한국심리학회지 : 인지 및 생물
The Korean Journal of Cognitive and Biological Psychology
2010, Vol. 22, No. 4, 573-588

구성적 선다형 검사 방식에서 배점 조작의 효과*

박 주 용†
서울대학교 심리학과

구성적 선다형 검사 방식은 선다형 방식의 문제점을 보완하기 위해 개발된 새로운 컴퓨터화 검사 방식이다. 이 방식에서는, 하나의 발문에 대해 먼저 단답식으로 반응하도록 한 다음, 다시 동일한 발문에 대해 선다형으로 반응을 하도록 한다. 따라서 선다형에서 정답을 선택했더라도 단답식 반응을 통해 정답을 알고 선택했는지 모르고 선택했는지를 확인할 수 있다. 이 방식을 사용한 선행 연구에서, 구성적 선다형으로 본 집단과 단답식으로만 본 집단의 단답식 수행을 비교하였을 때, 단답식으로 본 집단이 더 높은 점수를 얻는다는 것이 반복적으로 관찰되었다. 본 연구는 구성적 선다형의 단답식 수행과 단답식으로만 보았을 때의 수행간 동등성을 확보하기 위해, 구성적 선다형의 단답식 점수 배점을 높이는 조작이 실제 수행에 어떻게 영향을 주는지를 알아보기 위해 수행되었다. 초등학교 6학년생 227명을 대상으로 3개의 실험이 수행되었다. 실험 1과2에서는, 무선적으로 나뉘어진 두 집단을 대상으로, 한 집단은 구성적 선다형 방식으로, 다른 집단은 단답식으로 보게 한 다음 수행을 비교하였다. 각각 따로 채점된다고 지시를 준 실험 1에서는 구성적 선다형에서의 단답식 반응이 단답식으로만 반응하게 한 경우보다 낮았다. 그렇지만 단답식과 선다형의 배점을 90% 대 10%로 했을 때에는 두 집단간에 단답식 점수의 차이가 사라졌다. 실험 3에서는, 점수 배점이 90% 대 10%일 때, 구성적 선다형 집단과 선다형 집단의 수행을 비교하였다. 그 결과 오히려 구성적 선다형 집단의 선다형 점수가 높아짐을 발견하였다. 이 결과는 구성적 선다형에서 단답식의 비중을 크게 하면, 두 가지 반응을 하게 하더라도 그로 인해 따로 따로 반응하게 했을 때에 비해, 수행이 저하되지는 않음을 시사한다.

주요어 : 컴퓨터화 검사, 구성적 선다형 검사 방식, 반응 동등성, 단답식, 선다형

* 이 논문은 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음 (KRF-2007-B00130).

† 교신저자 : 박주용, 서울대학교 사회과학대학 심리학과, 서울시 관악구 관악로 599번지 16동 M508
E-mail : jooypark@snu.ac.kr

서 론

인지과정은 눈에 보이지 않기 때문에, 자극에 대한 반응시간이나 정확성과 같은 검사결과를 바탕으로 추론된다. 교육장면에서의 성취수준도 검사 결과에 근거하여 추정된 인지과정의 하나로 볼 수 있다. 성취수준을 측정하기 위한 검사 방식은 크게 구성형(constructed response format)과 선택형(selection format)으로 나뉘는데, 통상 학교 교육장면에서 널리 쓰이는 것은 단답식(short answer format)과 선다형(multiple-choice format)이다. 단답식은 제시된 발문(stem)에 수험자가 스스로 답을 산출하고 이를 표현하는 데 반해, 선다형에서는 이미 만들어진 답지를 이용하여 반응한다. 동일한 발문일 경우 단답식은 선다형에 비해 출제가 용이하다. 답지를 만들 필요가 없기 때문이다. 그렇지만 그에 따른 대가가 있다. 그 대가란 수험자의 답안을 읽어본 다음 출제자가 생각한 정답과 의미적으로 일치하는 지의 여부를 일일이 따져보아야 한다는 것이다. 선다형의 경우는 그럴듯한 오답지를 만들어내는 일이 어렵지만, 일단 만들어 놓으면 채점이 쉽다. 광학판독기를 이용하면 기계가 엄청난 수의 답안지를 짧은 시간 안에 다 채점을 해낼 수 있다. 채점의 신뢰도도 단답식 보다 높다(Wainer & Thissen, 1993; Downing, 2006). 채점상의 이런 효율성에 힘입어 선다형은 현재, 교육 장면이든 진단 장면이든 상관없이, 가장 널리 쓰이는 검사 방법으로 인정받고 있다.

선다형이 효율적이어서 많이 쓰이기는 하지만 선다형 방식에도 문제가 있다. 문제와 함께 답이 제시되기 때문에 잘 모르는 상태에서도 답지를 잘 읽어보면 정답을 찾아낼 가능성이 있다. 문제와 함께 답지가 주어지는 것은 실제 장면에서는 거의 일어나지 않는 일이기에 인위적이라는 비난도 있다(Veloski, Rabinowitz, & Young, 1999). 또 다른 비판은 과정 정보가 없다는 점이다. 정답을 고르긴 했는데 정말 알고 골랐는지 아니면 우연히 맞추었는지가 구분이 되지 않는다는 것이다.

선다형의 이런 문제를 해결하기 위해 다양한 시도가 이루어져 왔다. 한 문항에 여러 개의 빈칸을 할당한 다음 이들을 이용하여 답을 하게 한 그리드 문항(grid item)이나 답지수를 엄청나게 많게 한 상태에서 답을 고르게 하는 확장된 메뉴에서 찾기(extended matching)는 그 좋은 예이다. 전자의 경우 한 문항에 예를 들면 3칸을 배정하여 만일 답이 1/3이라면 첫째 칸에는 1, 둘째 칸에는 /, 셋째 칸에는 3을 표시하여 1/3을 수험자가 직접 쓰게 하는 방식이다. 최근에는 직접 숫자를 써넣게 하고 이를 컴퓨터로 읽는 방식을 취하기도 한다. 그렇지만 이 방식은 숫자에 국한된다는 문제가 있다. 또 다른 방식인 확장된 메뉴에서 찾기에서는 수험자에게 수천 개의 항목이 순서대로 배열된 책자를 제공하고 이를 이용하여 답을 하게 하는 방식이다. 예를 들어 정답이 심장인 경우 심장 앞에 붙여진 #3537을 써넣으면 된다.

