

교육과정평가연구

The Journal of Curriculum and Evaluation

2026, Vol. 29, No. 3, pp.395~419

DOI: <https://doi.org/10.29221/jce.2026.29.3.395>

대학수학능력시험 영어 고배점 지문의 텍스트 복잡도 종단 분석: 절대평가 전후비교 및 SAT Digital 지문과의 다층적 대조

이혜영 (상인고등학교 교사)*

요약

본 연구는 2010~2025학년도 대학수학능력시험(이하 수능) 영어 영역 고배점(3점) 읽기 지문 97편을 대상으로 텍스트 복잡도(text complexity)의 종단적 변화를 분석하였다. 분석 지표로는 Flesch-Kincaid 학년 수준(Flesch-Kincaid Grade Level, 이하 FK Grade), 이동평균 어형-토큰 비율(Moving-Average Type-Token Ratio, 이하 MATTR), 연결어 밀도, 상상가능성, 명사화 지수, 담화 결속성, 종합 텍스트 복잡도 지수(Composite Text Complexity Index, 이하 CTCI)의 7가지를 적용하였으며, College Board가 공개한 SAT Digital 지문 13편을 동일 지표로 비교 분석하였다. 분석 결과, 수능 영어 고배점 지문은 지난 16년간 단순히 '어려워진' 것이 아니라 '추상화'되는 방향으로 구조적으로 변화하고 있었다. CTCI는 16개년 최강 상승 추세를 보이며($\rho=+.831$, $p<.001$), 상상가능성($\rho=-.608$, $p=.010$)과 연결어 밀도($\rho=-.566$, $p=.033$)는 유의미하게 감소하였다. 2018학년도 절대평가 전환 이후에는 지문 가독성(FK Grade)과 전반적 복잡도(CTCI)가 더욱 상승한 반면, 연결어 밀도와 상상가능성이 감소하여 절대평가의 취지와 상반되는 방향으로 지문 복잡도가 심화되었음을 확인하였다. FK Grade는 1등급 비율과 반직관적인 양의 상관을 보인 반면($\rho=+.743$, $p=.002$), 다층 복잡도를 종합한 CTCI는 유의미한 음의 상관을 나타내어($\rho=-.501$, $p=.048$) 단일 지표 의존 방식의 한계가 실증되었다. 수능-SAT 비교에서는 담화 결속성에서만 유의미한 차이가 나타났으며($p=.021$, $r=.397$), SAT 지문의 담화 결속성이 유의미하게 높았다. 이상의 결과를 바탕으로 다층 가독성 지수(MRI)를 출제 가이드라인에 통합하고, 에듀데이터서비스시스템(Education Data Service System, 이하 EDSS)기반 연도별 복잡도 모니터링 체계 구축을 정책적으로 제언하였다.

주제어 : 수능 영어, 텍스트 복잡도, FK Grade, Coh-Metrix, 절대평가, SAT 비교, 종단연구

* 제1저자, miru9921045@korea.kr

I. 서 론

수능 영어 영역은 2018학년도부터 절대평가로 전환되었다. 교육부(2017)는 이 전환의 목적으로 ‘점수에 따른 성취 수준별 등급 제공’과 ‘영어 학습 부담 완화’를 명시하였다. 그 교육적 의도는 분명하다. 지문과 문항의 복잡도를 연도별로 일관되게 유지함으로써, 일정 수준의 영어 능력을 갖춘 학생이라면 연도에 관계없이 같은 등급을 안정적으로 획득할 수 있는 공정한 평가 환경을 조성하는 것이다.

그러나 현실은 이 교육적 의도와 상당한 거리가 있는 것으로 보인다. 2024학년도 수능 영어 1등급 비율은 4.69%로 급락하였고 2026학년도에는 3.11%로 역대 최저를 기록하였다. 사교육걱정없는세상(2025)의 분석에 따르면 2026학년도 수능 영어 독해지문 중 약 40%가 미국 11학년 이상의 읽기 수준에 해당하며, 수능 시험 범위인 영어Ⅱ 교과서 최고 수준(FK Grade 기준 8.45~11.05)을 최대 5학년 초과한다. 이는 교과서 중심 수업만으로는 수능 지문에 대비하기 어려운 구조적 문제를 시사하며, 고등학교 영어 교육과 수능 평가 간의 간극이 사교육 수요를 구조적으로 심화시키는 요인이 될 수 있다는 점에서 교육정책적으로 중요한 문제이다.

기존 수능 영어 난이도 연구들은 Flesch-Kincaid 학년 수준(Flesch-Kincaid Grade Level, 이하 FK Grade)이나 어휘 빈도와 같은 단일 지표에 의존하여 지문 복잡도를 분석하는 경향이 있었다(김미란, 2022; 이해진·이제영, 2018). 그러나 단일 지표는 텍스트의 표면적 어휘·구문 특성만을 포착할 뿐, 수험생이 실제로 체감하는 독해 부담—특히 문장 간 논리 흐름의 불명시성, 추상적 개념어의 밀도, 담화 구조의 복잡성—을 충분히 설명하지 못한다는 한계가 있다(Crossley et al., 2011). 수능 영어의 교육적 문제를 정확히 진단하고 적절한 정책적 개입 방향을 찾기 위해서는, 텍스트 복잡도를 어휘·구문뿐 아니라 담화·심리언어 층위까지 포함하는 다층적 틀로 분석하는 연구가 필요하다.

이에 본 연구는 두 분석 틀을 통합한다. 첫째, 2010~2025학년도 수능 영어 고배점(3점) 지문 97편을 대상으로 7개 다층 복잡도 지표를 종단 분석한다. 고배점(3점) 지문을 선정한 이유는 수능 영어에서 등급 변별이 실질적으로 이 지문들에 집중되며, 교과서 수준과의 간극 문제가 가장 심각하게 드러나는 층위이기 때문이다. 둘째, 탐색적 외적 타당화를 위해 SAT Digital 지문(n=13)을 동일 지표로 비교한다. SAT Digital을 비교 준거로 선정한 이유는 College Board의 공식 자료로 신뢰성이 보장되며 장르 분포가 명시되어 있어 체계적 비교가 가능하기 때문이다.

본 연구의 연구 문제는 다음과 같다.

- (1) 16개년간 수능 고배점 지문의 복잡도 지표는 어떠한 종단적 추세를 보이는가?
- (2) 절대평가 전환 전후로 복잡도 지표에 통계적으로 유의한 차이가 있는가?
- (3) 텍스트 복잡도 지표와 1등급 비율 간 순서 상관이 존재하는가?
- (4) 수능과 SAT Digital의 담화·심리언어적 복잡도는 어떻게 다른가?

II. 이론적 배경 및 선행연구

1. 텍스트 복잡도의 다층구조

텍스트 복잡도(text complexity)는 독자가 텍스트를 처리하는 데 요구되는 인지적·언어적 부담의 총체로, 어휘·구문·담화·심리언어적 속성이 상호작용하는 다층 구조이다(McNamara et al., 2010). 본 연구에서 텍스트 복잡도는 지문 자체의 언어적 속성을 가리키며, 선지 구성이나 추론 유형에 의해 결정되는 문항 난이도(item difficulty)와 개념적으로 구별된다. 수능 영어처럼 고부담 평가에서 지문의 교육적 적절성을 독립적으로 평가하기 위해서는 이 개념적 구분이 전제되어야 한다.

Graesser et al.(2004)이 개발한 Coh-Metrix는 100여 가지 텍스트 지표를 통합하는 권위 있는 분석 도구로, 본 연구의 분석 틀을 제공한다. 본 연구에서 활용하는 7개 지표와 수능 영어 지문과의 관련성은 다음과 같다.

Flesch-Kincaid 학년 수준(Flesch-Kincaid Grade Level, 이하 FK Grade; Kincaid et al., 1975)는 평균 문장 길이(Average Sentence Length, 이하 ASL)와 어절당 평균 음절 수(Average Syllables per Word, 이하 ASW)에 기반한 가독성 지수이다($FK\ Grade = 0.39 \times ASL + 11.8 \times ASW - 15.59$). 수능 영어 지문에서 FK Grade가 높을수록 문장이 길고 다음절어가 많아 표면적 독해 부담이 증가하나, 연결어 생략과 같은 담화 층위 복잡도는 포착하지 못한다는 한계가 있어 단독 사용 시 체감 난이도를 과소추정할 위험이 있다.

