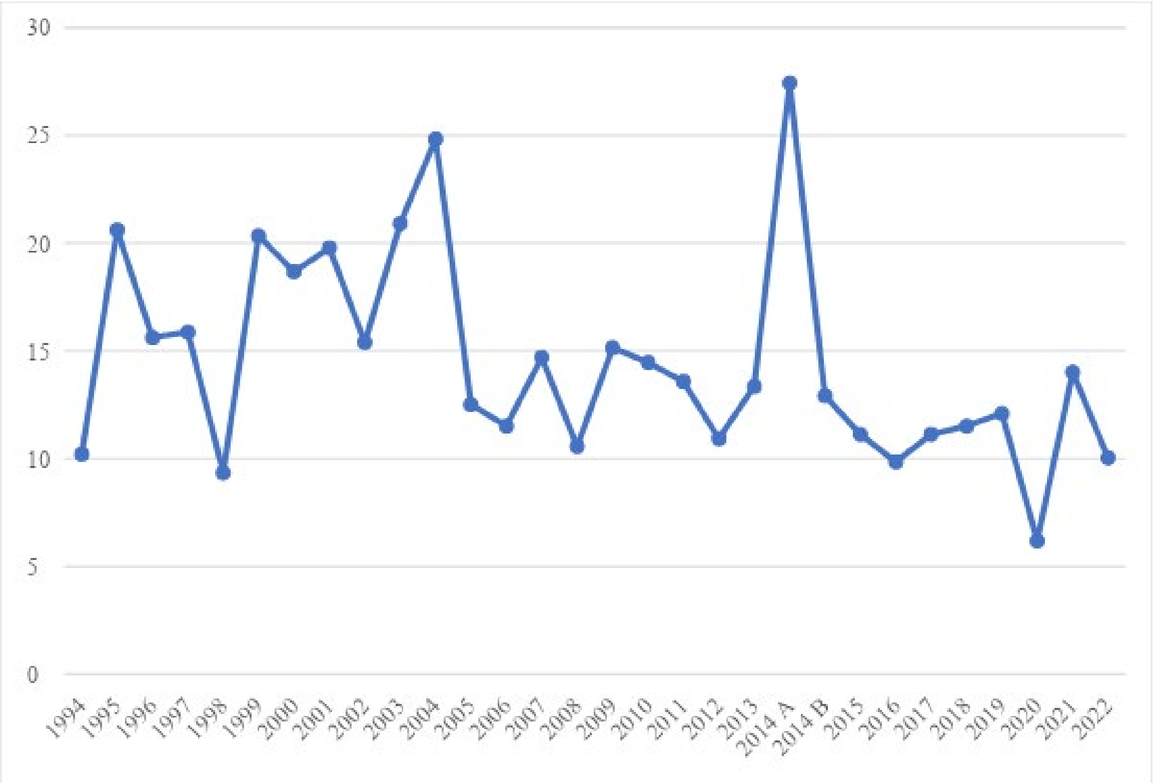
Referential Cohesion in the CSAT English Reading Section over Time

Year	Referential Cohesion Percentile Score
1994	10.2
1995	20.61
1996	15.62
1997	15.87
1998	9.34
1999	20.33
2000	18.67
2001	19.77
2002	15.39
2003	20.9
2004	24.83
2005	12.51
2006	11.51
2007	14.69
2008	10.56
2009	15.15
2010	14.46
2011	13.57
2012	10.93
2013	13.35
2014 A	27.43
2014 B	12.92
2015	11.12
2016	9.85
2017	11.12
2018	11.51
2019	12.1
2020	6.18
2021	14.01
2022	10.03

Results on referential cohesion appear especially erratic. These data demonstrate how difficult it is to predict how referentially cohesive a test might be from one year to the next. The tests during the mid and late 2000s that scored quite low in syntactic simplicity also suffered in this area. Tests from 1999–2001 which featured high syntactic simplicity also rated highly here. One surprising detail is the high cohesiveness of the 2014A test. In the next chapter’s discussion, I investigate further why this may be the case. Interestingly, the 2022 test scored almost exactly the same as the 1994 test, which was also true in the results on word concreteness.

Figure 3

Referential Cohesion in the CSAT English Reading Section over Time



I proceed now to what is perhaps the least meaningful component to apply to a standardized test, but I opted to include it nevertheless as a sort of control group. Deep cohesion, the results of which are shown in Table 4 and Figure 4, concerns the degree to which a text builds meaningful causal and intentional relationships across its length. Standardized tests scarcely accomplish this because of their fragmentary composition.

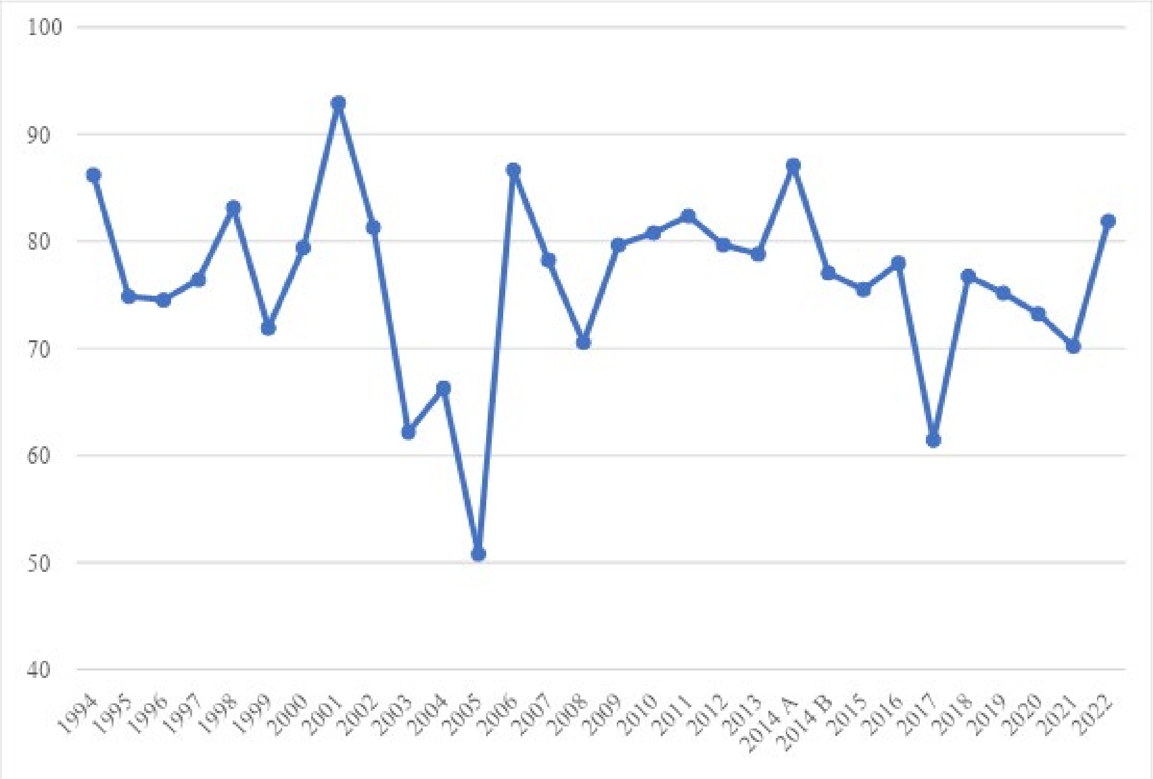
Table 4

Deep Cohesion in the CSAT English Reading Section over Time

Year	Deep Cohesion Percentile Score
1994	86.21
1995	74.86
1996	74.54
1997	76.42
1998	83.15
1999	71.9
2000	79.39
2001	92.92
2002	81.33
2003	62.17
2004	66.28
2005	50.8
2006	86.65
2007	78.23
2008	70.54
2009	79.67
2010	80.78
2011	82.38
2012	79.67
2013	78.81
2014 A	87.08
2014 B	77.04
2015	75.49
2016	77.94
2017	61.41
2018	76.73
2019	75.17
2020	73.24
2021	70.19
2022	81.86

Figure 4

Deep Cohesion in the CSAT English Reading Section over Time



As expected, there is little discernible trend in the data. There are scarcely any immediately observable similarities to my previous data sets either, perhaps except for the curious similarity of the 1994 and 2022 scores. During the section of this chapter on analytical statistics, I find out whether there is indeed any correlation between these data and those of the other principal components and, in turn, the BNC lexical analysis data. It is perhaps noteworthy here that both components relating to coherence display much greater irregularity than the other two components reviewed. I turn to this observation once more in the discussion chapter.

Next are the unidimensional readability measures produced by the Coh-Metrix evaluations: Flesch-Kincaid Grade Level, Flesch Reading Ease, and Coh-Metrix L2

Readability, the results of which are shown in Table 5 and Figure 5. As mentioned before, the first two measures are based on mean sentence length and mean word length, whereas the latter is based on content word overlap, sentence syntactic similarity, and word frequency. I expect some correlation between these measures and the dimensions explored by the principal components in Tables 1–4 and Figures 1–4.

First, Flesch-Kincaid Grade Level assesses the reading difficulty of texts on a scale that is thought to map approximately to the school grade levels of L1 English-speaking students. As mentioned before, Flesch-Kincaid Grade Level is among the measures that are heavily relied upon in studies of the CSAT. A simple formula, it is based exclusively on sentence length and word length (McNamara et al., 2014). Greater sentence length is thought to be an index of greater syntactic complexity, and word length is thought to positively correlate with a higher proportion of lesser-known and lower-frequency vocabulary.

The increase in grade level over the course of 29 years shown in Table 5 and Figure 5 is striking. Because a single integer represents such a great change in content difficulty, an increase in grade level from 7 in 1994 all the way up to 11 in 2022 indicates a significant transformation in content. It is almost as if middle schoolers of the present were being tested on the same level as juniors or seniors in high school were in the past. Viewing similar data in the Flesch-Kincaid Grade Level measure in their three-year study, Lee and Lee (2018) observed that the mere fact of testing Korean high schoolers with content appropriate for L1 English-speaking high schoolers represents, in itself, a severe challenge to the validity of the CSAT English Section. There is more detailed explication of this very clear trend in the discussion chapter.

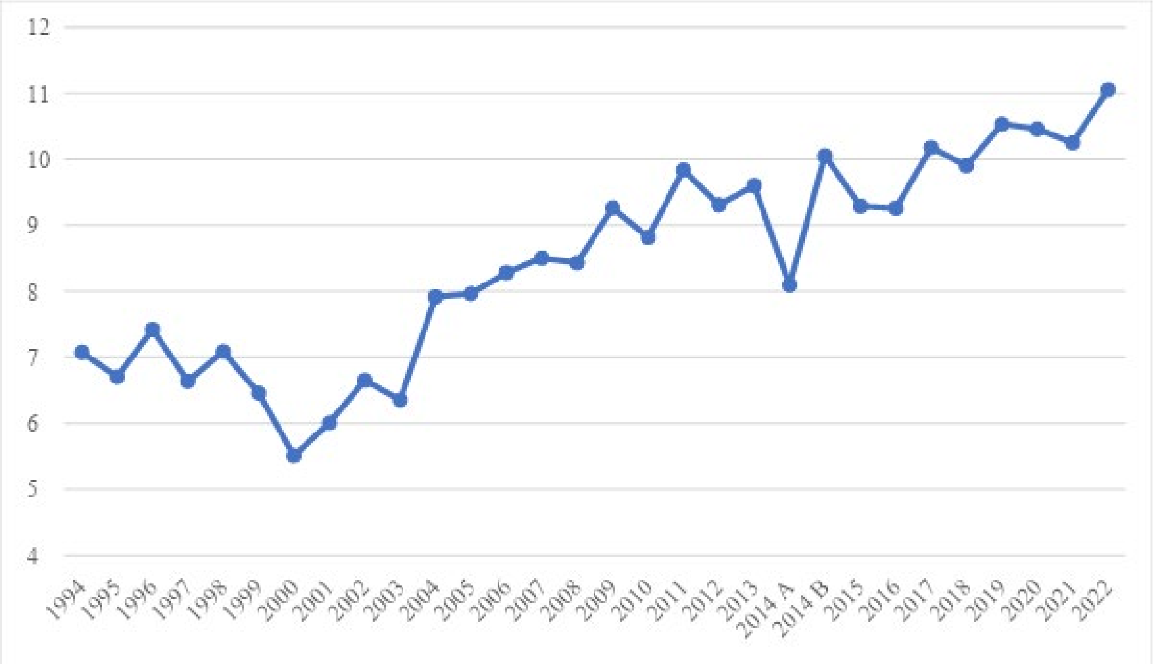
Table 5

Unidimensional Readability Measures of the CSAT English Reading Section over Time

