교육과정평가연구
The Journal of Curriculum and Evaluation
2013, Vol. 16, No. 1, pp. 137~160

기초학력 진단평가 서답형 문항의 자동채점 가능성 탐색[1)]

김 명 화(한국교육과정평가원 연구위원)*
노 은 희(한국교육과정평가원 연구위원)**
심 재 호(한국교육과정평가원 연구위원)

≪ 요 약 ≫

본 연구의 목적은 초등학교 3학년 기초학습 진단평가의 서답형 문항을 대상으로, 2012년에 한국교육과정평가원에서 개발한 한국어 단어·구 수준 서답형 자동채점 프로그램을 적용하여 자동채점 가능성을 탐색하려는 것이다. 이를 위해 읽기, 쓰기 서답형 문항 중 11문항을 선정하고 각 506~929명의 답안 자료를 대상으로 자동채점 프로그램을 적용하여 자동채점 단계별 정답 수, 오답 수, 미채점 수, 채점 비율을 계산하였다. 이와 함께 각 문항별로 Kappa계수와 상관계수를 계산하여 제시하였다.

한국어 서답형 자동채점 프로그램을 활용하여 채점한 결과 초3 기초학습 진단평가의 단어·구 수준의 단답형 문항은 대부분 자동채점이 가능한 것으로 나타났다. 즉, 초3 기초학습 진단평가의 단어·구 수준의 한국어 서답형 문항에 대한 자동채점 프로그램의 채점 비율과 채점자와의 일치도는 적정한 수준이었고, 일부 채점 오류가 있었으나 그 비율은 적은 편이었다. 채점 오류 중 가장 많은 것은 철자 오류이고, 나머지는 유사어를 인지하지 못하거나 다른 기호나 용어가 포함되어 있는 경우가 대부분이었다.

초3 기초학습 진단평가에 자동채점 프로그램을 적용할 경우, 담임교사들이 직접 채점하므로 맞춤형 교수학습이 가능하도록 학생별 정·오답 처리 결과를 피드백할 수 있는 기능을 추가하고, 교사들이 자동채점 프로그램을 쉽게 활용할 수 있도록 편의성이 높은 인터페이스를 추가로 개발할 것을 제안하였다.

주제어 : 기초학력 진단평가, 서답형 문항, 자동채점, 대규모 평가, 초등학교 3학년 기초학습 진단평가, 읽기, 쓰기

1) 이 연구는 한국교육과정평가원에서 수행한 '대규모 평가를 위한 서답형 문항 자동채점 방안 연구'(노은희 외, 2012)의 일부 내용을 수정·보완한 것임.

* 제1저자, hwa@kice.re.kr
** 교신저자, noro@kice.re.kr

Ⅰ. 서 론

초등학교 3학년 기초학습 진단평가(이하 초3 진단평가)는 기초학력 향상을 위한 지원 체제의 일환으로, 학습 부진 학생에게 맞춤형 프로그램을 제공하고 모니터링 체제를 구축하기 위하여 학습 부진 학생을 선별하려는 목적으로 시행되고 있다(김명화 외, 2011). 즉, 학습 부진 학생의 심각한 학습 결손과 누적을 예방하고 학생의 특성에 맞는 처방을 제공하기 위하여, 진단평가를 통하여 학생의 상태를 진단하고 그 결과에 따라 보정 학습을 하도록 하려는 것이다. 따라서 초3 진단평가에서는 초등학교 3학년을 대상으로 학생들의 기초학습 성취 여부를 판별하고 맞춤형 보정 학습이 가능하도록, 각 학생별 진단 정보를 제공하도록 하고 있다. 2002년 이후 시행되고 있는 초3 진단평가는 최근 학생들의 창의성과 핵심 역량을 강조하는 세계적인 추세를 반영하여 선택형 외에 서답형 또는 수행평가를 병행하여 출제하고 있다. 그러나 서답형 문항을 좀 더 적극적으로 활용하기 위해서는 채점의 부담을 경감하고 채점의 신뢰도 문제를 해결할 필요가 있다. 초3 진단평가의 경우 시험 시행 후 담임교사가 채점하도록 하고 있으므로 채점 비용은 크게 문제가 되지 않으나, 서답형 문항의 경우 채점 시간이 많이 걸리고 채점 기준표에 의해 채점하더라도 채점 결과가 일관되지 않을 가능성이 있다.

한편, 교육과학기술부에서는 스마트 교육의 일환으로 초3 진단평가를 2014년부터 온라인 평가로 전환하기 위해 준비하고 있다(반재천, 김선, 2012). 온라인 평가의 장점은 즉각적인 피드백과 개인에 맞는 맞춤형 피드백을 제공할 수 있다는 것인데, 여기서 온라인 초3 진단평가의 효율성을 제고하기 위해서는 서답형 문항의 자동채점 프로그램 도입이 요구된다. 만일 자동채점이 되지 않는다면 온라인 평가 시행 후 교사들은 각 학생의 답안을 보고, 채점하여 그 결과를 입력하는 과정을 거쳐야 하므로 채점 과정에 시간이 많이 걸리게 된다. 자동채점 프로그램이 도입된다면 이러한 비효율적인 측면을 해소할 수 있다. 이와 더불어 학생들의 서답형 답안을 미리 입력된 기준(코드)에 따라 분석하여 학생들에게 오답노트를 제공한다면 학생들의 부진 요인과 오답 원인도 쉽게 파악할 수 있다는 측면에서 온라인 평가의 목적을 실현하고 효율성을 높일 수 있다는 장점도 있다.

최근 국내외에서는 서답형 문항의 자동채점 및 프로그램 개발에 관한 연구가 활발히 진행되고 있다. 미국에서는 1960년대부터 에세이 자동채점에 대한 연구가 진행되어 왔고, 일부 프로그램은 대규모 평가에서 인간 채점자를 보조하는 용도로 사용되고 있다. 이에 비하여, 우리나라의 경우 국어의 특성상 자연어 처리를 위한 지식베이스 연구가 미진하여 자동채점화 방안을 본격적으로 논의하지 못하고 있는 실정이다. 한국어 자동채점 프로그램 개발과 관련하여 정동경(2001), 조우진(2006), 강원석(2011) 등과 같은 연구가 있으나, 이는 대학교 기말고사를 대상으로 한 소규모 실험 연구로 타당도와 신뢰도가 검증되지 않았고 주로 컴퓨터 프로그램밍 관

련 내용으로 다른 영역에 일반화하여 사용하기는 어렵다. 한편 성태제 외(2010)의 연구에서는 국가수준 학업성취도 평가의 자동채점 가능성을 탐색하기 위하여 자동채점이 가능한 문항과 그렇지 않은 문항을 구별하고 문항 특성에 대하여 분석한 바 있으나 본격적으로 프로그램을 개발하지는 못했다(성태제 외, 2010).

2012년에 한국교육과정평가원에서는 대규모 평가에서 서답형 문항 채점에 소요되는 인력과 채점 시간, 채점 비용 등을 절감할 수 있도록 현재의 기술 상황이 허락하는 한도에서 단어·구 수준의 단답형 답안이라도 컴퓨터로 자동채점할 수 있는 프로그램을 개발하였다(노은희 외, 2012). 본 연구에서는 한국교육과정평가원에서 개발한 한국어 서답형 문항 자동채점 프로그램을 활용하여 초3 진단평가 서답형 문항을 채점하고, 자동채점 가능성을 탐색하려 한다.

2011년 초3 진단평가의 서답형 문항을 정리하면 〈표 1〉과 같다. 초3 진단평가는 동등화를 위하여 동형 검사를 개발하므로 매 해 개발되는 서답형 문항 수와 형태는 크게 변화가 없다. 또한 서답형 문항 중 상당 부분이 한 문장 이내의 단어·구 수준 정도의 답안을 요구하고 있고 향후 온라인 컴퓨터 기반 평가를 시행하기 위하여 준비 중이므로, 초3 진단평가에 자동채점 프로그램이 활용될 경우 효용성이 높을 것으로 판단된다. 이에 우선 현재의 기술로 해결 가능한 짧은 답안부터라도 자동으로 채점할 수 있는 프로그램 개발을 시도하고 채점 프로그램의 타당성을 탐색해 볼 필요가 있다.