선다형의 문제를 해결하는 또 다른 방식은 혁신적인 컴퓨터화 기술을 사용하는 것이다. 그 중 대표적인 것은 자동화된 채점 기법을 이용하는 것이다. 영어의 자연어 처리에 대한 연구를 기반으로 다양한 자동화된 논술채점방식이 개발, 사용되고 있다(예, 최윤정, 성태제,

2006; Dikli, 2006; Wang & Brown, 2007). 그렇지만 이 시스템들은 대개 문항마다 전문가의 주석이 붙은 방대한 데이터 베이스를 구축해야 하는 어려움이 있어 누구나 쉽게 사용하기가 어렵다. 논술이 아닌 단답식을 자동적으로 채점할 수 있도록 한 c-Rater(Leacock & Chodorow, 2003)도 구축해야 할 데이터 베이스의 규모는 작지만 동일한 문제를 갖고 있다.

본 연구에서는 위에서 언급된 선다형의 문제점을 어느 정도 완화시키면서도, 그리드 문항, 확장된 메뉴에서 찾기, 혹은 자동화된 채점 기법보다 사용이 용이한 새로운 컴퓨터화 검사 방식을 소개하고, 이 방식의 타당성을 검증하는 연구를 수행하고자 한다. 본 연구에서 다루는 새로운 방식은 Park(2010)에 의해 구성적 선다형으로 명명되었다. 이 이름은 구성형 방식 중의 하나인 단답식과 답지가 주어지는 선다형이 결합된 특징을 드러내기 위해 붙여졌다.

구성적 선다형 방식에서 수험자는 동일한 발문에 대해 두 번 반응을 하게 된다. 먼저 단답식으로 반응을 하고 난 다음, 다시 선다형을 푸는 것이다. 이렇게 함으로써 얻어지는 장점은, 그리 많은 시간을 들이지 않으면서, 동일한 피험자를 대상으로 두 가지 다른 방식의 반응을 쉽게 얻을 수 있다는 것이다. 문항형식간의 차이를 알아보기 위한 선행 연구에서는 동일한 피험자들에게 시간을 두고 같은 검사를 두 번 실시하였다. 그러나 구성적 선다형에서는 그럴 필요가 없다. 단답식으로 반응한 다음, 아직 기억에 남아 있을 때 다시 한 번 선다형을 푸는 것이기 때문이다(그림 1 참조). 이 방식에서는 먼저 단답형으로 응답한 다음, 자신이 없을 경우 일단 저장을 한 다음 다른 문제를 풀 수 있고, 아니면 버튼을 눌러 컴퓨터에 신호를 보내 선다형 선택지가 제시되도록 할 수 있다. 일단 선다형답지가 제시되면 단답식 응답을 더 이상 수정할 수 없고 선다형의 답지만을 고를 수 있다. 선다형에 대한 반응이 끝나면 그 문항에 대한 모든 반응이 종료된다.

Park(2010)은 이 방식을 이용하여 초등학생과 대학생을 대상으로 검사를 실시하였다. 그 결과, 동일한 발문에 대해 단답식과 선다형으로 반응하게 했을 때, 이 두 방식간의 상관이 비교적 높았으며, 정답률은 선다형에서 더 높음을 발견하였다. 비록 그 비율이 낮기는 했지만, 또 다른 흥미로운 결과는 선다형 방식의 경우 매력적인 오답지가 있을 경우, 단답식에서는 정답을 쓰더라도 선다형에서는 오답을 선택하는 것도 관찰되었다. 박주용과 민경석(2009)은 Park(2010)이 얻은 결과를 2모수 문항 반응 모형을 사용하여 단답식과 선다형을 비교하였다. 그 결과 선다형 문항과 비교하여 단답형 문항에서 문항 난이도와 변별도가 높아지며, 문항 및 검사 전체에서 정보량이 단답형 문항에서 더 커짐을 발견하였다. 이에 반해 선다형 문항은 낮은 능력 수준의 피험자를 변별하는데 상대적으로 강점을 보임을 발견하였다.

구성적 선다형은 이처럼 단답식과 선다형간의 차이를 알아보기 위한 도구로 활용될 수 있을 뿐만 아니라 이를 이용하여 문항 개발 도구로 사용할 수도 있다. 구성적 선다형 방식으로 예비 검사를 하여, 선다형과 단답식에서 큰 차이를 보이지 않는 문항은 선다형으로

