교육과정평가연구
The Journal of Curriculum and Evaluation
2015, Vol. 18, No. 1, pp. 185~208

국가수준 학업성취도 평가 서답형 문항에 대한 자동채점의 실용성 분석[1)]

이 상 하(한국교육과정평가원 연구위원)*
노 은 희(한국교육과정평가원 연구위원)
성 경 희(한국교육과정평가원 부연구위원)**

≪ 요 약 ≫

서답형 문항은 선택형 문항에 비하여 고차원적인 사고를 측정하기 용이하다는 교육적인 장점이 있음에도 불구하고, 채점에 소요되는 시간과 비용으로 인해 대규모 지필평가에서는 선택형 문항의 비중이 오히려 높은 편이다. 대규모 교육평가의 제한된 예산 범위 내에서 서답형 문항의 다양성과 비중을 확대하기 위해서는 컴퓨터 기반 평가와 자동채점을 적극적으로 도입할 필요가 있다.

이 연구의 목적은 단어・구 수준의 한국어 서답형 답안을 처리할 수 있는 자동채점 프로그램을 활용하여 지필평가의 서답형 문항을 채점하는 방법의 실용성을 분석하는 것이다. 이를 위하여 2014년 국가수준 학업성취도 평가의 사회 교과에 응시한 중학교 3학년 표집학생 7,442명의 서답형 답안을 합숙채점과 자동채점의 두 가지 방법으로 채점하였다. 서답형 문항에 대한 자동채점의 실용성 평가는 채점 비용의 효율성과 채점 결과의 정확성 측면에서 기존의 합숙채점과 비교하는 방법으로 이루어졌다.

자동채점은 합숙채점에 비하여 채점 비용의 60% 이상을 절감할 수 있고, 자동채점에 최적화된 채점 방식을 사용할 경우 채점 비용의 80% 이상을 절감할 수 있는 것으로 분석되었다. 또한, 합숙채점 최종점수와 자동채점 점수 간의 상관계수는 0.97~1, 완전일치도는 97.76%~99.99%, 근사일치도는 98.15%~100%, 카파계수는 0.94~1, 일차가중카파계수와 이차가중카파계수는 0.96~1인 것으로 나타났다. 이와 같은 상관계수와 일치도 수준은 자동채점과 채점 전문가의 점수가 매우 유사하다는 것을 의미한다. 결론적으로, 이 연구는 채점 비용의 효율성과 채점 결과의 정확성 측면에서 지필평가의 서답형 답안에 대한 자동채점의 실용성을 확인하였다. 또한, 단어・구 수준 한국어 자동채점 프로그램이 인간 채점자의 일부 또는 전부를 대체하거나 인간 채점자의 채점 오류를 관리할 수 있는 수준까지 도달했다는 것을 확인하였다.

주제어: 자동채점, 기계채점, 한국어 자동채점, 단답형 문항 채점, 서답형 문항 채점

1) 이 논문은 '한국어 서답형 문항 자동채점 프로그램 개발 및 실용성 검증' 한국교육과정평가원 연구보고(RRE 2014-6)의 일부 내용을 수정・보완하여 재구성한 것임.

* 제1저자, sangha@kice.re.kr
** 교신저자, kelly9147@kice.re.kr

Ⅰ. 서론

유럽과 북미 국가들을 중심으로 초·중등 학생들의 미래사회 핵심역량을 강화하기 위한 교육이 강조되면서, 고차원적 핵심역량을 평가하기 위해 선택형 문항에서 벗어나 서답형 문항과 수행형 문항의 비중을 확대하기 위해 노력하고 있다(이상하 외, 2014). 선택형 문항은 학생들이 주어진 답안 중에서 정답을 선택하도록 요구하지만, 서답형 문항은 학생들이 스스로 답안을 작성하도록 요구한다. 즉, 서답형 문항은 선택형 문항에 비하여 학생들의 답안 작성이 자유롭다는 측면에서 고차원적 사고를 측정하는 데 유리한 점이 있다. 학교 교육이 지향하는 고차원적 사고를 평가하기 위해서는 다양한 형식의 서답형 문항들을 사용하는 것이 바람직하지만, 지필평가 형식으로 시행되는 대규모 학업성취도 평가의 경우 서답형 문항의 채점에 소요되는 시간과 비용 등은 현실적으로 큰 부담이 된다. 예를 들어, 우리나라 국가수준 학업성취도 평가에 포함되는 서답형 문항의 형식과 비중은 매우 제한적이지만 채점에 참여하는 인력과 채점에 소요되는 시간과 비용은 적지 않다. 대규모 학업성취도 평가에서 다양한 형태의 서답형과 수행형 문항을 부담 없이 사용하기 위해서는, 서답형 문항과 수행형 문항을 효율적으로 시행하고 채점할 수 있는 컴퓨터 기반 평가와 자동채점을 도입할 필요가 있다.

미국의 NAEP(National Assessment of Educational Progress), 영국의 NCA(National Curriculum Assessment), 호주의 NAPLAN(National Assessment Program - Literacy and Numeracy), 국제 학업성취도 평가인 PISA(Programme for International Student Assessment)와 TIMSS(Trends in International Mathematics and Science Study) 등의 경우에, 전체 문항 수 중 서답형 문항의 비율이 30%~60%정도이다(노은희 외, 2012). 그런데 현재 우리나라 중등학교 학생들을 대상으로 시행되는 국가수준 학업성취도 평가의 경우, 서답형 문항의 비중이 약 20%정도로 외국에 비하여 다소 낮은 편이다. 한편, 국제 수준의 대규모 학업성취도 평가인 PISA에서도 2006년 과학 소양, 2009년 디지털 읽기 소양, 2012년 문제해결능력 평가 등에서 컴퓨터 기반의 평가 방식을 통해 학생들의 기본 소양을 측정하였고 2015년에는 컴퓨터 기반 평가를 모든 영역으로 확대할 예정이다(송미영 외, 2013). 이러한 국제적인 교육평가 동향에 힘입어 우리나라도 국가수준 교육평가에서 컴퓨터 기반 평가 체제의 도입 가능성을 타진한 바 있다(김경희 외, 2013). 이에 노은희 외(2012, 2013, 2014)는 국가수준 학업성취도 평가를 지필평가에서 컴퓨터 기반 평가로 전환하고 서답형 문항의 비중을 확대하면서도 채점 부담을 크게 줄일 수 있도록, 2012년부터 한국어 서답형 문항 자동채점 프로그램을 개발하는 연구를 진행하였다.

미국은 차세대 학력평가 2.0을 통해, 2015년부터 단순지식을 측정하는 선택형 문항의 지필평가를 지양하고 고차원적 핵심역량을 평가할 수 있는 서답형 및 수행형 문항이 강화된 컴퓨터 기반 평가로 전환하는 계획을 추진하였다(Educational Testing Service, 2012). 미국 교육

부는 차세대 학력평가를 개발하기 위해 46개 주정부가 참여하는 PARCC(Partnership for the Assessment of Readiness for College and Careers)과 SBAC(Smarter Balanced Assessment Consortium) 협력체에 모두 3억 3천만 달러를 지원하였다(U.S. Department of Education, 2010). 컴퓨터 기반 평가에서 서답형 및 수행형 문항의 비중을 확대하면서 비용을 크게 증가시키지 않는 방법 중의 하나는 인간 채점자 대신에 공학 도구를 활용하여 서답형 및 수행형 문항을 채점하는 것이다. 미국의 차세대 학력평가 개발 및 시행을 위한 비용분석의 결과는 서답형 및 수행형 문항을 채점하기 위해 인간 채점자를 활용할 경우 채점 비용이 전체 예산의 60% 이상을 차지하는 것으로 나타났으며, 서답형 문항의 특성 및 채점 환경 등과 같은 제반 조건에 따라 차이가 있지만 자동채점의 비용은 인간 채점 비용의 25%~80% 정도인 것으로 추정되었다(Topol, Olson, & Roeber, 2014). 즉, 컴퓨터 기반 평가로 시행된 서답형 문항에 대한 자동채점은 서답형 문항에 대한 채점의 부담을 획기적으로 줄임으로써 대규모 교육평가에서 서답형 문항의 비중을 확대할 수 있는 가장 현실적인 수단이 될 수 있다.

한국어 서답형 자동채점 프로그램은 컴퓨터 기반 평가 환경에서 서답형 문항을 효율적으로 처리하기 위한 목적으로 개발되기 시작하였다. 현재 단어·구 수준의 서답형 답안을 처리할 수 있는 자동채점 프로그램 개발이 어느 정도 완료되었고, 문장 수준의 서답형 답안을 처리할 수 있는 자동채점 프로그램을 개발하는 연구가 진행될 예정이다. 그런데 국가수준 학업성취도 평가는 여전히 지필평가로 시행되고 있으며, 가까운 시일 내에 지필평가에서 컴퓨터 기반 평가로 전환할 것이라는 구체적인 계획이 아직은 없다. 따라서 이미 개발되어 있는 한국어 서답형 자동채점 프로그램의 활용성을 높이기 위해, 지필평가로 시행되고 있는 국가수준 학업성취도 평가의 서답형 답안을 채점하는 방안을 고려할 수 있다. 자동채점 프로그램의 장점이 지필평가 환경에서 많이 반감되지만, 자동채점 프로그램을 활용하여 지필평가의 서답형 문항을 채점하는 것이 기존의 합숙채점 방식보다는 많은 장점이 있을 수 있다.

