인지과학, 제21권 제4호
Korean Journal of Cognitive Science
2010, Vol. 21, No. 4, 627～653.

풀이 과정을 답지로 이용한 시험 방식의 학습 효과

임 정 만
세종대학교 교육학과

박 주 용†
서울대학교 심리학과

본 연구에서는 수학 학습을 위한 평가 도구로 풀이 과정을 답지로 이용한 시험 방식이 제안되었고, 그 효과가 검증되었다. 풀이된 예제 연구에 따르면, 학생들에게 문제에 대한 단계적인 해결책인 풀이된 예제를 문제 사이에 제시할 경우, 문제만 풀 때보다 효과적으로 학습한다. 그러나 풀이된 예제의 단순한 제시는 학습 효과가 제한적이라는 최근의 발견들이 있었다. 이에 따라 본 연구에서는 학습자가 풀이된 예제를 더 적극적으로 탐구하게 하는 방법으로 풀이 과정을 선다형의 답지로 제시하였다. 이 시험 방식은 컴퓨터화 시험으로 구현되었으며, 학생들이 컴퓨터에서 단답형으로 문제를 풀고 나서 답지를 요청하면 선다형 답지가 제시되었다. 이 때 실험 집단은 답지가 풀이 과정으로, 비교 집단은 전통적 선다형과 같이 최종 정답으로 구성되었다. 초등학교 6학년 학생들을 대상으로 실험이 수행되었다. 사후 검사 결과 실험집단의 평균 점수가 비교집단의 평균 점수보다 높았다. 이 결과는 풀이 과정을 답지로 이용한 시험 방식이 교실현장에서 학습을 촉진하기 위한 도구로 활용될 수 있음을 시사한다. 끝으로 후속 연구의 방향성이 논의되었다.

주제어 : 컴퓨터화 평가 시스템, 학습을 위한 평가, 풀이된 예제

---

* 이 논문은 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음 (KRF-2007-B00130).
임정만, 세종대학교 교육학과, 연구 세부 분야 : 학습과 기억, 인지심리 연구의 교육적 적용, 컴퓨터화 시험, E-mail : jmanlim@gmail.com
† 교신저자: 박주용, 서울대학교 심리학과
E-mail : jooypark@snu.ac.kr

서 론

최근 수학적 힘(mathemetical power)이 수학 교과과정의 주요한 목표로 강조되고 있다. 수학적 힘은 문제해결, 논리적 분석, 추론 등을 효과적으로 적용하는 능력을 가리킨다 (Schoenfeld, 1992; Santos-Trigo, 2007). 이를 위해 교수 방법 및 교과과정을 개선하려는 노력이 진행되어왔다. 본 연구에서는 교수 방법과 교과과정 대신, 평가를 개선함으로써 수학 학습을 향상시킬 방법을 찾고자 하였다. 이를 위해 풀이 과정을 선다형 답지로 사용하는 새로운 컴퓨터화 시험 방식을 제안하였고, 그 효과를 검증하고자 하였다.

평가를 통해 수학 학습을 향상시킬 방법을 찾고자 한 이유는 평가가 학습과 밀접한 관련이 있기 때문이다(Boud & Falchikov, 2007; Tillema, 2009). 학습을 위한 교실 평가는 학생들의 학업성취도를 향상시킬 수 있으며(예, Black & William, 1998), 시험은 복습보다 학습을 강화한다(예, Carpenter, Pashler, & Cepeda, 2009; Roediger & Karpicke, 2006). 평가가 학습과 밀접한 관련이 있지만 학습자가 교수자의 적절한 안내 없이 문제 풀이를 시도할 경우 학습에 어려움을 겪을 수 있다는 문제가 있다 (Sweller, 2006).

이에 대한 대안으로 학생들에게 문제 사이에 풀이된 예제(worked examples)를 제시하는 교수 방법이 제안되었고, 여러 실험 연구는 이 방법이 문제풀이 전략보다 효과적이라고 보고하였다(예, Große & Renkl, 2006; Sweller & Cooper, 1985). 그러나 후속 연구에서 풀이된 예제 효과가 학습자의 능력과 상호작용한다는 사실이 밝혀졌다(Kalyuga, Chandler, & Sweller, 2001). 즉, 사전 지식이 높은 학생들은 풀이된 예제로부터 학습의 이득을 보지 못했다. 이 문제를 해결하기 위한 후속 연구 중 Große와 Renkl(2007)은 학습자들에게 올바른 예제와 틀린 예제를 함께 제시하였는데, 사전 지식이 높은 학생들의 학업성취도가 증가됨을 발견하였다.

본 연구는 이에 착안하여 풀이 과정을 선다형의 답지로 제시하고 이를 단답형 문제 사이에 끼워 넣음으로써, 단답형 문제를 풀고서 같은 문제에 대해 답지가 풀이 과정으로 구성된 선다형 문제를 푸는 시험 방식을 개발하였다. 이때 학생들이 단답형을 푸는 동안 선다형 답지로 제시되는 풀이 과정을 볼 수 없게 해야 하는데, 이를 위해 박주용과 김용국(2010)에 의해 제안된 초등학생용 컴퓨터화 개방형

수학 시험 방식이 활용되었다. 이 시험 방식에서 학생들은 수식 입력기를 이용해 단답형을 먼저 풀고서 선다형 답지를 요청하면 답지가 제시되어 선다형에 응답하게 된다. 본 연구에서는 이 시험 방식에서 선다형 답지를 올바른 풀이 과정과 틀린 풀이 과정으로 제시하였다.

틀린 풀이 과정 사이에서 올바른 풀이 과정을 찾는 추가적인 활동은 문제 사이에 하나의 풀이된 예제를 제시하는 것보다 부가적인 인지 자원을 요구하게 된다. 그러나 이는 학습자가 올바른 해결책을 찾을 때까지 임의의 탐색 활동을 계속하게 하기보다는 몇 가지 한정된 답지 내에서의 탐색 활동이다. 또한, 선다형에서 오답지를 가려내고 정답을 찾을 확률을 높이는 추측 전략은 학생들에게 익숙한 방법이다. 따라서 풀이 과정으로 구성된 답지 내에서의 탐색 활동이 인지 과부하를 가져오기보다는 학습 내용의 제시 방법이 효과적일 때 나타나는 적절한 인지부하(germane cognitive load)를 촉진하여, 학습자의 수행에 도움을 줄 것으로 예상하였다(Chandler & Sweller, 1991; Sweller, Merriënboer, & Paas, 1998).