이동평균 어형-토큰 비율(Moving-Average Type-Token Ratio, 이하 MATTR; Covington & McFall, 2010)은 지문 전체에 걸쳐 이동 창(moving window) 방식으로 분석하여 어휘다양성을 측정한다. MSTTR(Mean Segmental TTR)과 달리, MATTR은 창 이동 방식이므로 지문 길이가 달라도 안정적인 비교가 가능하다. 수능 고배점 지문은 매년 일정한 어휘 다양성 수준을 유지하는지, 아니면 특정 시기에 어휘 집중도가 높아지는지를 파악할 수 있어 출제 경향 분석에 유용하다.

담화·심리언어 지표(Category B)로는 다음 5가지를 활용하였다.

① 연결어 밀도는 인과·대조·첨가 연결어의 비율로, 낮을수록 문장 간 논리 관계를 독자 스스로 추론해야 하는 인지 부담이 증가한다. 수능 영어에서 연결어 생략 출제 경향이 있어 이 지표가 체감 난이도를 직접 반영할 것으로 기대된다.

② 상상가능성은 MRC 심리언어 데이터베이스(MRC Psycholinguistic Database, 이하 MRC DB; Coltheart, 1981)의 이미지화 용이성 점수(100~700 범위)로, 낮을수록 추상적인 텍스트이다. 외국어로서의 영어(English as a Foreign Language, 이하 EFL) 학습자는 추상 텍스트에서 L1·L2 이중 처리 부담을 더 강하게 경험하므로(Crossley et al., 2011) 수능 수험생에게 특히 민감한 지표이다. 특히 상상가능성이 낮은 학술적 추상 텍스트는 고등학교 영어 교과서의 구체적 소재 중심 텍스트와 질적으로 다른 독해 전략을 요구하므로, 교과서 기반 수업과 수능 대비 사이의 질적 간극을 드러내는 지표로서 중요한 교육적 의미를 지닌다.

③ 명사화 지수는 -tion, -ness, -ment, -ity 접미사를 포함한 어휘의 비율로, 높을수록 논리 구조가 압축·은폐된다.

④ 담화 결속성은 인접 문장 간 의미적 일관성 종합 지수로, 낮을수록 추론 부담이 증가한다. ⑤ 종합 텍스트 복잡도 지수(Composite Text Complexity Index, 이하 CTCI)는 위 6개 지표를 min-max 정규화 후 방향성에 맞게 집계한 합성 지수(0~1 척도)이다.

2. 수능 영어 텍스트 난이도 선행연구

수능 영어 지문의 언어적 특성을 분석한 선행연구는 어휘·가독성 관점과 담화 구조 관점으로 구분된다. 이들 연구의 성과와 한계를 본 연구와의 관련성 중심으로 정리하면 다음과 같다.

김미란(2022)은 2015~2022학년도 수능 영어 독해지문 136편의 어휘 다양성과 가독성을 분석하여 지문 간 가독성 불균질성을 체계적으로 기술한 최초의 연구라는 의의를 지닌다. 그러나 절대평가 전환을 가로지르는 종단적 변화와 담화 층위 지표를 포함하지 않아, 본 연구의 핵심 질문을 다루지 못하였다.

이혜진과 이제영(2018)은 Coh-Metrix 기반 이독성 분석으로 수능 영어 지문이 일본·중국 대입 지문보다 어휘·구문 수준이 높음을 국제 비교로 실증한 선구적 연구이나, SAT Digital 비교와 담화·심리언어 지표를 포함하지 않았다.

황이수와 이제영(2020)은 수능 및 모의고사 지문 299편에서 담화 결속성이 높을수록 정답률이 상승함을 확인하여 담화 지표의 교육적 유효성을 실증하였으나, 절대평가 전후 종단 비교는 수행하지 않았다.

최민주와 김정렬(2017)은 수능 문항 유형별 응집력과 어휘정보 차이를 코퍼스 기반으로 체계적으로 분류하였으나, 절대평가 전후의 시간적 변화를 추적하지 않았다.

Crossley et al.(2011)은 단일 가독성 지수보다 다층 지표가 EFL 학습자의 독해 수행을 더 잘 예측함을 입증하여, 수능처럼 EFL 환경의 고부담 시험에 다층 지표를 적용할 이론적 근거를 제공한다.

이상의 선행연구들은 단일 지표 의존, 단기 관찰 기간, 담화·심리언어 지표 미포함이라는 공통 한계를 지닌다. 본 연구는 이를 극복하여 16개년 종단 데이터에 7개 다층 지표를 적용하고, 절대평가 전후 비교와 SAT Digital 비교를 통합하였다는 점에서 차별성을 갖는다.

III. 연구 방법

1. 분석 대상 및 자료 출처

분석 대상은 2010~2025학년도 수능 영어 고배점(3점) 읽기 지문 97편으로, 절대평가 이전(2010~2017학년도, n=36)과 이후(2018~2025학년도, n=61)로 구분하였다. 고배점 지문을 선정한 교육적 근거는 세 가지이다. 첫째, 수능 영어에서 등급 변별이 실질적으로 고배점 문항에 집중되므로 절대평가 취지와 의 정합성 문제를 가장 직접적으로 드러낸다. 둘째, 매 학년도 출제 수가 일정하여 종단 비교에 적합한 데이터를 형성한다. 셋째, 황이수와 이제영(2020)에서 담화 결속성과 정답률의 유의한 상관이 확인된 층위이다. 자료는 한국교육학술정보원(이하 KERIS)이 운영하는 에듀데이터서비스시스템(EDSS; <https://edss.neies.or.kr>)과 한국교육과정평가원(이하 KICE) 공개 기출 PDF를 활용하였다.

SAT 비교 표본은 College Board 공식 SAT Digital Practice Test(PT5·PT6·PT10; <https://satsuite.collegeboard.org>)에서 독해 지문 13편을 수집하였다. PT5·PT6는 2023년, PT10은 2024년에 공개된 자료이다. SAT Digital을 비교 준거로 선정한 이유는 College Board의 공식 자료로 신뢰성이 보장되며 문학(LIT, n=3), 역사·사회(HSS, n=5), 과학(SCI, n=5)의 장르 분포가 명시되어 있어 구조적 비교가 가능하기 때문이다. 다만 수능 고배점 지문과 SAT 공개 Practice Test 지문 간의 비교 설계 비대칭성은 연구의 한계로 VI-3절에 명시한다.

1등급 비율 데이터는 EDSS 공개 학교별 성적 자료(시행연도 기준 CSV파일)를 참조

하되, 학교별 단순평균이 전국 실제 비율을 과소추정하는 경향이 있어(예: 2018학년도 단순평균 5.99% vs. · KICE 공식 10.03%, 약 4%p 차이) 한국교육과정평가원(KICE)의 연도별 공식발표값을 최종 분석 기준값으로 사용하였다. KICE 공식값은 전국 응시자 전체 기준의 실제 등급비율이며, 학교별 단순평균보다 대표성이 높다.

2. 분석 지표 및 전처리

전처리는 Python 3.12(spacy 3.7)로 수행하였다. 전처리 절차는 ① HTML 태그 및 주석 제거, ② spacy ‘en_core_web_sm’ 파이프라인을 이용한 문장 분리(sentencizer), ③ 품사 태깅 및 표제어 추출(lemmatizer), ④ FK Grade 산출용 음절 수 산정(Carnegie Mellon Pronunciation Dictionary 기반), ⑤ MRC DB 매핑을 통한 상상가능성 점수 추출, ⑥ 명사화 접미사(-tion, -ness, -ment, -ity) 포함 어휘 추출의 6단계로 진행하였다.

본 연구에서 사용한 7개 지표는 두 범주로 구분된다. 어휘·구문 복잡도 범주(Category A)는 FK Grade와 MATTR로, 담화·심리언어 복잡도 범주(Category B)는 연결어 밀도, 상상가능성, 명사화 지수, 담화 결속성, CTCI로 구성된다. Category B 지표는 Graesser et al.(2004)의 Coh-Metrix 프레임워크를 참조하여 Python으로 직접 구현하였으며, 상상가능성은 MRC DB의 imageability 점수(100~700)를 지문 내용어 평균값으로 산정하였다. CTCI는 7개 지표를 min-max 정규화 후 방향성에 맞게 집계한 합성 지수(0~1 척도)이다. 각 지표의 정의 및 계산 방법은 표 1 및 표 2에 상세히 제시하였다.