대규모 교육평가의 서답형 문항에 대한 자동채점 방식의 도입 여부를 결정하기 위해서는 크게 채점 비용의 효율성과 채점 결과의 정확성 측면으로 실용성을 판단할 필요가 있다. 교육평가의 관점에서 서답형 문항은 선택형 문항에 비해 많은 장점을 가지고 있지만, 답안을 채점하는 데 소요되는 비용으로 인해 대규모 평가에서 사용하기 어려운 점이 있다. 따라서 자동채점 도입 여부를 판단할 때 채점 비용의 효율성은 중요한 고려사항이 되어야 한다. 또한, 평가의 결과가 학생 개인에게 미치는 영향을 고려할 때, 채점 비용의 효율성을 이유로 채점 결과의 정확성을 포기할 수는 없다. 특히, 고부담 시험의 경우 자동채점 프로그램이 갖는 채점의 효율성으로 채점 결과의 부정확성을 정당화할 수는 없다. 따라서 채점 결과의 정확성은 자동채점 도입 여부를 판단하는 또 다른 중요한 고려사항이 되어야 한다.

이 연구의 목적은 단어·구 수준의 한국어 답안을 처리할 수 있는 자동채점 프로그램을 활용하여 국가수준 학업성취도 평가의 서답형 답안을 채점하는 방식이 채점 비용의 효율성 측면과 채점 결과의 정확성 측면에서 실용성이 있는지를 확인하는 것이다.

Ⅱ. 단어·구 수준 서답형 문항 자동채점 프로그램

서답형 문항에 대한 '자동채점'이라는 용어는 서답형 문항을 선택형 문항만큼 기계가 자동으로 정확하게 채점할 수 있을 것 같은 오해를 불러일으키는 경우가 많다. 서답형 문항에 대한 자동채점은 기계가 채점할 수 있도록 인간이 채점 기준에 해당하는 정답 템플릿을 작성하고, 기계가 처리하지 못하는 답안의 경우 인간이 직접 채점하는 과정 등을 포함한다. WCAL(n.d)에 따르면, 다양한 자동채점 프로그램의 시스템이 있지만, 대개 ① 채점 기준표 제공, ② 샘플 답안을 이용하여 모범 답안 템플릿 생성, ③ 샘플 답안과 비교하여 학생 답안 분석 및 채점, ④ 채점 과정을 확인하기 위한 수작업 조정 등의 과정을 거친다. 이러한 절차 속에서 채점 관리자가 개입하여 정답 템플릿을 생성하거나 조정하는 것이며, 프로그램은 이를 보다 간단하고 효율적으로 진행할 수 있도록 지원한다. 즉, 기본적으로 자동채점 프로그램이라고 해서 컴퓨터가 완전하게 100% 처리하는 것이 아니라 채점 관리자가 주요 단계마다 개입하여 채점 과정을 진행하는 것이고, 이러한 인간 개입 과정을 최소화하는 과정이 곧 자동화 과정인 것이다(노은희, 2014).

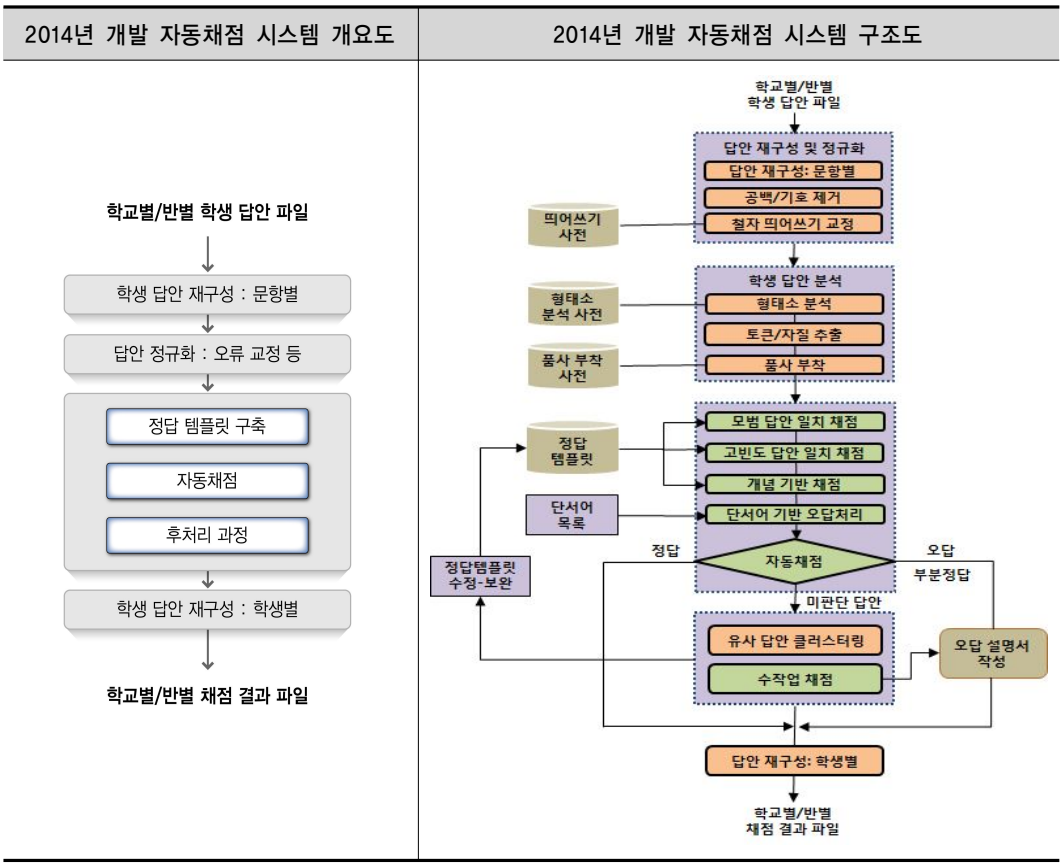
〔그림 1〕은 2013년에 개발한 단어·구 수준 자동채점 시스템을 토대로 2014년에 수정·보완한 자동채점 프로그램의 구조도이다. 2013년에 개발한 자동채점 프로그램은 단어·구 수준의 답안은 95% 이상 자동채점 처리가 가능하도록 단계별 시스템을 갖추고, 문항별로 채점의 정확성과 효율성이 제고되도록 지원 도구들을 보강·개발하였다. 그런데 자동채점 프로그램이 실제성을 갖고 운영되기 위해서는, 단일 교과의 전체 서답형 문항을 대상으로 일련의 통합된 채점 처리 과정을 지원하고 학생별로 최종 채점 결과를 보고하여 교수·학습 과정에 실질적인 피드백을 제공할 수 있는 시스템을 구축할 필요가 있다. 따라서 학생별로 작성된 답안 파일(입력 파일)을 문항별로 분리하고 각 문항에 대한 채점이 완료된 후에 다시 각 학생에 대한 답안 파일(출력 파일)로 작성하는 과정이 별도로 필요하다.

한편, 2013년 개발한 시스템은 채점 전문가가 개입해야 하는 과정이 자동채점 전(前) 정답 템플릿 준비 단계, 자동채점 단계, 후(後)처리 과정인 수작업 채점 단계로 분산되어 있었다. 이러한 개입 과정은 채점 옵션을 전문가의 판단에 따라 결정하는 단계에서 나타나는데, 보통 특수기호와 불용어[2] 제거, 철자 및 띄어쓰기 오류의 교정, 유의어를 적용한 유사 답안 채점과 같이 학생 답안을 정규화하는 과정과 연관되어 있다. 예를 들어, 자동채점 과정에서 실수로 띄어쓰기 무시 옵션을 잘못 설정하거나 정답 템플릿을 구축할 때 유의어 적용 오류가 있었다면 자동채점 및 후처리 과정에서 철자 오류 교정 확인이라는 수작업을 반복해야 하는 비효율성이 발생한다.

2) 불용어(不用語, stop word)는 색인을 작성할 때 표제어로 하지 않는 단어나 관계없는 분야의 용어로, 예를 들어, 학생 답안 중 채점에 필요하지 않은 이모티콘, 감탄사, 욕 등이 이에 해당한다.

이에 인간 개입을 최소화하는 방식으로 자동채점 프로그램의 효율성을 제고할 필요가 있다.

따라서 〔그림 1〕의 좌측에 기술한 시스템 개요도에서 나타난 바와 같이, 2014년에 수정·보완한 자동채점 프로그램은 학생 답안 파일을 문항별로 분리하여 답안 정규화 모듈을 전처리 과정으로 독립시켰다. 또한 〔그림 1〕의 우측 그림은 전체 시스템의 각 모듈을 상세하게 표현한 것이다. 자동채점 시스템은 크게 '답안 분석 및 정규화' 과정, '자동채점' 과정, '후처리 수작업 채점'의 과정으로 구성된다.



〔그림 1〕 2014년 단어·구 수준 자동채점 시스템 구조도

먼저 '답안 분석 및 정규화' 과정에서는 문항의 채점 요건에 맞게 학생 답안을 정규화(normalization)하는 작업을 수행한다. 문장부호 및 기호, 띄어쓰기, 맞춤법 사용의 오류를 점검해야 하는 문항이 있는 반면, 그렇지 않은 문항도 존재한다. 채점자가 문항의 특성에 맞게 정규화 옵션을 설정하고 학생 답안을 불러오도록 하면 시스템은 불러온 학생 답안을 대상으로 정규화 작업을 수행한다.