〈표 1〉 초3 진단평가의 단어·구 수준 서답형 문항 현황 (2011년 기준)

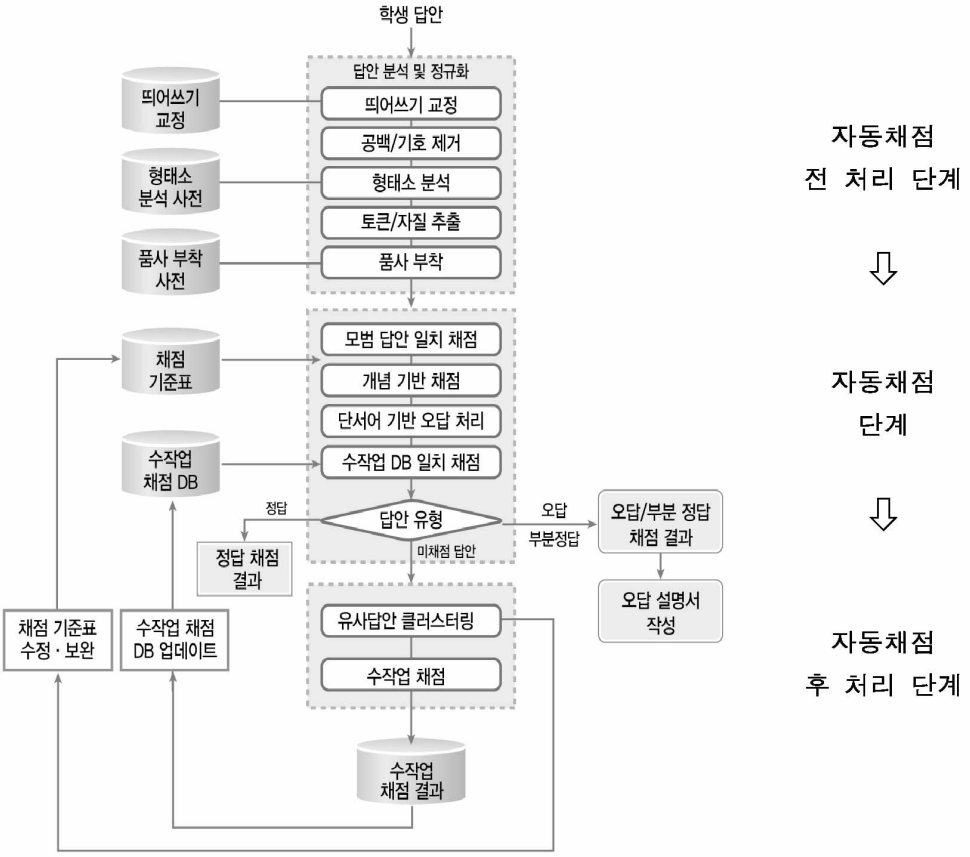
하위 영역	서답형 문항 수	
	문항 수	단어·구 수준 답안
쓰기	16문항	15문항
읽기	7문항	2문항

Ⅱ. 한국어 서답형 문항 자동채점 프로그램 개요[2)]

한국어 서답형 자동채점 프로그램의 구조 및 채점 방법을 간략히 소개하면 다음과 같다. 한국어 자동채점 프로그램은 평가 문항과 관련하여 미리 제시된 채점 기준표의 요구 사항을 컴퓨터가 이해할 수 있는 용어로 변환하기 위하여 자연언어 처리 기술을 활용하여 정답과 유사 답안,

2) 본 프로그램은 한국교육과정평가원 '대규모 평가를 위한 서답형 문항 자동채점 방안 연구'의 일환으로 개발한 것이다(노은희 외, 2012년 참조). 프로그램 개발은 국민대 강승식 교수 자연언어처리 연구실에서 맡아 진행하였다.

오답 등의 정보를 자동채점에 적합한 형식으로 기술하고 이를 기반으로 한국어 서답형 문항을 자동으로 채점하는 시스템이다. 자동채점 프로그램의 전체 구조도를 제시하면 〔그림 1〕과 같다.



〔그림 1〕 한국어 서답형 문항 자동채점 프로그램 구조도

본 채점 프로그램은 크게 자동채점 전 처리 단계, 자동채점 단계, 자동채점 후 처리 단계로 구분할 수 있다. 자동채점 전 처리 단계는 자동채점을 하기 전 학생 답안을 분석하고 정규화하는 단계이다. 자동채점 단계는 채점 기준표에 따라 작성된 정답 템플릿을 통해 초기 일부 학생 답안에 대하여 채점을 수행하고, 정답과 오답으로 판정하기 어려운 답안에 대해서는 미채점 답안으로 보류하는 단계이다. 자동채점 후 처리 단계는 미채점 답안들의 클러스터링과 수작업 채점[3]을 통해 미채점된 답안의 채점을 마무리하고 이에 따라 새로운 답안들의 자동채점 비율을

3) 수작업 채점이란 철자 오류 등으로 인하여 자동채점 단계에서 미보류로 된 답안에 대해 사람이 채점하여 처리하고, 그 결과를 자동채점 기준에 추가하는 것이다. 뒤에 설명할 인간채점은 자동채점을 전혀

높이기 위해 채점 기준표를 수정·보완하고 수작업 채점 DB를 업데이트하는 단계이다. 이 프로그램은 일부 학생 답안의 채점 결과 분석 및 후처리 단계를 거쳐 정답 템플릿을 보완하여 더 많은 학생 답안의 자동채점을 반복적으로 가능하게 만드는 사이클형 구조로 설계되어 있기 때문에 대규모 평가의 자동채점 프로그램으로 적합성이 높다. 이 프로그램의 각 단계별 특징을 좀 더 구체적으로 설명하면 다음과 같다.

자동채점 전 처리 단계에서는 학생 답안이 자동채점기에 입력되면 자동채점을 수행하기 전 학생 답안에 대하여 분석 및 정규화하는 언어 처리 단계를 거치게 된다. 이 과정에서 채점 옵션에 따라 띄어쓰기 교정, 공백 또는 기호 등이 제거된다. 이후 채점 기준표에 제시된 모범 답안에 대해 형태소 분석, 토큰/자질 추출, 품사 부착 단계를 거치게 되는데 이를 답안 분석 및 정규화 처리라고 한다.

자동채점 단계에서는 답안 정규화 후 정답 템플릿으로 기술된 채점 기준표 내용에 따라 학생 답안에 대하여 모범 답안 일치 채점, 개념 기반 채점, 단서어(cue word) 기반 오답 처리, 수작업 DB 일치 채점 등 총 4단계를 거쳐 자동채점을 수행하고 점수를 부여하게 된다. 모범 답안 일치 채점이란 문자열 일치 점검을 통해 학생 답안이 모범 답안과 완전히 일치하는 경우에만 정답으로 처리하는 것이다. 개념 기반 채점은 고빈도 학생 답안들을 정답 템플릿으로 기술하여 그것과 관련하여 채점하는 것으로, 만일 정답으로 인정할 수 있는 개념이 4개인 경우 이 개념들을 템플릿으로 기술한 후 이 개념과의 일치 및 유사성 여부를 판단하여 채점하는 것이다. 단서어 기반 오답 처리란 단서어 목록에 있는 단어가 존재할 경우 미채점하고, 이 단서어가 없을 경우 오답으로 처리하는 것이다. 만일의 경우 개념에 포함되는 단서어는 정답 또는 부분 정답이 될 가능성이 있기 때문에 그 단서어가 출현하기만 하면 미채점으로 보류하는 것이다. 단서어 기반 채점은 자칫 개념만으로 채점하여 정답 또는 오답으로 잘못 처리되는 것을 막기 위한 장치라 할 수 있다. 오답 채점 및 미채점 답안에 대한 수작업 채점 과정을 통해 확인된 내용을 토대로 채점 전문가가 오답 설명서[4]를 작성하도록 설계하였다. 오답 설명서는 수작업 채점에 도움을 주기 위한 목적으로 사용할 수도 있고, 오답의 여러 유형에 대해 각각 코드를 부여함으로써 학생들의 오개념이나 대안 개념들을 고려하여 교수·학습의 개선에 도움을 주기 위한 것으로 활용될 수도 있다.