<표 1> Category A 분석 변수 정의 및 계산 방법

변수명(약어)	계산식 / 측정 방법	해석 범위	본 연구 활용	CSAT 2023-25(n=15)
Flesch-Kincaid 학년수준 (FKGrade)	$FK = 0.39 \times (W/S) + 11.8 \times (Syl/W) - 15.59$ W=총토큰수, S=총문장수, Syl=총음절수	미국 학년 수준 FK10~12=고교 고학년 FK12~14=대학입문	Category A 주요 지표	54.0
이동평균 어형-토큰 비율 (MATTR)	$MATTR = (1/N) \times \sum TTR(wi \sim wi+49)$ 50토큰이동창을 1토큰씩 이동하며 각창의 TTR산출후 평균	0~1 척도; 높을수록 어휘 다양성 높음; 이동창방식으로 길이편향보정	Category A 보조 지표	20.0

주. W=총 어절(토큰) 수; S=총 문장 수; Syl=총 음절 수; TTR=Types/Tokens.

<표 2> Category B 분석 변수 정의 및 계산 방법

변수명(약어)	계산식 / 측정 방법	방향 (↑=어려움)	본 연구 활용	CSAT 2023-25 (n=15)
연결어 밀도 (Conn.Density)	인과(because, thus), 대조(however), 첨가(moreover)등 연결어 토큰수 ÷전체토큰수	↓ 어려움 (낮을수록 논리단서 적어 추론부담증가)	Category B 핵심 지표	54.0
상상가능성 (Imagery)	MRC Psycholinguistic Database (Paivioetal.,1968)기반 각내용어imageability점수 (100~700) 지문평균	↓ 어려움 (낮을수록 추상적, 이해부담증가)	Category B 핵심 지표	20.0
명사화 지수 (Nominalization)	-tion / -ness / -ment / -ity 접미사포함어휘수 ÷전체토큰수	↑ 어려움 (높을수록 명사화심화, 인지부담증가)	Category B 보조 지표	5.4
담화 결속성 (DiscourseCoherence)	지문 전체 의미적 일관성 종합 지수; 인접문장간내용어중첩·연결어· 명사참조등복합산출	↓ 어려움 (낮을수록 담화 흐름 불명확)	Category B 핵심 지표	0.32
종합 텍스트 복잡도지수 (CTCI)	연결어 밀도, 상상가능성, 명사화, 담화결속성, FKGrade,MATTR을 min-max정규화후 방향성 맞게집계; 0~1척도	↑ 어려움 (높을수록 전반적 복잡도 높음)	Category B 통합 지표	20.7

주: CTCI 산출 시 낮을수록 어려운 지표(연결어 밀도, 상상가능성, 담화 결속성, MATTR)는 1-x' 로 방향 반전 후 합산.

어휘 층위 특성의 보완적 분석을 위해서 CEFR(Common European Framework of Reference for Languages; Council of Europe, 2001) 어휘 분류 체계를 적용하였다. CEFR은 언어 능력을 A1·A2(기초), B1·B2(자립), C1·C2(숙달)의 6단계로 구분하며, 본 연구는 Cambridge Learner Corpus(4,200만 토큰)를 기반으로 각 어휘의 CEFR 습득 단계를 실증적으로 규명한 Cambridge English Vocabulary Profile(EVP; Capel, 2012)을 어휘 분류 준거로 활용하였다. 미등재(UNK) 어휘—고유명사·전문술어·약어—는 별도 분류하였다. 분석 결과는 표 3에 제시하였다.

<표 3> CEFR 레벨별 어휘 출현 비율 기간별 비교 (% , 평균)

지표	CSAT 2010-12 (n=7)	CSAT 2013-17 (n=29)	CSAT 2018-22 (n=22)	CSAT 2023-25 (n=15)	SAT Digital (n=13)
A1+A2 합계(%)	53.8	54.6	52.6	54.0	43.9
B1 어휘(%)	21.8	20.8	21.6	20.0	28.6
B2 어휘(%)	4.9	4.7	5.0	5.4	7.3
C1+C2 합계(%)	0.10	0.14	0.12	0.32	0.23
미등재(UNK, %)	19.2	19.7	20.7	20.7	18.7
유의성(CSAT vs. SAT)	—	—	—	—	A1+A2: p<.001*** C1+C2:ns

주: CEFR 레벨별 분류 준거: Cambridge EVP(Capel, 2012). 수능 출현 비율은 기간별 단순 평균값. *** p<.001, ns=not significant; Mann-Whitney U, 양측 검정.

표 3에서 확인되듯, 수능 고배점 지문은 A1+A2(기초) 어휘 비율이 절대평가 이전부터 일관되게 50% 이상을 유지하며 SAT Digital(43.9%)보다 유의미하게 높은 반면 ($p<.001$), C1+C2(학술) 어휘 비율은 매우 낮다. 이는 수능 고배점 지문의 어려움이 어휘 수준보다 담화 구조와 추상성에서 비롯됨을 보완적으로 뒷받침한다.

3. 통계 분석

통계 분석은 다음과 같이 연구 문제별로 설계하였다.

RQ 1(종단 추세): 연도별 평균값($n=16$ 개 연도)에 Spearman 순위상관(ρ)과 Mann-Kendall 추세 검정을 병행하였다.

RQ 2(절대평가 전후 비교): 비정규 분포 데이터에 Mann-Whitney U 검정을 적용하고 효과크기 r 을 산출하였다[소($r\geq.10$), 중($r\geq.30$), 대($r\geq.50$); Cohen, 1988].

RQ 3(복잡도-1등급 비율 상관): 연도별 집계값($n=16$)에 7개 지표 각각과 1등급 비율 간 Spearman ρ 를 산출하였다.

RQ 4(수능-SAT 비교): Mann-Whitney U 검정과 사후 검정력 분석(G*Power)을 병행하였다.

모든 분석에서 유의수준 $\alpha=.05$ (양측)를 적용하였다. CTCI 가중치 가정의 강건성 확인을 위해 대안 가중치 모형 두 가지에 대한 민감도 분석을 병행하였으며, 코로나 19 대응 출제로 이상값을 형성한 2020학년도를 제외한 민감도 분석을 추가로 수행하였다.

IV. 연구 결과

1. 분석 대상 지문 기술통계

분석 대상 97편의 기술통계는 표 4에 제시하였다. 전반적으로 수능 고배점 지문은 미국 고등학교 후기~대학 입문 수준의 가독성을 나타내었으며(전체 평균 FK Grade 11.47), 절대평가 전환 이후 지문의 가독성 수준이 약 1.5 학년 상승하였다(절대평가 이전 M=10.57 → 이후 M=12.04). 연도 간 편차도 확대되어(SD: 2.10→3.49), 지문 복잡도의 연도별 일관성이 저하된 것으로 나타났다. 특히 지문의 구체성 지표인 상상가능성이 절대평가 이전(M=347.2) 대비 이후(M=272.4)에 약 74.8점 감소하여 지문이 더욱 추상적으로 변화하였다. SAT Digital은 FK Grade(M=16.79)가 수능보다 높지만 지문 길이가 짧고(M=91어절), 담화 결속성(M=0.060)이 수능 절대평가 이후 평균(M=0.035)의 약 1.7배에 달하였다.

<표 4> 분석대상 기술통계 (n=97 + SAT n=13)

구분	n	M 토큰	[A]FK Grade M(SD)	[A]MATTRM(SD)	[B]연결어 밀도 M	[B]상상 가능성 M	[B]CTCI M(SD)
수능 (CSAT) 전체	97	152.3	11.47 (3.49)	0.763 (0.196)	0.023	300.2	0.680 (0.074)
절대평가 전 (2010-2017)	36	165.2	10.57 (2.10)	0.799 (0.045)	0.028	347.2	0.655 (0.077)
절대평가 후 (2018-2025)	61	174.1	12.04 (3.49)	0.742 (0.230)	0.021	272.4	0.695 (0.072)
SAT Digital	13	91.0	16.79 (1.92)	0.817 (0.048)	0.018	292.0	0.702 (0.058)

주: 표 4의 [A] = Category A(어휘·구문 복잡도), [B] = Category B(담화·심리언어 복잡도). FK Grade = Flesch-Kincaid 학년 수준; MATTR=Moving -Average Type-Token Ratio; CTCI = 종합 텍스트 복잡도 지수. M 토큰 = 지문당 평균 어절 수(기술통계 참고용; 복잡도 지표 산출에 미사용).