다음으로 '자동채점' 과정에서는 채점기준표에 따라 작성된 정답 템플릿을 통해 초기 일부 학생 답안에 대하여 채점을 수행하고, 정답과 오답으로 판정하기 어려운 답안은 미판단 답안으로 보류하여 후처리 작업으로 넘겨진다.[3] 모범 답안 일치 채점, 고빈도 답안 일치 채점, 개념 기반 채점, 단서어 기반 오답 처리의 순서로 진행되며 각 과정에서 모든 답안을 처리하고 미채점된 답안들만 다음 단계로 넘어간다. 모범 답안과 고빈도 답안 일치 채점은 정답 템플릿의 해당 정보와 학생 답안이 완전하게 일치하는 경우에 점수를 부여한다. 가령, 모범 답안인 "아름이를 생포하였다."와 완전 일치하거나, 다수의 학생들이 작성한 것으로 정답으로 인정되는 답안인 "아름이를 산 채로 붙잡았다."와 완전 일치하는 경우 점수가 부여된다. 개념 기반 채점은 학생 답안의 일부 혹은 전체가 정답으로 인정할 수 있는 개념과의 일치 및 유사성 여부 정도를 판단하여 점수를 부여하는 단계이다. 예를 들어, "반달가슴곰을 생포하였다."라는 답안에서 '아름이'와 '반달가슴곰'을 동일 개념으로 보고 점수를 부여하게 된다. 개념 기반 채점으로 걸러지지 않는 답안에 대해서는 '단서어 목록'을 작성하고 답안에 단서어가 출현하지 않으면 오답으로 처리하는 과정이 추가된다. 단서어는 정답 또는 부분 정답이 될 수 있는 단어이다. 단서어 기반 오답 처리는 정답 템플릿에 단서어가 존재하는 경우에만 수행되며 학생 답안에 단서어 목록의 단어가 하나도 존재하지 않으면 오답으로 처리한다. 즉, '아름이', '반달가슴곰', '반달곰' 등의 단어가 출현하게 되면 수작업 채점을 할 수 있도록 미판단 답안으로 남겨두고, 출현하지 않게 되면 오답으로 처리한다. 이는 수작업 채점을 최소화할 뿐만 아니라 자동채점의 오류 가능성을 줄이기 위한 단계라 할 수 있다.

마지막으로 '후처리 수작업 채점' 과정에서는 자동채점 프로그램으로 채점이 되지 않은 미판단 답안에 대해 수작업으로 채점을 수행한다. 이때, 수작업 채점 도구를 지원하여 유사한 답안들을 모아서 한꺼번에 채점할 수 있도록, 클러스터링 작업을 수행하여 정답 템플릿을 갱신하거나 빈도가 낮은 답안 유형들을 수작업으로 채점한다.

Ⅲ. 분석 방법

현행 국가수준 학업성취도 평가는 지필평가의 형식으로 시행되고 있으며 선택형 문항과 서답형 문항으로 구성되어 있다. 자동채점 프로그램을 활용하여 국가수준 학업성취도 평가의 서답형 답안을 채점하는 것이 어느 정도 유용한지를 분석하기 위해 합숙채점과 자동채점의 비용을 비교하고, 자동채점의 결과와 인간채점 결과가 일치하는 정도를 추정하는 분석 방법을 사용하였다.

3) 정답 템플릿 생성에 대한 상세한 설명은 '대규모 평가를 위한 서답형 문항 자동채점 프로그램 정교화 및 시범 적용' 한국교육과정평가원 연구보고(RRE 2013-5)를 참조할 수 있다.

1. 분석 대상

단어·구 수준 자동채점 프로그램의 실용성을 분석하기 위해, 2014년 국가수준 학업성취도 평가의 중3 사회 교과 서답형 문항을 합숙채점과 자동채점의 두 가지 방식으로 채점하였다. 중3 사회 교과는 총 7,442명의 학생을 대상으로 표집평가로 시행되었으며, 서답형 문항은 9개의 대문항(14개의 소문항)으로 구성되어 있다. 〈표 1〉은 분석 대상 문항에 대한 정보를 정리하여 나타낸 것이다.

〈표 1〉 2014년 중3 사회 자동채점 대상 서답형 문항 정보

구분 / 문항 번호	답안 작성 유형	답안 유형 수 (/1,000)	모범 정답	배점 분포
1	1단어 명사형	276 (37)	경도	0, 1, 2
2	1문장	2413 (324)	빙하가 침식 작용을 하였다.	0, 1, 2, 3
3-㉠	1단어 명사형	730 (98)	성비	0, 1
3-㉡	1단어 술어형	333 (45)	크다.	0, 1
4-㉠	1단어 명사형	602 (81)	원효	0, 1
4-㉡	1단어 명사형	306 (41)	불교	0, 1
5-㉠	1단어 명사형	481 (65)	임진왜란	0, 1
5-㉡	2문장	3,359 (451)	명이 쇠퇴하고, 후금이 성장하였다.	0, 1, 2
6-㉠	1단어 명사형	696 (94)	루이 14세	0, 1
6-㉡	2단어 명사형	795 (107)	베르사유 궁전	0, 1
7-㉠/㉡/㉢	3단어 명사형	1,880 (253)	국회(입법부), 정부(행정부), 법원(사법부)	0, 1, 2, 3
8-㉠	2단어 명사형	246 (33)	평등권(평등권적 기본권)	0, 1
8-㉡	2단어 명사형	431 (58)	청구권(청구권적 기본권)	0, 1
9*	4단어 명사형	1219 (164)	문화 상대주의	0, 1, 2

* 모범 답안은 '문화 상대주의'로 2단어 명사형이나, 실제 학생들의 답안은 '다른 문화를 인정하는 자세', '문화적 차이 받아들이기' 등과 같이 풀어쓴 경우가 많아 4단어 명사형으로 분류하였다.

사회 교과는 소문항을 기준으로 답안 작성 유형을 살펴보면 1단어 명사형이 6문항, 2단어 명사형이 3문항, 1단어 술어형이 1문항으로 나타난다. 또한 3단어 명사형이 1문항(채점 연동으로 실제 3문항), 4단어 명사형이 1문항이고, 문장형 문항이 2문항으로 나타난다. 한편 답안 1,000개당 서로 다른 답안의 유형 수를 살펴보면, 1~2단어 명사형/술어형 문항은 33~107개로 학생 답안 간 편차가 비교적 작게 나타난 반면, 3~4단어 명사형 문항은 164~253개, 문장형 문항은 324~451개로 학생 답안 간 편차가 상대적으로 크게 나타난다. 즉, 2014년 학업성취도 평가 중3 사회 교과는 문장형 답안을 요구하는 2번과 5-㉡번 문항을 제외하고 나머지 12

개의 문항이 모두 학생 답안 간 편차가 비교적 작은 단어·구 수준 답안을 요구하고 있어서, 단일 교과로서 단어·구 수준 자동채점 프로그램을 사용하여 채점하기에 적절하였다. 다만, 문장형 2문항(2번, 5-㉡번)의 경우는 구문 분석 기술까지 요구하여 단어·구 수준 프로그램에서 처리하기에는 까다로운데, 문장 수준 프로그램 개발을 위한 점검 차원에서 실험적으로 대상 문항에 포함하였다.

2. 실용성 평가 준거

단어·구 수준 자동채점 프로그램의 실용성을 검증하기 위해, 채점 비용의 효율성과 채점 결과의 정확성 측면으로 나누어 분석하였다. 자동채점 프로그램의 효율성은 현행 학업성취도 평가에서 시행하는 온라인 인간채점 방식과 비교하여 평가하는데, 채점에 소요되는 비용(인력 포함)을 어느 정도 줄일 수 있는지를 분석한다. 이를 위하여, 2014년 학업성취도 평가의 사회 교과 서답형 문항을 채점하는 데 사용한 채점 비용을 산출하여 이를 효율성 분석의 준거로 삼았다. 한국어 서답형 문항 자동채점 프로그램은 컴퓨터 기반 평가의 환경을 전제로 하므로, 학생 답안을 수작업으로 입력하여 채점할 경우 자동채점 프로그램의 효율성은 크게 떨어질 수밖에 없다. 다만, 지필평가의 환경에서도, 자동채점 프로그램이 기존 방식에 비해 유용한지를 비용 절감 차원에서 점검하고자 한다. 이에 따라 '채점자들이 합숙을 하면서 온라인상에서 서답형 답안을 채점하는 현행 방식'과 '서답형 답안을 수작업으로 입력하여 자동채점하는 방식'의 비용을 비교하여 분석한다.