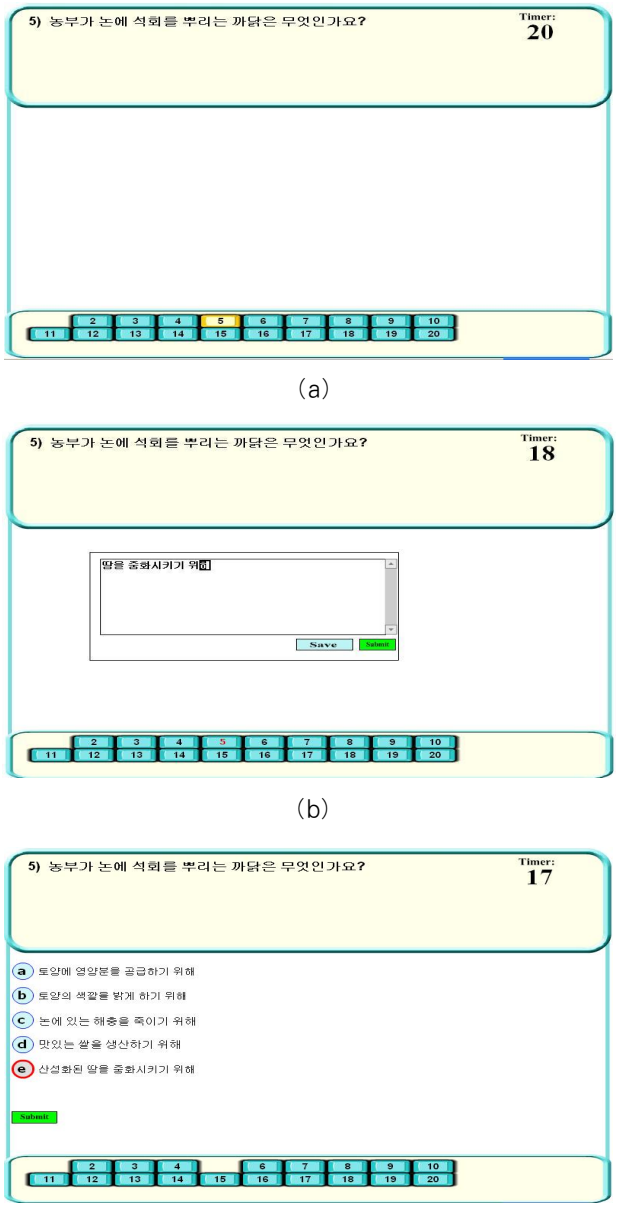

그림 1. 구성적 선다형 방식의 작동

(a) 1번 문항은 단답형과 선다형을 다 풀었기 때문에 하단의 문제제시 박스가 사라졌고, 현재 5번 박스에 놓여 있기 때문에 5번 문제가 제시되어 있다. (b) 5번 문제제시 박스를 클릭하면 단답형 반응을 할 수 있는 반응지가 제시된다. 피험자는 이 반응지에 답을 입력한다. 입력이 끝나면 일단 저장해두고 다른 문제를 풀거나 아니면 submit 버튼을 눌러 바로 선다형 문제의 선택지를 제시하도록 할 수 있다. (c) 일단 선다형 선택지가 제시되면 그 문제를 풀어야 한다(박주용과 민경석, 2009, [그림 1], 교육평가연구, 22(2), 458쪽에서 재인용).

출제하고, 큰 차이를 보이는 문항은 단답식으로 출제하거나 경우에 따라서는 답지를 조절할 수 있다.

그런데 아직 검증되지 않은 한 가지 경험적 문제는 과연 구성적 선다형 방식으로 얻은 두 반응이 하나씩 따로 얻었을 때의 결과와 동등하다고 할 수 있는가이다. 반응의 동등성 문제는 여러 장면에서 논의되어 왔다. 그 중 하나는 문항 형식간의 구인 동등성 문제이다(Messick, 1989; Rodriguez, 2003). Rodriguez는 동일한 발문을 사용한 연구들을 메타분석한 결과, 단답식과 선다형 방식간의 상관은 거의 1에 가까움을 발견하였다. 그렇지만, Rodriguez가 분석한 선행 연구들에 사용된 문항들은 대개, 기존의 선다형 문항을 단답형으로 변형한 것이기에 발문이 같을 경우, 선다형과 단답식이 동등하다고 결론짓기에는 아직은 성급해 보인다.

또 다른 동등성 문제는 지필식 검사와 컴퓨터 검사간의 동등성 문제이다(예, 박도순, 김종필, & 양길석, 2002; Bennett, Braswell, Oranje, Sandene, Kaplan, & Yan, 2008). 지필식이나 컴퓨터화 검사의 어느 한 방식으로 볼 경우 크게 문제가 되지 않지만, 만일 동일한 검사를 지필식 혹은 컴퓨터화 검사 중에서 선택할 수 있는 상황에서는 두 검사 방식간의 동등성 확보는 필수적이다. 일반적으로 동일한 검사일 경우, 큰 차이가 없지만, 컴퓨터로 볼 경우에는 컴퓨터 사용 능력에 영향을 받는다는 연구결과들이 있다(예, Bennett et al., 2008).

본 연구에서 다루는 동등성은, 컴퓨터화 검사 장면이라는 점에서는 동일하지만, 한 발문에 대해 두 가지 다른 방식으로 반응하게 했을 때와, 한 가지로만 반응하도록 했을 때의 동등성이다. 다시 말해 구성적 선다형에서 먼저 단답식으로 반응을 한 다음, 선다형으로 반응을 하는 경우와 단답식으로만 반응을 할 경우 단답식 반응을 비교했을 때 차이가 있는가이다. 또 다른 경우는 구성적 선다형으로 반응하게 한 집단과 선다형으로만 반응을 하게 한 집단을 비교했을 때 선다형 반응간에 차이가 있는가이다. 만일 구성적 선다형 방식을 통해 얻은 반응이 따로 따로 얻은 반응과 다르다면 이 방식으로 얻은 결과를 따로 얻은 결과로 일반화하는 데 문제가 있다. 본 연구는 이 가능성을 알아보기 위해 수행되었다.

구성적 선다형 방식에서 얻어진 반응이 단답식이나 선다형 방식의 어느 하나의 방식을 사용했을 때와 차이가 있는지를 알아보기 위해 초등학교 6학년 학생들을 대상으로 3개의 실험이 수행되었다. 실험 1, 2에서는 구성적 선다형과 단답식간의 비교가 이루어졌고, 실험 3에서는 구성적 선다형과 선다형 방식의 비교가 이루어졌다.

실험 1

실험 1에서는, 구성적 선다형과 단답식 방식의 비교가 이루어졌다. 실험집단의 학생들은 구성적 선다형 검사 방식에 익숙하도록 훈련을 시킨 학생들을 대상으로, 단답식과 선다형이 따로 채점되니, 두 방식에 대해 각각 최선을 다하도록 지시한 다음 검사를 실시하였다. 이에 반해 통제 집단의 학생들은 특별한 지시 없이 단답식으로 반응하게 하였다.

연구방법

실험 대상 6학년이 모두 8학급인 서울 강남에 위치한 O 초등학교의 학생 227명 전원이 컴퓨터 수업 시간의 일부로 본 연구에 참여했다. 이 중 남학생은 115명, 여학생은 112명이었다. 8개의 학급을 수업 시간표를 고려하여 무선적으로 4개 반씩 2집단으로 나눈 다음, 한 집단은 단답식으로 다른 집단은 구성적 선다형으로 무선 배정하였다.

실험 도구 자극재료. 초등 6학년 사회 1, 2 단원으로부터 20문항이 출제되었다. 이 문항들은 연구자가 교사의 자문을 받아 만든 문항들이었다.

컴퓨터 프로그램. FLASH MX로 만들어진 2개의 프로그램이 사용되었다. 구성적 선다형 방식은 서론에서 소개된 것과 동일한 절차와 화면으로 작동하는 방식이다. 단답식 방식은 구성적 선다형 방식과 동일하지만, 선다형이 제시되지 않는다는 것만 차이가 있다. 수험자는 한 문항을 여러 번 풀 수 있으며, 모든 문제를 다 풀었을 때에는 검사를 종료할 수 있도록 하는 종료버튼을 누르도록 하였다.