이를 검증하기 위한 실험이 초등학교 6학년 학생들을 대상으로 수행되었다. 실험집단은 단답형에서 문제를 풀고 올바른 풀이 과정과 틀린 풀이 과정으로 구성된 답지에서 올바른 풀이 과정을 선택하도록 지시받았다. 반면, 비교집단은 단답형에서 문제를 풀고 나서 전통적 선다형처럼 최종적인 해(solution)로 정답과 오답지가 구성된 답지에서 정답을 선택하도록 지시받았다.

이론적 배경

학습을 위한 평가

평가의 목적은 크게 두 가지로 구분할 수 있다. 하나는 학생들이 얼마나 배웠는지를 확인하고자 하는 목적으로, 평가 결과는 학생들의 진학과 취업에서 선별을 위한 자료로 사용된다. 이는 총괄 평가(summative assessment)의 역할로 알려졌다. 다른 하나는 학생들의 학습을 촉진하고자 하는 목적으로, 평가 결과는 학생들의 수행 수준을 진단하여 앞으로의 학습 방향을 제시하고, 교사들은 이를 통해 교수를

개선한다. 이는 형성 평가(formative assessment)의 역할로 알려졌다 (Boud & Falchikov, 2006).

최근 평가의 강조점이 총괄 평가에서 형성 평가로, 그리고 학습 결과를 보고하기 위한 목적의 학습에 대한 평가(assessment of learning)에서 학습을 촉진하기 위한 목적의 학습을 위한 평가(assessment of learning)로 이동하고 있다 (Earl, 2003; Lee, 2007). 이는 두 가지 사실에서 확인된다. 먼저, 최근에 교실기반 평가에 대한 연구가 급속히 증가하였을 뿐만 아니라 학습을 위한 평가가 여러 학회에서 핵심적인 의제로 자리 잡기 시작했다 (Boud & Falchikov, 2007). 다음으로, 세계 각국의 교과과정 정책에서 학습을 위한 평가가 강조되고 있다. 이러한 변화는 영국, 호주, 그리고 홍콩에서 특히 두드러진다 (Lee, 2007).

학습을 위한 평가가 강조되는 배경에는 Black과 William(1998)의 연구가 있다. 이들은 평가와 학습을 연결 짓는 250개 이상의 연구를 개관하였다. 그 결과 학습을 위한 교실평가가 학생들의 학업성취도를 향상시킨다는 결론이 도출되었다. 이는 특히 교실에서 상대적으로 학업성취도가 낮은 학생들에게 효과적인 것으로 밝혀졌다. 그럼에도 Kvale(2007)에 따르면, 여러 고등교육 기관들이 평가를 통해 학습을 촉진해야 한다고 명시하고 있지만 이는 여전히 부차적인 역할에 머물고 있다. 또한, 그는 평가가 학습을 촉진하는 목적을 달성하기 위해 심리학에서 발견된 효과적인 학습 원리를 따를 필요가 있다고 보았다.

학습을 위한 평가의 강조와 더불어 시험이 그 자체로 학습을 강화한다는 경험적 증거들이 있다(예, Glover, 1989; Roediger & Karpicke, 2006). 이는 평가를 통해 학습을 촉진할 수 있다는 주장을 뒷받침한다.

시험 효과

시험에 관한 여러 연구는 학생들이 배운 내용을 다시 읽을 때보다 시험을 볼 때 그 내용을 더 오래 기억함을 보여주었다. 이러한 현상은 시험 효과(testing effect)로 불린다(Glover, 1989; Roediger & Karpicke, 2006, Butler & Roediger, 2007). 시험 효과에 대한 Roediger와 Karpicke(2006)의 메타분석에서 대부분의 연구 결과는, 피드백이 없는 경우에도 시험이 복습보다 장기적으로 인출을 더 촉진한다고 보고하였다.

한 예로, Carpenter, Pashler와 Cepeda(2009)는 교실 현장에서 장기간의 파지간격을 두고 시험 효과를 검토하기 위해 미국의 차터 스쿨(charter school) 8학년 학생들을 대상으로 실험을 수행하였다. 학생들은 미국 역사에 대한 수업이 끝나고 세 집단에서 각각 다른 활동을 수행하였다. 첫 번째 집단은 시험을 보고 바로 피드백을 받았다. 두 번째 집단은 배운 내용을 다시 공부했다. 세 번째 집단은 수업만 들었다. 이때 시험 집단과 복습 집단은 수업이 끝나고 각자가 속한 집단의 활동을 시작하는 시점에 따라 다시 두 집단으로 구분되었다. 각 집단에서 절반의 학생들은 1주일 후에, 그리고 나머지 절반의 학생들은 16주 후에 각자가 속한 집단의 활동을 수행하였다. 이는 똑같은 시간을 공부하더라도 공부 간격을 분산시키거나 복습 시점을 지연시킬수록 효과적이라는 분산 효과(spacing effect)가 시험을 통해서도 발견되는지를 확인하기 위함이었다(예, Rohrer & Taylor, 2006).

사후검사는 학생들의 활동이 완료되고 나서 36주, 즉 아홉 달이 지나서 시행되었다. 그 결과 1주일 후에 시험을 본 학생들보다 16주 후에 시험을 본 학생들의 점수가 더 높았다. 마찬가지로 1주일 후에 복습을 한 학생들보다 16주 후에 복습을 한 학생들의 점수가 더 높았다. 즉 시험을 통해서도 분산 효과가 확인되었다. 또한, 1주일 후와 16주 후 모두에서 시험을 본 학생들이 복습을 한 학생들보다 더 많은 학습 내용을 회상했다.