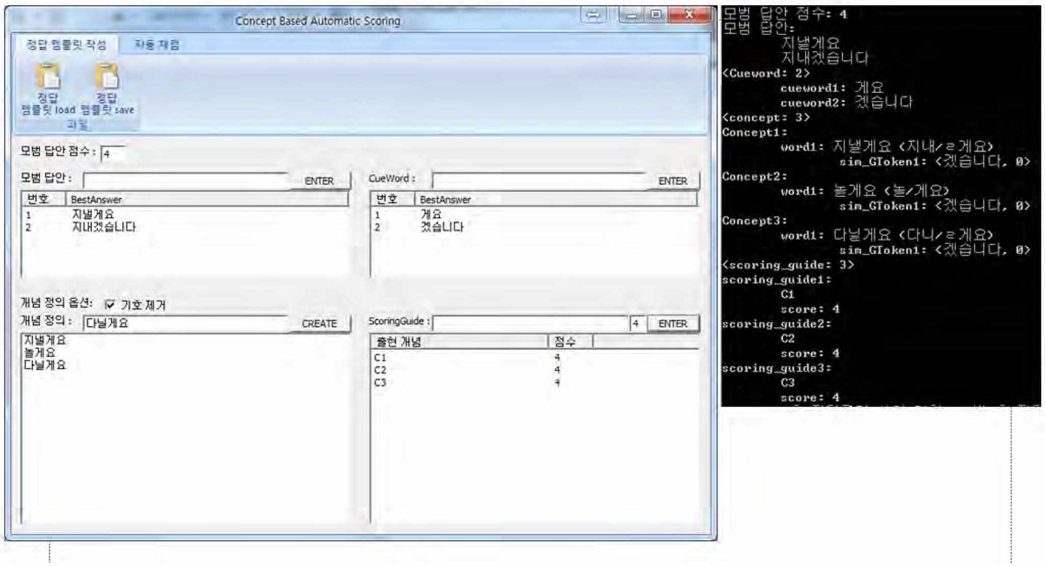
자동채점 후 처리 단계에서는 미채점 답안을 유사 답안끼리 묶어서 수작업 채점을 수행하고, 수작업 채점에서 얻은 새로운 정답, 오답 정보를 토대로 정답 템플릿을 업데이트할 수 있도록 하였다. 이와 같은 순환 과정을 통해 학생 답안에서 얻은 새로운 정답, 오답 유형들을 채점 기준에 추가하여 채점 기준표를 완성하고 수정된 채점 기준표를 적용하여 채점이 보류되었거나 새

하지 않고 오직 사람이 채점하는 것을 말한다.

4) 이와 같은 오답 설명서 작성 기능은 현재의 프로그램에는 실현되어 있지 않으나 향후 한국어 서답형 자동채점 프로그램이 정교화되는 장기적인 개발 계획에는 이러한 기능이 보완될 예정이다.

롭게 적용할 학생 답안들에 대해 자동채점을 수행한다.

본 연구의 이해를 돕기 위해 초3 진단평가 쓰기 (가)형 24번의 정답 템플릿의 예시를 제시하면 다음과 같다. 이 템플릿에는 모범 답안, 개념 정의, 단서어, 각 개념에 대한 점수가 부여되어 있다. 채점 기준표에서 이 문항의 정답은 '지낼게요.', '앞으로는 친구들과 사이좋게 지낼게요.'로 제시되었다. 채점자의 경우 철자가 다소 틀리더라도 정답으로 채점한 경우들이 있었지만 자동채점에서는 철자가 틀릴 경우 오답으로 처리한 것에 차이점이 있었다.



〔그림 2〕 초3 진단평가 쓰기 (가)형 24번 정답 템플릿

Ⅲ. 연구 방법

1. 분석 자료 및 분석 방법

가. 분석 문항

분석 대상 문항은 2011년 초3 진단평가의 서답형 중 읽기 4문항((가)형 2개, (나)형 2개) 쓰기 7문항((가)형 3개, (나)형 4개)으로 총 11문항이다. 초3 진단평가 (가)형, (나)형은 동형으로 구성되어 있어서 내용은 다르나 형태는 같다. 분석 대상 문항은 연구진이 답안 작성 유형

(1~3단어), 답안 유형 수(40% 이하, 즉 답안지가 1000개일 경우 정답·부분 정답·오답의 총 유형 수가 400개 이하), 정답 패턴(P1~P4)[5] 등을 복합적으로 고려하여 자동채점 가능성을 중심으로 선정하였다. 답안 작성 유형 및 답안 유형 수에 따른 분석 대상 문항은 〈표 2〉와 같다.[6]

〈표 2〉 2011년 초3 진단평가 분석 대상 문항

교과	문항 번호	답안 작성 유형	답안 유형 수*	정답 패턴	자동채점 가능성
읽기 (가)형	6	2단어 명사형	76	P2	높음
	19	1단어 술어형	84	P3	보통
읽기 (나)형	6	2단어 명사형	69	P2	높음
	19	1단어 술어형	84	P3	보통
쓰기 (가)형	11	1문장	64	P2	매우 높음
	24	1단어 술어형	112	P3	매우 높음
	29	2단어 술어형	315	P4	보통
쓰기 (나)형	5	1단어 술어형	48	P1	매우 높음
	11	1문장	37	P2	매우 높음
	14	1문장	243	P3	보통
	26	1문장	587	P4	낮음

* 전체 답안 수는 읽기 (가)형 809개, 읽기 (나)형 502개, 쓰기 (가)형 929개, 쓰기 (나)형 916개임.

나. 분석 대상

분석 자료는 2011년 초3 진단평가 표집 대상 학생의 답안이다. 표집한 답안 자료는 과목과 형태에 따라 차이가 있는데 읽기 (가)형은 809개, 읽기 (나)형은 502개, 쓰기 (가)형은 929개, 쓰기 (나)형은 916개이다. 표집 자료의 채점자 점수는 합숙 채점에서 2명의 채점자가 복수 채점하여 일치하는 점수를 표기한 것이다. 채점자들은 국어를 전공한 초등학교 경력 교사들이

5) 정답 작성 유형은 서답형 답안 중 간략하게 기호나 숫자로 쓰는 경우, 단어, 구, 문장으로 기술하는 경우, 그림 또는 그래프로 작성하는 경우로 크게 나누고, 단어·구 중에서도 명사형인지 술어형인지에 따라 구분하였다. 정답 패턴은 자동채점의 용이성에 따라 구분하였는데, 한 단어로 제시한 모범 답안에 거의 완전 일치할 경우에만 정답으로 처리하는 문자열 완전 일치(스트링 매치) 패턴 P1, 자동채점을 위해 형태소 분석이 필요한 경우, 구문 분석이 필요한 경우, 의미/담화 분석이 필요한 경우 등에 따라 정답 패턴을 P2에서 P6까지 구분하였다. 이 중 단어, 구 수준 답안을 대상으로는 스트링 매치, 형태소 분석 정도의 처리 기술만을 요구하는 것은 P3까지이고, 문장 수준 답안에서 구문 분석까지 요구되는 정답 패턴은 P4이다. 본고는 P3까지를 주요 분석 대상으로 하되, P4에 해당하는 쓰기 두 문항을 향후 프로그램 보완을 위해 실험적으로 자동채점 대상에 포함시켰다.

6) 문항에 대한 이해를 돕기 위한 예시로, 쓰기 (가)형 11번과 29번 문항을 뒷부분에 부록으로 수록한다.

다. 채점자들은 1박 2일의 합숙 채점을 하였는데, 채점자 훈련을 하고 예비 채점을 통하여 채점 기준을 공유하였다. 본 채점에서는 2명의 채점자가 독립적으로 채점하여 두 사람의 채점 일치도를 산출하였다.