2. 종단적 추세 분석

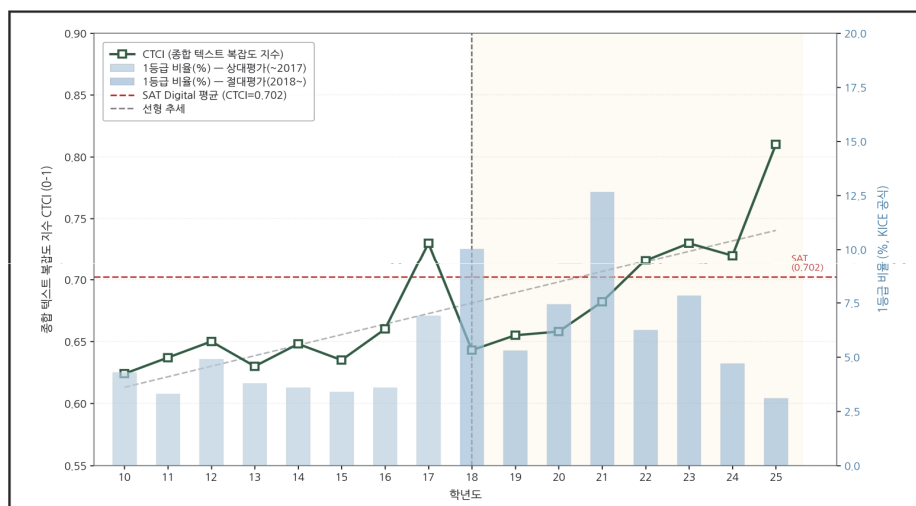
16개년에 걸쳐 수능 영어 고배점 지문은 전반적인 복잡도 심화 경향을 보였다. 교육적으로 가장 주목할 만한 변화는 두 가지이다. 첫째, 지문의 전반적 텍스트 복잡도(CTCI)가 16개년 동안 가장 강건한 상승 추세를 보여 수능 지문이 지속적으로 복잡해지고 있음을 실증하였다($\rho=+.831$, $p<.001$, 표 2). 이는 절대평가 도입 이후에도 지문

복잡도가 안정화되지 않고 있음을 의미한다. 둘째, 지문의 논리적 단서(연결어 밀도, $\rho = -.566$, $p = .033$)와 구체성(상상가능성, $\rho = -.608$, $p = .010$)이 유의미하게 감소하였다. 이는 수능 영어 지문이 연결어 없이 추론해야 하고 구체적 이미지 형성도 어려운 추상적 텍스트로 변해가고 있음을 의미하며, 고등학교 영어 교육에서 목표로 하는 실용적 독해 능력과 수능이 요구하는 독해 전략 사이의 거리가 구조적으로 확대되고 있음을 시사한다.

<표 5> 종단적 추세 분석 결과(Spearman + Mann-Kendall, n=16개 연도)

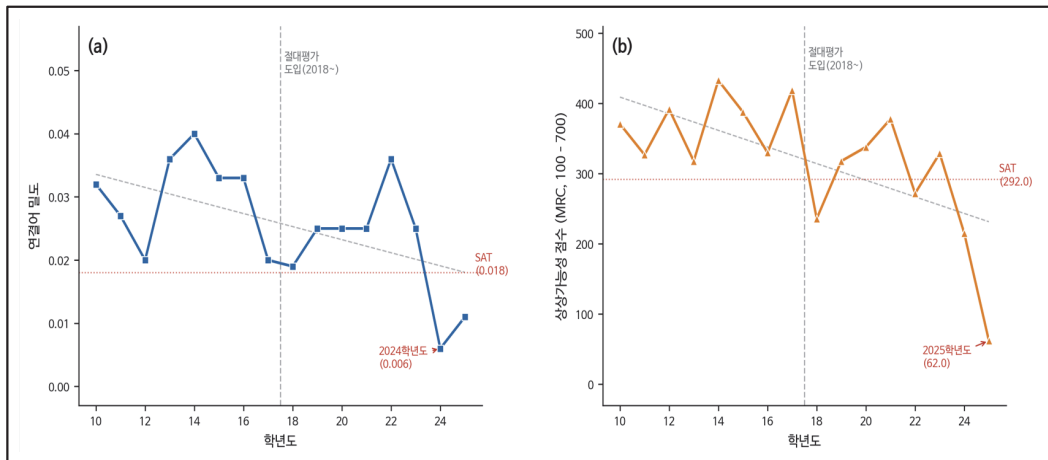
지표	Spearman ρ	p 값	M-K 방향	교육적 해석
FK Grade	+0.404	ns	상승 경향	표면 가독성 상승-연도별 분산 확대
MATTR	-0.125	ns	추세 없음	어휘 다양성 안정- 변별은 다른 층위에서
연결어 밀도	-0.566	*	감소 ★	논리 단서 감소-추론 부담 증가
상상가능성	-0.608	**	감소 ★★	추상화 가속-구체적 독해 단서 감소
명사화 지수	+0.228	ns	추세 없음	단독 추세 미확인
담화 결속성	-0.412	ns	감소 경향	결속 약화 경향 (유의성 미달)
CTCI	+0.831	***	상승 ★★★	16개년 최강 상승-지문 전반 복잡화 실증

주: M-K = Mann-Kendall; CTCI=종합 텍스트 복잡도 지수. * $p < .05$. ** $p < .01$. *** $p < .001$. ns = not significant.



주: 수직 파선 = 절대평가 도입(2018학년도~); 점선(회색) = 선형 추세; 점선(빨강) = SAT Digital 평균 (0.702); KICE 공식 1등급 비율 기준. † 2014학년도는 A·B형 원자료 가중 합산값. *** $p < .001$.

[그림 1] 종합 텍스트 복잡도 지수(CTCI) 및 1등급 비율 추이(2010학년도-2025학년도)



주: 실선 = 연도별 측정값; 점선(회색) = 선형 추세; 점선(빨강) = SAT Digital 평균. * $p < .05$. ** $p < .01$.

[그림 2] 연결어 밀도(a) 및 상상가능성(b) 추이(2010학년도-2025학년도)

그림 2는 연결어 밀도(좌)와 상상가능성(우)의 연도별 추이를 나란히 제시한다. 연결어 밀도는 절대평가 이후 지속적으로 낮아지며, 특히 2024학년도(0.006)와 2025학년도(0.011)에 급격히 감소하여 SAT Digital 평균(0.018)보다도 낮은 수준에 이르렀다. 상상가능성은 2025학년도($M=62.0$)가 MRC 척도 하한에 근접한 역대 최저 수준을 기록하였다. 두 지표 모두 절대평가 이후 SAT Digital 평균보다 낮은 값을 보인다.

3. 절대평가 전후 Mann-Whitney U 검정

절대평가 전환 이후 수능 영어 고배점 지문은 복잡도가 높아지는 동시에 독해 단서는 줄어들며, 교육적으로 우려되는 이중 변화를 보였다. 지문의 가독성 수준은 유의미하게 상승하였으며, FK Grade는 절대평가 이전 대비 약 1.5 학년 증가하였다($U=599$, $p=.001$, 효과크기 $r=.41$, 중간 효과) 이는 중간 수준의 실질적 효과에 해당하며(Cohen, 1988), 지문의 표면적 어휘와 구문 에 대한 부담이 절대평가 도입 이후 뚜렷이 심화되었음을 의미한다. CTCI 또한 통계적으로 유의미하게 증가하였다($U=745$, $p=.009$, $r=.32$, 소-중간 효과). 이와 동시에 연결어 밀도($U=1368$, $p=.044$, $r=.25$)와 상상가능성($U=1372$, $p=.041$, $r=.25$)이 유의미하게 감소하였다. 이는 절대평가 이후 지문이 더 어렵고 동시에 더 추상적으로 변화하였음을 실증하며, 매년 일관된 성취 수준 기준을 유지해야 한다는 절대평가의 취지에 부합하지 않는 방향의 변화임을 나타낸다. MATTR ($p=.864$)과 명사화 지수($p=.774$)는 유의한 변화를 보이지 않아, 변별 기제의 변화가 어휘 목록이 아닌 담화 구조 층위에서 이루어지고 있음을 시사한다.

<표 6> 절대평가 전후 텍스트 복잡도 비교(Mann-Whitney U)

지표	절대평가 전 M (n=36)	절대평가 후 M (n=61)	차이 (후-전)	U	p값	효과크기 r
FK Grade	10.57	12.04	+1.47	599	.001 **	+.41(중간)
MATTR	0.799	0.742	-0.058	1122	.864 ns	-.02(소 미만)
연결어 밀도	0.028	0.021	-0.007	1368	.044 *	-.25(소-중간)
상상가능성	347.2	272.4	-74.8	1372	.041 *	-.25(소-중간)
명사화 지수	0.044	0.043	-0.001	1137	.774 ns	-.04(소 미만)
담화 결속성	0.043	0.035	-0.008	1258	.235 ns	-.15(소 미만)
CTCI	0.655	0.695	+0.040	745	.009 **	+.321

주: CTCI = 종합 텍스트 복잡도 지수 절대평가 = 절대평가, 효과크기 r: 소 $\geq .10$, 중 $\geq .30$, 대 $\geq .50$ (Cohen, 1988). * $p < .05$. ** $p < .01$.