실험 절차 실험은 컴퓨터 수업시간을 활용하여, 반 단위로 진행되었다. 모든 학생들은 이미 구성적 선다형 방식으로 2회 이상 경험하였기 때문에 특별한 지시가 주어지지 않았고, 단답식만으로 푸는 집단의 경우 구두로 단답식으로만 반응할 것이라는 지시가 주어졌다. 최선을 다해 풀 것을 지시받았고, 연구 목적을 알지 못하는 수업 담당 교사가, 연구 보조원의 도움을 받아, 검사를 진행하였다. 컴퓨터 검사가 종료되면 선다형 반응에 대한 점수만 피드백으로 화면에 제시되었다. 학생들이 컴퓨터에 입력한 반응은, 한 명의 채점자가 채점을 하였으며, 애매한 반응이 있을 경우 연구자와 협의하여 부분점수가 주어졌다. 채점은 두 방식 모두 100점 만점으로 이루어졌다.

결과와 논의

피험자들 중 3명은 검사에 참여하지 않았고, 7명은 종료 전에 검사를 끝내 그 결과가 기록되지 않았다. 구성적 선다형 집단 중 1명은 선다형을 전혀 풀지 않아 분석에서 제외되었다. 이들을 제외한, 4개의 반의 단답식 집단과 또 다른 4개 반의 구성적 선다형 집단에 대한 기술 통계는 표 1에 제시되었다.

본 연구에 앞서 실시된 학업성취도 평가 결과 중 사회과에 대한 점수가 비교되었는데, 단답식 집단과 구성적 선다형 집단의 평균은 86 대 82점으로 이들간에 통계적인 차이가 없었다(t (214) = 1.8, p = 0.07, MS_e = 288). 단답식 집단의 경우 점수의 범위는 0~90점이었

표 1. 실험 1의 주요 결과: 단답식과 구성적 선다형의 단답식 수행 비교 (괄호 안은 표준편차)

검사 방식	단답식	구성적 선다형의 단답식
피험자 수	108	108
사회 성취도 검사	86 (18)	82 (16)
평균	45 (23)	33 (18)
검사시간	886 (312)	879 (290)

고, 평균은 45점이었다. 학교에서 실시한 성취도 검사 결과와의 상관은 0.65였다. 구성적 선다형 집단의 경우, 선다형 점수의 범위는 15~90점이었고, 평균과 표준편차는 각각 59, 17점이었다. 단답식의 범위는 0~70점이었고 평균은 33점이었다. 단답식과 선다형 점수 간의 상관은 .76이었다. 이 결과는 선행 연구에서 얻어진 상관 정도와 비슷하였다. 성취도 검사와 단답형, 그리고 성취도 검사와 선다형 점수 간의 상관은 각각 .64, .63으로 서로 비슷하였을 뿐만 아니라 단답식만으로 본 경우의 상관과도 비슷하였다.

본 실험의 주 관심사인 단답식 점수에서는 두 집단간에 큰 차이가 있음이 관찰되었다(45 대 33, t (214) = 4.3, p < .001, MSe = 429). 비록 통계적인 차이가 없었지만, 성취도 검사에서 4점 정도의 차이가 났기 때문에, 공분산분석(Analysis of Covariance)이 수행되었고, 그 결과는 표 2에 제시되었다. 성취도 검사를 통제했을 때, 단답식 집단은 구성적 선다형 집단보다 8.7점이나 더 높은 평균 점수를 얻었는데, 이 차이는 유의미하였다(p < .01). 이 결과는 초등학생 6학년을 대상으로 사회와 과학 문제로 수행된 이전의 예비 연구에서도 관찰되었는데, 단답식으로 반응하고 나서 다시 선다형으로 반응하게 했을 때보다 단답식으로만 반응하게 했을 때, 평균 점수가 더 높음을 보여준다. 검사 시간의 경우 구성적 선다형에서는 두 가지 다른 반응을 했어야 했음에도 불구하고, 단답식으로 반응하게 한 집단에 비해, 검사시간 상에서는 차이가 관찰되지 않았다(886 대 879, t<1).

이상의 결과는 구성적 선다형에서의 단답식에 대한 반응이 단답식으로만 반응하게 한 경우와 동등하지 않다는 점에서, 동일한 발문에 대해 두 가지 다른 반응을 쉽게 얻으려는 구성적 선다형 방식의 개발 의도와 맞지 않는다. 구성적 선다형에서 단답식에 대한 반응이, 단답식으로만 볼 때보다 상대적으로 부진하기 때문이다.

그렇다면 도대체 왜 구성적 선다형에서 단답식에 대한 반응이 부진할까? 이에 대한 한 가능한 설명은, 수험자들이 단답식으로는 답을 잘 할 수 없지만, 선다형에서 이를 만회할 수 있다고 생각했기 때문일 수 있다. 이에 반해 단답식으로만 볼 때는 이런 만회의 기회가 없기 때문에 끝까지 최선을 다했을 수 있다. 이 가능성을 지지하는 한 증거는, 두 가지 다른 반응을 하게 한 구성적 선다형의 검사 시간과 단답식의 검사 시간상에서 차이가 없음에서 볼 수 있다. 단답식 한 방식으로만 이루어진 검사에서는 단답식에 대해서만 최선을

표 2. 공분산 분석 결과

변수	비표준화 계수		표준화 계수	
	β	표준 오차	β	t
상수	-23.6	5.71	-4.13	
사회 성취도 검사	.79	.064	.63	12.39
집단	-8.74	2.17	-.20	-4.03

다할 수 밖에 없다. 하지만 구성적 선다형방식의 경우 단답식으로 생각해보다가 답을 생각해낼 수 없으면 끝까지 최선을 다하는 대신, 빨리 포기하고 선다형 방식에 치중했기 때문에 검사 시간상에서 차이가 없었던 것으로 보인다. 구성적 선다형에서 단답식에 대한 반응이 부진했던 또 다른 이유는, 단답식 반응에 이어 제시되는 선다형 문항이 일종의 피드백 역할을 하는데서 찾을 수 있다. 구성적 선다형에서는 단답식으로 반응한 다음 선다형 답지가 주어지는데, 이 답지를 보면 자신이 쓴 답이 출제의도에 맞는지를 어느 정도 판단할 수 있다. 따라서 어느 정도 확신하지 않으면 단답식으로 반응하는 것을 부담스럽게 느낄 수 있다. 실제로 구성적 선다형으로 검사가 이루어질 때, 자신이 쓴 답이 선다형의 답지와 거리가 있을 때 당혹스러워하는 장면이 자주 관찰되었다.