Year	Flesch-Kincaid Grade Level	Flesch Reading Ease	Coh-Metrix L2 Readability
1994	7.076	68.251	17.84
1995	6.703	70.999	22.469
1996	7.418	68.653	19.562
1997	6.638	70.143	21.827
1998	7.088	69.566	18.351
1999	6.457	71.639	21.269
2000	5.509	77.124	22.467
2001	6.008	74.047	20.887
2002	6.652	71.725	18.672
2003	6.352	74.015	20.703
2004	7.915	65.958	19.285
2005	7.964	64.914	15.984
2006	8.28	65.116	15.013
2007	8.503	63.296	14.448
2008	8.429	63.812	14.564
2009	9.264	61.371	14.046
2010	8.815	61.246	14.69
2011	9.839	56.693	14.672
2012	9.307	58.186	13.266
2013	9.599	57.265	13.698
2014 A	8.093	66.465	18.232
2014 B	10.053	55.166	12.735
2015	9.289	57.963	14.407
2016	9.256	57.672	13.289
2017	10.176	51.776	12.393
2018	9.901	53.57	12.239
2019	10.533	49.985	12.382
2020	10.458	49.425	10.545
2021	10.25	50.182	14.098
2022	11.051	47.018	10.611

Figure 5

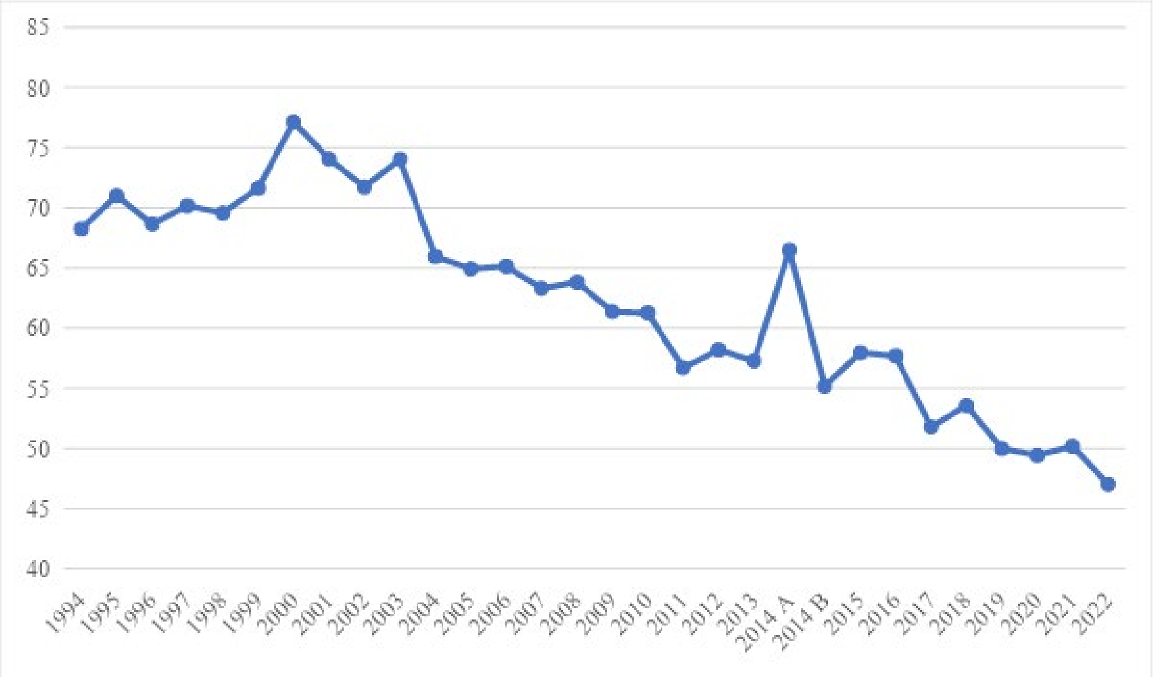
Flesch-Kincaid Grade Level of the CSAT English Reading Section over Time



Flesch Reading Ease also uses mean sentence length and mean word length as proxies for text difficulty, but it adjusts the formula in a way that produces a figure of easability rather than difficulty. Again, in Figure 6, I see a strong trend indicating a precipitous decline in reading ease, and therefore a corresponding rise in difficulty, of the CSAT English Reading Section material. With these two measures in conceptual agreement, it is clear to see that mean sentence length and mean word length are increasing in the CSAT English Reading section over time, to the detriment of readability according to this conceptualization.

Figure 6

Flesch Reading Ease of the CSAT English Reading Section over Time

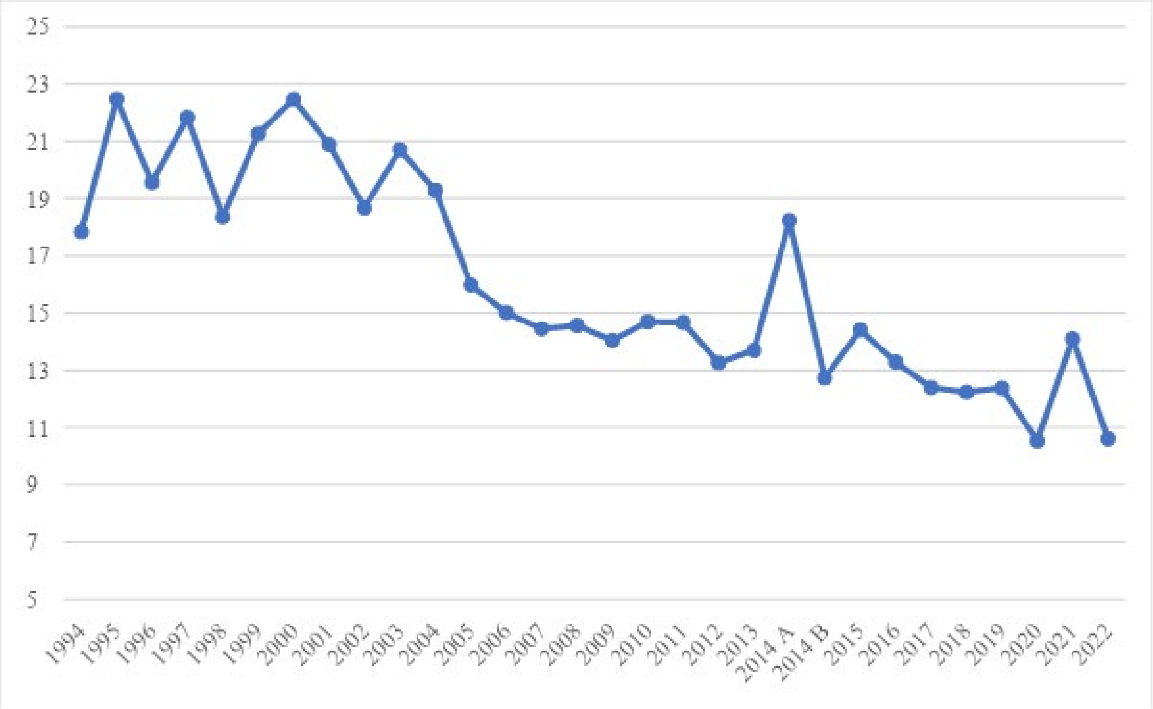


Coh-Metrix L2 Readability represents an attempt to expand the notion of readability formulae to look at factors other than mere word and sentence length (McNamara et al., 2014). These factors include content word overlap, sentence syntactic similarity, and word frequency. Despite the fact that Coh-Metrix and RANGE draw upon different corpora, I anticipate in advance that this last factor of word frequency should also interact with subsequent data drawn from the BNC corpus word frequency calculation tool RANGE. Word overlap and sentence syntactic similarity are related to cohesion, so I could also reasonably expect some degree of correlation between the Coh-Metrix L2 Readability scores and the referential cohesion scores of the CSAT year over year.

In Figure 7, representing Coh-Metrix L2 Readability scores over time, there is an observable downward trend in this score, which directly measures readability for L2 students. This dimension of readability for L2 English students specifically, not general readability for all students, additionally distinguishes the formula from the Flesch measures in a way that entails more targeted implications for a text like the CSAT English Reading Section, which is intended for L2 students exclusively.

Figure 7

Coh-Metrix L2 Readability of the CSAT English Reading Section over Time



Spikes in readability also align with my expectations based on the trends in the principal component scores surveyed in the previous section. Concentrated around the years 1999–2001 and the 2014A test are high readings of text easability and correspondingly low readings in difficulty. These observations are also supported by the Flesch-Kincaid Grade Level and Flesch Reading Ease data sets. So far, the visual representations of the various data sets have painted different pictures that nevertheless exhibit striking similarities. The analytical statistics of the final section of this chapter cuts straight through to those similarities and what, if any, insights they can provide.

The decreasing Coh-Metrix L2 Readability scores shows that the test has decreasing word overlap and syntactic similarity between passages. This could mean that the language in the test is more varied, that the genres represented are ones which entail these features, and simply that the reading material is at a higher grade level year over year. There is also the possible influence of decreasing word frequency overall, as Chon found (2014) and increasing incidence of hapax legomena, the high incidence of which in the CSAT was a key finding of Goh and Back (2010b). On that note, to the next section includes the results of lexical analysis with the RANGE tool.

Results of Lexical Analysis of CSAT Text with RANGE

In Table 6, the relevant data from RANGE is arrayed in a composite format. The first column shows the year, and the second column shows the total percentage of the word families in the CSAT English Reading Section that belong to the first, or highest-frequency, band of word families of the BNC. The third column shows the percentage of the word families in the CSAT that belong to the first through the fifth frequency bands of the BNC. Finally, the fourth column shows the percentage of the word families that belong to the lower frequency bands of six through 10. Although there are word families present in the CSAT that fall in the 11th up through the 14th band, these numbers are so small as to be unhelpful for the present analysis.

Table 6

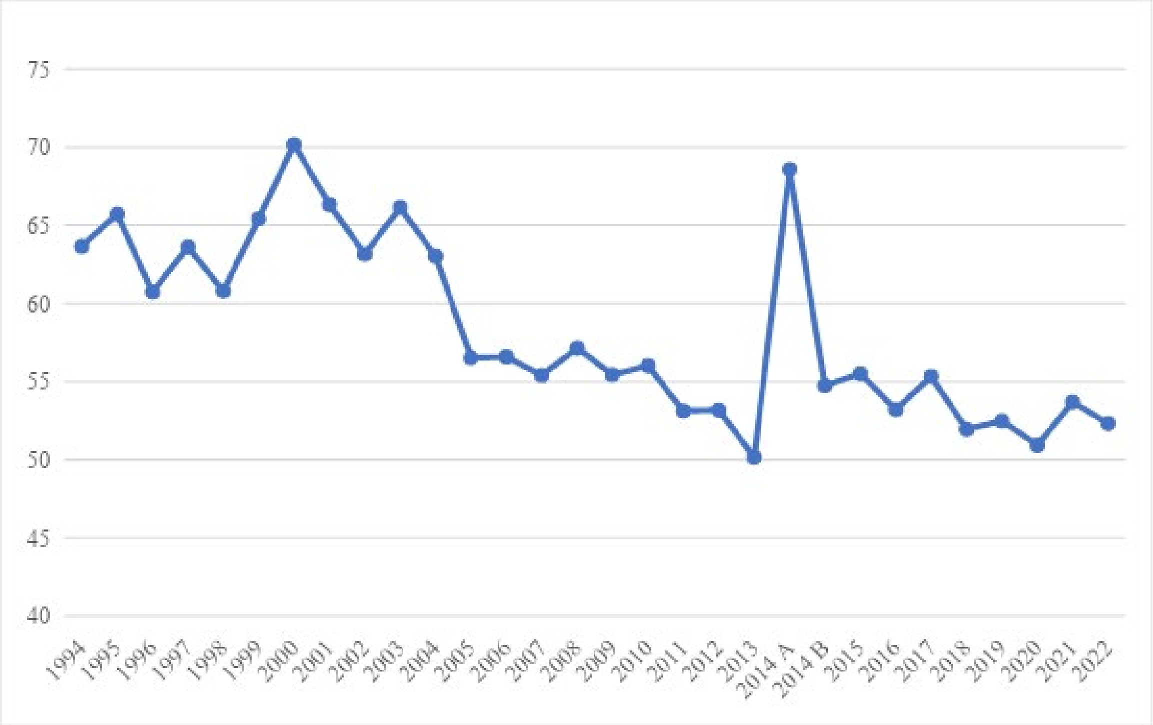
Percentage of Total Word Families in the CSAT in Each BNC Freq. Band