다. 분석 방법

자동채점을 실시하고 그 결과를 분석하기 위해 정답 템플릿을 설계한 과정은 다음과 같다. 먼저 컴퓨터가 답안을 인식할 수 있도록 학생들이 수기로 작성한 답안을 엑셀 프로그램을 활용하여 컴퓨터에 입력하였다. 채점자들이 사용한 채점 기준표를 활용하여 정답 템플릿 초안을 작성하였다. 채점 기준표와 학생 답안에서 응답 비율이 높은 것을 토대로 개념(concept), 단서어(cue word), 채점 점수를 결정하였다. 이를 토대로 정답 템플릿 초안을 만들어 입력한 후 전체 답안 중 임의 추출한 250개의 답안에 대해 사전 테스트용(기계 훈련용)으로 자동채점을 실시하였다. 사전 테스트를 통해 자동채점된 정답 유형 및 수, 오답 유형 및 수, 미채점 결과 등의 정보를 토대로 개념(concept), 단서어(cue word), 채점 점수 등을 수정하여 정답 템플릿을 만든 후 기계 훈련용으로 사용한 250개의 답안을 제외하고, 나머지 답안을 이용하여 자동채점을 실행하였다. 이후 자동채점 결과에 대해 인간 채점자와 일치도를 산출하고 채점 오류를 검토하였다.

2. 자동채점 프로그램의 채점 신뢰도 검증을 위한 일치도 계수

자동채점 프로그램의 신뢰도를 검증할 때 일반적으로 인간채점과 자동채점 결과를 비교하여 일치도와 상관계수가 높을 경우 신뢰도가 높다고 판단한다. 일반적으로 서답형 문항은 전문가나 교사가 채점하고, 고도로 훈련된 전문가가 채점할 경우 채점 결과의 타당성과 신뢰성이 높다고 판단한다. 이런 이유로 대부분의 자동채점 프로그램 검증 연구에서는 채점자와 자동채점 프로그램의 상관계수나 일치도를 계산하여 검증하고 있다(Chung & Baker, 2003; Jiang & Wei, 2012). 한국어 서답형 문항 자동채점 프로그램도 인간채점 결과를 기준으로 하여 자동채점과 인간채점 결과의 일치도와 채점의 정확성을 분석할 필요가 있다.

자동채점 프로그램의 실제 활용 가능성은 정확한 채점의 비율이 어느 정도인지에 따라 달라질 수 있다. 또한 자동채점 프로그램은 피험자의 오타나 사소한 실수 등에 대하여 인간과 같이 유연한 판단을 하기가 어려우므로 정답을 오답으로 또는 오답을 정답으로 채점할 가능성이 있다. 따라서 채점 오류가 얼마나 적은가 그리고 이런 오류를 보완할 수 있는 시스템을 갖추고 있는가가 무엇보다 중요하다.

일치도는 특정 문항에 대하여 채점자와 자동채점 결과의 일치하는 정도(비율)를 의미하는데 일치도 계수의 계산 공식은 다음과 같다.

$$P_A = \frac{N_{11} + N_{22} + \cdots + N_{JJ}}{N} \quad \cdots\cdots (1)$$

N_{JJ}: 두 관찰자가 일치되게 평정한 피험자 수

N: 사례 수

그러나 이런 일치율(P_A)은 과대 추정되는 경향이 있어서 Cohen이 개발한 Kappa계수를 사용하기도 한다(Bejar, 1991; Burstein, 2001; Clauser, et al., 1997;, Williamson, Bejar, & Hone, 1999). Kappa계수는 두 채점자의 평정 결과가 일치하는 대각선 부분에 우연에 의해 피험자가 포함되어 과대 추정되는 것을 막기 위해 우연에 의한 확률을 제거한 지수로, 일반적으로 Kappa계수가 0.75 이상 되어야 의미 있다고 판단한다(박도순 외, 2007). Kappa계수 계산 공식은 다음과 같다.

$$K = \frac{P_A - P_C}{1 - P_C} \quad \cdots\cdots (2)$$

P_A : 일치율

P_C: 우연에 의해 일치될 확률, $P_C = \frac{N_c}{N}$

N_C: 우연히 두 채점자에 의해 일치된 평정을 받은 피험자 수, $N_C = \sum_{j=1}^{j} \frac{N_{.j} \times N_{j.}}{N}$

N_{jc}: $J \times J$분할표의 대각선에 있는 칸에 분류되는 사례 수

일반적으로 채점자 간 신뢰도를 분석할 때 Pearson 적률 상관계수를 사용하는데 자동채점 프로그램의 신뢰도를 검증할 때도 사용할 수 있다(Yang, et al, 2002). 상관계수가 높다면 채점자는 동일한 채점 기준에 의해 채점한 것으로 해석할 수 있고, 상관계수가 낮으면 각 다른 채점 기준에 의해 채점한 것으로 해석할 수 있다. Pearson 적률 상관계수 계산 공식은 다음과 같다.

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} \quad \cdots\cdots (3)$$

S_{XY} : 두 변수의 공분산, $S_{XY} = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{n}$

S_X, S_Y : X, Y변수의 표준편차

여러 선행 연구에서 제시된 자동채점 프로그램과 채점자의 상관계수는 0.7에서 0.9 정도인데, 0.8에서 0.85 정도이면 에세이 채점에서 피험자의 나이나 에세이 주제와 관계없이 한 명의 채점자를 대체할 수 있는 근거로 인정된다(Burstein and Chodorow, 1999; Landauer, Laham, & Foltz, 2001; Nichols, 2004; Page, 2003).

Ⅳ. 연구 결과

본 장에서는 한국어 서답형 문항 자동채점 프로그램을 활용하여 2011년 초3 진단평가 읽기, 쓰기 자료를 채점하고, 그 채점 결과를 제시하고자 한다. 문항별로 자동채점 프로그램의 채점 단계별 채점 비율과 채점 오류 내용을 분석하고, 채점자 점수의 평균과 자동채점 프로그램 점수의 Kappa계수와 상관계수를 분석한다. 한국어 서답형 문항 자동채점 프로그램은 모범 답안 일치 채점, 개념 기반 채점, 단서어 기반 오답 처리, 수작업 채점의 4단계로 이루어진다. 자동채점 프로그램에서 채점이 진행되는 각 단계별로 채점 비율의 변화를 살펴보고, 채점 오류 비율 및 오류 원인 분석을 할 것이다. 초3 진단평가 읽기, 쓰기 서답형 문항 중 연구 목적에 따라 11문항을 선정하여 506~929명의 답안 자료에 대해 자동채점 단계별 정답 수, 오답 수, 미채점 수, 채점 비율을 계산하고, 각 문항별로 Kappa계수와 상관계수를 계산하여 제시한다.

1. 읽기

초3 진단평가 읽기 (가)형, (나)형 서답형 문항은 총 14문항이 출제되었다. 이 중 기호로 답하는 것이 2문항, 단어나 구 수준의 1~3개의 단어로 작성하는 것이 4문항, 선긋기로 답을 기술하는 것이 8문항이었다. 단어나 구 수준으로 답안을 작성하는 (가)형 2문항, (나)형 2문항을 분석하였다.

읽기의 자동채점 단계별 채점 비율 및 채점 오류 비율을 문항별로 제시하면 〈표 3〉~〈표 6〉과 같다.

〈표 3〉 읽기 (가)형 6번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	337	7	344	344	215	61.5
개념 기반 채점	55	0	55	399	160	71.4

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
단서어 기반 오답 처리	0	160	160	559	0	100.0
수작업 DB 채점	0	0	0	559	0	100.0
합계	392	167	559			

읽기 (가)형 6번은 정답·부분 정답·오답을 포함한 답안 유형 수가 전체 809개 중 76개로 응답 자유도가 낮은 편이고, 2단어 명사형의 P2 유형으로 자동채점 가능성이 높은 문항이다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 337, 오답 수 7) 단계에서 약 62% 정도가 채점되었다. 개념 기반 채점(정답 수 55, 오답 수 0) 단계에서 채점 비율이 약간 증가하였고, 단서어 기반 오답 처리(정답 수 0, 오답 수 160) 단계에서 채점 비율이 약 29% 증가하여 전체 자동채점 비율이 100%가 되었다. 이 문항의 경우 모범 답안 일치 채점과 단서어 기반 오답 처리에 의해 대부분의 답안이 채점되었다.