그렇다면 어떻게 하면 구성적 선다형 방식에서 학생들로 하여금 두 가지 반응에 대해 모두 최선을 다하게 할 수 있을까? 그리고 그 결과로 구성적 선다형에서의 단답식 반응과 단답식으로만 볼 때의 수행 수준에서의 차이를 없앨 수 있을까? 한 가지 간단한 방법은 단답식의 배점을 높이는 것이다. 이 가능성은 다음 실험에서 탐색되었다.

실험 2

실험 2에서는, 실험 1에서와 마찬가지로 구성적 선다형과 단답식 방식의 비교가 이루어졌는데, 실험 1과의 가장 큰 차이는 지시문이었다. 구성적 선다형 방식으로 검사가 실시된 실험집단의 학생들에게, 단답식과 선다형이 따로 채점되지만, 단답식에 90% 따라서 4.5점이, 선다형에서는 10%, 즉 0.5점으로 각각 배점이 된다고 지시를 한 다음 검사를 실시하였다. 통제 집단의 학생들은 특별한 지시없이 단답식으로 검사를 실시하였다.

연구 방법

실험 대상 실험 1에 참여한 동일한 학생들이 참여하였다. 실험 1이 실시된 후 2주 후에 실시되었기 때문에, 충분한 문항수를 확보하기 위해 사회 대신 과학 과목에서 출제하였다. 수업시간을 고려하여 총 8개 학급을 다시 무선적으로 4반씩 두 집단으로 나누었다.

실험 도구와 절차 자극재료. 초등 6학년 과학 1, 2 단원으로부터 20문항이 출제되었다. 이 문항들은 연구자가 교사의 자문을 받아 만든 문항들이었다. 그밖에 실험에 사용된 프로그램과 실험 절차는 실험 1과 동일하였다. 단답식에 90%이 배점되었다는 지시문과 상관없이 채점은 각 방식에서 100점 만점으로 이루어졌다.

결과와 논의

피험자들 중 1명은 검사에 참여하지 않았고, 3명은 종료 전에 검사를 끝내 그 결과가 기록되지 않았다. 이들을 제외한, 4개의 반의 단답식 집단과 또 다른 4개 반의 구성적 선다형 집단에 대한 기술 통계는 표 3에 제시되었다.

실험 1에서처럼 이전에 실시된 학업성취도

표 3. 실험 2의 주요 결과: 단답식과 구성적 선다형의 단답식 수행 비교 (괄호 안은 표준편차)

검사 방식	단답식	구성적 선다형의 단답식
피험자 수 (명)	110	113
과학 성취도 검사	84 (12)	81 (17)
평균	58 (17)	56 (19)
검사시간 (초)	618 (186)	729 (280)

평가 결과 중 과학에 대한 점수가 비교하였다. 그 결과 단답식 집단과 구성적 선다형 집단의 평균은 각각 84, 81점으로 이들간에 통계적인 차이가 없었다(t (221) = 1.6, p > .1). 단답식 집단의 경우 점수의 범위는 5~95점이었고, 평균은 58점이었다. 학교에서 실시한 성취도 검사 결과와의 상관은 0.64였다. 구성적 선다형 집단의 경우, 선다형 점수의 범위는 25 ~100점이었고, 평균과 표준편차는 각각 75, 11점이었다. 단답식의 범위는 5~90점이었고 평균은 56점이었다. 단답식과 선다형 점수 간의 상관은 이전의 결과보다는 다소 낮은 .70이었다. 실험 1에서보다 상관이 낮아진 이유는 실험 2의 과학 선다형 문제가 다소 쉬워져서 변별력이 떨어져서 인 것으로 보인다. 성취도 검사와 선다형 점수 간의 상관은 0.6이었지만, 성취도 점수와 단답형 점수간의 상관은 0.68로 더 높았다. 성취도 점수와 단답형 점수간의 상관은 구성적 선다형 방식과 단답식으로만 보는 방식 간에 차이가 없었다.

본 실험의 주 관심사인 단답식 점수에서는, 실험 1에서의 결과와 달리, 두 집단간에 큰 차이가 없었다(58 대 56; t (217) < 1, p = .5). 성취도 검사에서 3점 정도의 차이가 났기 때문에, 실험 1에서와 같은 공분산분석(Analysis of Covariance)이 수행되었고, 그 결과는 표 4에 제시되었다. 과학의 성취도 검사 결과가 동일하게 했을 경우, 구성적 선다형 집단의 단답식 평균이 단답식의 평균보다 0.67점 높았는데 이 차이는 통계적으로 유의미하지 않았다. 실험 1에서 관찰된 집단간 평균 점수의 차이가 사라지는 대신, 실험 1에서 차이가 없던 검사 시간의 경우 단답식으로 본 집단보다 구성적 선다형으로 본 집단에서 더 길어졌다 (618 대 732, t (221) = 4.1, p < .001, MSe = 44302). 이 결과는 구성적 선다형 검사 방식의 수험자들이, 단답식에서 최선을 다한 다음 선다형을 풀었을 가능성을 높여준다.

이상의 결과는 구성적 선다형에서 단답식 반응을 더 잘 하도록 하기 위해서 단답식의 배점을 상대적으로 높이는 것이 효과적임을 보여준다. 단답식의 배점을 높임으로써, 각각 따로 채점되어 단답식에서 틀리더라도 선다형

표 4. 공분산 분석 결과

변수	비표준화 계수		표준화 계수	
	β	표준 오차	β	t
상수	-8.4	5.31	-1.58	
사회 성취도 검사	.79	.06	.66	12.91
집단	.67	1.85	.02	.33

에서 맞추어 보충할 수 있다고 생각할 여지가 적어지고, 이로 인해 단답식만으로 시험을 보게 한 집단과 평균 수행 면에서 차이가 나타나지 않은 것으로 보인다. 게다가 시험 시간의 경우, 두 방식간에 차이가 없던 실험 1과는 달리, 구성적 선다형 방식에서 유의미하게 길어졌다. 따라서 구성적 선다형에서 단답식과 선다형으로 반응하게 하였을 때, 적어도 단답식 반응이 단답식만으로 보게 할 경우와 차이가 없도록 하는 한 방법은, 구성적 선다형에서 단답식에 대한 배점을 상대적으로 높이는 것이겠다. 그렇지만 구성적 선다형 방식에서 단답식의 배점을 높임으로써 생길 수 있는 문제가 있다. 그것은 구성적 선다형 방식의 수험자들이 상대적으로 배점이 낮은 선다형에 최선을 다하지 않을 수 있다는 것이다. 이 가능성을 알아보기 위해 실험 3이 수행되었다.