4. 텍스트 복잡도와 1등급 비율의 순서 상관

단일 지표와 다층 지표는 1등급 비율과 정반대 방향의 상관을 나타내었다. FK Grade는 1등급 비율과 반직관적인 양의 상관을 보인 반면($\rho = +.743$, $p = .002$), 다층 복잡도를 종합한 CTCI는 유의미한 음의 상관을 나타내었다($\rho = -.501$, $p = .048$). 이 상반된 결과는 단일 가독성 지표(FK Grade)와 다층 복잡도 지표(CTCI)가 서로 다른 측면을 측정하고 있으며, 단일 지표만으로 체감 난이도를 판단하는 것이 얼마나 위험한지를 실증한다. 7개 지표 전체와 1등급 비율 간 순서 상관은 표 7에 제시하였다.

<표 7> 7개 복잡도 지표와 1등급 비율 간 Spearman 순위상관 (n=16개 연도)

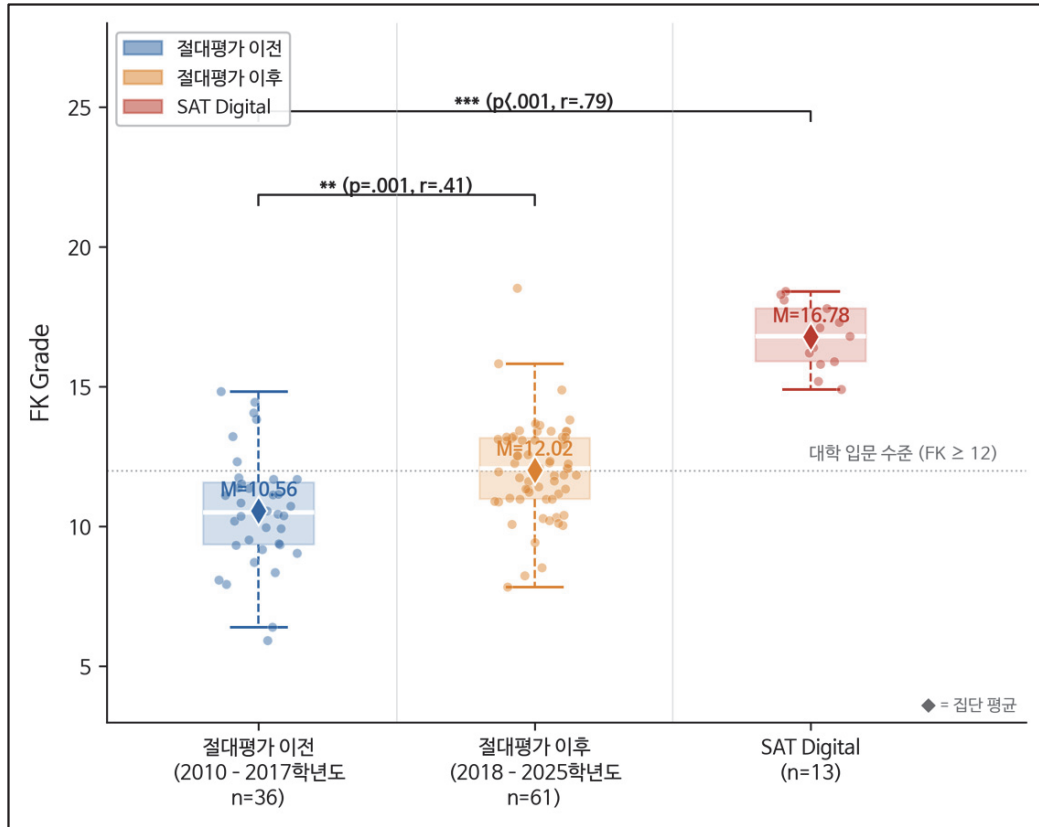
지표	Spearman ρ	p값	교육적 해석
FK Grade	+.743	.002 **	반직관적 양적 상관—단일 지표의 한계
MATTR	+.312	.239 ns	유의하지 않음
연결어 밀도	-.646	<.05 * (2020 제외 시)	2020 제외 시 유의한 음적 상관 회복
상상가능성	+.289	.279 ns	유의하지 않음
명사화 지수	-.198	.464 ns	유의하지 않음
담화 결속성	+.341	.196 ns	유의하지 않음
CTCI	-.501	.048 *	유의한 음적 상관—복잡도 ↑ → 1등급 ↓

※ 2020학년도(코로나19 이상값) 제외 시 연결어 밀도 $\rho = -.646$ ($p < .05$) 회복

주: CTCI = 종합 텍스트 복잡도 지수. * $p < .05$. ** $p < .01$.

2020학년도를 제외한 민감도 분석에서 연결어 밀도가 1등급 비율과 유의미한 음의 상관을 회복하였으며($\rho = -.646$, $p < .05$) 2025학년도 FK Grade(12.38)가 2020학년도(12.87)보다 낮음에도 1등급 비율이 역대 최저(3.11%)를 기록하였는데, 이는 담화 결

속성(0.031)과 상상가능성(M=62.0) 급락이 체감 난이도를 주도하였기 때문으로 해석된다.



주: ◆ = 집단 평균; 박스 = IQR(25~75 백분위); 수염 = $1.5 \times \text{IQR}$; 각 점 = 개별 지문. ** $p < .01$. *** $p < .001$ (Mann-Whitney U).

[그림 3] FK Grade 집단별 분포-Box+Jitter(절대평가 전, 후, SAT Digital)

5. 수능(CSAT) vs. SAT Digital 담화-심리언어 복잡도 비교

수능과 SAT Digital을 비교한 결과, 두 시험은 담화 결속성에서만 교육적으로 의미 있는 구조적 차이를 나타내었다. SAT Digital 지문의 담화 결속성은 수능보다 약 58% 높아($p=.021$, $r=.397$, 중간 효과), SAT 지문의 논리 흐름이 더 명시적으로 제시됨이 확인되었다. 사후 검정력 분석 결과($\text{power}=.986$)는 이 발견이 SAT 표본($n=13$)의 소규모와 무관하게 신뢰할 수 있음을 보여준다. FK Grade는 두 시험 간 큰 차이를 보였으나 ($p<.001$), 이는 지문 길이와 장르 구성의 차이를 반영하는 것으로 난이도 직접 비교에는 주의가 필요하다.

<표 8> 수능 vs. SAT Digital 텍스트 복잡도 비교(Mann-Whitney U)

변수	증감(난도)	수능 전체	SAT (n=13)	p값	효과크기 r
담화 결속성	↓ 어려움	0.038	0.060	.021 *	+.397
연결어 밀도	↓ 어려움	0.023	0.018	.211 ns	-.215
상상가능성	↓ 어려움	300.2	292.0	.791 ns	+.046
명사화 지수	↑ 어려움	0.043	0.056	.172 ns	+.235
CTCI	↑ 어려움	0.680	0.702	.429 ns	+.136
FK Grade †	↑ 어려움	11.47	16.79	<.001 ***	+.791

주: † 지문 길이 차이(수능 M = 152어 vs. SAT M = 91어) 반영 — 난이도 직접 비교 주의. ↓ 어려움=수치가 낮을수록 어렵고, ↑ 어려움=수치가 높을수록 어려움. * $p < .05$. *** $p < .001$.

6. 종합 텍스트 복잡도 지수 가중치 민감도 분석

CTCI는 6개 변수에 동일 가중치($w=1/6$)를 부여하여 산출하였다. 두 가지 대안 가중치 모형과 주요 결론을 비교한 결과는 표 9에 제시하였다. 세 가중치 모형 모두에서 CTCI의 절대평가 이후 유의한 증가(+0.040~+0.041)와 강건한 종단 상승 추세($\rho = +.812 \sim +.845$)가 일관되게 확인되어, 가중치 선택에 대한 방법론적 민감도가 낮음을 입증한다.

<표 9> 가중치 민감도 분석: CTCI 추세 및 절대평가 전후 비교

가중치 모형	절대평가 전 M	절대평가 후 M	차이	Spearman ρ (시계열)	p값(절대평가 전후)
동일 가중치 (기본 모형: $w=1/6$)	0.655	0.695	+0.040	+.831	.009 **
모형 A: Cat.A 1/3, Cat.B 1/6×4	0.648	0.688	+0.040	+.812	.012 *
모형 B: Cat.A 1/6, Cat.B 5/24×4	0.659	0.700	+0.041	+.845	.007 **

주: CTCI = 종합 텍스트 복잡도 지수. 모든 모형에서 종단 상승 및 절대평가 이후 유의 증가 결론 유지됨.

* $p < .05$. ** $p < .01$.