또한, 자동채점 프로그램의 정확성을 판단하려면 평가의 준거가 필요하다. 단어·구 수준 서답형 문항의 경우, 다른 서답형 문항에 비해 채점자의 주관성이 덜 개입하는 편이라 할 수 있다. 동일한 채점 기준에 근거하더라도 답안에 대한 채점자의 해석과 판단의 차이에 따라 채점 결과가 달라질 수 있기 때문에, 자동채점 결과의 정확성을 확인할 수 있는 준거가 사실은 명확하지 않은 것이다(Shermis & Hamner, 2013). 따라서 자동채점 프로그램에 관한 대부분의 연구에서는 자동채점의 결과가 전문가 또는 채점자들이 채점한 결과와 얼마나 일치하는지를 분석한다(Shermis, 2010; Zhang, 2013). 이에 본 연구에서도 개별 문항에 대한 자동채점의 결과와 채점자들의 채점 결과가 어느 정도 일치하는지를 통해 채점의 정확성을 분석하였다. 채점 점수는 두 명의 채점자 점수, 채점자들의 최종점수, 자동채점 점수를 사용한다. 여기서 두 명의 채점자 점수는 1라운드에서 채점자 두 명이 제출한 점수이고, 최종점수는 복수의 채점자가 3라운드에 걸쳐 확정한 채점자들의 최종점수이다. 국가수준 학업성취도 평가의 서답형 문항 채점에서는 채점자 간에 점수 차이가 있는 경우 추가적인 채점 절차를 통해 최종적으로 하나의 점수를 확정한다. 그리고 자동채점 점수는 단어·구 수준 자동채점 프로그램을 사용하여 채점한 점수이다.

3. 통계 분석

채점의 정확성을 확인하기 위한 통계 분석 방법으로 상관계수(correlation coefficient), 완전일치도(exact agreement), 근사일치도(exact and adjacent agreement), 카파계수(kappa coefficient), 일차가중카파계수(linear weighted kappa coefficient), 이차가중카파계수(quadratic weighted kappa coefficient)를 추정하였다.

'상관계수'는 점수들 간의 선형적인 관계를 나타내는 척도로서 두 가지 방법으로 채점한 점수들이 정확하게 일치하지 않더라도 동일한 양상으로 변화한다면 높게 나타날 수 있다. 한편, '완전일치도'는 두 가지 방법으로 채점한 점수들이 정확하게 일치하는 경우의 비율을 나타내고, '근사일치도'는 두 가지 방법으로 채점한 점수들 간 차이가 1점 이하인 경우의 비율을 나타낸다. 단어·구 수준의 답안이라고 하더라도 채점자의 주관적인 판단이 개입하는 것을 완전히 배제할 수는 없다. 즉, 채점자는 채점 과정에서 글씨와 그 의미를 판단하고 채점 기준에 따라 점수를 부여하게 되는데, 두 채점자의 판단이 다른 경우들 중에는 어느 한 사람의 판단이 다른 사람의 판단보다 옳다고 확실하게 결론을 내리기 어려운 경우들이 있다. 근사일치도는 채점자의 주관적 판단을 존중하여 채점자 간의 점수 차이가 1점 이하인 경우에 일치하는 것으로 간주하는 일치도라고 할 수 있다.

완전일치도와 근사일치도는 우연에 의해 두 점수가 일치한 것도 포함하기 때문에 일치도가 과장되게 나타나는 경향이 있다. '카파계수'는 두 점수가 우연에 의해 일치한 경우를 제외하고 계산한 완전일치도라고 할 수 있다. (식 1)은 카파계수($\hat{\kappa}$)를 계산하는 공식인데, 두 가지 점수가 우연에 의해 일치하는 확률(P_e)을 제외한 나머지 부분에서 두 가지 점수가 완전히 일치한 확률(P_o)의 비로 정의된다.[4]

$$\hat{\kappa} = \frac{(P_o - P_e)}{(1 - P_e)} \qquad \text{(식 1)}$$

카파계수는 두 범주가 완전히 일치하지 않는 경우는 모두 불일치한 것으로 간주하는데, 가중카파계수는 두 범주가 일치하지 않는 정도에 따라 가중치를 부여하여 계산한 카파계수이다. 일반적으로, 가중카파계수는 일치하는 정도에 따라 1~0의 가중치를 부여하게 되는데, 카파계수는 두 점수가 완전히 일치하면 가중치가 1이 되고 일치하지 않으면 가중치가 0이 되는 가중카파계수의 특수한 경우라고 할 수 있다. 가중카파계수($\hat{\kappa}_w$)는 (식 2)와 같이 정의된다.[5]

---

4) (식 1)에서 P_o는 두 개의 점수가 완전히 일치하는 비의 합을 의미하며 $P_o = \sum_i p_{ii}$이고, P_e는 두 개의 점수가 일치하지 않는 두 범주의 주변 확률(marginal probability)의 곱을 합한 것으로 $P_e = \sum_i \sum_j p_{i.} p_{.j}$와 같이 정의된다.

$$\hat{\kappa}_w = \frac{(P_{o(w)} - P_{e(w)})}{(1 - P_{e(w)})} \quad \text{(식 2)}$$

가중카파계수는 가중치를 부여하는 방식에 따라 일차가중카파계수와 이차가중카파계수로 나뉜다. 일차가중카파계수에 부여되는 가중치 w_{ij}는 (식 3)과 같이 정의된다.[6] 예를 들어, 서답형 문항의 배점이 0, 1, 2, 3인 경우에 점수 범주는 모두 4개라고 할 수 있으며, 일차가중카파계수를 계산할 때 두 점수의 차이에 따른 가중치는 $w_{11} = 1$, $w_{12} = w_{23} = w_{34} = 0.67$, $w_{13} = w_{24} = 0.33$, $w_{14} = 0$와 같이 된다. 즉, 두 점수가 일치하는 경우에는 가중치가 1이지만, 그 점수 차이가 커질수록 가중치는 점점 작아지게 된다.

$$w_{ij} = 1 - \frac{|C_i - C_j|}{C_C - C_1} \quad \text{(식 3)}$$

이차가중카파계수는 일차가중카파계수와 마찬가지로 분류의 범주가 불일치하는 정도에 따라 가중치를 부여하는데, 두 점수의 범주 i 와 j에 대해 (식 4)와 같이 w_{ij}의 가중치를 부여한다. 마찬가지로, 서답형 문항의 배점이 0, 1, 2, 3인 경우에 두 점수의 차이에 따른 가중치는 $w_{11} = 1$, $w_{12} = w_{23} = w_{34} = 0.89$, $w_{13} = w_{24} = 0.56$, $w_{14} = 0$와 같이 된다. 즉, 두 점수가 일치하는 경우에는 가중치가 1이지만, 그 점수 차이가 커질수록 가중치는 점점 작아진다.

$$w_{ij} = 1 - \frac{(C_i - C_j)^2}{(C_C - C_1)^2} \quad \text{(식 4)}$$

일차가중카파의 가중치는 점수 차이에 비례하고, 이차가중카파의 가중치는 점수 차이의 제곱에 비례한다. 따라서 점수 차이에 비례하여 일치하는 정도를 낮게 보는 경우에는 일차가중카파를 사용하고, 점수 차이가 커짐에 따라 일치하는 정도를 일차가중카파 이상으로 낮게 보는 경우에는 이차가중카파를 사용한다.

---

5) (식 2)에서 $P_{o(w)}$는 두 점수 범주의 비 p_{ij}에 가중치 w_{ij}를 곱하여 모두 합한 $P_{o(w)} = \sum_i \sum_j w_{ij} p_{ij}$이고, $P_{e(w)}$는 두 개의 점수가 일치하지 않는 두 범주의 주변 확률(marginal probability) $p_{i.}$와 $p_{.j}$, 그리고 가중치 w_{ij}를 곱하여 합산한 것으로 $P_{e(w)} = \sum_i \sum_j w_{ij} p_{i.} p_{.j}$와 같이 정의된다.

6) (식 3)과 (식4)에서 C_i는 점수 범주 i의 점수, C_j는 점수 범주 j의 점수, C_C는 점수 범주의 최고 점수, C_1는 점수 범주의 최하 점수이다.

Ⅳ. 분석 결과

자동채점 프로그램을 활용하여 국가수준 학업성취도 평가의 서답형 문항을 채점하는 것이 현행 합숙채점 방식에 비하여 어느 정도 유용한지를 분석하였다. 합숙채점 방식과 자동채점 방식의 채점 과정을 비교하여 비용 차이를 산출하였고, 인간채점 결과와 자동채점 결과의 상관관계와 일치도를 추정하였다.

1. 채점 비용의 효율성

국가수준 학업성취도 평가의 서답형 문항을 합숙채점으로 진행하는 경우 채점자 및 지원인력의 인건비와 숙박비가 채점 비용의 대부분을 차지한다. 즉, 합숙채점에 참여하는 채점 인력의 규모, 채점 기간, 채점 시간 등은 모두 채점 비용에 직접적으로 영향을 주는 요소들이다. 따라서 기존의 합숙채점 방식과 자동채점 프로그램을 활용한 채점 방식에서 비용 차이가 발생할 수 있는 채점 절차와 인적 구성을 우선 비교하고 채점에 소요되는 비용을 산출하였다.

가. 인간채점과 자동채점의 절차 및 인적 구성 비교

2014년 중학교 국가수준 학업성취도 평가의 사회 교과 서답형 답안을 기존의 합숙채점 방식으로 채점한 경우와 자동채점 프로그램을 활용하여 채점한 경우의 채점 절차와 인적 구성을 비교하였다. 2014년 국가수준 학업성취도 평가의 서답형 문항에 대한 채점은 온라인 인간채점 방식으로 진행되었다. 특히, 사회 교과의 표집채점은 2박 3일 동안 합숙 장소에 채점자들을 소집하여 채점자 훈련을 시키고, 학생 답안을 2명의 채점자에게 임의로 할당하여 독립적으로 채점하게 하는 방법으로 이루어졌다. 이때 채점자들은 학생들의 답안지에 직접 수기로 점수를 기입하는 것이 아니라, 학생 답안의 스캐닝 이미지를 온라인으로 할당받아 점수를 직접 입력하였다. 채점자들이 1차 채점을 완료한 후에 점수 차이가 있는 답안에 대해서는 2차 채점을 진행하였고, 2차 채점을 완료한 후에도 점수 차이가 있는 답안에 대해서는 3차 채점이 이루어졌다. 3차 채점에서도 점수 차이가 있는 답안에 대해서는 기획위원이 최종점수를 결정하였다.