Year	$^{\%}$ 1st	$^{\%}$ 1st-5th	$^{\%}$ 6t^{h-}10th
1994	63.6598	95.8763	3.22165
1995	65.7258	96.7742	3.0914
1996	60.7355	96.3227	3.08422
1997	63.6245	95.9583	2.86832
1998	60.8028	96.222	3.18772
1999	65.4182	97.3783	2.37203
2000	70.1731	98.5353	1.1984
2001	66.3324	97.851	1.7192
2002	63.1649	97.3404	2.65957
2003	66.1333	97.7333	2
2004	63.0263	97.5	2.10526
2005	56.5264	94.7141	4.74649
2006	56.5789	94.4079	4.93421
2007	55.404	95.5929	4.3022
2008	57.1429	95.5162	3.54536
2009	55.4379	94.9952	4.23484
2010	56.0123	95.5807	3.08325
2011	53.1191	92.8166	6.23819
2012	53.1754	92.1327	6.54028
2013	50.177	93.2743	4.86726
2014 A	68.5756	96.9163	2.64317
2014 B	54.7492	91.889	6.72359
2015	55.5094	94.3867	4.26195
2016	53.2101	94.2607	5.15564
2017	55.3171	93.7561	5.17073
2018	51.9531	93.2617	5.66406
2019	52.4857	92.065	6.69216
2020	50.9416	92.8437	5.64972
2021	53.6853	92.6295	6.0757
2022	52.3153	93.399	5.51724

It is now possible to extend Chon's (2014) methodology up until the most recent test to see if the trends she found in 2014 have continued. In the following three graphs, the data from columns two through four are represented visually over time. The first graph, Figure 8, shows the change in proportion over time of word families in the highest frequency band of the BNC.

Figure 8

Percentage of Total Word Families in the CSAT in the 1st BNC Freq. Band

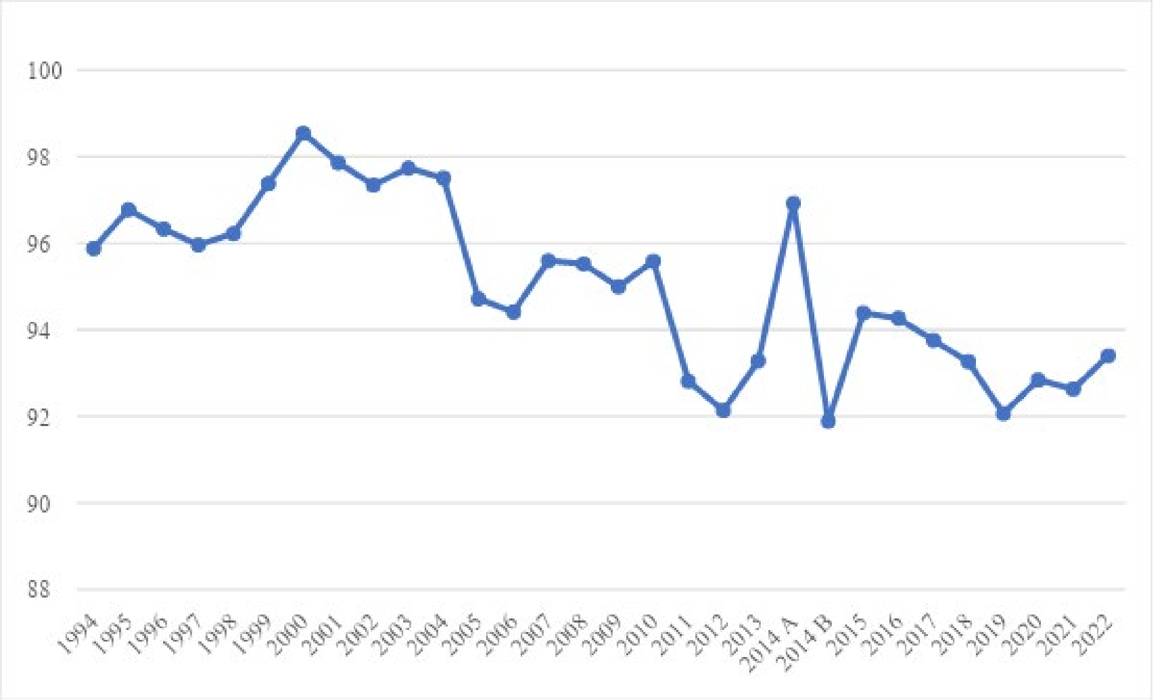


This frequency band is important and deserving of its own treatment because, as discussed previously, the first frequency band includes words that serve as cohesive cues in establishing connections and relationships between elements of the text. Cohesion indices should be directly correlated to these data if the Coh-Metrix 3.0 components are indeed valid measurements. Throughout my observations of the visible trends in the data on Coh-Metrix principal components and readability and easability measures, I found certain noteworthy similarities. Yet again, I found a downward overall trend in this very critical index of text cohesion, and, by extension, readability. Less cohesive texts are, by nature, more difficult, especially for low-level learners. Again, tests administered during the early 2000s seem to be the easiest, whereas the latest stretch of the 2010s has been, by far, the most difficult, with the only exception being the 2014A test.

Figure 9 groups together the first five word family frequency bands of the BNC. These five bands collectively represent the vast majority of the vocabulary contained in the CSAT English Reading Section. Any salient trend in this data over time would be cause for alarm due to the very large number of word families captured by even a few percentage points of change.

Figure 9

Percentage of Total Word Families in the CSAT in the 1st–5th BNC Freq. Band

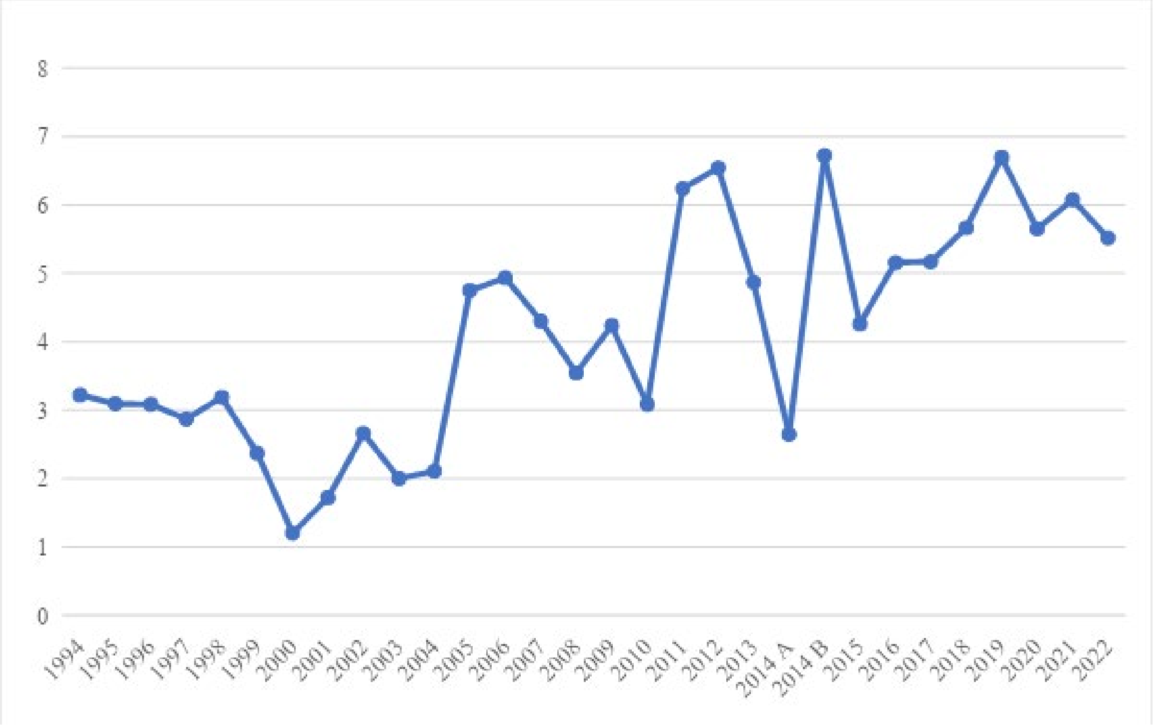


If words occupying the first frequency band are an index of readability in large part because of their functional capacity as connectives and coordinators, the broader subset of bands one through five represents the most common, and therefore the most easily recognizable body of vocabulary. Here, as before, the trend is one of decreasing proportion of word families in this subset over time, with recent tests from the 2010s onward providing the lowest figures. The 2000 and the 2014A tests once again represent the peaks of their respective decades.

The final figure, Figure 10, indicates the percentage of total word families occupying the lowest-frequency bands of the BNC that provide sufficient numbers of occurrences for statistical analysis to be conducted. The 10th-band threshold is in alignment with the methodology of Chon (2014). Although she did include reference to the 14th band, the bulk of the analysis in her article is concentrated on the first 10 for this same reason (Chon, 2014).

Figure 10

Percentage of Total Word Families in the CSAT in the 6th–10th BNC Freq. Band



Just as word families in the high-frequency bands declined over the 29-year period, word families in the low frequency band increased in proportion relative to word families in other bands. This means that the proportion of unfamiliar vocabulary is steadily increasing in the CSAT English Reading Section. Unsurprisingly. The line in Figure 10 looks almost like a mirror image of the one in Figure 9, with the main trough concentrating around the year 2000 and absolute peaks rising in the lattermost decade.

Results of Analyses of Correlations between Data Sets

This statistical analysis of the data reported thus far strives to accomplish two objectives: first, to test the extent to which there is a statistically significant positive or negative trend in each area of inquiry from the preceding section, and second, to understand whether, and to what extent, there is a statistically significant strong correlation between Coh-Metrix components and measures, on the one hand, and, on the other, between Coh-Metrix components and measures and the lexical profile of the CSAT according to the BNC lexical analysis using RANGE. Each subsection under this heading handles these objectives, and the sub-objectives upon which they depend, in this order. Results in this section are provided in prose format with all relevant statistics as described in the introduction to this chapter.

Results of Analysis of Linear Regressions and Significance

Under this subheading, I examine the extent to which the results of each data set from both Coh-Metrix evaluation and lexical analysis display a statistically significant trend over time. In order to accomplish this, Pearson’s Correlation Coefficient is used whereby the X–axis data are represented as time, in years from 1994–2022, and the Y–axis data field is occupied by the numerical results from each table and figure in the preceding section. Pearson’s coefficient is most often used to frame the correlation between variables with data generated as the output of research, but in this case one of the two variables is time, a natural phenomenon. The test releases once, yearly, with the exception of 2014 being a year that saw two tests. The units of the X–axis can be understood to be denominated, by necessity, as even integers. One new year means exactly one new test, so I can focus exclusively on the activity of the lines represented by each figure in the preceding section and whether or not they can be relied upon as a basis for generalizations about the CSAT English Reading Section's change over time. As a foundation for added insight, I calculated the linear regression equation of each of the variables to adjust my understanding for the influence of outliers such as the 2014A test.

Syntactic Simplicity. For syntactic simplicity, an overall downward trend was discernible in Figure 1, with the steepest decline being in the early 2000s. This was the only figure in which there was no obvious spike in the 2014A test. The Pearson's Coefficient calculation for this data set produced an *r* value of -.62, which is indicative of a moderate negative correlation. I can conclude that there is a tendency for the text to have reduced syntactic simplicity over time. With a *p* value of .000274, these results are significant at $p < .05$. The linear regression calculation gave a stronger impression of linearity than the *r* value, however. The linear regression calculation, formulated as $\hat{y} = bX + a$, yielded an equation $\hat{y} = -0.60X + 1276.63$, which shows that despite the scattering of points at the extremities of the plot, the line of best fit is rather steep indeed.