이 연구 결과는 본 연구와 관련하여 두 가지 측면에서 중요하다. 하나는 실제 교실 환경에서 매우 긴 파지 간격으로 다시 한 번 시험 효과를 확인했다는 것이고, 다른 하나는 시험이 분산 학습과 같은 효과적인 학습 원리와 결합할 경우, 그 효과를 극대화할 가능성을 보여주었다는 것이다. 이는 풀이된 예제 효과를 통해 시험의 학습 이득을 극대화할 수 있음을 시사한다.

인지부하이론

인지부하이론(cognitive load theory)은 풀이된 예제 효과의 기초가 되는 이론으로 가장 주요하게는 작업기억의 제한된 능력을 고려하며, 작업기억과 장기기억 간의 관계에 주목한다. 작업기억의 능력은 처리용량과 지속시간에서 매우 제한되어 있지만, 장기기억의 용량은 거의 무한하기 때문에 인지부하이론 연구자들은 학습의

궁극적인 목적을 장기기억의 변화로 본다(Clark, Nguyen, & Sweller, 2006).

특히 스키마 습득이 강조되는데, 정보가 장기기억에 스키마 형태로 저장되어 자동화 될 때 학습자는 제한된 작업기억의 능력과 관계없이 복잡한 정보를 처리할 수 있기 때문이다 (Kalyuga, Chandler, & Sweller, 2001). 따라서 인지부하이론은 작업기억에 과도한 부하를 일으키는 불필요한 인지부하를 줄이고, 효과적인 스키마 습득을 위해 적절한 인지부하를 촉진하는 교수 설계 및 학습 방법을 탐구한다. 이를 위해 연구자들은 인지부하의 종류를 세 가지로 구분하였다.

먼저, 본질적 인지부하(intrinsic cognitive load)는 과제 또는 학습 내용의 복잡성과 관련이 있다. 이는 학습자가 어떤 학습 내용을 이해하기 위해 동시에 처리해야 하는 요소의 수와 관련이 있는 부하이다(Renkl & Atkinson, 2003). 다음으로, 비본질적 인지부하(extraneous cognitive load)와 적절한 인지부하(germane cognitive load)는 교수 설계에 의해 부과되는 인지부하이다. 비본질적 인지부하는 과제 또는 학습 내용을 제시하는 방법이 비효율적일 때 부과되는 인지부하이며, 적절한 인지부하는 비본질적 인지부하처럼 과제 또는 학습 내용을 제시하는 방법에서 비롯되지만, 학습을 촉진하는 부하이다 (Sweller, Merriënboer, & Paas, 1998). 수학과 물리학처럼 잘 구조화된 영역에서 학습자에게 풀이된 예제를 제공하는 방법은 적절한 인지부하를 촉진하는 것으로 알려진 교수 설계 중 하나이다.

풀이된 예제를 통한 학습

풀이된 예제는 학습자에게 문제해결을 위한 단계적인 논증을 제시하는 것으로 1) 문제, 2) 문제에 대한 해결 단계, 그리고 3) 최종적인 정답으로 구성되어 있다 (Moreno, 2006). 풀이된 예제 효과는 Sweller와 Cooper(1985)의 고전적 연구에서 밝혀졌다. 이들은 대수학에서 풀이된 예제와 함께 문제를 제시받은 학생들이 문제만 풀었던 학생들보다 더 좋은 수행을 보임을 발견했다. 그러나 후속 연구에서 풀이된 예제 효과가 학습자의 사전 지식 및 경험과 상호작용한다는 결과가 관찰되었다.

Kalyuga, Chandler와 Sweller(2001)는 풀이된 예제 효과에서 학습자의 사전 지식 및 경험의 역할을 검토하기 위한 실험을 수행하였다. 이 연구의 실험 1은 지름과 원

주 전환에 대한 도표 학습에서 탐구기반 교수 접근과 풀이된 예제 교수 접근을 비교하였다. 실험 결과, 풀이된 예제 교수 방법은 탐구 기반 교수 방법보다 근소하게 더 효과적이었다. 그러나 더 복잡한 과제를 사용한 실험 2에서는 좀 더 분명한 결과가 나타났다. 과제에 익숙하지 않은 초기 학습 기간에는 풀이된 예제 교수 방법이 효과적이었으나, 이후 학습자들이 과제에 친숙해짐에 따라 풀이된 예제 효과는 사라지고 탐구 기반 교수 방법이 더 효과적이었다. 즉 학습자의 경험이나 숙달 수준이 증가하면 풀이된 예제의 학습 이득은 감소되었다(Kalyuga, Chandler, & Sweller, 2003). 이 현상은 연구자들에 의해 전문반전효과(expertise reversal effect)로 지칭되었다(Kalyuga, Chandler, & Sweller, 1998).

Renkl(1999)은 이를 “이해의 착각(illusion of understanding)”으로 설명하였다. 사전 지식 및 경험이 많은 학습자들은 풀이된 예제를 보고 쉽사리 이해한 것으로 착각하여 문제해결 방법에 대해 충분히 숙고하지 않을 수 있다는 것이다. 따라서 학습자들이 풀이된 예제를 보다 적극적으로 처리할 수 있는 추가적인 활동이 요구될 필요가 있다. 전문반전효과를 고려한 후속 연구들은 학습에 도움을 주는 부하인 적절한 인지부하를 촉진하기 위해 풀이된 예제의 새로운 제시 방법 및 추가적인 활동을 탐구하였다.

후속 연구를 통해 제안된 새로운 제시 방법 및 추가적인 활동은 크게 네 가지로 분류될 수 있다. 첫째는, 문제-예제 쌍의 제시 순서를 변화시켰다(예, Reisslein, Atkinson, Seeling, & Reisslein, 2006). 둘째는, 풀이된 예제의 일부 또는 전부가 점차 사라지도록 했다(예, Reisslein, Atkinson, Seeling, & Reisslein, 2006; Salden, Aleven, Renkl, & Schwonke, 2009). 셋째는, 학습자에게 풀이된 예제를 스스로 설명하도록 했다(예, Catrambone & Yuasa, 2006; Große & Renkl, 2006, 2007). 마지막으로, 다양한 예제를 제시하거나 틀린 예제를 함께 제시했다(예, Große & Renkl, 2006, 2007). 이러한 시도들은 대개 또는 제한적으로 학습에 긍정적인 효과를 보여주었다.