2014년 중학교 사회 교과의 서답형 문항에 대한 합숙채점에서 〈표 2〉와 같이 채점팀이 구성되었다. 사회 교과는 하위 영역인 지리, 역사, 일반사회로 나누어 각각 기획위원 1명, 채점자 11명이 배속되어 채점을 진행하고 이를 평가위원 1명과 채점 보조 1명이 지원하여, 총 38명으로 채점팀이 구성되었다.

〈표 2〉 2014년 사회 교과 합숙채점 인원 구성

(단위: 명)

영역(문항 번호)	기획위원	채점자	평가위원	채점 보조	총원
지리 (1, 2, 3)	1	11	1	1	38
역사 (4, 5, 6)	1	11			
일사 (7, 8, 9)	1	11			

한편, 2014년 중학교 사회 교과의 서답형 답안에 대해 자동채점 방식을 적용하기 위하여, 합숙채점 장소에서 입력팀을 구성하여 학생 답안의 이미지를 보고 답안을 입력하고 추후 교과 전문가 1명이 자동채점 프로그램을 활용하여 전체 답안을 채점하였다. 이때 교과 전문가가 합숙채점의 채점 기준을 그대로 적용하기 위하여 합숙채점의 기획위원 3명과 별도의 협의를 진행하였다. 〈표 3〉은 자동채점팀의 인적 구성으로, 총 11명이 참여하였다.

〈표 3〉 2014년 사회 교과 자동채점 인원 구성

(단위: 명)

영역(문항 번호)	입력자	입력팀장	채점자	총원
지리 (1, 2, 3)	6 (2명 3개조)	1	4 (기획위원 3명 포함)	11
역사 (4, 5, 6)				
일사 (7, 8, 9)				

자동채점을 위해 필요한 인력과 인간채점에 필요한 인력을 단순하게 비교하였을 때, 자동채점 인력은 인간채점 인력의 $\frac{1}{3}$ 이하 수준인 것으로 나타났다. 그리고 자동채점의 경우, 채점자 4명을 제외하고 다른 인원에게는 별도의 교과 전문성을 요구하지 않는다. 그러나 인간채점의 경우 박사급 기획위원 및 평가위원 4명, 채점자인 중등교사 33명, 보조인력 1명으로 구성되어 있어서 채점에 필요한 인력과 그 수준에 큰 차이가 있다.

나. 채점자 훈련

2014년 학업성취도 평가의 사회 서답형 문항에 대한 합숙채점의 경우, 채점자들이 본 채점에 들어가기 전에 약 5시간 정도 채점자 훈련을 받았다. 이 채점자 훈련은 다수의 채점자들이 동일한 채점 기준에 의거하여 답안을 채점할 수 있도록 채점자들을 훈련하는 것을 목적으로 한다. 대규모 시험에서는 다수의 채점자들이 답안을 나누어 채점하게 되므로, 채점자들이 채점 기준을 다르게 해석할 경우 동일한 답안을 다르게 채점할 수 있다. 따라서 채점자 훈련은 채점자들 간의 채점의 일관성을 최대한 유지하기 위해 이루어지는 사전 조치라고 할 수 있다.

일반적으로, 채점자 훈련은 채점자들이 채점 기준을 공유하는 과정과 연습 채점을 통해 채점자 신뢰도를 확보하는 과정으로 나누어 볼 수 있다. 학업성취도 평가의 채점자 훈련도 이를 따르고 있다. 우선, 채점자들은 출제자가 제시한 채점 기준에 대해 논의하고 이를 수정하거나 명료화하면서 공유하는 시간을 가진다. 그리고 채점자들은 훈련받은 채점 기준에 의거하여 학생 답안 10개를 채점하고, 그들의 점수가 기준점수와 어느 정도 일치하는지를 점검받는다. 채점자 신뢰도가 일정 수준에 이르지 못할 경우, 채점자는 채점 기준에 대해 다시 한 번 훈련을 받은 후에 새로운 학생 답안 10개를 채점하고 다시 채점자 신뢰도를 점검받는다. 채점자들은 채점자 신뢰도가 일정 수준에 이를 때까지, 채점 기준에 대한 훈련과 연습 채점을 최대 3회 반복한다. 대부분의 채점자들은 채점자 훈련 단계에서 매우 높은 채점자 신뢰도를 보여주었다. 그럼에도 불구하고, 채점자들은 본 채점 과정에서 예상하지 못한 다양한 답안을 처리하기 위해 채점을 중단하고, 채점 기준을 수정하거나 명료화하기 위해 협의하는 과정을 반복하였다. 즉, 제한된 학생 답안을 사용하여 채점자 훈련을 철저하게 시행한다고 하더라도, 다양한 형태의 답안을 처리해야 하는 본 채점에서 채점 기준을 수정하거나 명료화해야 하는 경우들이 빈번하게 발생할 수 있다. 특히, 새로운 답안을 처리하는 과정에서 채점 기준을 수정하게 되는 경우 이미 채점한 답안을 다시 찾아서 채점해야 하는 부담이 생기는 경우도 발생하였다.

한편, 자동채점의 경우에는 합숙채점에서 사용한 채점 기준을 동일하게 적용하기 위하여 교과 전문가 1명이 합숙채점 기획위원 3명과 협의를 하였다. 자동채점의 경우에는 다수의 채점자가 아닌 교과 전문가가 자동채점 프로그램으로 답안들을 내용별·순위별로 분류하고 한꺼번에 채점하는 절차를 밟는다. 따라서 자동채점에서는 채점자들 간에 채점 기준을 일치시키기 위한 별도의 훈련이 필요하지 않으며, 채점 과정에서 새로운 답안 때문에 채점 기준을 수정하지는 않는다. 요컨대, 인간채점은 다수의 채점자들이 동일한 채점 기준에 의거하여 채점할 수 있도록 본 채점에 들어가기 전에 반드시 채점자 훈련을 실시해야 하지만, 자동채점은 전체 답안을 내용별·순위별로 분류하여 한꺼번에 채점할 수 있어서 별도의 채점자 훈련을 시행할 필요가 없다. 즉, 자동채점은 다수의 채점자를 훈련시키는 과정을 생략할 수 있다는 점에서 채점에 소요되는 시간, 인력, 비용을 절감할 수 있을 뿐만 아니라, 채점 도중에 변경되는 채점 기준으로 인한 재채점의 부담을 덜 수 있다.

다. 인간채점과 자동채점의 비용 분석

2014년 중학교 학업성취도 평가의 사회 교과 서답형 답안은 기존의 합숙채점 방식에 따라 채점이 이루어졌고, 이와 함께 답안지에 대한 보안 문제로 동일한 합숙 장소에서 자동채점이 별도로 이루어졌다. 자동채점 방식에 맞도록 채점 본부가 운영된다면 합숙비용의 상당 부분은 불필요하나, 여기서는 학업성취도 평가의 합숙채점 방식을 유지한 채 사회 교과의 서답형 답안에 대해 자동채점을 적용하는 경우에 한하여 인간채점과 자동채점의 비용을 비교하여 분석하였다.

합숙채점에 소요되는 주요 비용은 채점자들의 인건비와 숙식비라고 할 수 있다. 〈표 4〉는 사회 교과의 합숙채점팀의 인건비, 숙박비, 식비를 정리한 것이다. 사회 교과의 합숙채점팀은 모두 38명으로 2박 3일 동안 7,442명의 학생 답안을 채점하였다. 채점본부 운영 및 장비 대여 등의 기타 비용을 제외할 경우, 합숙채점팀 운영에 들어간 비용은 참여 인력의 3일 인건비, 3일 숙박비, 7회 식비라고 할 수 있으므로 총 비용은 모두 35,400,000원 정도로 볼 수 있다.

〈표 4〉 2014년 사회 교과 합숙채점 비용

(단위: 천 원)

직책	인원 (명)	인건비			숙박비			식비			합계
		수당	일	소계	단가	일	소계	단가	식	소계	
기획위원	3	270	3	2,430	70	3	630	15	7	315	3,375
평가위원	1	270	3	810	70	3	210	15	7	105	1,125
채점자	33	200	3	19,800	70	3	6,930	15	7	3,465	30,195
채점보조	1	130	3	390	70	3	210	15	7	105	705
합계	38			23,430			7,980			3,990	35,400

자동채점에 소요되는 비용은 답안 입력자 및 채점자의 인건비와 숙식비라고 할 수 있다. 〈표 5〉는 사회 자동채점팀의 인건비, 숙박비, 식비를 정리한 것이다. 답안을 입력하고 채점에 참여한 인력은 모두 11명으로 입력팀장 1명과 입력자 6명이 5박 6일에 걸쳐 7,442명의 학생 답안을 입력하였다. 사회 서답형 답안을 입력하는 비용은 입력팀장과 입력자의 6일 인건비, 6일 숙박비, 16회의 식비라고 할 수 있다. 그리고 자동채점은 합숙 장소가 아닌 별도의 회의실에서 교과 전문가 1명이 합숙채점 기획위원 3명과 함께 채점 기준에 대한 협의를 포함하여 3일 동안 진행하였다. 따라서 자동채점에서 채점자 인건비와 식비는 3일 동안 전문가 협의회에 참석한 것으로 인정하여 경비를 계산하였다. 이에 따라 자동채점의 전체 비용은 〈표 5〉에 제시된 13,500,000원 정도로 산출된다.