V. 논의 및 결론

본 연구는 2010~2025학년도 수능 영어 고배점 지문 97편의 텍스트 복잡도를 7개 다층 지표로 종단 분석하고, SAT Digital 지문과의 비교를 통해 외적 타당성을 검토하였다. 이하에서는 네 가지 핵심 분석 결과를 이론적·교육적 맥락에서 논의하고, 이를 토대로 정책적 제언과 연구의 한계를 제시한다.

1. 종합 텍스트 복잡도(CTCI) 상승과 추상화 가속

CTCI는 16개년 동안 가장 강건한 상승 추세를 보이며($\rho = +.831$, $p < .001$), 상상가능성($\rho = -.608$)과 연결어 밀도($\rho = -.566$)가 유의미하게 감소하였다.

이러한 추세는 학술 논문 유형의 지문 출처 비중 확대와, 원문 편집 과정에서 연결어를 의도적으로 삭제하거나 구체적 서술을 추상 명사구로 변환하는 관행에서 비롯된 것으로 추정된다. 2025학년도 상상가능성 평균($M=62.0$)이 MRC DB 척도 하한(100)에 근접하는 역대 최저치를 기록한 것은 이 편집 관행이 극대화되고 있음을 시사한다.

이 결과는 김미란(2022)이 보고한 가독성 불균질성을 16개년 종단 추세로 확장하여 복잡도 심화가 구조적 현상임을 입증한다. Paivio(1986)의 이중부호화 이론에 따르면 추상 텍스트는 심상 형성이 어려워 인지 부담이 가중되며, EFL 학습자는 이를 모국어 화자보다 더 강하게 경험한다(Crossley et al., 2011). 특히 McNamara et al.(2010)은 텍스트 복잡도가 어휘·구문 층위만이 아닌 담화·심리언어 층위의 다층 구조로 결정됨을 강조한 바 있으며, 본 연구의 종단 결과는 수능 고배점 지문이 바로 그 담화 층위에서 구조적으로 심화되고 있음을 16개년 자료로 최초 실증한다.

교육 현장에서 이 결과가 갖는 함의는 명확하다. 현재 고등학교 영어Ⅱ 교과서의 FK Grade 범위(8.45~11.05)와 수능 고배점 지문의 평균(절대평가 이후 $M=12.04$) 사이의 간극은 교과서 수업만으로는 수능 지문 대비가 어려운 구조를 만들고 있다. 교사는 교육 과정에 따라 교과서를 중심으로 수업을 설계하지만, 현재 수능 고배점 지문은 교과서 수준을 최대 5학년 초과하고 있고, 이로 인해 발생한 간극을 학교 수업에서 메우려면 교사가 교육과정 이탈을 감수해야하는 구조적 딜레마가 발생한다. 나아가 연결어 생략과 추상화 심화는 교사가 수업에서 지도하기 어려운 ‘수능 전용 독해 전략’을 요구하게 되어 사교육 의존도를 심화시키는 요인이 될 수도 있다. 어휘 다양성(MATTR)이 안정적으로 유지된 채 담화 층위 지표만이 선택적으로 악화된 패턴은, 출제 가이드라인에서 담화 구조 복잡도를 명시적으로 관리하지 않음을 시사하며 가이드라인 보완의 시급성을 뒷받침한다.

가중치 민감도 분석에서 결론이 가중치 설정과 무관하게 유지되어 결과의 강건성이 확인되었다.

2. 절대평가 취지와 구조적 괴리

절대평가 이후 FK Grade($p=.001$)와 CTCI($p=.009$)가 상승하고, 연결어 밀도($p=.044$)와 상상가능성($p=.041$)이 감소하였으며, 어휘 다양성(MATTR)은 안정적으로 유지되었다($p=.864$). 이 패턴은 중요한 출제 경향을 시사한다. 어휘 목록 자체를 유지하면서 담화 구조(연결어 삭제, 추상화)를 변별 기제로 활용하는 방향으로의 전환이다. 장지연(2019)은 절대평가 도입 전후 통사적 복잡성 변화를 보고하였으나 담화 층위 지표는 다루지 않았다. 본 연구는 그 공백을 채워 통사·어휘 층위에서 안정적으로 보이는 지문이 담화 층위에서는 유의미하게 복잡해지고 있음을 실증한다.

수능 영어를 준비하는 교사와 학생 입장에서, 어휘 수준 기반 학습만으로는 담화적 복잡도에 대응하기 어렵다. 특히 추상적 텍스트 독해 능력은 풍부한 독서 경험이나 사교육을 통해 계발되는 경향이 있어, 수능 지문의 상상가능성이 급락하고 연결어가 줄어들수록 사교육 접근성이 낮은 학생이 불리해지는 구조적 불평등 문제가 발생할 수 있다(황이수·이제영, 2020). 대부분의 학생들이 어휘와 구문 중심으로 영어를 학습하지만, 수능 고배점 지문에서 실제로 요구하는 능력은 연결어 없이 문장 간 논리를 추론하고 추상 명사구의 의미를 문맥에서 재구성하는 것이다. 이처럼 학습 방식과 시험 요구 방식 사이의 근본적 차이가 수험생의 체감 난이도를 높이는 결과로 이어진다. EFL 학습자로서 학생들은 모국어 화자에게도 어려움을 추상 텍스트를 L2로 추론해야 하는 이중 인지 부담을 안고 있으며(Crossley et al., 2011), 이에 대한 교육적 논의가 더욱 필요하다. 따라서 담화 구조 독해 전략을 명시적으로 지도하는 교수·학습 방안을 교육과정 개선 논의에 포함시키는 것이 요청된다.

교육부(2017)의 절대평가 원칙(‘점수에 따른 성취 수준별 등급 제공’)에 비추어, 특정 복잡도 지표만이 선택적으로 상승하는 현상은 출제 가이드라인의 공백에서 비롯된 것으로 해석된다. 절대평가 도입의 본래 취지를 실현하기 위해서는, 어휘·구문 층위만이 아니라 담화 층위를 포괄하는 복합적 지표 가이드라인이 제도적으로 기능해야 할 필요성이 제기된다.

3. FK Grade의 반직관적 상관: 단일 지표의 한계

FK Grade는 1등급 비율과 반직관적인 양의 상관을 보인 반면($\rho=+.743$, $p=.002$), 다층 복잡도를 종합한 CTCI는 유의미한 음의 상관을 나타내었다($\rho=-.501$, $p=.048$). FK Grade가 반직관적 결과를 보이는 이유는, 이 지수가 문장 길이와 음절 수만을 측정하므로 연결어 생략 등 담화 층위 복잡도를 포착하지 못하기 때문이다(Kincaid et al.,

1975). 황이수와 이제영(2020)은 담화 결속성이 체감 난이도를 더 잘 반영한다는 본 연구의 결과와 맥락을 같이하며, Crossley et al.(2011) 또한 단일 가독성 지수보다 다층 지표가 EFL 독해 수행을 더 잘 예측함을 실증한 바 있어 본 연구의 방향성을 뒷받침한다.

출제 검토 과정에서 FK Grade만을 근거로 지문 난이도를 판단하는 것은 구조적 오류를 내포한다. FK Grade를 다층 지표의 하나로만 활용하고, 담화 결속성·연결어 밀도·상상가능성을 함께 검토하는 다층 검토 체계가 마련되어야 한다.

4. CSAT-SAT 담화 결속성 차이: 출제 철학의 구조적 차이

SAT Digital 지문의 담화 결속성이 수능보다 약 58% 높아($p=.021$, $r=.397$, 중간 효과), SAT 지문의 논리 흐름이 더 명시적으로 제시됨이 확인되었다. 사후 검정력 분석 결과($\text{power}=.986$)는 이 결과가 SAT 표본($n=13$)의 소규모에도 불구하고 신뢰할 수 있음을 보여준다. 이 차이는 두 시험의 출제 철학의 차이를 반영한다. SAT는 College Board의 Evidence-Based Reading 원칙에 따라 텍스트 논리 흐름을 명시적으로 제시하는 반면, 수능은 연결어 생략을 통해 독자의 추론을 유도하는 방향을 취한다.

수능 영어에만 특화된 국내 고등학생들은 ‘연결어 없이 추론하는 전략’에 치우쳐 담화 흐름이 명시적인 텍스트 독해 경험이 부족할 수 있으며, 이는 대학 입학 후 영어 원서 독해와의 간극을 야기할 수 있다는 점에서 교육적으로 주목할 필요가 있다. SAT 비교는 탐색적 목적에서 시작되었으나, 담화 결속성이라는 단일 발견만으로도 수능 출제 방향 재검토의 근거를 제공한다.