실험 3

실험 3에서는, 실험2에서처럼 구성적 선다형으로 보는 집단에게 단답식과 선다형의 배점이 각각 90% 대 10%임을 알려준 다음 구성적 선다형과 전통적 선다형 시험 방식을 비교하였다. 단답식 점수가 전체 점수의 90%나 되게 되면 상대적으로 선다형에 대한 배점이 작아지고 따라서 수험자들이 선다형에 최선을 다해 반응하지 않을 수 있다. 실험 3은 과연 그런 일이 일어나는지를 확인하기 위해 수행되었다.

연구 방법

실험 대상 실험 2에 참여한 동일한 학생들이 참여하였고, 실험 2에서 단답식으로 본 학생들이 선다형으로 시험을 보도록 하였다.

실험 도구와 절차 자극재료. 초등 6학년 과학 3, 4단원으로부터 20문항이 출제되었다. 이 문항들은 연구자가 교사의 자문을 받아 만든 문항들이었다.

컴퓨터 프로그램. 구성적 선다형 방식은 실험 1, 2에서 사용된 것과 동일하였고, 선다형 방식은 일단 답지를 보면 그 문제를 풀어야 하는 구성적 선다형과는 달리, 여러 번 풀 수 있다는 점에서 차이가 있었다. 모든 문제를 다 풀고 더 이상 수정하기를 원치 않으면 종료 버튼을 눌러 시험을 끝내도록 하였다. 그 밖에 실험 절차는 실험 1과 동일하였다.

결과와 논의

피험자들 중 1명은 시험을 보지 않았고, 1명은 종료 전에 검사를 끝내 그 결과가 기록되지 않았다. 이들을 제외한, 4개의 반의 선다형 집단과 또 다른 4개 반의 구성적 선다형 집단에 대한 기술 통계는 표 5에 제시되었다.

실험 2에서와 동일하게 과학 학업성취도 평가 결과를 비교하였다. 그 결과 선다형 집단과 구성적 선다형 집단의 평균은 각각 83, 80점으로 이들간에 통계적인 차이가 없었다($t(223) = 1.6$, $p > .1$). 선다형 집단의 경우 점수의 범위는 32~92점이었고, 평균은 73점이었다. 학교에서 실시한 성취도 검사 결과와의 상관은 0.57이었다. 구성적 선다형 집단의 경우, 단답식의 범위는 4~80점이었고 평균과 표

표 5. 실험 3의 주요 결과: 선다형과 구성적 선다형의 선다형 수행 비교 (괄호 안은 표준편차)

검사 방식	선다형	구성적 선다형의 선다형
피험자 수	114	111
과학 성취도 검사	83 (12)	80 (17)
평균	73 (13)	75 (14)
검사시간	436 (126)	700 (235)

준편차는 각각 46, 18점이었다. 선다형 점수의 범위는 36~100점이었고, 선다형과 성취도 검사와의 상관은 0.68, 단답식과 성취도 검사간의 상관은 0.72로 단답식이 다소 높았다. 단답식과 선다형 점수 간의 상관은 .80였다.

실험 3의 주 관심사인 선다형 점수의 비교에서는, 두 집단간에 큰 차이가 없었다(73 대 75). 그렇지만, 성취도 검사에서 3점 정도의 차이가 났을 뿐만 아니라, 선다형 시험에서는 반대 방향으로 2점 정도 차이가 나타났기 때문에, 이전 실험에서처럼 공분산분석(Analysis of Covariance)이 수행되었다. 그 결과 표 6에 제시된 것처럼, 과학 성취도 검사를 동일하게 했을 때, 구성적 선다형 집단의 평균 점수가 선다형으로 본 집단의 평균에 비해, 3.7점이 높아졌는데, 이 차이는 통계적으로 유의하였다($p < .01$). 구성적 선다형 방식으로 검사를 실시했을 때, 평균이 더 높아졌을 뿐만 아니라, 검사 시간에서도 선다형 집단에 비해 더 길어졌다(437 대 700, t (223) = 10.5, $p < .001$, MSe = 35223).

이 결과는 구성적 선다형에서 단답식 반응을 더 잘 하도록 하기 위해서 단답식의 배점을 90%로 높이는 동시에 선다형에 대한 배점을 낮추더라도, 그로 인해 선다형에 대한 수행이 떨어지지 않음을 보여준다. 수행이 떨어지기보다는 오히려 선다형에 대한 반응이 향상되었다. 이 결과는 구성적 선다형 방식에 대한 반응과 선다형 방식 간의 반응 동등성을 지지하지 않는데, 수행이 저하되는 대신 향상되었다는 점에서 예상치 않은 결과이다. 따라서 이에 대한 반복 검증이 시급히 이루어질 필요가 있다. 구성적 선다형 방식에서 선다형 반응 점수가 더 높은 이유에 대한, 현재로서 가능한 사변적 설명은, 구성적 선다형에서 단답식으로 먼저 푸는 과정이 보다 신중하게 선다형에 임하게 했을 수 있다는 것 정도이다. 본 연구에 참여한 초등학생들은 경쟁적으로 열심히 공부하는 지역의 학생들이라 나름대로 최선을 다했겠지만, 성적에 반영되는 검사가 아니었기에, 구성적 선다형에서보다 선다형에서 상대적으로 적당히 반응을 했었을 수 있다. 이로 인해 그냥 선다형을 푼 집단의 학생들보

표 6. 공분산 분석 결과

변수	비표준화 계수		표준화 계수	
	β	표준 오차	β	t
상수	25.57	4.18	6.12	
사회 성취도 검사	.57	.05	.619	11.70
집단	3.69	1.42	.138	2.61

다 문제 해결을 위해 더 깊은 생각이 촉발되었을 수 있고, 이것이 수행에 영향을 주었을 수 있다. 이런 가능성을 시사하는 한 연구는 Berg와 Smith(1994)에서 볼 수 있다. 이들은 그래프가 정보를 '상징'하는 것임에도 불구하고 그 보다는 단순히 '그림'으로 잘못 이해하고 있음을 보여주었다. 예를 들면, 학생들에게 한 지점으로부터 벽까지 걸어갔다가 되돌아 올 때의 거리와 시간의 관계 그래프를 고르도록 하였다. 많은 경우 단순히 걸어갔다가 되돌아 올 때의 경로를 나타내는 그래프를 고르거나 갔다가 되돌아 온 거리의 총 길이를 나타내는 그래프를 선택하였다. 놀랍게도 인터뷰를 하거나 학생들로 하여금 직접 그리게 하였을 때는 오히려 그래프의 상징적인 측면을 더 잘 이해하고 있음을 발견하였다. 요컨대, 자신이 이해한 대로 반응하게 하면 어느 정도 해결할 수 있는 문제라도, 매력적인 오답지가 답지로 제시되는 선다형에서는 그로 인해 오히려 수행이 떨어질 수 있다는 것이다.