Word Concreteness. For word concreteness, it was more difficult to see a clear downward trend in Figure 2, but the right half of the data did seem to be a bit lower than the left half, which stayed well above the absolute minimum. 2014A was a significant outlier in this graph and, surprisingly, the absolute peak of word concreteness, an index of text easability, was the year 2009, not the early 2000s as expected. The Pearson's

Correlation Coefficient calculation produced an *r* value of -.40, which is indicative of a high-weak negative correlation. It is difficult to conclude at this time that there is a tendency for the text to have reduced word concreteness over time. With a *p* value of .028, these results are significant at $p < .05$. The linear regression calculation produced the equation $\hat{y} = -0.57X + 1194.40$, which shows a gradual slope and a plot very centrally scattered with significant outliers.

Referential Cohesion. For referential cohesion, a moderate downward trend seemed to manifest in certain sections of Figure 3, but an overall trend was even harder to discern than in Figure 2's graph of word concreteness over time. Among the several scattered spikes in referential cohesion, there was a surprising absolute peak in the 2014A test. Curiously, the first test from 1994 and the last test from 2022 achieved roughly the same score. The Pearson's Correlation Coefficient calculation also produced an *r* value of -.40, which is again a high-weak negative correlation. It is difficult to conclude at this time that there is a tendency for the text to have reduced referential cohesion over time. With a *p* value of .031, these results are significant at $p < .05$. The linear regression calculation $\hat{y} = -0.22X + 450.33$ shows a line-of-best-fit slope quite similar to the one calculated for word concreteness, with a similar distribution of scattered peaks and troughs in the data plot.

Deep Cohesion. It was expected at the outset that the accuracy of measures of cohesion will suffer in the analysis of standardized tests because of the nature of their composition. This is partly demonstrated by the inconclusive figures seen in the referential cohesion subsection, which defied the expectations that are elsewhere primed by very clear trends in text easability and readability measures. Deep cohesion tells a much more dramatic story. In a text that otherwise exhibits clear disparities over time in the variables underlying reading difficulty, an evaluation of deep cohesion shows nothing at all. The Pearson Correlation Coefficient calculation yielded an *r* value of -.09, which is indicative of practically no correlation at all. Therefore, it cannot be said using the tools available in this methodology that the text is either increasing or decreasing in deep cohesion over time. On the other hand, the *p* value of the significance test was .62, which is in fact not significant at $p < .05$. It will not, therefore, be possible to draw any conclusions whatsoever about deep cohesion in the CSAT English Reading Section using Coh-Metrix automated text evaluation as a methodology. The line of best fit for this data is nearly horizontal, with a linear regression equation of $\hat{y} = -0.090X + 257.48$.

Flesch-Kincaid Grade Level. The formula underlying this measure produces a figure that is thought to approximately correspond to reading grade level. In the previous section, I found that the grade level of the CSAT English Reading Section material had increased under this measure quite significantly over 29 years. There was a striking perception of correlation between this figure and the passage of time. The Pearson Correlation Coefficient calculation produced an *r* value of .9, which is indicative of a very strong correlation. It can certainly be said that the Flesch-Kincaid grade level of the CSAT English Reading Section has been increasing steadily over time. With a *p* value of $< .00001$, these results are significant at $p < .05$. The linear regression calculation produced a line of best fit with a regression equation of $\hat{y} = 0.16X - 311.84$., indicative of a strong upward trend of almost forty-five degrees.

Flesch Reading Ease. Just as the measure above shows grade level, and therefore large degrees of increasing reading difficulty, this measure shows the inverse. According to the Pearson Correlation Coefficient calculation, this data set has an *r* value of -.9, indicative of a very strong negative trend. It can be confirmed that the reading ease of the CSAT English Reading Section according to this measure is decreasing progressively over time. These results are significant at p < .05 with a *p* value of < .00001. The linear regression calculation produced a line of best fit with an almost forty-five-degree negative slope and a regression equation of ŷ = -0.86X + 1796.93.

Coh-Metrix L2 Readability. The graph of Coh-Metrix L2 Readability very closely mirrored that of the Flesch Reading Ease data. The Pearson Correlation Coefficient calculation resulted in an *r* value of -.86, indicating a strong correlation. It is safe to say that the L2 readability of the CSAT English Section is decreasing over time. These results are significant at P < .05 with a *p* value of < .00001. The linear regression equation produced a line of best fit that very closely resembles the one for the Flesch Reading Ease data, accompanied by a linear regression equation of ŷ = -0.36X + 732.74.

Lexical Profile according to RANGE. As shown in the first half of this chapter, the RANGE results were split into three columns in Table 6 based on word family frequency band or ranges of word family frequency. These included the first frequency band in the second column, the first through fifth frequency bands in the third column, and the sixth through tenth frequency bands in the fourth column. Each of the resultant graphs from these data sets, which are displayed in Figures 8, 9, and 10, displayed apparent linearity and clear correlations in the initial observation, even though there were significant outliers, such as the 2014A test. The Person Correlation Coefficient calculation for column 2 produced an *r* value of -.75, indicative of a high-moderate negative correlation. This means that it is highly likely that the proportion of words in the first, or highest, frequency band of the BNC present in the CSAT English Reading Section is decreasing over time. These results are significant at $p < .05$ with a *p* value of $< .00001$. The linear regression calculation produced a line of best fit with a noticeable negative slope and a linear regression equation of $\hat{y} = -0.49X + 1051.40$.

The third column includes the data on proportion of word families in the first through fifth frequency bands of the BNC. The Pearson Correlation Coefficient calculation of this data set produced an *r* value of -.76, which indicates a high-moderate negative correlation. This means that it is highly likely that the proportion of word families in the largest, and most high-frequency segment of vocabulary in the CSAT English Reading Section is decreasing in proportion relative to other, lower-frequency word family frequency bands over time. With a *p* value of $< .00001$, these results are significant at $p < .05$. The linear regression calculation produced a very similar line of best fit to the one for the second-column data, with a linear regression equation of $\hat{y} = -0.17X + 434.82$.

The fourth column includes the data on proportion of word families in the sixth through tenth frequency bands of the BNC. The Pearson Correlation Coefficient calculation of this data set produced an *r* value of .74, indicating a high positive correlation. This means that it is highly likely that the proportion of word families in the lowest-frequency segment of vocabulary in the CSAT English Reading Section is increasing over time relative to other, higher-frequency segments. With a *p* value of $< .00001$, these results are significant at $p < .05$. The linear regression calculation produced a line of best fit that shows a distinct positive slope and a linear regression equation of $\hat{y} = 0.14X - 268.75$. Having completed the subsection on the temporal correlation analyses of all of my data sets, I now proceed to the results on internal correlations between Coh-Metrix components and measures.

Results of Analysis of Correlation between Coh-Metrix Components and Measures

This subheading explores whether, and to what extent, the findings of individual Coh-Metrix components and measures produced through automated evaluation of the text of the CSAT English Reading Section support one another in terms of what trends they display and what interpretations they invite. Once again, Pearson's Correlation Coefficient is employed as a test of the relationship between two data sets. Significance tests are conducted to produce a *p* value.

Syntactic simplicity and Readability-Easability Measures. Among the three Coh-Metrix principal components, syntactic simplicity offered the clearest picture of a linear change, and my calculations found a moderate correlation between the data gathered and the succession of time, framed as new yearly iterations of the CSAT English Reading Section. With the various readability-easability measures occupying, in turn, the X–axis, which was previously occupied by time, in my Pearson Correlation Coefficient calculation, the relationship between these variables was tested. The first test, looking at Flesch-Kincaid Reading Level and syntactic simplicity, produced an *r* value of -.73, which indicates a high-moderate negative correlation significant at p < .05 with a *p* value of < .00001. This means that it is highly likely that as the grade level of the CSAT English Reading Section according to this measure increases, syntactic simplicity decreases.

The second test, looking at Flesch Reading Ease and syntactic simplicity, produced an *r* value of .63, indicating a moderate positive correlation significant at p

< .05 with a *p* value of .00021. This means that it is likely that as the reading ease of the CSAT English Reading Section according to this measure increases, so does syntactic simplicity as measured by Coh-Metrix. Finally, the third test, looking at Coh-Metrix L2 Readability, produced an *r* value of .72, indicating a high-moderate positive correlation significant at p < .05 with a *p* value of < .00001. This means that it is highly likely that as L2 Readability decreases, syntactic simplicity also decreases.

Word Concreteness and Readability-Easability Measures. Word concreteness was first among the Coh-Metrix principal components to begin to show inconsistent results in terms of linearity over time, where I found merely a high-weak correlation according to the Pearson Coefficient. The first test puts Flesch-Kincaid Grade level against word concreteness. The result from the Pearson Coefficient calculation is an *r* value of -.44, indicating a high-weak negative correlation, which is significant at p < .05 with a *p* value of .015. It can be understood from this that it is possible, although not demonstrable, that as the Flesch-Kincaid Grade Level of the text increases, so does the word concreteness decrease. The second test, which opposes Flesch Reading Ease with word concreteness, produced an *r* value of .54, indicating a moderate positive correlation which is significant at p < .05 with a *p* value of .002. Thus, it is likely that as the Flesch Reading Ease of the text decreases, so does word concreteness increase. The final test putting Coh-Metrix L2 Readability against word concreteness produced an *r* value of .32, which is not significant at p < .05 with a *p* value of .09. It is not possible to draw conclusions about this correlation at this time.

Referential Cohesion and Readability-Easability Measures. Evaluation of referential cohesion failed to produce insightful results in my initial observation of the data, but now I turn to analytical statistics with the hopeful objective of gaining more from this data set. The results of the first test are more promising than my previous examination of the possible correlation between increased referential cohesion and time passed. Looking at Flesch-Kincaid Grade Level and referential cohesion, I produced an *r* value of -.55 from the Pearson Correlation Coefficient calculation, indicative of a moderate negative correlation, which is significant at p <.05 with a *p* value of .002. Thus, it is likely that as the Flesch-Kincaid Grade Level of the text increases, so does its referential cohesion decrease. The second test putting Flesch Reading Ease against referential cohesion produced an *r* value of .59, indicating a moderate positive correlation, which is significant at p < .05 with a *p* value of .0007. This means that it is likely that as the Flesch Reading Ease of the text decreases, so does its referential cohesion. The final test, which pairs Coh-Metrix L2 Readability with referential cohesion, produced an *r* score of .70, showing a high-moderate correlation significant at p < .05 with a *p* score of .00002, from which I understand that it is very likely that as the text decreases in Coh-Metrix L2 Readability, so does it also decrease in referential cohesion.

Deep Cohesion and Readability-Easability Measures. Deep cohesion was always a mere footnote to this analysis, but it still bears inclusion as a null data set from which I expect to see no effect or correlation. The first test of Flesch-Kincaid Grade Level and deep cohesion produced an *r* score of -.10, a weak negative correlation. A negative correlation would mean that as the Flesch-Kincaid Grade Level goes up, deep cohesion goes down. Nevertheless, these results are not significant at p < .05 with a *p* value of .62. A similar story emerged with the other two tests, which provided *r* values of

0.12 and .05, respectively, neither of which were statistically significant. Deep cohesion results are truly all over the map because of the CSAT English Section's variable genre distribution and lack of meaningful connections between passages.

Results of Analysis of Correlation between Coh-Metrix Components and Measures and Lexical Data from RANGE

The final synthesis of findings from automated text evaluation through Coh-Metrix includes data on the lexical profile of the CSAT English Section assigned by the BNC corpus word family frequency tool RANGE. This subsection proceeds with a structure similar to the one preceding it. As before, the Pearson Correlation Coefficient is used to calculate an *r* score for each pair of data sets, with the RANGE data for each year occupying the X–axis and the relevant Coh-Metrix data set occupying the Y–axis. Each subheading below this one provides the results of statistical analysis for each correlation.