채점 오류 수는 14개로 채점자가 정답 처리하였으나 자동채점 프로그램은 오답 처리한 경우이다. 정답은 '거북이와 고양이(라는 영화)' 인데 '거북이와 공양이', '거북북이와고양이', '거북이와과고양이', '거북이야고양이영화', '거북 이와 고양이' 등의 철자 오류와 띄어쓰기 오류가 주원인이고, 일부 '영화', '고양이 영화', '이' 등과 같은 것을 정답 처리한 채점자 실수도 있었다.

〈표 4〉 읽기 (가)형 19번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	418	19	437	437	122	78.1
개념 기반 채점	7	0	7	444	115	79.4
단서어 기반 오답 처리	0	114	114	558	1	99.8
수작업 DB 채점	0	0	0	558	1	99.8
합계	392	167	559			

읽기 (가)형 19번은 답안 유형 수가 전체 809개 중 84개로 응답 자유도는 낮지만, 1단어 술어형의 답안을 요구하는 P3 유형이어서 자동채점 가능성이 보통 수준인 문항이다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 418, 오답 수 19) 단계에서 약 78% 정도가 채점되었다. 개념 기반 채점(정답 수 7, 오답 수 0) 단계에서 채점 비율이 약간 증가하였고, 단서어 기반 오답 처리(정답 수 0, 오답 수 114) 단계에서 채점 비율이 약 20% 증가하여 전체 자동채점 비율이 99.8%가 되었다. 이 문항의 경우 모범 답안 일치 채점과 단서어 기반 오답 처리에 의해 대부분의 답안이 채점되었다.

채점 오류 수는 2개로 채점자가 정답 처리하였으나 자동채점 프로그램은 오답 처리한 경우가

1개이고, 채점자가 오답 처리하였으나 자동채점 프로그램이 정답 처리한 경우가 1개이다. 이 문항의 정답은 원래 '안전(하게)'인데, '안위험'이라는 답안에 대해 '안전'을 '안위험'으로 쓴 것을 채점자는 유연하게 정답 처리하였으나, 자동채점 프로그램은 오답 처리하여 의미상 맞는 것을 반영하지 못한 경우이다.

〈표 5〉 읽기 (나)형 6번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	135	2	137	137	115	54.4
개념 기반 채점	7	0	7	144	108	57.1
단서어 기반 오답 처리	0	100	100	244	8	96.7
수작업 DB 채점	0	0	0	244	8	96.7
합계	142	102	244			

읽기 (나)형 6번은 답안 유형 수가 전체 502개 중 69개로 응답 자유도가 낮고, 2단어 명사형 답안을 요구하는 P2 유형이어서 자동채점 가능성이 높은 문항이다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 135, 오답 수 2) 단계에서 약 54% 정도가 채점되었다. 개념 기반 채점(정답 수 7, 오답 수 0) 단계에서 채점 비율이 약간 증가하였고, 단서어 기반 오답 처리(정답 수 0, 오답 수 100) 단계에서 채점 비율이 약 40% 증가하여 전체 자동채점 비율이 96.7%가 되었다. 이 문항의 경우 모범 답안 일치 채점과 단서어 기반 오답 처리에 의해 대부분의 답안이 채점되었다.

채점 오류 수는 8개로 채점자가 정답 처리하였으나 자동채점 프로그램은 오답 처리한 경우가 7개이고, 채점자가 오답 처리하였으나 자동채점 프로그램이 정답 처리한 경우가 1개이다. 정답은 '(말벌의) 벌집'인데, 채점 오류 답안은 '말벌들을집', '꿀법집' 등으로 철자 오류, 띄어쓰기 오류 등이 주원인이었다. '구멍', '숲속', '낮선숲속'의 답안의 경우에 자동채점은 오답 처리하였으나 채점자는 정답 처리하였는데, 이는 채점자의 실수라 할 수 있다. 또한 '벌집에 갇히게 되었다', '말벌 벌집', '벌집의 감옥', '벌집의 구멍'과 같은 답안은 정답 이외에 이름이나 다른 말이 들어간 경우로 자동채점 프로그램은 미채점으로 남겨두었다.

〈표 6〉 읽기 (나)형 19번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	161	5	166	166	86	65.9
개념 기반 채점	0	0	0	166	86	65.9
단서어 기반 오답 처리	0	86	86	252	0	100.0
수작업 DB 채점	0	0	0	252	0	100.0
합계	161	91	252			

읽기 (나)형 19번은 답안 유형 수가 전체 502개 중 84개로 응답 자유도는 그리 높지 않으나, 1단어 술어형의 답안을 요구하는 P3 유형이어서 자동채점 가능성이 보통 수준인 문항이다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 161, 오답 수 5) 단계에서 약 66% 정도가 채점되었다. 개념 기반 채점(정답 수 0, 오답 수 0) 단계에서 채점 비율이 증가하지 않고, 단서어 기반 오답 처리(정답 수 0, 오답 수 86) 단계에서 채점 비율이 약 34% 증가하여 전체 자동채점 비율이 100%가 되었다. 이 문항의 경우 모범 답안 일치 채점과 단서어 기반 오답 처리에 의해 모든 답안이 채점되었다.

채점 오류 수는 4개로 채점자가 정답 처리하였으나 자동채점 프로그램이 오답 처리한 경우이다. 채점 오류 답안은 '닫습니다', '답씁니다'와 같이, 이 문항의 원래 정답 '닫습니다'와 비교할 때 대체로 철자 오류에 기인하여 나타난 것이다.

2. 쓰기

초3 진단평가 쓰기 (가)형, (나)형 서답형 문항은 총 36문항(소문항 기준)이 출제되었다. 이 중 기호로 답하는 것이 2문항, 1단어 명사형 답안 16문항, 1단어 술어형 답안 6문항, 2단어 술어형 답안 2문항, 1문장 이상 답안 10문항이 출제되었다. 초3 진단평가 쓰기의 경우 일반적으로 단어나 구 수준으로 답하는 문항이 많다. 이 중 자동채점 가능성이 보통 이상인 문항인 (가)형 3문항, (나)형 4문항의 총 7문항을 분석하였다. 쓰기의 자동채점 단계별 채점 비율 및 채점 오류 비율을 문항별로 제시하면 〈표 7〉~〈표 13〉와 같다.

〈표 7〉 쓰기 (가)형 11번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	549	3	552	552	127	81.3
개념 기반 채점	0	0	0	552	127	81.3
단서어 기반 오답 처리	0	127	127	679	0	100.0
수작업 DB 채점	0	0	0	679	0	100.0
합계	549	130	679			

쓰기 (가)형 11번 문항은 답안 유형 수가 전체 929개 중 64개로 응답 자유도가 낮고, 1문장의 답안을 요구하지만 P2 유형이어서 자동채점 가능성이 매우 높은 문항이다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 549, 오답 수 3) 단계에서 약 81% 정도가 채점되었다. 개념 기반 채점(정답 수 0, 오답 수 0) 단계에서 채점 비율이 증가하지 않고, 단서어 기반 오답 처리(정답 수 0, 오답 수 127) 단계에서 채점 비율이 약 19% 증가하여 전체 자동채

점 비율이 100%가 되었다. 이 문항의 경우 모범 답안 일치 채점과 단서어 기반 오답 처리에 의해 모든 답안이 채점되었다.

이 문항은 띄어쓰기 문항으로 정답이 '다음∨방학에도∨갈게요'인데 미채점 답안은 하나도 없어서 모든 답안이 채점되었고, 채점 오류도 없었다. 즉, 자동채점 프로그램이 679개의 답안을 오류 없이 완벽하게 채점하였다.