〈표 5〉 2014년 사회 교과 자동채점 비용

(단위 : 천 원)

직책	인원 (명)	인건비			숙박비			식비			합계
		수당	일	소계	단가	일	소계	단가	식	소계	
입력팀장	1	270	6	1,620	70	6	420	15	16	240	2,280
입력자	6	130	6	4,680	70	6	2,520	15	16	1,440	8,640
채점자	4	200	3	2,400				20	3	180	2,580
합계	11			8,700			2,940			1,860	13,500

이와 같이 사회 교과의 합숙채점과 자동채점의 인력 운영과 관련된 예산을 비교한 결과, 자동채점 비용이 합숙채점 비용의 38% 이하인 것으로 나타났다. 그리고 자동채점 비용의 약 50%는 현행 합숙채점과 같은 방식으로 답안을 입력했기 때문에 발생한 것으로, 이를 고려하면 실제로는 더욱 절감될 것으로 예상된다. 서답형 답안을 입력하는 것은 별도의 전문성이나 훈련을 필요로 하지 않기 때문에 합숙채점과 같은 형식의 체제가 필요하지 않다. 따라서 서답형 문항에 대한 자동채점 방식을 본격적으로 적용한다면, 자동채점 비용은 현행 합숙채점 비용의 20% 이하 수준으로 떨어질 수 있다.

2014년 학업성취도 평가에서 표집학생들의 서답형 답안들을 채점하기 위해 합숙채점 본부를 운영하였고, 그 기간 동안 중학교 5개 교과, 고등학교 3개 교과에 대한 채점을 동시에 진행하면서 채점자들을 전국 단위의 온라인 채점을 관리할 수 있도록 훈련하였다. 따라서 현행 합숙채점 방식에서 합숙채점 본부의 관리 인력과 지원 인력의 인건비 및 숙식비뿐만 아니라 각종 장비 대여료 등을 포함하는 기타 비용의 크기를 무시할 수 없다. 만약, 모든 교과의 서답형 답안 채점을 합숙채점 방식에서 자동채점 방식으로 전환할 경우, 합숙채점 본부의 운영과 관련된 상당 부분의 비용도 절감할 수 있을 것으로 보인다. 즉, 국가수준 학업성취도 평가의 중학교 5개 교과와 고등학교 3개 교과의 채점 방식을 모두 자동채점으로 전환할 경우, 자동채점 비용은 합숙채점 비용의 약 15% 정도가 될 것으로 추정해볼 수 있다.

2. 채점 결과의 정확성

자동채점 결과의 정확성에 대한 분석은 기본적으로 인간채점 점수와 자동채점 점수 간의 상관관계 및 일치도 결과를 토대로 이루어진다. 여기서 인간채점 점수는 채점자 2명이 1차 채점에서 부여한 점수, 그리고 2명의 채점자 점수가 일치하지 않을 경우 재채점을 통해 최종적으로 확정한 점수를 포함하고 있다. 즉, 채점자1과 채점자2 점수, 최종점수, 자동채점 점수 간의 상관계수, 완전일치도, 근사일치도, 카파계수, 일차가중카파계수, 이차가중카파계수를 비교하여 자동채점 결과의 정확성을 분석하였다.

〈표 6〉은 사회 서답형 14개 문항에 대한 채점자 점수, 최종점수, 자동채점 점수 간의 상관계수를 정리한 것이다. 서답형 14개 문항에 대한 채점자1 점수와 채점자2 점수의 상관계수는 0.97~1.00 사이로 나타났다. 채점자 점수와 자동채점 점수 간의 상관계수를 살펴보면, 채점자1 점수와 자동채점 점수의 상관계수는 0.96~1.00의 범위에 있고, 채점자2 점수와 자동채점 점수 간 상관계수의 범위도 동일하게 나타났다. 이와 같은 수준의 상관계수는 인간채점과 자동채점 사이에 별 차이가 없음을 보여준다. 학업성취도 평가의 최종점수는 채점자들이 여러 회기를 걸쳐 확정한 점수이므로 거의 완전한 기준점수에 가깝다고 할 수 있다. 따라서 최종점수와의 비교는 자동채점의 정확성을 확인하는 데 중요한 준거가 될 수 있다. 최종점수와 자동채점 점수

간의 상관계수는 0.97~1.00 범위인데, 11개 문항에서 두 점수의 상관계수가 1인 것으로 나타났다. 이와 비교하여 최종점수와 채점자1 점수 그리고 최종점수와 채점자2 점수 간의 상관계수는 각각 0.99~1.00, 0.98~1.00인 것으로 나타나 그 차이가 크지 않았다. 결국, 2번 문항을 제외하면 자동채점 점수는 인간채점 점수보다 최종점수와의 상관계수가 같거나 높은 것으로 나타났다.

〈표 6〉 2014년 사회 문항 채점 결과의 상관계수

문항 번호	최종-자동	인간1-자동	인간2-자동	인간1-인간2	최종-인간1	최종-인간2
1	1.00	1.00	1.00	1.00	1.00	1.00
2	0.97	0.96	0.96	0.99	0.99	0.99
3-㉠	1.00	0.99	0.99	0.99	1.00	0.99
3-㉡	1.00	1.00	0.99	0.99	1.00	1.00
4-㉠	1.00	0.99	1.00	0.99	1.00	1.00
4-㉡	1.00	0.99	0.99	0.99	1.00	0.99
5-㉠	1.00	1.00	1.00	0.99	1.00	1.00
5-㉡	0.98	0.98	0.97	0.97	0.99	0.98
6-㉠	1.00	1.00	0.99	0.99	1.00	0.99
6-㉡	1.00	0.99	0.99	0.99	0.99	0.99
7	1.00	1.00	0.99	0.99	1.00	0.99
8-㉠	1.00	1.00	0.99	0.99	1.00	0.99
8-㉡	1.00	1.00	0.99	0.99	1.00	0.99
9	0.99	0.98	0.98	0.98	0.99	0.99

〈표 7〉은 사회 서답형 14개 문항의 채점자 점수, 최종점수, 자동채점 점수 간의 완전일치도를 정리한 것이다. 점수들 간 완전일치도의 범위는 97.29%~99.99%로 나타났다. 최종점수와 자동채점 점수가 완전히 일치하는 비율은 2번 문항을 제외하고 모두 99% 이상인 것으로 나타났으며, 문항 2번의 경우에도 97.76%로 매우 높은 편이었다. 최종점수와 채점자1 점수의 완전일치도는 모든 문항에 대해 99% 이상이었으며, 최종점수와 채점자2 점수의 완전일치도는 98% 이상인 것으로 나타났다. 최종점수와 자동채점의 완전일치도는 2번 문항을 제외한 다른 문항에서 최종점수와 채점자 점수의 완전일치도 평균보다 모두 높은 것으로 나타났다. 이는 자동채점이 인간채점에 비해 결코 떨어지지 않음을 의미한다.

〈표 7〉 2014년 사회 문항 채점 결과의 완전일치도

문항 번호	최종-자동	인간1-자동	인간2-자동	인간1-인간2	최종-인간1	최종-인간2
1	99.99	99.89	99.91	99.83	99.91	99.92
2	97.76	97.47	97.29	98.68	99.40	99.27
3-㉠	99.97	99.80	99.65	99.50	99.83	99.68
3-㉡	99.88	99.83	99.66	99.73	99.95	99.79
4-㉠	99.95	99.76	99.81	99.68	99.81	99.87
4-㉡	99.87	99.68	99.57	99.52	99.81	99.70
5-㉠	99.97	99.84	99.79	99.68	99.87	99.81
5-㉡	99.11	99.07	98.93	97.93	99.09	98.79
6-㉠	99.99	99.84	99.80	99.66	99.85	99.81
6-㉡	99.97	99.66	99.80	99.49	99.69	99.80
7	99.62	98.79	98.44	97.80	99.01	98.76
8-㉠	99.99	99.95	99.88	99.85	99.96	99.89
8-㉡	99.99	99.97	99.91	99.91	99.99	99.92
9	99.34	98.71	98.70	98.41	99.23	99.17

* 최종 : 3라운드에 걸쳐 복수의 채점자가 채점한 인간채점 점수
* 인간1, 인간2 : 1라운드에서 채점한 채점자1과 채점자2의 인간채점 점수

〈표 8〉은 사회 서답형 14개 문항 중에서 배점이 2점 이상인 5개 문항에 대해 근사일치도를 계산한 것이다. 점수들 간의 근사일치도의 범위는 98.15%~100%인데, 앞의 완전일치도가 매우 높아서 근사일치도는 크게 증가하지 않았다. 문항 2번을 제외한 나머지 문항에서 최종점수와 자동채점 점수 간의 근사일치도가 최종점수와 채점자 점수 간 근사일치도보다 높게 나타났다. 또한, 2번 문항을 제외한 나머지 문항에서 자동채점 점수와 채점자 점수 간 근사일치도가 채점자 점수들 간의 근사일치도보다 근소하지만 높게 나타났다. 이와 같은 근사일치도 차이는 자동채점이 인간채점에 비해 떨어지지 않음을 보여준다.