본 연구와 관련하여 중요한 연구는 Große와 Renkl(2007)에 의해 수행되었다. 이전의 풀이된 예제 효과 연구들과는 달리, 이들은 학습자들이 오류를 통해 보다 깊이 있게 학습할 수 있다는 점에 주목하여 학습자들에게 올바른 예제와 틀린 예제를 함께 제공하였다. 틀린 예제를 함께 제시하는 것의 학습 효과를 검토하기 위해 이들은 대학생들을 대상으로 총 여섯 가지 조건에서 확률 문제를 풀게 하였다. 가

장 주요한 요인은 풀이된 예제의 제시 방법으로, 첫 번째 조건에서는 올바른 예제와 틀린 예제가 함께 제공되었는데, 틀린 예제에서 오류 사항이 강조되었다. 두 번째 조건에서는 오류의 강조 없이 틀린 예자가 올바른 예제와 함께 제공되었다. 마지막 조건에서는 올바른 예제만 제공되었다. 이와 함께 또 다른 요인으로 학습자의 자기설명 유무에 따른 학습 효과가 검토되었다.

사후검사는 피험자들에게 제시되었던 풀이된 예제의 구조와 동형으로 만들어진 근전이(near transfer) 문항 4개와 구조적 특징은 같지만 정확히 동일한 구조는 아닌 원전이(far transfer) 문항 11개로 구성되었다. 사후검사 결과 학습자의 사전 지식과 학습 결과 사이의 상호작용이 발견되었다. 먼저, 원전이 문항 점수에서 유의미한 차이가 발견되었다. 사전지식이 높은 학습자들은 올바른 예제만 제시된 조건보다 올바른 예제와 틀린 예제가 함께 제시된 조건에서 더 높은 점수를 받았다. 반면, 사전 지식이 낮은 학습자들은 올바른 예제와 틀린 예제가 함께 제시된 조건보다 올바른 예제만 제시된 조건에서 더 높은 점수를 받았다. 근전이 문항 점수에서는 유의미한 차이가 발견되지 않았다. 다음으로, 틀린 예제에서 오류 강조 유무에 따른 차이는 근전이 문항 점수에서 발견되었다. 사전지식이 높은 학습자들은 오류가 강조되지 않은 틀린 예제를 통해 더 높은 점수를 받은 반면, 사전지식이 낮은 학습자들은 오류가 강조된 틀린 예제를 통해 더 높은 점수를 받았다. 자기설명에 따른 차이는 발견되지 않았다.

이상의 결과는 올바른 예제와 틀린 예제를 함께 제공할 때 사전 지식이 높은 학습자들의 전이 수행을 촉진할 수 있음을 보여준다. 그러나 사전 지식이 낮은 학습자들은 올바른 예제만 제시되었을 때 학습의 이득을 얻었으므로 그 효과는 여전히 제한적이다. 두 가지 가능성이 제기될 수 있는데, 사전 지식이 낮은 학습자들이 제시된 예제들을 적극적으로 탐색하지 않았기 때문일 수 있다. 또한, 틀린 예제의 제시가 사전 지식이 낮은 학습자들에게 작업기억의 과부하를 가져왔기 때문일 수 있다. 틀린 예제 제시의 학습 효과를 평가에서 활용하는 한편, 사전 지식이 낮은 학습자들에게 나타난 제한적인 학습 효과를 극복하기 위한 방안으로 본 연구에서는 선다형의 특징을 활용하였다.

선다형 검사

시험의 형식은 크게 선다형(multiple-choice)과 구성형(constructed-response)으로 구분된다. 이 두 시험 형식의 결정적인 차이는 미리 정해진 답지가 있는지 여부이다. 즉 선다형에서 학습자는 답을 선택하는 반면, 구성형에서 학습자는 답을 스스로 만들어야 된다(Bible, Simkin, & Kuechler, 2008). 다양한 교육 현장과 학문 영역에서 가장 널리 사용되어온 시험 방식은 선다형이다. 실제로 대부분의 교사와 학생들은 선다형을 다른 시험 형식보다 더 선호한다(Betts, Elder, Hartley, & Trueman, 2009; Simkin & Kuechler, 2005).

교사들이 선다형을 선호하는 이유는 기계가 채점하면 채점 정확성을 증가시킬 수 있고(Holder & Mills, 2001), 그래서 경제적일뿐더러 채점 절차가 객관적이라는 인상을 줄 수 있다(Becker & Johnson, 1999). 그렇기 때문에 학생들은 선다형 시험을 다른 시험 형식보다 객관적인 것으로 받아들이며(Simkin & Kuechler, 2005), 교사는 광범위한 영역에서 많은 문항을 출제할 수 있다 (Delgado & Prieto, 2003). 그러나 학생들은 답지에서 틀린 것을 가려냄으로써 추측을 통해 올바르게 응답할 확률을 높일 수 있고 (Frary, 1988; Bush, 2001), 이러한 전략 때문에 내용을 깊이 있게 이해할 필요가 없어서 선다형에 우호적이다(Entwistle & Entwistle, 1992). 이로 인해 선다형 시험은 피상적인 학습을 촉진한다는 비판을 받아왔다(예, Nicol, 2007; Williams & Clark, 2004). 그럼에도 Marsh, Roediger, Bjork과 Bjork(2007)은 선다형의 학습 효과에 대한 선행 연구를 검토한 후 학습을 촉진하는 선다형의 긍정적인 학습 효과가 부정적인 학습 효과를 압도한다고 결론지었다.