선다형의 수행을 향상시킬 수 있다는 점에서 이 가능성은 그 자체로 흥미로운 연구 주제가 될 수 있다. 그렇지만, 이보다 더 시급한 것은 앞에서도 언급했던 것처럼, 이런 결과를 좀 더 다양한 상황에서 반복 검증하는 일일 것이다. 지금으로서는 적어도 구성적 선다형 방식에서 단답식에 더 많은 배점을 하면, 그 때문에 선다형에 대한 수행이 떨어지지는 않는다는 결론 정도가 안전해 보인다. 사실, 단답식에서 최선을 다해 나름대로 답을 찾은 다음, 배점이 작다고 선다형을 대충 푸는 일은, 상상하기 어렵다.

전체 논의

구성적 선다형 시험방식은 하나의 발문에 대해 먼저 단답식으로 반응하고 나서 선다형으로 한 번 더 반응하게 한다. 이 방식은 같은 검사를 두 번 실시하게 하는 것의 실제적 어려움을 겪지 않으면서도, 결과적으로는 두 가지 다른 방식의 반응을 얻도록 하기 위해 고안되었다. 본 연구는 구성적 선다형 방식으로 시험을 보았을 때 얻어지는 자료가, 과연 두 방식을 따로 따로 풀게 하였을 때 얻는 결과와 동등한지를 알아보기 위해 수행되었다. 실험 1과2에서는 구성적 선다형으로 보았을 때와 단답식으로 시험을 보게 하였을 때의 차이를 알아보았다. 실험 1과 2의 차이는 구성적 선다형에서 단답식과 선다형에 대한 배점이었다. 실험 1에서는 특별한 비율 대신 각각 따로 채점된다는 지시문이 주어졌지만, 실험 2에서는 단답식에 90% 선다형에 10%라는 구체적인 숫자가 제시되었다. 그 결과 실험 1에서는 단답식과 구성적 선다형의 단답식 점수간에 큰 차이가 관찰되었지만, 실험 2에서는 그 차이가 사라졌다. 평균 점수상에서의 이런 변화와 함께, 시험 시간의 변화도 주목할 만하다. 실험 1에서는 구성적 선다형 방식과 단답식 방식간에 시험 시간상에서 차이가 없었지만, 실험 2에서는 유의미한 차이가 발견되었다. 수험자들은 단답식을 풀기 위해 더 많은 시간을 소비하였으며, 단답식에 이어 제시되는 선다형을 풀려고 하였다. 이 결과는 구성적 선다형에서 단답식에 대한 반응이, 단답식으로만 보게 한 집단과 비슷한 수준이 되게 하려면 구성적 선다형의 단답식 배점을 높임

으로써 가능함을 보여준다.

실험 3에서는 구성적 선다형 방식에서 단답식과 선다형의 배점이 90%-10%라는 지시문이 주어진 상태에서 선다형 방식간의 비교가 이루어졌다. 구성적 선다형에서 선다형에 대한 배점이 상대적으로 작기 때문에 단답식에서는 나름대로 최선을 다하지만, 선다형에서는 최선을 다해 답을 하지 않는지를 알아보려고 하였다. 그 결과, 구성적 선다형의 평균이 오히려 선다형으로 보았을 때의 평균보다 상대적으로 높아짐을 발견하였다. 이 결과는 선다형만으로 풀 때보다는, 구성적 선다형으로 풀 때 먼저 단답식으로 반응하는 과정이 추가되기 때문에 더 깊은 사고를 유도했기 때문일 수 있다. 그렇지만 통계적인 차이에도 불구하고, 상대적으로 효과의 크기가 작았기 때문에 이 가능성은 후속 연구를 통해 재검토될 필요가 있다. 현 단계에서는 적어도 구성적 선다형에서 단답식의 배점을 높이고 선다형의 배점을 낮게 하더라도 그로 인해 선다형에 대한 수행이, 선다형으로 볼 때와 비교하여, 더 떨어지지는 않는다고 하는 것이 안전해 보인다.

이상의 결과를 요약하자면, 구성적 선다형과 선다형 방식의 동등성에 대해서는 후속 연구에서 좀 더 명료해질 문제가 남아있지만, 구성적 선다형에서 단답식의 비중을 높임으로 반응 동등성을 확보할 수 있는 것으로 보인다. 본 연구를 통해 다시 한 번 확인하게 된 것은 수험자들은 가능하면 최소의 노력으로 좋은 점수를 얻으려는 전략가로 볼 수 있다는 것이다. 단답식은 교사에게는 채점에 대한 부담이 있지만, 수험자들에게도 답을 스스로 산출해야 한다는 부담이 따른다. 따라서 선다형 보다 많은 인출 노력이 요구되는 단답식으로 반응을 하게 하려면 그에 상응하는 댓가가 필요하다. 본 연구에서는 점수 배점을 다르게 하여 수행에 영향을 줄 수 있음을 확인하였다. 그렇지만 구성적 선다형 방식의 특성에 대한 연구는 아직은 초기 단계이고, 본 연구도 다음과 같은 제한점이 있으므로 앞으로의 연구를 통해 충분히 재검토될 필요가 있다.