〈표 8〉 2014년 사회 문항 채점 결과의 근사일치도

문항 번호	최종 -자동	인간1-자동	인간2-자동	인간1-인간2	최종-인간1	최종-인간2
1	100.00	99.95	99.96	99.91	99.95	99.96
2	98.29	98.15	98.16	99.40	99.69	99.68
5-㉡	99.91	99.89	99.77	99.19	99.61	99.61
7	99.97	99.96	99.85	99.81	99.96	99.88
9	99.95	99.88	99.88	99.85	99.93	99.93

〈표 9〉는 사회 서답형 문항에 대한 채점자 점수, 최종점수, 자동채점 점수 간의 카파계수를 정리한 것인데, 카파계수의 범위는 0.92~1.00인 것으로 나타났다. 최종점수와 자동채점 점수 간의 카파계수는 0.94~1.00으로 나타났으며, 2번과 5-㉡번 문항을 제외한 나머지 문항에서 최종점수와 채점자 점수 간 카파계수보다 같거나 높게 나타났다. 그리고 채점자1 점수와 자동채점 점수, 채점자2 점수와 자동채점 점수 사이의 카파계수는 0.92~1.00으로, 2번과 5-㉡번 문항을 제외한 나머지 문항에서 채점자 점수들 간의 카파계수보다 같거나 높게 나타났다.

〈표 9〉 2014년 사회 문항 채점 결과의 카파계수

문항 번호	최종-자동	인간1-자동	인간2-자동	인간1-인간2	최종-인간1	최종-인간2
1	1.00	1.00	1.00	1.00	1.00	1.00
2	0.96	0.96	0.95	0.98	0.99	0.99
3-㉠	1.00	0.99	0.99	0.99	1.00	0.99
3-㉡	1.00	1.00	0.99	0.99	1.00	1.00
4-㉠	1.00	0.99	1.00	0.99	1.00	1.00
4-㉡	1.00	0.99	0.99	0.99	1.00	0.99
5-㉠	1.00	1.00	1.00	0.99	1.00	1.00
5-㉡	0.94	0.93	0.92	0.95	0.98	0.97
6-㉠	1.00	1.00	0.99	0.99	1.00	0.99
6-㉡	1.00	0.99	0.99	0.99	0.99	0.99
7	0.99	0.98	0.98	0.97	0.99	0.98
8-㉠	1.00	1.00	0.99	0.99	1.00	0.99
8-㉡	1.00	1.00	0.99	0.99	1.00	0.99
9	0.98	0.96	0.96	0.95	0.98	0.97

분류의 범주가 2개인 경우 일치와 불일치로만 구분되기 때문에 카파계수와 가중카파계수는 동일한 값을 갖게 된다. 따라서 가중카파계수는 문항의 배점이 2점 이상인 5개 문항에 대해서만 제시하였다. 〈표 10〉는 배점이 2점 이상인 5개의 문항에 대해 채점자 점수, 최종점수, 자동채점 점수 간의 일차가중카파계수를 정리한 것인데, 전체적으로 일차가중카파계수의 범위는 0.95~1.00인 것으로 나타났다. 최종점수와 자동채점 점수 간의 일차가중카파계수는 0.96~1.00으로 나타났으며, 2번과 5-㉡번 문항을 제외한 나머지 문항에서 최종점수와 채점자 점수 간의 일차가중카파계수보다 같거나 높게 나타났다. 그리고 자동채점 점수와 채점자 점수 간의 일차가중카파계수는 0.95~1.00으로 나타났으며, 2번과 5-㉡번 문항을 제외한 나머지 문항에서 채점자 점수들 간의 일차가중카파계수보다 같거나 높게 나타났다.

〈표 10〉 2014년 사회 문항 채점 결과의 일차가중카파계수

문항 번호	최종 -자동	인간1-자동	인간2-자동	인간1-인간2	최종-인간1	최종-인간2
1	1.00	1.00	1.00	1.00	1.00	1.00
2	0.96	0.96	0.96	0.98	0.99	0.99
5-㉡	0.96	0.96	0.95	0.96	0.98	0.98
7	1.00	0.99	0.99	0.98	0.99	0.99
9	0.99	0.97	0.97	0.97	0.98	0.98

〈표 11〉은 배점이 2점 이상인 5개의 문항에 대해 채점자 점수, 최종점수, 자동채점 점수 간의 이차가중카파계수를 정리한 것인데, 전체적으로 이차가중카파계수의 범위는 0.96~1.00인 것으로 나타났다. 최종점수와 자동채점 점수 간의 이차가중카파계수는 0.96~1.00으로 나타났으며, 2번과 5-㉡번 문항을 제외한 나머지 문항에서 최종점수와 채점자 점수 간의 이차가중카파계수보다 같거나 높게 나타났다. 그리고 자동채점 점수와 채점자 점수 간의 이차가중카파계수는 0.95~1.00으로 나타났으며, 2번과 5-㉡번 문항을 제외한 나머지 문항에서 채점자 점수들 간의 이차가중카파계수보다 같거나 높게 나타났다.

〈표 11〉 2014년 사회 문항 채점 결과의 이차가중카파계수

문항 번호	최종 -자동	인간1-자동	인간2-자동	인간1-인간2	최종-인간1	최종-인간2
1	1.00	1.00	1.00	1.00	1.00	1.00
2	0.96	0.96	0.96	0.98	0.99	0.99
5-㉡	0.96	0.96	0.95	0.96	0.98	0.98
7	1.00	0.99	0.99	0.98	0.99	0.99
9	0.99	0.97	0.97	0.97	0.98	0.98

요약하면, 2014년 중학교 국가수준 학업성취도 평가의 사회 교과 서답형 14개 문항에 대한 최종점수와 자동채점 점수 간 상관계수는 0.97~1.00으로, 3개 문항을 제외하면 두 점수 간 상관계수는 1인 것으로 나타났다. 서답형 14개 문항에 대한 최종점수와 자동채점 점수 간 완전일치도는 97.76%~99.99%, 카파계수는 0.94~1.00, 그리고 배점이 2점 이상인 5개 문항의 근사일치도는 98.29%~99.99%, 일차가중카파계수와 이차가중카파계수는 모두 0.96~1.00인 것으로 나타났다. 비교적 자동채점 정확성이 떨어지는 서답형 문항 2번과 5-㉡번 문항도 최종점수와 자동채점 점수 간의 일치 정도와 상관관계가 결코 낮은 것이 아니다. 이들 문항을 제외할 경우, 완전일치도는 99.34% 이상, 카파계수는 0.98 이상, 상관계수는 0.99 이상으로 나타났다. 따라서 전반적으로 단어·구 수준 서답형 문항에 대한 자동채점 프로그램의 정확도는 매우 높은 편이라고 할 수 있다.

Ⅴ. 결론 및 제언

본 연구는 2012년부터 개발하기 시작한 단어·구 수준 자동채점 프로그램을 활용하여 지필평가 서답형 문항을 채점하는 방식의 실용성을 확인하기 위해 채점 비용의 효율성과 채점 결과의 정확성 측면으로 나누어 분석하였다. 이를 위하여, 2014년 중학교 국가수준 학업성취도 평가의 사회 교과 서답형 문항을 대상으로 기존의 합숙채점 방식과 자동채점 방식을 적용하여 채점하고 그 결과를 비교하였다.

자동채점 프로그램을 이용한 채점 방식은 기존의 합숙채점 방식에 비하여 인력은 약 $\frac{1}{3}$수준으로 절감할 수 있었으며, 채점에 필요한 인력의 수준도 채점이 가능한 전문성보다는 답안을 정확하고 빠르게 입력할 수 있는 능력 정도를 요구하였다. 또한, 자동채점 방식과 기존 합숙채점 방식의 인건비와 숙식비를 비교하였을 때, 자동채점은 합숙채점 비용의 60% 이상을 절감할 수 있는 것으로 나타났으며 자동채점에 최적화된 채점 방식을 적용할 경우 합숙채점 비용의 80% 이상을 절감할 수 있는 것으로 나타났다. 따라서 단어·구 수준의 지필평가 답안에 대한 자동채점 방식은 기존의 합숙채점 방식에 비하여 채점 비용의 효율성이 있는 것으로 확인되었다. 또한, 합숙채점 최종점수와 자동채점 점수 간의 상관계수와 일치도가 매우 높게 나타났다. 2014년 학업성취도 평가 중3 사회 서답형 문항 14개에 대한 최종점수와 자동채점 점수 간 상관계수는 0.97~1.00, 완전일치도는 97.76%~99.99%, 카파계수는 0.94~1.00인 것으로 나타났고, 5개 문항에 대한 최종점수와 자동채점 점수 간 근사일치도는 98.29%~100%, 일차가중카파계수와 이차가중카파계수는 0.96~1.00인 것으로 나타났다. 이것은 자동채점의 결과가 채점 전문가들이 부여한 점수와 매우 유사하다는 것을 의미하며, 단어·구 수준 서답형 문항의 자동채점 결과가 채점 전문가만큼 정확하다는 것을 확인할 수 있었다. 결론적으로, 단어·구 수준의 답안을 요구하는 지필평가의 서답형 문항에 대해서 자동채점 프로그램을 활용하여 채점하는 것이 채점 비용의 효율성과 채점 결과의 정확성 측면에서 실용성이 있는 것으로 확인되었다.