선다형의 긍정적인 효과와 부정적 효과를 모두 검토한 연구는 Roediger와 Marsh(2005)에 의해 수행되었다. 이들은 실험에 참여한 대학생들에게 TOEFL과 GRE 시험의 지문을 읽게 했다. 이때 학생들은 지문을 읽은 집단과 읽지 않은 집단으로 구분되었다. 또한, 최초로 시행된 선다형 검사는 답지의 수가 2, 4, 6개로 차이가 있었다. 마지막으로, 시험을 보지 않은 집단이 설정되었다. 이들은 최초 선다형 검사에서 학생들에게 답지를 통한 추측을 하지 말 것을 강력하게 경고하였다. 최초 선다형 검사 결과 지문을 읽은 학생들의 수행은 지문을 읽지 않은 학생들보다 높았다. 또한, 2개의 답지로 시험을 본 학생들의 수행이 가장 높았다. 그러나 답지의

수가 각각 4개, 6개인 시험에서는 수행상의 차이가 발견되지 않았다. 이후 단서회상 검사가 사후검사로 시행되었다. 사후검사 결과, 먼저 지문을 읽었는지 여부와 관계없이 시험을 보지 않은 학생들보다 선다형 시험을 본 학생들의 점수가 높았다. 이는 선다형에서도 시험 효과가 발견됨을 보여준다. 흥미로운 결과는 학생들의 오류에 대한 측정에서 발견되었다. 지문을 읽었는지 여부와 관계없이 최초 선다형 시험의 답지 수에 따라 단서회상 검사에서 오류가 증가하였다. 이는 학생들이 추측 전략을 사용함으로써 오답지를 학습함을 보여준다.

이 연구 결과는 선다형의 특징을 활용하려는 본 연구에 중요한 시사점을 제공한다. 왜냐하면 추측 전략을 사용하지 말 것을 강력하게 경고한 상황에서도 학생들은 추측 전략을 사용하였기 때문이다(Roediger & Marsh, 2005). 따라서 Große와 Renkl(2007)의 연구에서 제한적인 효과가 발견된 올바른 예제와 틀린 예제의 제시를 선다형 답지의 형식으로 제공한다면, 학생들은 풀이된 예제로 구성된 선다형 답지에서 추측 전략을 통해 올바른 예제와 틀린 예제를 적극적으로 탐색할 것으로 기대할 수 있다. 더욱이 선다형은 학생들에게 매우 익숙할 뿐만 아니라 선호되는 방식이므로 다른 풀이된 예제 제시 방법보다 작업기억에 과부하를 초래할 가능성이 적을 것으로 예상하였다. 그러나 추측가능성의 문제가 남아있다.

풀이 과정을 답지로 이용한 시험 방식

선다형의 답지를 풀이 과정으로 제시하여 추측 전략을 통해 학습자가 풀이 과정을 적극적으로 탐색하게 할 때 교수자는 학생들의 능력을 과잉 추정할 수 있다. 그러나 적절한 교수학습은 학습자의 능력에 대한 정확한 판단을 통해 가능하기 때문에 과잉 추정을 방지할 필요가 있다. 이를 위해 박주용과 김용국(2010)에 의해 제안된 초등학생용 컴퓨터화 개방형 수학 시험 방식이 활용되었다. 이 시험 방식은 수학에서 문제 풀이 과정을 지필식 대신 컴퓨터로 풀 수 있도록 개발되었으며, 학생들은 단답형을 먼저 풀고 난 후 선다형에 응답해야 하기 때문에 단답형 응답과 선다형 응답을 동시에 수집할 수 있다 (박주용 & 김용국, 2010). 따라서 교수자가 학생들의 단답형 응답과 선다형 응답을 대조해본다면 추측을 통한 정답 여부를 확인할 수 있다.

또한, 평가 상황에서 문제 사이에 풀이 과정으로 구성된 답지를 제시하려면 학생들이 문제를 풀 때 풀이 과정을 볼 가능성을 차단해야 하는데, 지필식에서는 한계가 있다. 그러나 초등학생용 컴퓨터화 개방형 수학 시험 방식에서는 학생들이 답지를 요청할 경우에만 답지가 제시되기 때문에 평가 상황에서도 문제 사이에 풀이 과정을 제시할 수 있다. 본 연구에서는 이러한 특징을 활용하여 초등학생용 컴퓨터화 개방형 수학 시험 방식의 선다형 답지를 풀이 과정으로 제시하는 시험 방식을 제안하였다.

이 방식에서 학생들은 컴퓨터 화면을 통해 발문(stem)만 제시받는다. 학생들이 컴퓨터 상에서 문제를 풀 수 있도록 고안된 수식입력기를 이용해 단답형 응답을 완료하면, 풀이 과정이 선다형의 답지로 제시된다. 선다형 답지는 올바른 풀이 과정과 틀린 풀이 과정으로 구성되어 있고, 학습자는 하나의 올바른 풀이 과정을 선택해야 한다. 이때 학습자가 정답만을 비교하여 답지를 선택할 수 있기 때문에 풀이 과정에서 최종적인 정답은 제외하였다.

연구방법

표본 및 설계

서울 중랑구에 소재한 J 초등학교의 6학년 6개 학급이 실험에 참여하였다. 참여자는 총 190명이었다. 참여자 중 남학생은 98명, 여학생은 92명으로 남녀의 비율은 거의 동일하였다. 실험은 임의할당 사후검사 통제집단 설계로 수행되었다. 실험이 시작되기 전, 학급과 관계없이 모든 참여자들에게 임의로 번호를 부여한 다음, 임의화(randomization) 프로그램을 통해 각 집단의 구성원을 선정하여 실험집단(n=95)과 비교집단(n=95)으로 임의할당(random assignment)하였다.

독립변수는 선다형 답지의 구성이다. 실험집단은 단답형으로 문제를 풀고서 올바른 풀이 과정과 틀린 풀이 과정으로 구성된 답지에서 올바른 풀이 과정을 선택했다. 비교집단은 단답형으로 문제를 풀고서 최종적인 답으로 구성된 답지에서 올

바른 답을 선택했다. 사후검사는 처치검사와 동형으로 구성되었으며, 단답형으로 출제되었다. 이 설계를 도식화하면 <표 1>과 같다.