Syntactic Simplicity and Lexical Profile of the CSAT. Starting in order with the first Coh-Metrix principal component, the first correlation calculation pairing change in the proportion of word families in the first frequency band with syntactic simplicity produced an *r* value of .70, indicating a high-moderate positive correlation, which is significant at p < .05 with a *p* value of .00002. This means that it is very likely that as the proportion of the highest frequency word families in the CSAT English Reading Section decreases, so does syntactic simplicity also decrease. The second test pairing change in the proportion of word families in the first through fifth frequency bands with syntactic simplicity produced an *r* value of .62, showing a moderate positive correlation, which is significant at p <.05 with a *p* value of .0003. This means it is likely that as the proportion of high frequency word families in the text decreases, so does syntactic simplicity decrease. The third and final test pairing change in the proportion of word families in the sixth through tenth frequency bands with syntactic simplicity produced an *r* value of -.62, a moderate negative correlation, which is significant at p < .05 with a *p* value of .0003. This means that it is likely that as the proportion of low-frequency word families in the text increases, syntactic simplicity decreases. In light of my findings in the previous subsection correlating syntactic simplicity with readability-easability measures, I also predict that based on this result, there is an increased likelihood of readability-easability measures correlating with the lexical data as well.

Word Concreteness and the Lexical Profile of the CSAT. Calculating the coefficient of correlation between change in the first frequency band of word families in the CSAT and word concreteness produced an *r* value of only .38, a high-weak positive correlation, which is significant at p < .05 with a *p* value of .037. It is possible that as the proportion of the highest frequency word families in the text decreases, word concreteness also decreases. The second calculation using the first through fifth frequency bands resulted in an *r* value of .52, a moderate positive correlation, which is significant at p < .05 with a *p* value of .003. It is likely that as the proportion of word families in the high frequency bands in the text decreases, word concreteness also decreases. Finally, the calculation pairing change in the proportion of low-frequency words families with word concreteness produced an *r* value of -0.46, a high-weak correlation, which is significant at p < .05 with a *p* value of .01. It is possible that as the proportion of low-frequency word families in the text increases, word concreteness decreases. It is reassuring to have found a moderate correlation between word concreteness and word frequency in general, even though the more concentrated peripheral categories of the highest frequency word family bands and the low frequency word family bands produced less conclusive results. So far, the .52 *r* value of the second test represents the second strongest correlation with word concreteness so far, barely lower than the .54 *r* value of the calculation pairing it with Flesch Reading Ease.

Referential Cohesion and the Lexical Profile of the CSAT. The first calculation of the correlation between the first word-family frequency band and referential cohesion produced an *r* value of .74, indicating a high-moderate positive correlation, which is significant at p < .05 with a *p* value of < .00001. This means that it is very likely that as the proportion of the highest-frequency word families in the text decreases, referential cohesion also decreases. The second calculation pairing the data on the first through fifth word frequency bands with the data on referential cohesion produced an *r* value of .68, a figure indicative of a moderate positive correlation, which is significant at p < .05 with a *p* value of .00004. This means that it is likely that as the proportion of high frequency word families in the text decreases, referential cohesion also decreases. Finally, the third calculation including the data from the low-frequency word bands produced an *r* value of -.65, indicating a moderate negative correlation, which is significant at p < .05 with a *p* value of .0001. This means it is likely that as the proportion of low-frequency word families increases in the text, referential cohesion decreases. These findings on referential cohesion represent the strongest correlations related to this component thus far in my analysis. It is encouraging to find that there is a more convincing relationship between the lexical data and referential cohesion when other correlation analyses, except for the one connecting the latter with L2 readability, failed to show this.

Deep Cohesion and the Lexical Profile of the CSAT. If my assumptions about deep cohesion and the CSAT, elaborated at length elsewhere, hold true, there will be little additional insight to gain from these three calculations. First of all, the *r* value generated by the calculation of the coefficient of correlation between the proportions of word families in the first frequency band and the data on deep cohesion was .12, a very weak positive correlation and one which is not statistically significant at p < .05 with a *p* value of .05. The other two tests produced *r* values of .07 and .09, neither of which are statistically significant. Even though these figures are not usable for analysis, they do, at the very least, always fall on the right side of the expected correlation—that is, I never find a negative correlation in deep cohesion, an index of easability, when I expect to find a positive one.

Readability-Easability Measures and the Lexical Profile of the CSAT. This subheading includes nine calculations in all—one set of three structured as above for each of the three readability-easability measures of Coh-Metrix. Starting with Flesch-Kincaid Grade Level, the *r* value for the coefficient of correlation between the change in proportion of word families in the first frequency band and grade level is -.9, indicating a strong negative correlation, which is significant at p < .05 with a *p* value of < .00001. This means it can definitely be said that as the proportion of word families in the highest frequency band decreases, the Flesch-Kincaid Grade Level increases. Furthermore, the second calculation putting the data on the first through the fifth frequency bands up against the grade level data yielded an *r* value of -.9 once again, which is also significant with a *p* value of < .00001. The final test using the data on the low-frequency word family bands produced another very strong *r* value of .88, demonstrating a positive correlation between the increase in proportion of low-frequency word families and the increase in grade level of the text.

The next set of tests employs Flesch Reading Ease instead of Grade Level. I expect similarly strong correlations here because these two measures of text strongly correlated with one another elsewhere in this chapter. Using the first set of data in my word family frequency table, I calculated an *r* value of .88, a strong correlation that is significant at p > .05 with a *p* value of < .00001. The second calculation using the first through fifth word family frequency band data produced another strong correlation with Flesch Reading Ease at .90, again with a *p* value of < .00001. The final calculation using the data on low-frequency word families produced an *r* value of -.88, a strong negative correlation, once again significant with a *p* value of < .00001. It is possible to discern from this data in this paragraph that as the proportion of high-frequency vocabulary in the CSAT English Reading Section decreases, so does the reading ease of the text decrease and its difficulty increase according to these measures. Conversely, as the proportion of low-frequency vocabulary increases, so does the reading ease of the text decrease and its difficulty increase.

The third and final set of statistical tests shows the extent to which Coh-Metrix L2 Readability scores correlate with data from the RANGE lexical analysis. The first calculation incorporating the data on the proportion of words in the first, highest frequency band produced an *r* value of .92, indicating a very strong positive correlation, which, with a *p* value of .00001 is statistically significant at p < .05. The other two calculations show similarly strong results, with *r* values of .87 and -.86, both of which are significant at p < .05 with *p* values of < .00001. In short, the Coh-Metrix L2 Readability scores correlate very strongly with the word family frequency data, especially with regard to the reduction in high-frequency vocabulary in the CSAT over time resulting in lower L2 readability according to this measure.

CHAPTER V

DISCUSSION

This chapter proceeds first with a restatement of the overall objectives for this study, followed by a discussion organized according to the same structure as the results section for ease of reading. First discussed are the results of the automated evaluation of the CSAT English Section text, with an explication of what individual contributions they make to the research objectives. Next are the results of the lexical analysis of the CSAT text. Finally, the statistical correlation analyses, representing the synthesis of all the gathered data, occupy the bulk of the discussion. Relevant connections to the scholarly literature are notated throughout.

Discussion of Results of Automated Evaluation of CSAT Text with Coh-Metrix

This section refers to the figures and tables of data produced in the various Coh-Metrix indices and measures employed in this research. From the beginning, this study has taken the position that the status held by Coh-Metrix as the default research tool for studying the text of the CSAT English Section should not go unexamined. As with any piece of software, Coh-Metrix can only perform the tasks that it is programmed to perform and in the way it is programmed to perform them. In the case of the various indices and measures encoded into Coh-Metrix, there are significant interpretive decisions involved in the creation of the underlying formulae that constitute these indices and measures. These are decisions that are based not only on scholarly consensus but also on the judgment of the software's creators. Although the present study concedes the justifications for many of these decisions, it is important to maintain the perspective that the base assumptions in the coding of automated text evaluation tools should be subject to scrutiny and should be continually updated in light of new findings. Software is not ideologically neutral and does not attain a level of objectivity that is higher than that of its human creators. Regardless of its stage of development and sophistication, Coh-Metrix will always only represent a certain stage in the advancement of scholarly knowledge on the inner workings of text and discourse—a stage that is necessarily delimited to that body of knowledge as perceived and consciously selected by the creators of Coh-Metrix and thus incorporated into its formulae. With that caveat out of the way, the remainder of this chapter engages in discussion of the results of the data this study has produced.

Among the principal components of text easability, syntactic simplicity is the very second one enumerated by McNamara et al. (2014) because of its great importance. Texts higher in syntactic simplicity contain shorter sentences and more familiar syntactic structures “that are less challenging to process” (McNamara et al., 2014, p. 85). In their diachronic studies, J-R. Kim (2017) and Moon and Kim (2017) had found that syntactic complexity had increased sharply over time, and other studies covering a variety of time periods, and from referential standpoints, have confirmed that tests from the past few years show very low syntactic simplicity according to Coh-Metrix measures (Ahn & Bae, 2021; J. Chang, 2018a; J. Chang, 2019a; M. Choi, 2018; Hwang & J. Lee, 2020a; S. Kim, 2021; Koh & Shin, 2017; H. Lee, 2020; J. Lee, 2020; J. Park, 2021; Shin, 2019). In light of these findings, it is not especially remarkable to reproduce the same conclusion that syntactic simplicity is decreasing. What may be found, however, is a point of correlation with the text readability-easability measures Flesch-Kincaid Grade Level and Flesch Reading Ease, both of which are constructed in large part based on sentence length. It is perhaps evident that more syntactically complex texts would also have a greater proportion of low-frequency words, which connects conceptually to my lexical analysis, even though these two factors are not directly related to one another.

It was observed in the overview of Chon's (2014) study that she found the number of lexical items in the CSAT was increasing over time, meaning that the test was becoming longer and longer. Moreover, in my own process of data gathering, I observed that the size of the passages had increased from a modest average of 5.2 sentences in 1994 to 9.4 sentences in 2022. The total length of the test increased from 2,887 words to 3,940 in 2022. These data on syntactic complexity, along with the Flesch-Kincaid measures, confirm that the increase in text length has been accompanied by increases in sentence length, passage length, and total text length. I had hoped that by referring back to Chon's finding of a fixed TTR in the CSAT, I would be able to discover a further trend if the number of tokens, or total words, continued to steadily increase, but in fact I found that there has not been a significant increase in total words in the CSAT since 2014B, where Chon's study left off. This means that any changes in the aspects of the lexical profile of the CSAT that relate to difficulty must stem from factors intrinsic to the words themselves, such as frequency, which influences familiarity and thus ease of recognition. Finally, in keeping with the conclusions of J-R. Kim (2017), Koh and Shin (2017), Shin (2019), Chang (2018a; 2019a), and S. Kim (2021), none of the peaks or troughs in Figure 1 correlate with the major government interventions of 2011, integrating EBS materials into the CSAT, or of 2018, switching to criterion-referenced grading.

McNamara et al. (2014) provide grade-level norms calculated based on corpora in the three genre categories of language arts, social studies, and science, that correspond to the grade bands of the Common Core Standards, to serve as a benchmark for interpreting

Coh-Metrix scores in the various indices. These are standards constructed for reading material intended for L1 English-speaking students, so some adjustment in frame of reference must be made for the interpretation of these scores' implications for L2 English learners, who in EFL curricula are not prepared to perform on the same level as L1 learners. The 2022 CSAT English Reading Section's syntactic simplicity score of 67.36, which is nearly identical to the mean score of 67.19, falls in between the mean for grade-three and grade-four language arts texts, in between grade-five and grade-six social studies texts, and in between grade-10 and grade-11 level science texts. The 1994 score of 79.1, on the other hand, corresponds to approximately grade one–two language arts texts, grade three–four social studies texts, and grade five–six science texts. These scores represent at least a whole grade level increase based on comparisons with each category of indices' norms, and in the case of science text norms, an increase of four to five levels.

The next component up for discussion is word concreteness. Word concreteness declined, although the temporal trend was less instructive than expected. Correlations are expected to come through with the lexical data analysis because of the fact that concrete words are assumed to be higher in frequency than abstract words, which are more difficult to use and understand (McNamara et al., 2014). Word concreteness did not seem to interact in any notable way with the reforms in the test from 2011 and 2018, echoing the work of the earlier cited scholars (J. Chang, 2018a; 2019a; 2019b; S. Kim, 2020; Koh & Shin, 2017; Shin, 2019). The major outlier in this respect was the 2014A test, which was constructed using intentionally easier vocabulary (Chon, 2014). It is unknown at this time what institutional aim influenced the production of the two tests for 2014, and that will require additional research separate from the present study.