한국어 서답형 문항 자동채점 프로그램의 실용성을 확인하는 과정에서 향후 보완이 필요한 부분을 발견할 수 있었다. 첫째, 자동채점의 경우 인간채점 점수와 비교하여 문항 특성에 따라 채점 비율과 채점 오류에 편차를 보였다. 물론 채점자들도 문항 특성에 따라 채점 비율과 오류에 차이가 나타났지만 그 편차는 자동채점에 비해 크지 않았다. 즉, 자동채점은 특정 문항에 대해서는 오류가 거의 없이 완벽하게 기능했지만 문항의 답안 복잡성, 답안 유형 수, 배점 분포 등에 따라 채점 비율이 떨어지기도 한다. 따라서 자동채점에 적용하고자 하는 평가 도구별, 교과별, 문항별 특성을 살펴 프로그램의 기능을 보강할 필요가 있었다. 둘째, 현재의 한국어 자연언어처리 기술의 한계에 따라 채점 오류 가능성은 언제나 수반되어 나타날 수 있다. 먼저, 개념 기반 채점 단계나 의미 유사도에 의한 군집화 과정에서 형태소 분석, 유사어 처리, 언어 관계

처리, 각종 사전 등이 동원되는데 현재의 자연언어처리 및 지식베이스의 한계로 채점 오류가 나타날 개연성이 있다. 따라서 자연언어처리 기술을 고도화하고 지식베이스를 풍부하게 구축하면서 자동채점의 오류 가능성을 줄여가는 것과 함께, 프로그램 진행 과정 내에서 채점 전문가가 개입하여 자동채점 결과를 모니터링하고 피드백해 주는 현재의 반복적 순환 흐름을 보다 강화할 필요가 있다. 마지막으로, 컴퓨터 기반 평가가 시행되지 않고 있는 상황에서 학생 답안 입력의 오류가 채점 오류로 나타날 수 있다. 사람이 수기 답안을 입력하는 과정에서 답안 판독이나 타이핑 오류가 발생할 수 있는데, 이러한 문제는 현재 지필평가 체제에서 자동채점 프로그램을 활용하는 데 가장 큰 제약 사항이라고 할 수 있다. 따라서 지필평가 상황에서 자동채점 프로그램을 활용하려면 학생이 답안을 보조 단말기 등으로 직접 입력하거나 또는 복수의 개별 입력자가 학생 답안을 입력하고 교차 검토하는 방식 등을 통해 입력 오류를 최소화하는 방안을 강구해야 한다. 물론 이보다는 서답형 답안을 자동으로 인식할 수 있는 별도의 프로그램 개발 노력을 병행하는 것이 좀 더 적극적인 대처 방안일 것이다.

인간채점 점수와 자동채점 점수의 불일치는 상당 부분 채점자의 채점 오류에 기인하기도 한다. 채점자들도 자동채점과는 다른 이유로 채점 오류를 발생시킨다. 즉, 채점자 간 및 채점자 내 일관성 문제, 채점 피로 누적에 따른 채점 실수 등이 종종 발견된다. 이러한 문제에 대해 자동채점 프로그램은 복수의 채점자 중 하나로 대체·기능할 수 있음은 물론, 채점자의 채점 일관성 및 채점 오류 문제를 검토할 수 있는 유용한 보조적 도구가 될 수 있다. 또한 자동채점 프로그램은 매회 실시되는 시험의 학생 답안과 그 채점 정보를 누적적으로 기록·저장할 수 있다. 따라서 학생에게는 맞춤형 학습 정보를 즉시 제공하고 교사들에게는 학생의 반응이나 오개념 유형을 빈도별로 파악하게 함으로써 유용한 교수·학습 지원 도구로 기능할 수 있다. 이러한 자동채점 프로그램의 부가적인 이점으로 대규모 평가는 물론 학교 단위 평가에서도 그 활용 가능성이 높고, 특히 디지털교과서를 활용한 수업, 온라인 수업, 방과후 사이버 학습 등 새로운 학습 환경에서 더욱 유용할 것으로 판단된다.

참 고 문 헌

김경희, 김완수, 김동영, 김종훈, 김미경, 최인봉, 신동광, 박인용, 이인호, 신진아, 최인선, 송미영, 한정아, 김희경, 한경택, 박거도(2013). 컴퓨터 기반 국가수준 학업성취도 평가 도입 방안. 한국교육과정평가원 연구보고 CRE 2013-5.

노은희(2014). 국가수준 학업성취도 평가 국어 서답형 문항의 자동채점 결과 분석. **국어교육학연구, 49**(2), 85-111.

노은희, 심재호, 김명화, 김재훈(2012). 대규모 평가를 위한 서답형 문항 자동채점 방안 연구. 한국교육과정평가원 연구보고 RRE 2012-6.

노은희, 김명화, 성경희, 김학수(2013). 대규모 평가를 위한 서답형 문항 자동채점 프로그램 정교화 및 시범 적용. 한국교육과정평가원 연구보고 RRE 2013-5.

노은희, 이상하, 임은영, 성경희, 박소영(2014). 한국어 서답형 문항 자동채점 프로그램 개발 및 실용성 검증. 한국교육과정평가원 연구보고 RRE 2014-6.

송미영, 박혜영, 임해미, 최혁준(2013). 21세기 역량 평가를 위한 OECD PISA의 변화 방향과 대응 방안. 한국교육과정평가원 2013 KICE 이슈페이퍼 연구자료 ORM 2013-57-13.

이상하, 박도영, 박상욱, 최인봉, 구남욱, 이은경(2014). 미래사회 핵심역량 교수·학습 지원을 위한 교육평가 정책의 방향. 한국교육과정평가원 연구보고 RRE 2014-14.

Educational Testing Service (2012). *Coming Together to Raise Achievement: New Assessments for the Common Core State Standards*. Princeton, NJ: Center for K-12 Assessment & Performance Management. Retrieved from http://www.k12center.org/rsc/pdf/Coming_Together_April_2012_Final.PDF

Shermis, M. D. (2010). Automated essay scoring in a high stakes testing environment. In V. J. Shute., & B. J. Becker. (Eds.), *Innovative assessment for the 21st century* (pp. 167-185). New York, NY: Springer.

Shermis, M. D., & Hamner, B. (2013). *Contrasting state-of-the-art automated scoring of essays: Analysis*. Retrieved from http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

Topol, B., Olson, J., & Roeber, E. (2014). *Pricing study Machine scoring of student essays*. Retrieved from http://cdno4.gettingsmart.com/wp-content/uploads/2014/02/ASAP-Pricing-Study-Final.pdf

U.S. Department of Education (2010, September 2). *Beyond the Bubble Tests: The Next Generation of Assessments - Secretary Arne Duncan's Remarks to State Leaders at Achieve's American Diploma Project Leadership Team Meeting*. Retrieved from

http://www.ed.gov/news/speeches/beyond-bubble-tests-next-generation-assessments-secretary-arne-duncans-remarks-state-leaders-achieves-american-diploma-project-leadership-team-meeting

WCAL (n.d). *Short Anwser Marking Engines*. Retrieved from http://www.worldclassarena.net/doc/file5.pdf.1~6.

Zhang, M. (2013). *Contrasting automated and human scoring of essays*. Retrieved from http://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf

· *논문접수 : 2015-01-08/ 수정본접수 : 2015-02-05/ 게재승인 : 2015-02-23*

ABSTRACT

Contrasting Automated and Human Scoring for Short-Answer NAEA Questions

Sang-Ha Lee
(Research Fellow, Korea Institute for Curriculum and Evaluation)
Eun-Hee Noh
(Research Fellow, Korea Institute for Curriculum and Evaluation)
Kyung-Hee Sung
(Associate Research Fellow, Korea Institute for Curriculum and Evaluation)

This study aims to evaluate the costs and accuracy of automated scoring for short-answer questions of National Assessment of Educational Achievement (NAEA). To contrast automated and human scoring, both scoring methods were used to grade 14 short-answer questions of the NAEA Social Studies subtest that was taken by 7,442 ninth-grade students in 2014. We analyzed the effects of automated scoring on the costs and accuracy of scoring short-answer questions.

We found that more than 60% of the human scoring costs can be saved by only using the automated scoring program. In addition, more than 80% of the human scoring costs can be saved if the current scoring method of NAEA becomes optimized for automated scoring. Correlation coefficients between human and machine scores for 14 questions ranged from 0.97 to 1, exact agreement ranged from 97.76% to 99.9%, and kappa coefficients ranged from 0.94 to 1. Exact and adjacent agreement for 5 questions ranged from 98.15% to 100%, and linear kappa coefficients and quadratic kappa coefficients ranged from 0.96 to 1. Moreover, results showed that correlations and agreement rates between human and machine scores are as high as the ones between human scores. We concluded that automated scoring can reduce the costs of scoring short-answer NAEA questions without sacrificing the accuracy of scoring. It is suggested that the automated scoring system can be used to replace one or two human raters when scoring the short answer NAEA questions, to monitor human scoring, or to train human raters

Key Words : automated scoring, automatic scoring, machine scoring, scoring short-answer questions, scoring constructed-response items