〈표 8〉 쓰기 (가)형 24번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	504	6	510	510	169	75.1
개념 기반 채점	76	0	76	586	93	86.3
단서어 기반 오답 처리	0	62	62	648	31	95.4
수작업 DB 채점	13	0	13	661	18	97.4
합계	593	68	661			

쓰기 (가)형 24번 문항은 답안 유형 수가 전체 929개 중 112개로 응답 자유도가 낮은 편이고, 1단어 술어형 답안을 요구하는 P3 유형이어서 자동채점 가능성이 높은 문항이다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 504, 오답 수 6) 단계에서 약 75% 정도가 채점되었다. 개념 기반 채점(정답 수 76, 오답 수 0) 단계에서 채점 비율이 약 11% 증가하였고, 단서어 기반 오답 처리(정답 수 0, 오답 수 62) 단계에서 채점 비율이 약 9% 증가하였고, 수작업 DB 채점 단계(정답 수 13, 오답 수 0)에서 약 2%가 증가하여 전체 자동채점 비율이 97.4%이었다. 이 문항의 경우 채점의 4단계를 모두 거쳐 채점이 이루어졌다.

채점 오류 수는 32개(4.8%)로 채점자가 정답 처리하였으나 자동채점 프로그램은 오답 처리한 경우가 29개이고, 채점자가 오답 처리하였으나 자동채점 프로그램이 정답 처리한 경우가 3개이다. 정답은 '지낼게요', '지내겠습니다'로 높임법을 사용하면 정답으로 맞춤법은 고려하지 않는다. 채점 오류는 '지낼개요', '지낼꺼요', '지내갰습니다', '지내게 습니다', '지낼깨요', '진내게습니다', '지내께요현지아미안해', '지내도록 할께요.', '앞으로친구들과사이좋게지낼게요' 등과 같은 철자 오류, 띄어쓰기 오류가 대부분이었다. 또한 정답 이외에 다른 문자가 들어갔을 때 오류가 나거나 미채점되었다.

〈표 9〉 쓰기 (가)형 29번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	255	11	266	266	413	39.1
개념 기반 채점	310	0	310	576	103	84.8

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
단서어 기반 오답 처리	0	31	31	607	72	89.4
수작업 DB 채점	14	0	14	621	58	91.5
합계	579	42	621			

쓰기 (가)형 29번은 답안 유형 수가 전체 929개 중 315개로 응답 자유도는 중간 정도이고, 2단어 술어형 답안을 요구하는 P4 유형이어서 자동채점 가능성이 보통 수준인 문항이다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 255, 오답 수 11) 단계에서 약 39.1% 정도가 채점되었다. 개념 기반 채점(정답 수 310, 오답 수 0) 단계에서 채점 비율이 약 45%로 많이 증가하였고, 단서어 기반 오답 처리(정답 수 0, 오답 수 31) 단계에서 채점 비율이 약 5% 증가하였고, 수작업 DB 채점(정답 수 14, 오답 수 0) 단계에서 채점 비율이 약 5% 증가하여 전체 자동채점 비율이 91.5%이었다. 이 문항의 경우 채점의 4단계를 모두 거쳐 채점이 이루어졌다.

채점 오류는 19개(2.8%)이었는데 채점자는 정답 처리하였으나 자동채점은 오답 처리한 것이 18개이고, 채점자는 오답 처리했으나 자동채점은 정답 처리한 것이 1개였다. 정답은 '사진을 찍는다', '사진을 찍어요', '사진을 찍습니다'인데 채점 오류는 '코기리를 찍습니다', '코끼리를 찍고있습니다.', '코끼르를카메라로찍었다', '카메라을직습니다', '동물을 찍습니다' 등으로 철자 오류, 띄어쓰기 오류, 정답 이외에 이름이나 다른 단어가 들어간 경우였다. 이 문항의 경우 채점자의 채점 실수도 일부 있었다.

〈표 10〉 쓰기 (나)형 5번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	523	0	523	523	143	78.5
개념 기반 채점	0	0	0	523	143	78.5
단서어 기반 오답 처리	0	143	143	666	0	100.0
수작업 DB 채점	0	0	0	666	0	100.0
합계	523	143	666			

쓰기 (나)형 5번은 답안 유형 수가 전체 916개 중 48개로 응답 자유도가 매우 낮고, 1단어 술어형 답안을 요구하는 P1 유형이어서 자동채점 가능성이 매우 높은 문항이다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 523, 오답 수 0) 단계에서 약 79% 정도가 채점되었다. 개념 기반 채점(정답 수 0, 오답 수 0) 단계에서 채점 비율이 증가하지 않았으나 단서어 기반 오답 처리(정답 수 0, 오답 수 143) 단계에서 채점 비율이 약 22% 증가하여 전체

자동채점 비율이 100%가 되었다. 이 문항의 경우 모범 답안 일치 채점과 단서어 기반 오답 처리에 의해 모든 답안이 채점되었다.

채점 오류 수는 7(1.1%)개로 채점자가 정답 처리하였으나 자동채점 프로그램은 오답 처리한 경우가 6개이고, 채점자가 오답 처리하였으나 자동채점 프로그램이 정답 처리한 경우가 1개이다. 정답은 '부릅니다'인데, '불름니다', '부름니다', '불읍니다' 등의 답안에 대해서는 철자 오류가 있어 이를 제대로 처리하지 못하였다.

〈표 11〉 쓰기 (나)형 11번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	606	3	609	609	57	91.4
개념 기반 채점	0	0	0	609	57	91.4
단서어 기반 오답 처리	0	57	57	666	0	100.0
수작업 DB 채점	0	0	0	666	0	100.0
합계	606	60	666			

쓰기 (나)형 11번은 띄어쓰기 문항으로 답안 유형 수가 전체 916개 중 37개로 응답 자유도가 매우 낮고, 1문장 답안을 요구하나 P2 유형이어서 자동채점 가능성이 매우 높은 문항이다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 606, 오답 수 3) 단계에서 약 91% 정도가 채점되었다. 개념 기반 채점(정답 수 0, 오답 수 0) 단계에서 채점 비율이 증가하지 않았으나 단서어 기반 오답 처리(정답 수 0, 오답 수 57) 단계에서 채점 비율이 약 9% 증가하여 전체 자동채점 비율이 100%가 되었다. 이 문항의 경우 모범 답안 일치 채점과 단서어 기반 오답 처리에 의해 모든 답안이 채점되었다.

채점 오류 수는 16개(2.4%)로 채점자가 정답 처리하였으나 자동채점 프로그램은 오답 처리한 경우가 13개이고, 채점자가 오답 처리하였으나 자동채점 프로그램이 정답 처리한 경우가 3개이다. 정답은 '즐거운∨놀이가∨참∨많아요' 인데, '놀이가 참 많아요', '놀이가 참많아요', '놀이가 참맑아요' 등으로 정답 중 일부가 빠져 있거나 철자가 오류가 있는 경우 잘못 처리하였다.

〈표 12〉 쓰기 (나)형 14번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	136	4	140	140	526	21.0
개념 기반 채점	463	0	463	603	63	90.5
단서어 기반 오답 처리	0	34	34	637	29	95.6
수작업 DB 채점	8	0	8	645	21	96.8
합계	523	143	666			

쓰기 (나)형 14번은 답안 유형 수가 전체 916개 중 243개로 응답 자유도는 중간 정도이고, 1문장 답안을 요구하는 P3 유형이어서 자동채점 가능성이 보통 수준인 문항이다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 136, 오답 수 4) 단계에서 약 21.0% 정도가 채점되었다. 개념 기반 채점(정답 수 463, 오답 수 0) 단계에서 채점 비율이 약 69% 증가, 단서어 기반 오답 처리(정답 수 0, 오답 수 34) 단계에서 채점 비율이 약 5% 증가, 수작업 DB 채점(정답 수 8, 오답 수 0) 단계에서 약 1% 증가하여 전체 자동채점 비율이 96.8%이었다. 이 문항의 경우 채점의 4단계 모두를 거쳐 채점이 이루어졌다.