이상의 논의를 종합하면 다음과 같다. 첫째, 수능 영어 고배점 지문은 지난 16년간 단순히 ‘어려워진’ 것이 아니라 ‘추상화’되는 방향으로 구조적으로 변화하였으며, 이는 교과서-수능 간 간극 확대와 사교육 의존도 심화로 이어지는 교육적 문제를 안고 있다. 둘째, 절대평가 전환 이후에도 담화 층위 복잡도가 지속 상승하여 절대평가의 교육적 취지와 상반되는 방향의 출제 경향이 확인되었으며, 이는 출제 가이드라인의 공백을 시사한다. 셋째, 단일 지표(FK Grade)는 1등급 비율과 반직관적 상관을 보인 반면 다층 지표(CTCI)는 유의미한 음의 상관을 나타내어, 단일 지표 의존 방식의 근본적 한계가 실증되었다. 넷째, 수능과 SAT Digital은 담화 결속성에서만 유의미하게 달랐으며, 이는 두 시험 간 출제 철학의 구조적 차이를 반영한다. 이러한 결과는 다층 텍스트 복잡도 지표에 기반한 출제 가이드라인의 제도적 정비와 시급함을 공통적으로 시사한다.

5. 정책적 제언

위의 논의 결과를 바탕으로 다음의 세 가지 정책적 제언을 제시한다.

가) 다층 가독성 지수(MRI) 출제 가이드라인 도입

한국교육과정평가원(KICE)에서 수능 영어 출제 과정의 지문 선정 단계에 다층 가독성 지수(MRI)를 기반으로 한 텍스트 복잡도 사전 검토 절차를 의무화할 것을 제안한다. 구체적으로 FK Grade, CTCL, 연결어 밀도, 상상가능성, 담화 결속성 5개 지표를 동시 검토하는 체크리스트를 출제 초안 단계에 도입하고, 권고 기준(안)으로 FK Grade 상한 13.0, CTCL 상한 0.72, 연결어 밀도 하한 0.025, 상상가능성 하한 280, 담화 결속성 하한 0.045를 제안한다. 이 기준값은 절대평가 이전(2010~2017학년도) 수능 고배점 지문 평균치를 참조한 것으로, Crossley et al.(2011)이 제안한 다층 접근의 이론적 틀과도 부합하며, 절대평가 취지인 성취 수준 일관성 유지와 직결된다. 단일 지표 의존에서 벗어나 균형 잡힌 복잡도 관리를 가능하게 한다는 점에서 이 제도적 보완은 교육 공정성 확보 차원에서도 중요하다.

나) 원문 충실성(Source Fidelity) 원칙 명문화

KICE는 변별력 확보를 위해 연결어를 의도적으로 삭제하거나 구체적 서술을 추상 명사구로 변환하는 편집 관행을 제한하는 원문 충실성 원칙을 출제 지침에 명문화할 것을 제안한다. 이 원칙은 지문이 출처 텍스트의 자연스러운 담화 구조를 유지하는 범위 안에서 교육적으로 적절한 복잡도를 갖도록 보장하는 기준으로 기능할 것이며, 교과서 수업과 수능 지문 간의 질적 간극을 줄이는 데도 기여할 것이다.

다) EDSS 기반 연도별 복잡도 공개 모니터링 체계 구축

KICE와 한국교육학술정보원(KERIS)이 협력하여 에듀데이터서비스시스템(EDSS) 공동데이터를 활용한 연도별 텍스트 복잡도 추적 시스템을 구축하고, 매 수능 직후 주요 복잡도 지표를 공개할 것을 제안한다. 이 체계는 연구자·교육 당국·교사가 동일한 데이터를 기반으로 수능 영어의 복잡도 변화를 모니터링하고 교육정책을 공개적으로 논의할 수 있는 토대를 마련한다. 정보 공개를 통한 출제 책무성 강화는 절대평가 신뢰도 회복을 위한 제도적 기반이 될 것이다.

6. 연구의 한계 및 후속 연구 과제

본 연구의 한계는 다음과 같다. 첫째, Coh-Metrix 원도구 대신 Python으로 직접 구현한 지표를 사용하였으므로 TAACO·TAALES 등 공인 도구와의 교차검증이 필요하다. 둘째, SAT 표본이 소규모($n=13$)이며 수능 고배점 지문과의 비교 설계 비대칭성이 있다. 셋째, 본 연구는 지문 텍스트의 복잡도에 한정하였으며, 선지 구성·추론 유형·배점 구조 등 문항 층위 요인은 분석 범위에 포함되지 않았다. 지문 복잡도와 1등급 비율의 비선형 관계를 완전히 설명하기 위해서는 문항 층위 분석을 통합하는 후속 연구가 요청된다. 후속 연구 과제로는 SAT 표본 확대, 문항 유형별 비교, 교사·학생 대상 체감 난이도 조사와의 통합 분석을 제안한다.

사사: 본 연구는 한국교육학술정보원(KERIS) 에듀데이터서비스(EDSS) 공공데이터를 활용하였음을 밝힙니다. 본 논문의 작성 과정에서 Anthropic의 Claude(버전: Claude Sonnet 4.6, 활용 기간: 2026년 7~8월)를 표현 다듬기, 문헌 형식 확인 등의 보조 목적으로 활용하였다. 연구 설계, 데이터 수집·분석, 해석 및 결론 도출은 연구자가 직접 수행하였으며, AI 생성 내용의 정확성과 최종 내용에 대한 책임은 저자에게 있다.

참고문헌

- 교육부(2017). **2018학년도 수능 영어 절대평가 시행 기본계획**. 서울: 교육부.
- 김미란(2022). 대학수학능력시험 영어 독해지문의 어휘다양성 및 가독성 분석. **외국어 교육연구**, 36(4), 71-90.
- 사교육걱정없는세상(2025). 2026학년도 수능 영어 독해지문 공동분석 결과 발표. 국회 소
통관 보도자료. 자료출처(검색일 2026.4.27.): ([https://noworry.kr/policyarchive/?
bmode=view&idx=168974628](https://noworry.kr/policyarchive/?bmode=view&idx=168974628))
- 이혜진, 이제영(2018). 한·중·일 대학 입학시험 영어 지문의 이독성 비교 분석. **인문사회**
21, 9(5), 441-454.
- 장지연(2019). 절대평가 도입 전후 수능 영어 독해 지문의 통사적 복잡성 비교. **영어교과**
교육, 18(2), 161-188.
- 최민주, 김정렬(2017). 수능 영어 문항 유형간 응집력과 어휘정보 분석. **한국콘텐츠학회**
논문지, 17(12), 378-385.
- 한국교육과정평가원(2010-2025). **대학수학능력시험 영어 영역 기출 문항** [EDSS 공공
데이터]. <https://edss.neies.or.kr>(검색일: 2026.4.24.)
- 황이수, 이제영(2020). 수학능력시험 영어 읽기 지문의 응집성과 문항 난이도 간의 상관관계
분석. **한국콘텐츠학회 논문지**, 20(5), 344-350.
- Capel, A. (2012). A1-B2 vocabulary: Insights and issues arising from the English
Profile Wordlists project. *English Profile Journal*, 3, e4.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.).
Hillsdale, NJ: Lawrence Erlbaum Associates.
- College Board. (2023-2024). *SAT Digital Practice Tests* (PT5, PT6, PT10).
<https://satsuite.collegeboard.org/sat/practice-preparation/practice-tests>(검색일:
2026.4.24.)
- Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of*
Experimental Psychology, 33(A), 497-505.
- Council of Europe. (2001). *Common European Framework of Reference for Languages:*
Learning, Teaching, Assessment. Cambridge: Cambridge University Press.
- Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The

- moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94-100.
- Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1), 84-101.
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193-202.
- Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python [Computer software, Version 3.7]. <https://doi.org/10.5281/zenodo.1212303>
- Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas for Navy enlisted personnel* (Research Branch Report 8-75). Millington, TN: Naval Technical Training Command.
- McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, 27(1), 57-86.
- Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. Oxford: Oxford University Press.
- Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76(1, Pt. 2), 1-25.

논문접수 : 2026.7.2. / 수정본접수 : 2026.7.30. / 게재승인 : 2026.8.11.