<표 1> 실험 설계

실험집단	R	X_1	O
비교집단	R	X_2	O

R : 임의할당

X_1 : 틀린 풀이 과정 사이에서 올바른 풀이 과정 선택, X_2 : 정답 선택

O : 단답형 사후검사

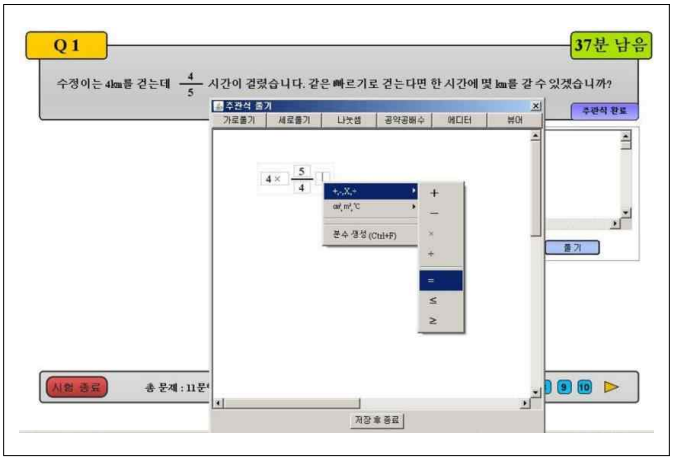
자료와 도구

처치검사와 사후검사 문항은 초등학교 4, 5학년 수학과목의 문제 푸는 방법 찾기 단원에서 4지 선다형으로 출제되었다. 수학 교과과정에서 문제 푸는 방법 찾기 단원은 상황을 단순화하거나 논리적 추론을 통해 문제 푸는 방법을 이해하고 문제를 해결하는 내용으로 구성되어 있다(교육과학기술부, 2008). 10개의 문항 중 거꾸로 생각하여 해결하는 문제가 4문항, 예상하고 확인하여 해결하는 문제가 2문항, 둘레에 심은 나무의 수를 구하는 문제가 2문항, 그리고 규칙을 찾아 수로 나타내는 문제가 2문항 출제되었다. 문항은 재직 경력 10년 이상의 현직 교사에 의해 내용 및 형식의 적설성이 검토되었다.

처치검사의 문항 정답률을 조정하고 오답지 구성을 위한 오류 형태를 확인하기 위해 실험에 참여하지 않는 1개 학급을 대상으로 사전검사가 시행되었다. 사전검사는 단답형 지필식으로 시행되었는데, 특별한 오류는 발견되지 않았으며, 문항의 정답률은 .6에서 .1이었다. 문항 정답률이 약 .3에서 .7일 때 피험자의 차에 관한 최대 정보를 제공하는 것을 고려하여(이종성, 1985), 문항 정답률이 .3 미만인 문항 3개를 보다 기초적인 수준의 문항으로 교체하였다. 이 중 풀이 과정으로 구성된 답지의 오답지는 학생들의 흥미를 끌 수 있도록 하기 위해 사전검사에서 회수된 시험지를 분석하여 학생들이 자주 범하는 오류로 구성하였다. 또한, 올바른 풀이

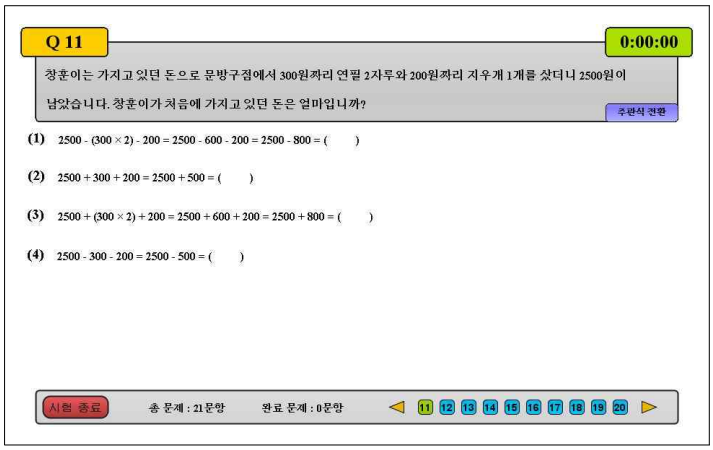
과정은 교과서와 시중 문제집에서 모범적인 풀이과정으로 제시된 예제를 사용하였다. 사후검사 문항은 처치검사 문항과 같은 구조의 발문에서 숫자만 변형하여 동형으로 구성하였다. 그러나 규칙을 찾아 수로 나타내는 문항의 경우(2개 문항) 계산하는 과정에서 수의 단위가 증가하였다. 사후검사의 신뢰도는 적절한 수준이었다(Cronbach's alpha: .72).

초등학생용 컴퓨터화 개방형 수학 시험은 수학 문제를 종이에 펜으로 푸는 것과 같이, 컴퓨터 화면을 보면서 키보드와 마우스로 풀도록 고안되었다. 프로그램은 자바(Java SE 1.4.2)로 개발되었으며, 시험 화면은 1024*768에 최적화되어 있다. 학생들이 컴퓨터실에 와서 로그인만 하면 시험을 시작할 수 있도록 필요한 프로그램(Java application) 설치 및 웹 접속은 미리 준비해두었다. 단답형에 응답할 수 있도록 고안된 수식입력기는 [그림 1]과 같이 제공되었다.

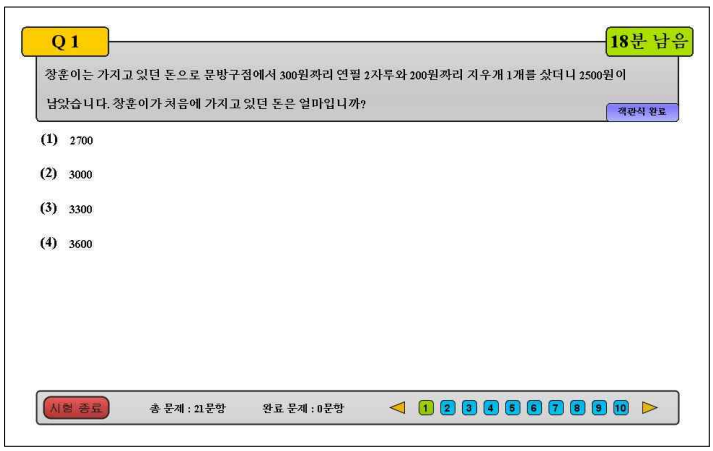


[그림 1] 단답형 응답

단답형 응답을 완료하고 나서 오른쪽 상단에 주관식 완료 버튼을 클릭하면, [그림 2, 3]과 같이 선다형 답지가 집단에 따라 각각 제시되었다. 선다형 답지가 제시되면 단답형으로 돌아갈 수 없도록 설계되었다.