우선 한 초등학교 학생들을 대상으로 한 연구였다. 따라서 동질 집단을 대상으로 지시문의 효과를 알아볼 수 있었지만, 외적 타당도가 낮다는 한계가 있다. 앞으로의 연구에서는 더 다양한 피험자 집단을 대상으로 반복 검증되어야 하고, 가능하다면 한 모집단을 대상으로 구성적 선다형, 단답식, 그리고 선다형을 동시에 비교하는 연구가 이루어질 필요가 있다. 또 다른 제한점은 본 연구가 성적에 반영되지 않는 검사였다는 점이다. 따라서 실제 성적에 반영되는 상황에서는 어떤 차이가 나타나는 지를 알아보는 연구가 필요하다. 이를 위해서는 많은 동질적인 피험자 집단과 동시 접속이 가능한 컴퓨터실이 확보되어야 하는데, 컴퓨터화 검사가 활성화되면 이런 연구가 실제 장면에서 쉽게 이루어질 것으로 예상된다. 이런 후속 연구를 통해 구성적 선다형 방식에서 얻은 두 가지 반응이, 단답식이나 선다형의 어느 한 방식으로 얻은 결과와 차이가 없게 하는 조건과 절차가 발견될 때, 한 번의 검사로 두 가지 다른 반응을 얻고자 하는 구성적 선다형 방식의 개발 의도가 달성될 것이다.

본 연구의 초점은 한 문항에 대해 두 가지 다르게 반응하게 할 때 그 둘간의 동등성 확

보 조건을 확인하는 것이었다. 그렇지만 설사 이런 동등성이 확보되지 않더라도 이 방식은 실제 평가 장면에서 활용될 가능성이 높다는 점이 지적될 필요가 있다. 우선 성적에 반영되는 시험일 경우, 본 연구에서처럼 점수 비중을 크게 하지 않더라도 단답식에 더 성실하게 반응할 가능성이 높다. 또 다른 활용방식은 단답식과 선다형 모두에서 맞을 때만 점수를 주는 것이다. 이 경우 채점자는 먼저 선다형 반응에서 맞은 문항에 대해서만 단답식의 답을 확인하면 되므로 단답식으로 볼 때보다 채점부담을 줄일 수 있다. 마지막으로 선다형 점수와 단답식 점수를 순차적으로 활용하여 평가하는 것도 생각해볼 수 있다. 통상 대학에서 이루어지는 5단계 평가를 예로 들면, C, D, F는 선다형 점수를 기준으로 결정한 다음, 나머지 학생들을 대상으로 단답식 점수를 기준으로 등급을 A, B로 나눌 수 있겠다. 이처럼 구성적 선다형 방식을 실제 교육 장면에서 어떻게 활용할 수 있는 방법은 다양하게 이루어질 수 있는데, 그 사례들은 후속 연구를 통해 보고될 것이다.

참고문헌

박도순, 김종필, 양길석 (2002). 컴퓨터검사와 지필검사의 점수 동등성에 관한 메타분석, 교육평가연구, 15(1), 247-272.

박주용, 민경석 (2009). 구성적 선다형 검사에서 선다형과 단답형의 문항 특성 비교. 교육평가연구, 22(2), 451-469.

최윤정, 성태제 (2006). 영어 논술 채점 컴퓨터 프로그램의 비교분석, 교육평가연구, 19(1), 145-160.

Bennett, R.E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B. & Yan, F. (2008). Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP. *Journal of Technology, Learning, & Assessment*. 6 (9). Retrieved July 25, 2009, from http://escholarship.bc.edu/jtla/.

Berg, C. A., & Smith, P. (1994). Assessing students' abilities to construct and interpret line graphs: Disparities between multiple-choice and free-response instruments. *Science Education, 78* (6), 527-554.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning,* & *Assessment*. 5 (1). Retrieved June 12, 2008, from http://escholarship.bc.edu/jtla/.

Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp.287-301). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Leacock, C., & Chodrow, M. (2003). C-rater: Automatied scoring of the short-answer questions. *Computers and the Humanities, 37*, 389-405.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (4th ed., pp.13-104). New York: Macmillan.

Park, J. (2010). Constructive multiple-choice testing system. *British Journal of Educational Technology, 41*(6), 1054-1064.

Rodriguez, M. C. (2003). Construct equivalence of

multiple-choice and constructed-response items: a random effects synthesis of correlations, *Journal of Educational Measurement, 40*(2): 163-184.

Veloski, J. J., Rabinowitz, H. K., Robeson, M. R., & Young, P. R. (1999). Patients don't present with five choices: An alternative to multiple-choice tests in assessing physician's competence. *Academic Medicine, 74,* 539-546.

Wainer, H., & Thissen, D. (1993). Combining multiple choice and constructed response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103-118.

Wang, J., & Brown, M. S. (2007). Automated Essay Scoring Versus Human Scoring: A Comparative Study. *Journal of Technology, Learning,* & *Assessment. 6*(2). Retrieved May 10, 2008, from http://escholarship.bc.edu/jtla/.

1 차원고접수 : 2010. 10. 8
2 차원고접수 : 2010. 11. 16
최종게재결정 : 2010. 11. 23

한국심리학회지 : 인지 및 생물
The Korean Journal of Cognitive and Biological Psychology
2010, Vol. 22, No. 4, 573-588

How differential credit assignment affects students' responses in the Constructive Multiple-choice Testing System

Jooyong Park

Seoul National University

The Constructive Multiple-choice Testing (CMT) system is a new computerized testing system developed to supplement the weaknesses of the multiple-choice (MC) format. The CMT system involves having the examinee respond to the stem first in the short answer format and then in the MC format. Therefore, one can see whether or not the examinee chose the correct option in the MC format because he or she actually knew the answer by checking the short answer portion of the examinee's response. The current study was carried out to examine whether there is any difference in the scores obtained from the CMT test as opposed to the short answer or the MC tests. Two hundred and twenty seven 6th graders in elementary school were randomly assigned to 2 groups. In Experiments 1 and 2, comparison was made between the performance of the group who took the test in the CMT format and that of the group who took the test in the short answer format. In Experiment 1, where the instruction for the CMT test was that the two portions would be graded separately, the mean of the short answer portion of the CMT group was lower than that of the short answer format group. However, in Experiment 2, the examinees were told that the short answer portion would be weighed 9 times as heavily as the multiple-choice portion (90% vs. 10%). There was no difference in the means of the short answer responses between the two groups. In Experiment 3, the means of the multiple-choice responses were compared between the CMT group and the multiple-choice group. The CMT group was given the same instruction as in Experiment 2. The result revealed that the mean of the MC portion of the CMT group was higher than that of the MC group. These results suggest that the performance is not affected by having the examinees respond twice if more points are allotted to the short answer portion of the CMT format.

Key words : Computerized Testing, Constructive Multiple-choice Testing System, Response Equivalence, Short Answer Format, Multiple-choice Format