ABSTRACT

Longitudinal Text Complexity Analysis of High-Point CSAT English Passages: Pre- and Post-Absolute Grading Comparison and Multi-Dimensional Textual Contrasts with U.S. SAT Passages

Haeyoung Lee

Teacher, Sangin High School

This study examines longitudinal changes in text complexity across 97 high-point (3-point) reading passages from the Korean College Scholastic Ability Test (CSAT) English section spanning 2010 to 2025. Seven indicators were applied: Flesch-Kincaid Grade Level (FK Grade), Moving-Average Type-Token Ratio (MATTR), Connective Density, Imagery Score, Nominalization Index, Discourse Coherence, and a Composite Text Complexity Index (CTCI). A subset of 13 SAT Digital passages was then analyzed using identical metrics. The findings reveal that CSAT high-point passages have not simply become ‘more difficult’ but have undergone structural ‘abstraction’ over 16 years. CTCI showed the strongest upward trend ($\rho=+.831$, $p<.001$), while Imagery Score ($\rho=-.608$, $p=.010$) and Connective Density ($\rho=-.566$, $p=.033$) declined significantly. Since the 2018 transition to absolute grading, FK Grade ($p=.001$, $r=.41$) and CTCI ($p=.009$, $r=.32$) rose further, even as Connective Density ($p=.044$) and Imagery Score ($p=.041$) fell—a direction contrary to the educational intent of absolute grading. FK Grade correlated positively with Grade-1 pass rates ($\rho=+.743$, $p=.002$), while CTCI showed a significant negative correlation ($\rho=-.501$, $p=.048$), demonstrating the limitations of single-metric approaches. In the CSAT-SAT comparison, only Discourse Coherence differed significantly ($p=.021$, $r=.397$), with SAT passages exhibiting markedly higher coherence. These results call for integration of a Multi-dimensional Readability Index (MRI) into test-development guidelines and an EDSS-based annual monitoring system.

Key Words: CSAT English, text complexity, FK Grade, Coh-Metrix, absolute grading, SAT comparison, longitudinal study

부 록(Appendices)

Appendix A. CEFR 어휘 레벨 분류 체계 및 어휘 선정 기준

CEFR(Common European Framework of Reference for Languages)은 유럽평의회(Council of Europe, 2001)가 개발한 언어 능력의 국제 공통 기준으로, A1·A2(기초), B1·B2(자립), C1·C2(숙달)의 6단계로 구분한다. 본 연구는 이 중 어휘 지식에 특화된 Cambridge English Vocabulary Profile(EVP)을 CEFR 어휘 분류의 준거로 활용하였다. EVP는 Cambridge Learner Corpus(4,200만 토큰)와 British National Corpus를 병용하여 각 어휘의 습득 단계를 실증적으로 규명한 데이터베이스이다(Capel, 2012).

<표 A1> CEFR 레벨별 정의 및 수능 출현 비율

레벨	명칭(한)	설명	목표 어휘량	수능 출현 비율(%)	예시 어휘
A1	입문	일상적 기초 표현	~500개	약 13.5	a, the, go, big, say
A2	초급	친숙한 주제의 표현	~1,000개	약 40.3	explain, recent, culture
B1	중급	주요 논점 처리 가능	~2,000개	약 25.0	analyze, significant, criteria
B2	중상급	복잡한 텍스트 이해	~4,000개	약 16.0	empirical, hypothesis, facilitate
C1	상급	학술 목적 정교한 어휘	~8,000개	약 0.15	articulate, conducive, nuance
C2	숙달	희귀·정교 어휘 포함	~16,000개+	약 0.01	obfuscate, perspicacious
미등재(UNK)	—	고유명사·전문술어·약어	—	약 20.0	(인명, 학술술어 등)

주: 수능 출현 비율은 본 연구 코퍼스(2009~2025학년도 독해 고배점 지문, n=97) 평균값. A1과 A2는 별도 제시. 미등재(UNK)는 EVP 사전 미수록 어휘로 난이도 분석에서 별도 처리.

<표 A2> 연구 적용 CEFR 어휘 분류 기준

분류 기준 항목	세부 내용
어휘 수록 기준	Cambridge EVP 공식 데이터베이스 기반; Cambridge Learner Corpus + BNC 병용
레벨 배정 원칙	단어의 '가장 중요한 의미'에 해당 레벨 배정; 다의어는 가장 낮은 CEFR 레벨 기준
분석 단위	문장 부호·숫자 제외 알파벳 토큰 기준; 표제어 변환 없이 표면형(surface form) 매칭
미등재(UNK) 처리	고유명사·전문술어·약어는 UNK 분류; 난이도 분석에서 별도 보고
통계 검증	CSAT vs. SAT: Mann-Whitney U 검정(양측, $\alpha=.05$); 효과크기 r 산출
한계	EVP는 일반 영어 맥락 기준 설계; 시험 특화 학술 어휘 미포괄 가능성 있음

<표 A3> CEFR 레벨별 어휘 출현 비율 — 기간별 비교(% , 평균)

지표	CSAT 2009-12 (n=7)	CSAT 2013-17 (n=29)	CSAT 2018-22 (n=22)	CSAT 2023-25 (n=15)	SAT Digital (n=13)	유의성 (CSAT 전체 vs. SAT)
A1 어휘(%)	13.5	11.3	11.7	12.0	9.8	—
A2 어휘(%)	40.3	43.3	40.9	42.0	34.1	—
B1 어휘(%)	21.8	20.8	21.6	20.0	28.6	—
B2 어휘(%)	4.9	4.7	5.0	5.4	7.3	—
C1 어휘(%)	0.10	0.12	0.12	0.27	0.23	—
C2 어휘(%)	0.00	0.02	0.00	0.05	0.00	—
A1+A2 합계(%)	53.8	54.6	52.6	54.0	43.9	p<.001 ***
C1+C2 합계(%)	0.10	0.14	0.12	0.32	0.23	ns
미등재(UNK,%)	19.2	19.7	20.7	20.7	18.7	—

주: *** p<.001, ns=not significant; Mann-Whitney U, 양측 검정. 수능 출현 비율은 기간별 단순 평균값.

Appendix B. 분석 변수 정의 및 계산 방법

본 연구에서 사용한 모든 텍스트 복잡도 변수의 정의, 계산식, 해석 범위를 아래에 상세히 제시한다

<표 B1> Category A 변수 정의

변수명	약어	계산식 / 측정 방법	해석 범위	본 연구 활용
Flesch-Kincaid Grade Level	FK Grade	$FK = 0.39 \times (W/S) + 11.8 \times (Syl/W) - 15.59$ (W=총 토큰수, S=총 문장수, Syl=총 음절수)	미국 학년 수준; FK 10~12=고교 고학년, FK 12~14=대학 입문	Category A 주요 지표
Moving-Average Text-Type Ratio	MATTR	50토큰 크기의 이동 창(moving window)을 1토큰씩 이동하며 각 창의 TTR 산출 후 평균: $MATTR = (1/N) \times \sum TTR(wi \sim wi+49)$ (N=총 이동 창 수, TTR=Types/Tokens)	0~1; 높을수록 어휘 다양성 높음; 이동 창 방식으로 길이 편향 보정	Category A 보조 지표
Type-Token Ratio	TTR	$TTR = Types / Tokens$	0~1; 지문 길이 증가 시 값 감소하는 한계 있음	성적 상관 분석 참고 지표

<표 B2> Category B 변수 정의

변수명	약어	측정 방법	방향 (↑=어려움)	본 연구 활용
연결어 밀도 (전체)	Conn. Density	인과(because, thus, therefore), 대조(however, whereas), 첨가(moreover, furthermore) 등 연결어 토큰 수 / 전체 토큰 수	↓ 어려움 (낮을수록 논리 단서 적어 어려움)	Category B 핵심 지표
대조 연결어 밀도	Contrast Conn.	대조 연결어 토큰 수 / 전체 토큰 수	↓ 어려움	Category B 세부 지표
상상가능성 점수	Imagery	MRC Psycholinguistic Database (Paivio et al., 1968) 기반 각 어휘의 상상 가능성 점수(100~700) 지문 평균	↓ 어려움 (낮을수록 추상적, 이해 부담 증가)	Category B 핵심 지표
명사화 지수	Nominaliza tion	-tion, -ness, -ment, -ity 접미사 포함 어휘 수 / 전체 토큰 수	↑ 어려움 (높을수록 명사화 심화, 인지 부담 증가)	Category B 보조 지표
담화 결속성	Discourse Coherence	지문 전체 의미적 일관성 종합 지수; 인접 문장 간 내용어 중첩, 연결어, 명사 참조 등 복합 산출	↓ 어려움 (낮을수록 담화 흐름 불명확)	Category B 핵심 지표
종합 텍스트 복잡도 지수(CTCI)	Composite Text Complexity Index	연결어 밀도, 상상가능성, 명사화, 담화 결속성을 min-max 정규화 후 방향성 맞게 집계; 0~1 척도	↑ 어려움 (높을수록 전반적 난이도 높음)	Category B 통합 지표