It is difficult to evaluate the temporal trend of word concreteness in the CSAT without consideration of additional variables, but interpreting the mean score in absolute terms produced alarming results. Mean word concreteness, which is thought to be an index of vocabulary difficulty, was exceptionally low in the CSAT, so much that it exceeded the capstone threshold of McNamara's (2014) grade level norms in every text genre category. The mean word concreteness score of 45.99 of the CSAT English Reading Section was lower than the thresholds for year 12, L1 English-speaking high-school students' reading material in language arts, which was 59.46, social studies, which was 51.25, and science, which was 50.67. This means that regardless of other correlations, the mean word concreteness of the CSAT is considerably lower than that of the reading material assigned to secondary-school students in the United States. This suggests that the English vocabulary difficulty of the CSAT is above the level of an L1 English-speaking high-school graduate, and perhaps more on par with a L1 English-speaking university student, in resounding confirmation of Chon's (2014) findings in the area of lexical thresholds, a thread seen again in discussion of the RANGE results and correlation analyses.

Referential cohesion, like word concreteness, did not produce as strong a visible temporal trend as expected in Figure 3, but it did continue to support the notion that little measurable change in the text of the CSAT coincided with the policy changes of 2011 and 2018. In both word concreteness and referential cohesion, the first and most recent tests scored very similarly. The underwhelming trend in word concreteness could perhaps be partly explained by M. Choi's (2018) findings that word concreteness is different strongly between question types in the test, which could mean that a change in distribution of question types, if there were one, would shift the word concreteness score separate from the influence of changes in other indices or lexical composition. Word concreteness could also be explained by an increase in overall text and possibly a change in genre distribution. If science texts increased in proportion to other texts as the length of the text was expanded, there could have been an increase in concrete words at the expense of abstract words that nevertheless still coincided with increased difficulty. Future research into genre distribution could tell us more about why certain Coh-Metrix indices respond to changes in lexical composition more robustly than others that elicit equivalent anticipation of correlation.

A lack of linearity in the decline in referential cohesion could be attributed to the fact that passage length increases progressively over time, and longer passages simply have a greater ability to establish the conceptual overlap that referential cohesion depends on, because a longer text creates more opportunities for overlap. More research is needed to confirm the sensitivity of referential cohesion to text length and to develop a method of adjusting for text length in calculations of cohesion. This confounding factor is perhaps partly why deep cohesion, which conceptually relies more than any other principal component on the opportunity to establish relationships over a long span of text, produced such scattered results. Like word concreteness, the inconsistency in referential cohesion and deep cohesion could also relate to genre, because different genres are different in cohesion, on average (McNamara et al., 2014). M. Choi's (2018) work also brings to light that, like word concreteness, referential cohesion varies depending on the question type, so a change in question type distribution could be a confounding variable for my study. Other scholars did not measure deep cohesion, but the operation of referential cohesion in the CSAT comes up in almost every Coh-Metrix study, excepting Moon and Kim (2017). No other scholar offers a picture of change in referential cohesion over the span of the entire test aside from J-R. Kim (2017), who concluded that the clearest changes in cohesion are observed in segments, where the first decade shows stably high referential cohesion, but as the 2010s approach, the results stay rather low with some fluctuations, except for the 2014A test, which is a dramatic outlier. What continues to be certain is that the policy changes in 2011 and 2018 did not have much of an impact, if any, on the composition of the test. Perhaps one clue to the erratic behavior of the referential cohesion data in my results is Hwang and J. Lee's (2020c) finding that whereas low argument overlap was observable in the 2014–2018 CSATs and CSAT mock tests and exhibited a statistically significant correlation with difficulty, data on semantic overlap, also known as latent semantic analysis, which is another driver of text cohesion, were inconclusive due to lack of statistical significance. It seems that cohesion itself may merit further study in the context of the CSAT English Section, which is a text with very unique properties unlike other compositions commonly evaluated with Coh-Metrix.

In consultation with the grade norms supplied by McNamara et al. (2014), I am astonished to see that the CSAT English Reading Section's mean referential cohesion from 1994–2022, a strong index of the text's easability, is far lower than the norm for the highest grade level in every text genre category. At 14.48, this mean percentile score is more than twenty percentile points below the norms for language arts and social studies texts at grades 11–12, which are 38.67 and 39.60 respectively, and more than forty percentile points below the norm for science texts at this grade level, which is 61.83. This suggests that the CSAT English Reading Section is much lower in referential cohesion and, therefore, much more difficult to read, than the reading material of L1 English speakers reaching adulthood. The finding of J-R. Kim (2017) that there was an overall sharp decrease in referential cohesion in the most recent decade of the CSAT is certainly upheld in my data, but this additional comparison with an external corpus bracketed by grade level provides a novel lens for interpreting the data in absolute terms.

Regardless of the extent to which the CSAT is or is not significantly changing in referential cohesion over time, it is clear that the CSAT is low in referential cohesion on any given year; the peak figure in Table 3 and Figure 3 is only 27.43. The validity of such a comparison to grade-level reading material could be called into question based on my own assumption throughout this work that a standardized test resists automated evaluation in the indices of cohesion because of its fragmentary nature. Nevertheless, I hope to lend support to the validity of this index as an instrument capable of accurately measuring referential cohesion in the CSAT when I discuss the correlation analysis between this index and the corpus-based data from RANGE, which, at the very least, can answer whether or not referential cohesion in the CSAT, as measured by Coh-Metrix 3.0, can be significantly correlated to the properties of corpora.

A valuable direction for my research would be a comparative investigation of referential cohesion that controls for the fact of the text being a standardized assessment. A continuation of the work of Lee and Lee (2018), who compared the CSAT English Reading Section to the English tests in the Chinese and Japanese college entrance exams in both descriptive indices such as word length and unidimensional easability-readability measures such as Flesch-Kincaid grade level, finding that the CSAT was the most difficult of the three, could further build upon the methodological framework of automated text evaluation of standardized tests by including more indices. Indeed, one of the suggestions made by these authors for continued research is to explore indices of cohesion and syntactic complexity among these three tests. I hope to take up this call for additional research in a future study. Also included in a future study would be examinations of mean scores and their trends of referential cohesion of all the reading passages in each CSAT over time.

Turning now to unidimensional text easability-readability measures, I may observe the very clear trends in increasing content grade level, along with decreases in readability, that seem to be only getting worse over time, confirming the stark predictions of Moon and Kim (2017), Chon (2014), Koh and Shin (2017), and many others that predicted the CSAT English Section was increasing in difficulty over time. The Flesch-Kincaid Grade Level results in Figure 5 and the Flesch Reading Ease results in Figure 6 show such a clear picture of change over time perhaps in part because they use such simply constructed formulae based on word and sentence length. On the other hand, Coh-Metrix L2 Readability also shows a very clear trend, and it is a much more complex formulation including factors like content word overlap, sentence syntactic similarity, and word frequency. Both Flesch-Kincaid Grade Level and Flesch Reading Ease are formulated based on an assumption of a reader who is an L1 English speaker. Coh-Metrix L2 Readability, in contrast, hopes to account for the perspective of an L2 English learner. McNamara et al. (2014) conveniently provide grade level indices norms for these three unidimensional text easability-readability measures, but even in the case of Coh-Metrix L2 Readability, they should still be approached based on the model of a L1 English learner, not that of an L2 learner, because the corpora informing the calculation of the norms consists of material bracketed for Common Core grade levels, not hypothetical L2 grade levels. A potentially valuable area of investigation for future research, therefore, would be the construction of grade-level norms for English L2 students in Korea.

The grade-level norms for Flesch-Kincaid Grade Level are presented in an expected manner by McNamara et al. (2014). For language arts texts, the grade level 11–12 norm for this measure is 12.24, while for social studies it is 11.43 and for science it is 10.35. The Flesch-Kincaid Grade Level of the 2022 CSAT English Reading Section is 11.05, the highest it has ever been, meaning that the material is approximately at the reading level of a L1 English-speaking high-school senior according to this measure. As Lee and Lee (2018) noted in their study of the CSAT's reading difficulty compared to similar exams from other countries, this English reading level is excessively high for a college entrance test designed for L2 learners and undermines the CSAT's validity as an assessment of L2 English competency. The Flesch Reading Ease of the CSAT in 2022 is 47, the lowest it has ever been, in comparison to the 11–12 grade level norms of 51.09 for language arts texts, 49.06 for social studies texts, and 52.16 for science texts. Regardless of category, the CSAT English Reading Section is lower in reading ease than the reading material of an L1 English-speaking high-school senior according to this measure. Finally, the Coh-Metrix L2 Readability of the 2022 CSAT is 10.61, the second lowest it has ever been after the 2020 score of 10.55, compared to the 11–12 grade level norm of 12.24 for language arts texts, 11.43 for social studies texts, and 10.35 for science texts. The scores are similar to, although slightly lower than the three norms in the tables provided by McNamara et al. (2014). The readability of the CSAT English Reading Section is, therefore, very similar to that of the reading material of a L1 English-speaking high-school junior or senior according to this measure.

This exploration of the first third of my results brought to light the fact that indications of the CSAT's increasing reading difficulty over time are borne out in the form of clear trends in unidimensional easability-readability measures up through the current exam year of 2022. My updated diachronic study of the CSAT provides a methodological design and data that other researchers may employ in their own correlation analysis and comparisons in the future. This discussion also introduced a point of comparison in McNamara et al.'s (2014) table of indices norms that allows for deeper qualitative insight into the meaning behind the numerical scores. Following from Lee and Lee's (2018) comparative work, I hope to see future possibilities explored in the continued refinement of the methodology of automated text evaluation of standardized L2 English assessments.

Discussion of Results of Lexical Analysis of CSAT Text with RANGE

This section discusses the results of the lexical analysis of the CSAT English Reading Section text conducted using the BNC tool RANGE. In keeping with Chon's (2014) methodological design, the three relevant data sets for this research are the percentage of total word families in the CSAT text appearing in the first BNC word family frequency band, the percentage appearing in the first through fifth bands, and the percentage in the sixth through 10th bands. These data are arranged in three columns in Table 6 and in the three graphs of Figure 8, Figure 9, and Figure 10.

Throughout these representations, the overall trends demonstrated in Chon's (2014) investigation of word family frequencies are observed to have continued through 2022. These include a decreasing proportion of words in the high-frequency word family bands and an increasing proportion of words in the low-frequency word family bands of the BNC. This means that the lexical burden of students taking the CSAT seems to be getting higher and higher each year as the familiarity of the vocabulary in the passages is decreasing. Chon's (2014) work was conducted before the policy that changed the evaluation of the CSAT from norm-referenced to criterion-referenced, so her work did not include reference to this policy change. Nevertheless, this extension of her methodology up to the present has allowed me to observe that there has been no noticeable change in the underlying trends of the text's composition since the revision, a finding which supports the many automated text evaluation studies that have shown continued trends in Coh-Metrix and Lexical Complexity Analyzer indices (J. Chang, 2018a; 2019a; 2019b; C. Kim, 2020; S. Kim, 2020; Shin, 2019). Om (2021) showed using LCA indices of lexical variables that lexical sophistication is strongly correlated with item difficulty, operationalized as percentage of correct answers per question item, as did Hwang and J. Lee (2020b) with Coh-Metrix indices of word information such as word concreteness. In light of these previous findings and proceeding from the data gathered in the present study, I hope to bring to light the essential interrelation between word family frequency and indices that are predictors of passage difficulty in the final section of this chapter, which discusses the correlations demonstrated using analytical statistics.

As stated before, I found that the total volume of text in the CSAT is not increasing every year as it was before, so there is no basis to assume that the number of types per lexical item in the test is increasing since Chon's (2014) work demonstrated such a stable type-token ratio up through the 2014B test. Additionally, J. Chang (2019a) found a type-token ratio that was very consistent with Chon's (2014) data when tabulating basic lexical information about the 2016–2019 tests. Therefore, if the correlations between the RANGE data and Coh-Metrix measures of text difficulty hold up, it is likely that word frequency is the one change in lexical profile that is most operative in driving the difficulty of the CSAT English Reading Section, since type-token ratio and word count and, therefore, absolute number of types per test, has become quite stable, whereas many of the Coh-Metrix indices continue to show change.