[그림 2] 실험집단의 선다형 답지



[그림 3] 비교집단의 선다형 답지

절차

실험은 컴퓨터 수업시간을 활용하여 동일한 컴퓨터실에서 시행되었다. 학생들은 컴퓨터실에서 연구자, 연구 보조원, 그리고 담임교사의 지도아래 실험에 참여하였

다. 먼저 학생들은 새로운 프로그램으로 수학 시험을 보기 위해 연습이 필요했다. 이를 위해 3주간 30분씩 연습이 시행되었다. 연습은 학생들의 수업 진도에 맞추어 출제한 문항으로 실험집단과 비교집단의 구분 없이 문제를 풀고 나서 정답을 선택하는 시험 방식으로 시행되었다. 모든 연습시험 및 처치검사는 오전에 시행되었다. 처치검사가 시행되기 이전에 실험에 참여하지 않는 학급을 대상으로 문항 오류 및 정답률 조정을 위한 사전검사가 시행되었다. 사전검사는 교실에서 담임교사에 의해 시행되었다.

3주 동안의 연습시험이 끝나고 나서 참여자들은 학급과 관계없이 실험집단과 비교집단으로 임의할당 되었다. 처치검사는 학생들에게 사전 공지 없이 연습시험이 끝나고 1주일 후에 시행되었다. 컴퓨터실에서 학생들은 실험집단과 비교집단에 관계없이 정해진 자리에 앉았다. 컴퓨터에는 실험을 위한 공지사항을 알리는 웹페이지에 미리 접속해두었는데, 이 웹페이지에는 1～10번까지 풀어야 하는 학생들과 11～20번까지 풀어야 하는 학생들의 명단이 구분되어 있었다. 이때 1～10번 문항은 문제를 푼 후 정답을 선택하는 비교집단이었고, 11～20번까지 풀어야 하는 학생들은 문제를 풀고서 틀린 풀이 과정 사이에서 올바른 풀이 과정을 선택하는 실험집단이었다. 사전에 자신이 풀어야 하는 문제 이외의 문제를 풀면 안 된다는 것이 공지되었다. 학생들의 시험 기록을 확인한 결과, 다른 문제를 풀거나 비교집단의 학생들이 실험집단의 답지를 요구한 사례는 없었다.

또한, 단답형 문제의 배점이 높기 때문에 반드시 단답형을 풀어야 함을 강조하였다. 학생들은 미리 부여된 ID와 비밀번호로 로그인을 하여 자신이 풀어야 하는 문제를 풀었다. 총 시험시간은 40분이었으며, 정해진 시간이 끝나면 시험은 자동으로 종료되었다. 시험이 종료되면 점수가 바로 제시되었고, 문제를 다 푼 학생들은 시험 화면을 종료하고 조용히 타자 연습을 하게 하였다.

처치검사가 끝나고 사전 공지 없이 1주일 후에 사후검사가 실시되었다. 사후검사는 처치검사와 동형으로 출제된 문항으로 구성되었고, 단답형으로 출제되었다. 또한 각 학급에서 담임교사에 의해 실시되었으며, 시험시간은 40분으로 제한되었다.

연구 결과

본 실험의 참여자는 190명이었으나 통제되지 않은 손실이 발생하였다. 전학생 8명(남 2명, 여 6명), 처치검사 문제를 거의 풀지 않거나 결석을 한 학생 15명(남 10명, 여 5명), 그리고 사후검사 시험지를 공백으로 제출한 학생 3명(남 1명, 여 2명)이 최종 분석에서 제외되었다. 이로 인해 총 26(남 13, 여 13)명이 제외되어 164명의 자료가 분석되었다. 집단별로는 실험집단에서 15명, 비교집단에서 11명이 제외되었다. 모든 결과는 유의수준 .05에서 검증되었다. 충분한 수의 피험자를 대상으로 임의할당을 하였으나 실험집단에서 학업능력이 낮은 학생들이 더 많이 제외되었을 가능성을 확인하기 위해 처치검사의 선다형 점수를 비교하였다. 그 결과는 <표 2>와 같다.

<표 2> 처치검사 결과

구분	N	평균	표준편차	t값	P값
실험집단	80	51.12	23.05	.637	.525
비교집단	84	53.33	21.30		

처치검사의 선다형 점수에서는 집단 간 유의미한 차이가 발견되지 않았다 ($t(163) = .637$, $p>.05$). 오히려 비교집단의 점수 ($M = 53.33$, $SD = 21.3$)가 실험집단의 점수 ($M = 51.12$, $SD = 23.05$)보다 근소하게 높았다.

처치검사 점수의 비교에서 모든 피험자에게 동일한 형식으로 제시된 단답형 점수가 아닌 선다형 점수를 비교한 이유는, 본 연구의 시험 방식에서 단답형 응답을 위해 제공되는 수식입력기가 지필식 시험지의 공백처럼 학생들이 자유롭게 문제풀이를 시도하는 공간이기 때문이다. 따라서 이 공간에는 정답이나 암산을 통한 과정은 생략되어 있기 때문에 정답을 기준으로 채점을 하는 것이 불가능하다. 반면, 선다형 점수는 정확한 채점이 가능하지만 실험집단과 비교집단의 선다형 답지 구성이 다르기 때문에 점수의 비교가 불가능하다.