채점 오류 수는 16개(2.4%)로 채점자가 정답 처리하였으나 자동채점 프로그램은 오답 처리한 경우가 14개이고, 채점자가 오답 처리하였으나 자동채점 프로그램이 정답 처리한 경우가 2개이다. 정답은 '인사를 하였습니다.'인데, '"안녕하세요"라고말하였습니다.', '인사를 하 였습니니다', '선생님께 안녕하세요 한다.' 등의 답안에 대해서는 철자 오류, 띄어쓰기 오류, 정답 이외 단어 추가와 같은 문제가 있어서 잘못 처리하였다.

〈표 13〉 쓰기 (나)형 26번 문항의 자동채점 단계별 채점 비율

	정답 수	오답 수	채점 수	누적채점 수	미채점 수	채점 비율
모범 답안 일치 채점	65	9	74	74	592	11.1
개념 기반 채점	409	66	475	549	117	82.4
단서어 기반 오답 처리	0	88	88	637	29	95.6
수작업 DB 채점	1	0	1	638	28	95.8
합계	523	143	666			

쓰기 (나)형 26번은 답안 유형 수가 전체 916개 중 587개로 응답 자유도는 매우 높고, 1문장 답안을 요구하는 P4 유형이어서 자동채점 가능성이 낮은 문항이다. 다만, 이 문항의 경우 자동채점 가능성이 낮은 P4 유형임에도 향후 프로그램 개발 방향과 채점 가능성 정도를 탐색하기 위하여 실험적으로 분석하였다. 채점 단계별 채점 수를 살펴보면 모범 답안 일치 채점(정답 수 65, 오답 수 9) 단계에서 약 11% 정도가 채점되었다. 개념 기반 채점(정답 수 409, 오답 수 66) 단계에서 채점 비율이 약 71% 증가하였고, 단서어 기반 오답 처리(정답 수 0, 오답 수 88) 단계에서 채점 비율이 약 13% 증가하고, 수작업 DB 채점(정답 수 1, 오답 수 0) 단계에서 약 0.2% 증가하여 전체 자동채점 비율이 95.8%이었다. 이 문항의 경우 채점의 4단계를 모두 거쳐 채점이 이루어졌다.

채점 오류 수는 49개(7.4%)로 채점자가 정답 처리하였으나 자동채점 프로그램은 오답 처리한 경우가 29개이고, 채점자가 오답 처리하였으나 자동채점 프로그램이 정답 처리한 경우가 20개이다. 정답은 '달리기를 하다가 넘어졌습니다.'인데, '달리기하다가 넘엊졌다.', '넘어져서 일등

을 못했다', '달리기를 하였다.', '너머져다.', '넘어진 것', '그다리며운동회였다.', '달리기를 했지만 다쳤지만 최선을 다했다.' 등과 같이 철자 오류, 띄어쓰기 오류, 정답 이외 단어 추가라는 문제가 있었으나 일부 답안의 경우는 학생들의 다양한 유사 의미 답안을 처리하지 못한 경우도 있었다. 이는 향후 문장 이상의 답안을 요구하는 서답형 문항을 자동채점하기 위해서는 구문 분석, 의미/담화 분석의 언어 처리 기술까지 보완할 필요가 있다는 점을 시사한다.

〈표 14〉 초3 진단평가 인간채점과 자동채점 결과 간 Kappa 및 상관계수

교과	문항 번호	사례 수	Kappa계수	상관계수
읽기 (가)형	6	559	.94	.94
	19	559	.99	.98
읽기 (나)형	6	252	.87	.87
	19	252	.96	.97
쓰기 (가)형	11	679	1.00	1.00
	24	679	.56	.58
	29	679	.40	.50
쓰기 (나)형	5	666	.97	.97
	11	666	.84	.85
	14	666	.55	.54
	26	666	.71	.72

한국어 단어·구 수준 서답형 자동채점 프로그램을 초3 진단평가 읽기, 쓰기의 문항 자료를 활용하여 검증한 결과, 문항에 따라 채점의 정확도가 차이가 있으나 단어, 구 수준에서 응답의 자유도가 낮은 문항의 경우 Kappa계수 및 상관계수가 .80 이상으로 채점자의 채점 점수와 일치도가 높게 나타났다. 반면 일부 문장형 답안을 요구하거나 응답 자유도가 높은 문항은 Kappa계수가 낮았다. 읽기의 경우 대부분의 문항이 Kappa계수가 .85 이상으로 매우 높았다. 쓰기의 경우는 (가)형 11번, (나)형 5번, 11번은 자동채점 가능성이 높으나 (가)형 24, 29번, (나)형 14, 26번은 Kappa계수가 낮았다. 이와 같이 Kappa계수가 낮은 문항은 답안의 응답 자유도가 높거나 답안의 길이가 길어서 현재의 단어·구 수준의 한국어 자동채점 프로그램으로 채점하기는 어려운 것으로 나타났다. 즉, 현재 개발된 한국어 서답형 자동채점 프로그램은 응답의 자유도가 높고 4단어 이상의 기술을 요구하는 문항이나 구문 분석을 요구하는 문항은 채점의 정확도가 아직 낮은 것으로 볼 수 있다.

또한 프로그램 검증 시 채점의 일치도나 상관계수뿐만 아니라 채점 오류의 비율과 내용이 매우 중요하다. 채점 오류 수는 1개~49개로 문항에 따라 차이가 있었다. 채점 오류는 철자 오류로 인한 자동채점 프로그램의 오류가 많았으나 채점자들의 채점 오류도 일부 있었다. 그런데 검

증 자료의 수가 적어서 기계학습을 한 번만 하고 검증하였는데, 일부 문항의 경우 철자 오류가 반복된 것들이 있어서 여러 번의 반복된 수작업 DB 채점 단계를 추가한다면 채점 비율과 채점의 정확도가 더 높아질 가능성이 있다. 추후 연구에서는 답안 자료를 확보하여 기계학습을 1번, 2번, 3번으로 늘려서 반복 채점을 한 후 검증할 필요가 있다. 채점 오류 중 가장 큰 문제점은 철자 오류가 있을 경우 채점자들은 피험자의 실수로 인정하고 내용의 정합성이 인정되면 채점을 하였으나 자동채점 프로그램은 채점을 하지 않거나 채점 오류를 범하는 경우가 많았다는 것이다. 또한 채점 오류와 미채점이 발생한 문항은 유사어를 인지하지 못하는 경우와 다른 문자나 용어가 포함되어 있는 경우가 대부분이었다. 이런 부분은 아직 정답 템플릿 작성이 정교하지 않아 생기는 문제이므로 향후 이에 대한 방안을 마련해야 할 것이다.

Ⅴ. 결론

한국어 서답형 자동채점 프로그램을 활용하여 채점한 결과, 초3 진단평가의 단어·구 수준의 단답형 문항은 자동채점이 가능한 것으로 나타났으나 응답 자유도가 높은 문항과 문장 수준의 문항, 정답 패턴이 P4에 해당하는 문항은 자동채점 프로그램의 보완이 필요한 것으로 나타났다. 초3 진단평가의 단어·구 수준의 한국어 서답형 문항에 대한 자동채점 프로그램의 채점 비율과 채점자와의 일치도는 적정한 수준이었다. 채점 오류는 문항에 따라 차이가 있으나 단어·구 수준의 단답형 문항의 채점 오류 비율은 적은 편이다. 정답 템플릿 정교화와 수작업 채점을 보완한다면 채점 오류도 줄어들 가능성이 있다. 채점 오류 중 가장 많은 것은 철자 오류이고, 유사어 처리, 부정어 처리, 다른 문자나 용어가 포함되어 있을 때 처리하지 못하는 경우였다. 채점 오류는 피험자에게 잘못된 점수가 제공되므로 철저하게 검증하고, 채점 오류에 대한 시스템적 보완이 요구된다.