This conclusion is easy to accept in light of Chon's (2014) observation that the maximum lexical threshold of the NCE is 3,500 word families, but that the demands of the CSAT English Reading Section in 2014 were such that knowledge of 13,000 word families was necessary for 98% coverage of the test, which is an excessive vocabulary burden. My data here show that the situation is only getting worse. Moreover, I found by comparing to McNamara's (2014) indices norms that the word concreteness of the text of the CSAT was, on average, far lower than what would be expected for the reading material of a high-school senior who is fluent in English. This is strong evidence for a high proportion of difficult and unfamiliar vocabulary. The low referential cohesion could be convincingly explained by a dearth of high-frequency words which are known to function as cohesive cues (McNamara et al., 2014). Flesch-Kincaid Grade Level and Flesch Reading Ease use formulae that consist half of word length, and half of sentence length, so an increase in longer, lower frequency words makes sense there as well. This is information that can only be found with reference to an external corpus as a source of data on the frequency of each word family. In the upcoming final section of this chapter, I discuss the extent to which this information can help illuminate and validate the Coh-Metrix output, which I discussed in the first section, through analysis of correlations between data sets.

Discussion of Results of Analysis of Correlation between Data Sets

Syntactic simplicity displayed a moderate negative correlation with the passage of time, so I can conclude, as expected, that there was a progressive reduction in syntactic simplicity in the CSAT English Section from 1994–2022. The analysis of word concreteness and referential cohesion over time showed merely a weak correlation, meaning that absent any other variables, the results were quite erratic. Turning to correlations with Coh-Metrix measures of easability-readability, I found a high-moderate correlation between referential cohesion and Coh-Metrix L2 Readability, reflecting the conceptual overlap between these two measures, and moderate correlations between referential cohesion and Flesch Reading Ease. There was a moderate correlation between word concreteness and Flesch Reading Ease and moderate to high-moderate correlations between syntactic simplicity and each of the three measures. These correlations support my understanding of the interrelatedness of word length and syntactic simplicity and the co-occurrence of longer word length with more complex syntax. The high-moderate correlation between referential cohesion and Coh-Metrix L2 Readability is a promising demonstration that cohesion in standardized tests may indeed be measurable, although it remains to be seen how accurate the measurement can be. Measures of deep cohesion, as expected, produced no useful data because of the fragmentary nature of the CSAT text.

Turning now to the correlations between the output in Coh-Metrix indices and lexical data from RANGE, I can see that syntactic simplicity did show moderate and high-moderate correlations with the three columns of data on BNC word family frequency bands of the CSAT vocabulary. This is a good indication that as overall word frequency of the text decreases, syntactic simplicity also decreases. The strongest correlation here was between change in syntactic simplicity and change in proportion of word families in the first, highest-frequency band, meaning that a decrease in the highest-frequency words could be associated with increasingly complex syntax. This can be interpreted as resulting from a constrained inventory of function words, which are high-frequency by nature, a situation which produces more lexically dense sentences. This calls back to my superficial observation in the introduction that from the perspective of an L1 English speaker, reading CSAT English passages feels difficult because the sentences are overly dense with sophisticated vocabulary and unusual collocations, almost as if the test writers were trying to fit in as many content words into each sentence as possible.

Word concreteness showed only a moderate positive correlation with the data on the decreasing frequency of word families in the first through fifth bands and weak correlations with the other two columns. This means that it is possible that as the bulk of high-frequency words in the CSAT decreases, so does word concreteness. Word concreteness correlated the weakest with change in the proportion of word families in the first, highest frequency band of the BNC. Since function words are exempt from the calculations of word concreteness, which only looks at content words, this makes sense (McNamara et al., 2014). A category largely characterized by content words should not make a big impact on measures of word concreteness, so this weak correlation is not entirely out of step with my expectations. An increase in the proportion of low-frequency words did not seem to correlate with lower word concreteness, but this could be due to some unseen factor such as genre distribution, since, for example, science texts, although featuring difficult and low-frequency vocabulary, are characterized by greater word concreteness than narrative text, which may nevertheless have similar word family frequency distributions.

Changes in referential cohesion correlated to each of the three columns of word family frequency groupings to a high-moderate degree. This is a good confirmation that it is very likely that lower referential cohesion in the CSAT English Reading Test is associated with a lower proportion of high-frequency word families and a higher proportion of low-frequency word families. Indeed, the strongest of the three correlations is the positive association between the decline in referential cohesion and the decline in proportion of words occupying the first word family frequency band of the BNC, justifying my assumption that the cohesion of a text relies most on words in this category to serve as cohesive cues. These results from correlation with the lexical data strengthen the case that Coh-Metrix measures of referential cohesion in the CSAT are indeed valid measurements. Deep cohesion, in confirmation of my expectations, did not produce any correlation whatsoever. It seems either that that deep cohesion as a concept might not apply to the structure of a standardized assessment or that it does apply but simply cannot be measured with our current tools.

Each of the three correlations between the RANGE data and readability-easability measures displayed very high strong correlations with *r* values above .85. This is a resounding confirmation that word family frequency plays a huge role in determining reading ease as operationalized by the formulae underlying these three measures. Coh-Metrix 3.0 is largely upheld as a valid tool for measuring the properties of the CSAT English Reading Section text. Chon’s (2014) assessment of the excessive lexical burden deriving from the word family frequency distribution of the CSAT text, which, with reference to the 3,500 word family threshold of the NCE, directly calls into question the validity of the CSAT English Test, is irrefutably correlated with Coh-Metrix measures of the text that have demonstrated its reading difficulty to be excessive compared to other similar tests (Lee & Lee, 2018), compared to official English curriculum textbooks (Ahn & Bae, 2021; H. Lee, 2020; J. Park & D. Lee, 2021; J. Park, 2021), based on correct answer rate (Hwang & J. Lee, 2020b), and finally, in my current study, with reference to grade level indices norms provided by McNamara et al. (2014). The final piece of the puzzle is in place that connects all of these disparate research threads, which I hope will point the way forward towards a unified statement on the problems with validity that characterize the CSAT and possible solutions for remedying them.

CHAPTER VI

CONCLUSION

During the first two decades of research into the validity of the CSAT English Reading Section, methodologies for analysis of the text have existed in relatively separate camps, each devoted to a particular toolset and limited to a narrow research purpose. This vast body of research has been instrumental in how I have structured my own study, and it cannot be overstated the extent to which none of my own work would have been possible without the scholarly contributions to the field that have preceded mine. Now that methodologies of automated text evaluation of the CSAT with tools like Coh-Metrix and LCA are well-established in the field and have been validated by significant correlations with external operationalizations of text reading difficulty, it is time to consider deeply how these powerful tools can be leveraged to create the positive changes in CSAT policy that are unanimously called for among the scholarly community.

It is known that Chon's (2014) lexical thresholds of the CSAT can be shown to conflict with the requirements of the NCE and, therefore, directly compromise the CSAT English Test's evaluative purpose. It is also possible to show, as has been demonstrated in the work of several scholars, that Korean high-school English textbooks provide insufficient preparation for the high vocabulary level and grammatical sophistication in the test that determines so much of students' future livelihoods. It is possible to point to the fact that despite being much more difficult in so many ways than other similar college entrance exams, the CSAT English Test does little to foster any demonstrable good in Korean society, either in terms of improving the practice of language teaching or contributing to greater English proficiency among the population. While it may be true that Korean high schoolers and young adults are more proficient in English now than high schoolers and recent graduates of thirty years ago, it is also true that the former group benefited from language instruction at a much earlier age than the latter as a result of a policy unrelated to the CSAT, and additionally experienced much increased contact with English-language media and pop culture through the Internet, smartphones, and improved economic circumstances enabling a greater percentage of the population to travel and study internationally.

On the other hand, it can be shown that the CSAT does much harm to society in the extreme competitiveness it fosters, creating pressure on families to send their school-age children to cram schools at as early an age as possible, increasing the financial burden of raising children and possibly even contributing to a slowed birth rate. It exacerbates existing inequalities and ensures that the children of wealthy families have the easiest path to success. South Korea has one of the highest youth suicide rates in the world, which is a social problem that is often associated with great economic inequality. At the end of a student's high-school career, all of that time spent in cram schools may ensure successful completion of the CSAT English Test, but it is questionable to what extent a course of learning strongly oriented towards test-taking strategies actually fostered communicative competence in the language.

It is not certain to what extent all of these problems can be addressed by drawing upon the scholarly consensus in a future CSAT reform. Because the role of the CSAT in its current state is so entrenched in the education system, it hardly seems possible that a simple revision of the test to make it a valid assessment of the curriculum, which would practically eliminate the need for English cram schools overnight, would be accomplished with a clean transition. If the school English curriculum were actually sufficient to prepare students for the language of the CSAT, and there was suddenly no need for English cram schools, major economic changes would occur. Families at middle- and lower-income levels would suddenly have significantly reduced education expenses, it is possible that many cram schools would see dramatically reduced business or even go out of business, and universities would have to restructure the standards of admission to reflect a reality where performance on an ultra-difficult test is no longer a possible benchmark for admissibility. The practices of English teaching in high school, which, especially in the uppermost grades, have long oriented themselves towards the powerful gravitational force of the CSAT, would need an infusion of new pedagogical strategies based on something other than test wiseness—perhaps this would finally allow teachers the autonomy to develop the communicative classrooms that will foster greater English learning, but as with any great institutional change, there will doubtless be some lingering inertia and need for clear directions.

In consideration of these realities, it is hoped that future research can begin to tackle what a comprehensive reform of the CSAT English Test would look like and how to account for its inevitable dramatic impact on Korean education and society. The good news is that scholars have valid tools to make these kinds of prescriptions, and with each new publication the answers are coming into clearer view. I hope that the present work bringing word frequency back into the spotlight has made a contribution to this ongoing conversation, even if it is small, and I look forward to the opportunity to continue working in this exciting and important field.

REFERENCES

Ahn, H., & Bae, J. (2021). *2015 개정교육과정이 적용된 고등학교 영어I 과II 교과서, EBS 수능 연계 교재, 대학수학능력시험의 읽기 지문의 난이도 비교* [A study about the analysis of linguistic difficulty among English textbooks with 2015 Revised National Curriculum, EBS-CSAT prep books, and College Scholastic Ability Test]. *Secondary English Education*, *11*(4), 39-58. https://doi.org/10.20487/kasee.14.3.202108.39

British National Corpus. (2009, January 26). *British National Corpus (BNC)*. https://www.english-corpora.org/bnc/

Chang, B. (2009). Korea's English education policy innovations to lead the nation into the globalized world. *Journal of Pan-Pacific Association of Applied Linguistics*, *13*(1), 83-97. https://files.eric.ed.gov/fulltext/EJ921027.pdf

Chang, J. (2018a). A comparison and analysis of CSAT reading texts before and after the use of criterion-referenced assessment (2017 vs. 2018): A focus on lexical complexity. *KAFLE Annual Conference*, *2018*(1), 89-90. The Korea Association of Foreign Languages Education.

Chang, J. (2018b). A comparison of 2017-2018 CSAT reading passages via Coh-Metrix: Focusing on descriptive, readability, and easibility measures. *Foreign Languages Education*, *25*(4), 81-106. The Korea Association of Foreign Languages Education. https://dx.doi.org/10.14334/FLE.2018.25.4.81

Chang, J. (2019a). A Comparison of lexical complexity in 2016-2019 CSAT English reading passages by the types of assessment (norm-referenced assessment vs.

criterion-referenced assessment). *The SNU Journal of Education Research*, *28*(3), 85-110. https://doi.org/10.54346/sjer.2019.28.3.85

Chang, J. (2019b). A comparison of syntactic complexity in CSAT reading passages before and after the introduction of criterion-referenced evaluation. *Journal of the Korea English Education Society*, *18*(2), 161-188.https://doi.org/10.18649/jkees.2019.18.2.161

Cho, G. (2013). *EBS 수능 연계교재와 고교 영어교과서 어휘수준에 대한 코퍼스 기반 분석* [A corpus-based analysis of the vocabulary level of EBS CSAT books and high school English textbooks] (Unpublished master's thesis). Yonsei University.

Choi, M. (2018). *코메트릭스(Coh-Metrix)를 활용한 수능 영어 읽기문항의 응집성과 어휘정보 분석* [A Coh-Metrix analysis of cohesion and word information in reading items of CSAT] (Unpublished master's thesis). Korea National University of Education.

Choi, M. & Kim, J. (2017). *수능 영어 문항 유형간 응집력과 어휘정보 분석* [An analysis of cohesion and word information among English CSAT question types]. *The Journal of the Korea Contents Association*, *17*(12), 378-385.10.5392/JKCA.2017.17.12.378

Choi, G. (2015). *수능 외국어 영역과 고등학교 교과서 어휘 분석* [Official title translation unavailable] (Unpublished master's thesis). Dongguk University.