한국어 서답형 자동채점 프로그램은 모범 답안 채점, 개념 기반 채점, 단서어 오답 처리, 수작업 DB 채점의 4단계로 이루어지는데 이런 단계별 채점 설계는 타당한 것으로 판단된다. 대부분의 문항이 4단계의 채점 단계를 거치고, 단계에 따라 채점 비율도 향상되었다. 다만, 정답 템플릿을 정교화하기 위해서는 개념을 어떻게 잡고, 단서어는 무엇으로 할 것인가에 대한 추가적인 연구가 필요하다. 또한 수작업 채점을 몇 회로 할 것인가 또는 채점 오류에 대한 목표치를 얼마로 해야 하는가에 대한 연구도 이루어져야 할 것이다. 초3 진단평가의 경우 고부담 시험이 아니므로 채점 오류를 낮추는 것보다 채점 비율을 높이는 방향으로 개발을 진행할 수 있을 것이다.

초3 진단평가의 경우 담임교사들이 채점하므로 맞춤형 교수학습이 가능하도록 학생별 오답노트를 제공하는 것과 같은 피드백 제공 기능이 추가되어야 한다. 채점 결과를 분석하여 개별화된

분석과 반별 통계 자료를 제공하여 교사들이 쉽게 학생들의 학력 상태를 확인할 수 있도록 프로그램을 개발할 필요가 있다. 또한 교사들이 자동채점 프로그램을 쉽게 활용할 수 있도록 편의성이 높은 인터페이스를 개발하여 제공해야 할 것이다. 정답 템플릿 작성 시에도 유의어 사전을 활용하여 유사 단어 리스트를 제공하고, 학생 답안을 분석하여 자주 나오는 오답과 정답 형태를 제공하여 교사들이 채점에 대한 부담을 덜 수 있도록 프로그램이 보완되어야 할 것이다.

참 고 문 헌

강원석(2011). 질의문 유형 분석을 통한 서답형 자동채점 시스템. 한국콘텐츠학회논문지 11(2). 13-21.

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

김명화, 김도남, 권점례, 김완수, 황인우, 강태훈(2011). 3R's 기초학습 부진 선별 도구 개발 연구. 한국교육과정평가원 연구보고 CRE 2011-10.

박도순 외(2007). 교육평가-이해와 적용. 서울: 교육과학사.

성태제, 양길석, 강태훈, 정은영(2010). 학업성취도 평가 서답형 문항 컴퓨터 채점화 방안 연구. 한국교육과정평가원 연구보고 CRE 2010-1.

정동경(2001). 벡터 유사도와 시소러스를 이용한 주관식 답안의 채점 방법. 동국대학교 교육대학원 석사학위 논문.

조우진(2006). 의미 커널과 한글 워드넷에 기반한 지능형 채점 시스템. 한림대학교 대학원 석사학위 논문.

반재천, 김선(2012). 2011년 기초학력 향상도 평가의 분류일관성, 기초학력 도달 비율, 그리고 하위 단계 성취 정도와 학년말 기초학력 도달간 관계. 한국교육평가학회 추계학술대회 발표논문.

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*, 522-532.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. Proceedings Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, June 1999, College Park, Maryland. 68-75.

Burstein, J. C. (2001a). Automated essay evaluation with natural language processing. Paper Presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Chung, G. K. & Baker, E. L. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis, & J. Burstein(Eds.), *Automated essay scoring: A cross-disciplinary perspective*, Lawrence

Clauser, B. E., Ross, L. P., Clyman, S. G., Rose, K. M., Margolis, M. J., Nungester, R. J., et

al. (1997). Development of a scoring algorithm to replace expert rating for scoring a complex performance-based assessment. *Applied Measurement of Education, 10,* 345-358.

Jiang, J., & Wei, W. (2012). Automated scoring research over 40 Years: Looking back and ahead. *Journal of Artificial Intelligence, 5(1),* 56-63.

Landauer, T.K., Laham, D., & Foltz, P.W. (2001). The intelligent essay assessor: Putting knowledge to the test. Proceedings of the Association of Test Publishers Conference on Computer-Based Testing: Emerging Technologies and Opportunities for Diverse Application, February 26-28, 2001, Tucson, AZ., USA.

Nichols, P. D. (2004). Evidence for the interpretation and use of scores from an automated essay scorer. Proceedings of the Annual Meeting of the American Educational Research Association(AERA), April 12-16, 2004, San Diego, CA.

Page, E. B. (2003). Project essay grade: PEG. In M.D. Shermis, & J. Burstein(Eds.), *Automated essay scoring: A cross-disciplinary perspective,* Lawrence Erlbaum Associates, NJ, 43-54.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). "Mental Model" comparison of automated and human scoring. *Journal of Educational Measurement, 36,* 158-184.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement of Education, 15,* 391-412.

· *논문접수 : 2013-01-01/ 수정본접수 : 2013-02-04/ 게재승인 : 2013-02-22*

ABSTRACT

The Application of Automatic Scoring Program to Supply-Type Items of Basic Competency Test

Myoung-Hwa Kim
(Research Fellow, Korea Institute for Curriculum and Evaluation)
Eun-Hee Noh
(Research Fellow, Korea Institute for Curriculum and Evaluation)
Jae-Ho Sim
(Research Fellow, Korea Institute for Curriculum and Evaluation)

The purpose of this study is to explore possibility of automatic scoring supply type items of the Grade 3 National Diagnostic Assessment of Basic Competency (NDABC) to reduce scoring burden, to improve scoring efficacy and scoring reliability. This study presented scoring rates, scoring errors, and Kappa(correlation) coefficients of scores between human scoring and automatic scoring in order to ensure scoring reliability. We also analyzed the sources of scoring errors, where the automatic scoring program fails. We used automatic scoring program developed by the Korea Institute for Curriculum and Evaluation (KICE). The results showed that the scoring rate was very high(91.5~100%), and that the Kappa coefficients depend on items. The numbers of scoring error were 1~42. The sources of scoring errors were caused by spelling errors, the non-recognition of analogous terms and symbols.

This study presented two suggestions as following.

First an automatic scoring program for NDABC should be supplemented to give feedback and information about wrong answer to teachers and students. Second the program should focus on providing convenient interface for teachers.

Key Words : basic competency test, supply-type item, automatic scoring, large-scale assessment, Grade 3 National Diagnostic Assessment of Basic Competency, reading, writing

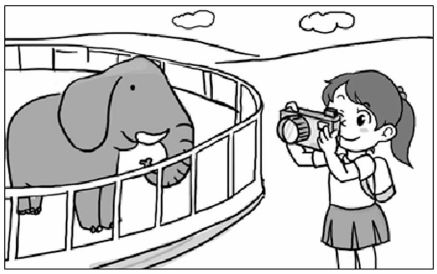
[부록 1] 초3 기초학습 진단평가 문항 예시

1. 자동채점 가능성이 높은 문항

• 문항	•정답 다음 ∨ 방학에도 ∨ 갈게요. 채점 기준 - 4점: 정답과 동일한 경우, 0점: 정답이 아닌 경우 ※ 맞춤법은 채점에 반영하지 않음 •유형 구분 띄어쓰기 문제로 1문장이지만 답이 명확하고, 간단한 형태소 분석으로 자동채점이 가능하여 P2에 해당한다.
11. 밑줄 친 ㉠을 바르게 띄어 쓰시오. 다 음 .	

2011년 초3 진단평가 쓰기 (가)형 11번

2. 자동채점 가능성이 보통인 문항

29. 그림을 보고, 밑줄 친 곳에 알맞은 말을 쓰시오.  보람이가 _____.	•정답 사진을 찍는다. 사진을 찍어요. 사진을 찍습니다. 모두 정답 처리 ※ '코끼리를 찍다'나 '사진기로 보다'의 의미가 드러나는 경우('사진기로 본다/보아요/봅니다' 등의 내용이면 정답으로 함.). 맞춤법은 채점에 반영하지 않음. •유형 구분 이 문항의 경우 단순 개념 분석을 통해 자동채점하는 P4에 해당한다. 하지만 학생들이 답안으로 기술하는 내용이 열려 있기 때문에 다양한 정답들이 출현할 가능성이 높은 문항이다.

2011년 초3 진단평가 쓰기 (가)형 29번