Chon, Y. (2014). Lexical threshold of LS reading in the Korean CSAT. *Institute of British & American Studies*, *31*, 341-376. http://builder.hufs.ac.kr/user/ibas/No31/14.pdf

Goh, G., & Back, J. (2010a). A corpus-based analysis of the vocabulary used in the CSAT English exam and two analogous tests. *English Language and Linguistics*, *16*(2), 1-26. https://doi.org/10.17960/ell.2010.16.2.001

Goh, G., & Back, J. (2010b). A corpus-based analysis of college entrance English exams in Korea, China, and Japan." *Studies in Modern Grammar*, *62*, 205-26.

Hwang, L., & Lee, J. (2020a). *수학능력시험 영어 읽기 지문의 응집성과 문항 난이도 간의 상관관계 분석* [Analysis of correlation between cohesion and item difficulty in English reading]. *The Journal of the Korea Contents Association*, *20*(5), 344-350. https://doi.org/10.5392/JKCA.2020.20.05.344

Hwang, L., & Lee, J. (2020b). *수학능력시험 영어 지문의 텍스트 요인과 문항 난이도의 상관관계 분석: 통사적 복잡성을 중심으로* [Correlation analysis between the text variables and item difficulty in CSAT: focusing on syntactic complexity]. *Studies in English Language & Literature*, *46*(1), 265-283. https://doi.org/10.21559/aellk.2020.46.1.013

Hwang, L., & Lee, J. (2020c). *수학능력시험 영어 영역 읽기 지문의 이독성 및 어휘 정보와 문항 난이도 간의 상관관계* [Correlation between readability/word information and item difficulty in CSAT English reading passage]. *The Journal of Humanities and Social Sciences 21*, *11*(2), 389-400. http://dx.doi.org/10.22143/HSS21.11.2.27

Jeon, M. (2011). Coh-Metrix 를 이용한 중학교 1 학년과 2 학년 개정 영어교과서 읽기 자료의 코퍼스 언어학적 연계성 분석 [A corpus-based analysis of the continuity of the reading materials in middle school English 1 and 2 textbooks with Coh-Metrix], *The Journal of Linguistic Science*, *56*, 210-218. The Linguistic Science Society.

Jeon, M., & Lim, I. (2009). 코메트릭스(Coh-Metrix)를 이용한 중학교 1 학년 개정 영어 교과서의 코퍼스 언어학적 비교 분석 [A corpus-based analysis of middle school English 1 textbooks with Coh-Metrix]. *English Language Teaching*, *21*(4), 265-292. https://doi.org/10.17936/pkelt.2009.21.4.012

Joo, H. (2008). *A corpus-based analysis of vocabulary in the BEWL and the lexical threshold of L2 reading in the Korean CSAT* (Unpublished master's thesis). Korea University.

Kim, C. (2020). A corpus-based comparative analysis of CSAT English from 2015 to 2020 on the basis of criterion-referenced assessment. *Multimedia-Assisted Language Learning*, *23*(2), 98-120. https://doi.org/10.15702/mall.2020.23.2.98

Kim, J-R. (2016). *말하기/쓰기 표현기능 강화 영어교육 방안. 올바른 서울영어교육을 위한 실용영어 정책 활성화 마련 정책토론회* [Official title translation unavailable]. Korea Institute for Curriculum & Education. https://www.kice.re.kr/filedown8.do?fileNM=RRI201100701.pdf&filePath=/research/20121227/1356597512519_522

Kim, J-R. (2017). *수능영어 읽기 지문에 대한 통시적 Coh-Metrix 분석* [A diachronic analysis of English KSAT reading passages]. *Elementary Education Research*, (27), 63-78.

Kim, J-W. (2016). *영어I·II교과서, 대학수학능력시험, EBS 교재 읽기 지문에대한 어휘 분석* [The corpus-based vocabulary analysis of the reading texts in the

English 1 and 2 textbooks, CSATs, and EBS materials] (Unpublished master's thesis). Korea National University of Education.

Kim, J., & Kim, H. (2021). *대학수학능력시험 영어 듣기와 읽기 지문의 어휘, 통사, 담화적 특성 비교* [A comparison of the lexical, syntactic, and discourse features between the listening and the reading texts in the College Scholastic Ability Test]. *Modern English Education*, *22*(1), 45-56. https://doi.org/10.18095/meeso.2021.22.1.45

Kim, J., & Yang, J. (2012). *Coh-Metrix 를 통한 초·중등 영어교과서 연계성 분석* [An analysis of the continuity of elementary and middle school English textbook using Cho-Metrix]. *English Teaching*, *67*(2), 319-341. https://doi.org/10.15858/engtea.67.2.201207.319

Kim, M. (2016). *코메트릭스(Coh-Metrix)를 이용한 중· 고등 영어교과서와 대학수능능력시험 읽기 지문의 난이도 분석* [An analysis of the difficulty of the reading materials in middle and high school English textbooks and the CSAT with Coh-Metrix] (Unpublished master's thesis). Jeju National University.

Kim, N. (2008). *대학수학능력시험 외국어(영어) 영역의 코퍼스 언어학적 어휘비교 분석* [A corpus-based lexical analysis of the foreign language (English) test for the college scholastic ability test (CSAT)]. *The English Teachers Association in Korea, 14*(4), 201-221.

Kim, S. (2021). *수능 영어영역 절대평가 코퍼스 기반 난이도 분석* [A corpus-based analysis of linguistic difficulty of the CSAT English section under absolute grading]. *English Language Assessment*, *16*(2), 205-227. https://doi.org/10.37244/

ela.2021.16.2.205

Koh, N., & Shin, J. (2017). *Coh-Metrix 를 이용한 수능 영어 읽기 영역 지문 난이도 비교: EBS-수능 연계 정책 전후* [A comparison of the level of difficulty in the English reading part of the CSAT: Before and after the EBS-CSAT linkage policy]. *Secondary English Education*, *10*(4), 3-24.

Korea Institute for Curriculum and Evaluation. (2022). *기출문제.* 한국교육과정평가원 대학수학능력시험. https://www.kice.re.kr/boardCnts/list.do?boardID=1500234&m=0403&s=suneung&searchStr=

Korean Englishman. (2020, November 10). *British high schoolers take Korea's SAT English exam!!* YouTube. https://www.youtube.com/watch?v=M_uGV2L5q3s

Kwon, O. (2015). A history of policies regarding the English section of Korea's College Scholastic Ability Test. *English Teaching*, *70*(5), 3-34. https://doi.org/10.15858/engtea.70.5.201512.3

Kwon, S., & Shin, D. (2014). EBS 연계정책에 따른 대학수학능력시험 영어영역 어휘 난이도 변화 분석 [The effects of the EBS books-CSAT linkage policy on vocabulary difficulty of the English section in the CSAT]. *Journal of the Korea English Education Society*, *13*(4), 97-121. http://scholar.dkyobobook.co.kr/searchDetail.laf?barcode=4010024445453

Larsen-Freeman, D. & Anderson, M. (2011). *Techniques and principles in language teaching* (3rd ed.). Oxford University Press.

Lee, C. (2010). *코퍼스 분석 프로그램을 활용한 대학수학능력시험 기출 문제 분석* [A study on the Korean SAT with corpus analysis program] (Unpublished master's thesis). Pusan University of Foreign Studies.

Lee, H. (2018). Unresolved issues in CLT and native-speakerism in Korean English language teaching contexts. *English 21*, *31*(1), 277-295. https://doi.org/10.35771/engdoi.2018.31.1.013

Lee, H. (2020). *코메트릭스(Coh-Metrix)를 이용한 고등학교3 학년 영어 교과서와 대학수학능력시험 읽기 지문의 난이도 비교* [An analysis of the difficulty of the reading materials in high school 3rd grader's English textbooks and the CSAT with Coh-Metrix] (Unpublished master's thesis). Kongju National University.

Lee, H., & Lee, J. (2018). *한·중·일 대학 입학시험 영어 지문의 이독성 비교 분석* [Readability analysis of the English reading texts in the college scholastic ability test of Korea, China, and Japan]. *The Journal of Humanities and Social Sciences 21*, *9*(5), 441-454. https://doi.org/10.22143/HSS21.9.5.32

Lee, J. (2009). Changes and improvement direction of the College Scholastic Ability Test (CSAT). The Second Curriculum Evaluation KICE Policy Forum. Korea Institute for Curriculum and Evaluation.

Lee, J. (2020). A comparison of text difficulty in 2015-2020 CSAT English reading passages between testing types and among question types. *Journal of the Korea English Education Society*, *19*(2), 21-43. https://doi.org/10.18649/jkees.2020.19.2.21

Lee, J. (2011). *대학수학능력시험 외국어영역 어휘의 코퍼스 분석* [A corpus-based study of English vocabulary in the College Scholastic Ability Test (CAST)] (Unpublished master's thesis). Sungshin Women's University.

Lu, X. (2022). Lexical complexity analyzer. Penn State University. Retrieved July 1, 2022, from http://www.personal.psu.edu/xxl13/downloads/lca.html

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix.* Cambridge University Press.

Moon, J., & Kim, H. (2017). *대학수학능력시험 영어 읽기 지문의 언어적 요소 분석* [An analysis of the linguistic elements of the text in the English reading section of the College Scholastic Ability Test]. *Modern English Education*, *18*(1), 193-211. http://dx.doi.org/10.18095/meeso.2017.18.1.09

Nation, I. S. P. (2022). *Vocabulary analysis programs*. Victoria University of Wellington. Retrieved July 1, 2022, from https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs

Om, H. (2021). Predicting the item difficulty of a simulated CSAT English test based on corpus analysis. *English Language Assessment*, *16*(1), 59-78. https://doi.org/10.37244/ela.2021.16.1.59

Park, J. (2021). *코메트릭스(Coh-Metrix)를 활용한 고등학교 영어 교과서와 수능 영어 읽기 지문의 난이도 분석* [An analysis of the difficulty of reading passages in the CSAT and high school English textbooks of the 2015 revised curriculum with Coh-Metrix] (Unpublished master's thesis). Korea National University of Education.

Park, J., & Lee, D. (2021). *Coh-Metrix 를 활용한 고등학교 영어 교과서와 대학수학능력시험 읽기 지문 난이도 비교 연구* [An analysis of the difficulty of reading passages in the CSAT and high school English textbooks of the 2015 revised curriculum]. *Studies in English Language & Literature*, *47*(3), 147-170. https://doi.org/10.21559/aellk.2021.47.3.008

Park, S. (2018, November 16). *Check your English ability with 'notorious' Korean College Entrance Exam.* The Korea Times. https://www.koreatimes.co.kr/www/nation/2018/11/177_258803.html

Shin, Y. (2019). *Coh-Metrix 와 VocabProfile 를 활용한 수능 영어 독해 지문의 분석: 2016-2019 년 4 년을 중심으로* [Analyzing CSAT reading passages by using Coh-Metrix and VocaProfile: focusing on four years from 2016 to 2019]. *Journal of Language Sciences*, *25*(4), 109-127. http://dx.doi.org/10.14384/kals.2019.26.4.109

Song, S. (2012). *A corpus-based study of make, get, and take collocations on the College Scholastic Ability Test of English.* (Unpublished master's thesis). Chungnam National University.

Yang, S., & Lee, D. (2019). *영어 교과서, EBS 교재, 대학수학능력시험의 읽기 지문에 대한 코퍼스 기반 소재별 어휘 사용 양상 분석* [A corpus-based analysis of the topic distribution and vocabulary level of textbooks, EBS Materials, and CSATs]. *Journal of Learner-Centered Curriculum and Instruction*, *19*(4), 711-729. http://dx.doi.org/10.22251/jlcci.2019.19.4.711

Yoo, H. (2016). An Acoustic Analysis of Korean EFL Learners' English Prosody: A Longitudinal Study. *Incheon National University*, *22*(1), 55–75. https://doi.org/10.17959/sppm.2016.22.1.55

Yoon, D. (2006). *수학능력시험 외국어영역과 고등학교 영어교과서의 어휘비교 분석* [A comparative analysis of the vocabularies in CSATs & high school English textbooks] (Unpublished master's thesis). Hongik University.