그러나 시험방식의 특성을 통해 최소한 실험집단이 비교집단보다 능력이 낮은

학생들이 더 많이 제외되지 않았음을 추론할 수 있다. 왜냐하면 선다형 답지가 풀이된 예제로 구성됨으로써 추측 전략을 사용할 수 있는 실험집단과 달리, 비교집단의 선다형 오답지는 정답과 유사한 임의의 숫자로 구성되어 있기 때문에 이를 제거하여 추측 가능성을 높이는 전략이 불가능하다. 일반적으로 선다형이 단답형보다 더 높은 점수를 받는데, 이는 확실하게 틀린 오답지를 제거함으로써 추측 가능성을 높일 수 있기 때문이다 (Betts, Elder, Hartley, & Trueman, 2009; Bush, 2001; Frary, 1988; Simkin & Kuechler, 2005). 따라서 실험집단이 비교집단과 달리 추측전략을 사용할 수 있음에도 처치검사에서 집단 간 유의미한 차이가 발견되지 않은 결과는 적어도 실험집단이 비교집단에 비해 능력이 낮은 학생들이 더 많이 제외되지는 않았음을 의미한다.

다음으로 사후검사 점수에서 집단 간 유의미한 차이가 있는지를 확인하였다. 그 결과는 <표 3>과 같다.

<표 3> 사후검사 결과

구분	N	평균	표준편차	t값	P값
실험집단	80	60.56	23.19	-1.989	.048*
비교집단	84	53.45	22.57		

*p<.05

문제를 푼 후 틀린 예제 사이에서 하나의 올바른 예제를 선택한 실험집단의 점수 (M = 60.56, SD = 23.19)가 문제를 푼 후 정답만 선택한 비교집단의 점수(M = 53.45, SD = 22.57)보다 높은 것으로 나타났으며, 이 차이는 통계적으로 유의미하였다($t(163) = -1.989$, p<.05).

선다형 답지 형태의 차이와 학습자의 사전 지식 간에 상호작용을 확인하기 위해 하위집단과 상위집단의 사후검사 점수 차이를 확인하였다. 이를 위해 Großе와 Renkl(2007)의 연구와 마찬가지로, 처치검사에서 -.85보다 낮은 z-점수 영역을 사전 지식이 낮은 학습자 집단으로, .85보다 높은 z-점수 영역을 사전 지식이 높은 학습자 집단으로 설정하였다. 하위집단과 상위집단의 사후검사 점수 차이는 <표 4>와

<표 4> 하위집단과 상위집단의 사후검사 결과

	구분	N	평균	표준편차	t값	P값
하위집단	실험집단	21	44.29	21.81	-.913	.367
	비교집단	18	37.78	22.64		
상위집단	실험집단	15	82.67	12.23	-2.862	.008*
	비교집단	16	66.88	17.78		

*p<.05

같다.

하위집단에서는 실험집단의 점수가 더 높기는 하였으나 통계적으로 유의미한 차이는 발견되지 않았다. 추가적으로 하위집단에서 처치검사 점수와 사후검사 점수의 변화를 확인하기위해 대응표본 t검정을 수행하였다. 그 결과 비교집단의 하위집단은 처치검사 점수(M = 28.33, SD = 3.83)에 비해 사후검사 점수(M = 37.78, SD = 22.64)가 상승하였으나 통계적으로 유의미한 차이는 아니었다(t(17) = -1.696, p>.05). 반면, 실험집단의 하위집단은 처치검사 점수(M = 21.43, SD = 7.93)에 비해 사후검사 점수(M = 44.29, SD = 21.81)가 상승하였으며 이는 통계적으로 유의미하였다(t(20) = -4.632, p<.05). 상위집단에서는 실험집단의 점수(M = 82.67, SD = 12.23)가 비교집단의 점수(M = 66.88, SD = 17.78)보다 유의미하게 높았다(t(30) = -2.862, p<.05). 이는 올바른 예제와 틀린 예제를 선다형 답지로 제시한 방법이 일반적인 선다형 답지보다 학습을 향상시켰음을 보여준다.

근전이 문항에서 틀린 예제의 제시를 통한 학습 효과가 발견되지 않은 Großе와 Renkl(2007)의 연구 결과와 달리 본 연구에서의 결과는 동형으로 제작된 사후 검사에서 긍정적인 학습 효과가 발견되었다. 그러나 하위 집단에서는 Großе와 Renkl (2007)의 연구와 마찬가지로 풀이 과정으로 구성된 답지가 처치검사 점수에 비해 사후검사 점수를 향상시키기는 하였으나 집단 간 사후검사에서 유의미한 차이는 없었다. 따라서 사전지식이 낮은 학생들에게 풀이 과정으로 구성된 답지가 풀이된 예제의 제시보다 효과적일지는 추가적으로 검증될 필요가 있다.

다음으로, 올바른 예제와 틀린 예제로 구성된 선다형 답지에서의 탐색활동은 추

가적인 시험시간을 필요로 했는지 확인하였다. 이를 위해 집단 간 처치검사 문항을 완료하는데 소요된 시간을 비교하였다. 그 결과는 <표 5>와 같다.

<표 5> 집단 간 처치검사에 소요된 시간

구분	N	평균	표준편차	t값	P값
실험집단	80	1365.41	483.94	-.243	.809
비교집단	84	1346.33	518.61		

실험집단(M = 1365초, SD = 483.94)은 비교집단(M = 1346초, SD = 518.61)보다 19초의 시간을 더 소요한 것으로 나타났다. 이 차이는 매우 적을뿐더러 통계적으로도 유의미하지 않았다($t(163) = -.243$, $p>.05$). 이는 본 연구에서 제안된 답지 구성이 추가적인 시간을 들이지 않고도 학습을 촉진할 수 있는 방식임을 보여준다.

마지막으로 본 연구에서 제안된 시험 방식이 측정도구로써도 의미 있는 도구인지를 확인하기 위해 처치검사와 사후검사 간에 상관관계를 확인하였다. 그 결과 실험집단과 비교집단 모두 컴퓨터로 실시된 처치검사와 지필식으로 실시된 사후검사 간에 정적 상관이 존재하였다. 각각의 시험 방식 중에서는 비교집단보다($r = .47$, $p<.001$) 실험집단이($r = .60$, $p<.001$) 지필식 사후검사와 더 강한 관계에 있음이 발견되었다. 물론 본 연구에서 제안된 시험 방식이 측정하는 구인이 무엇인지, 측정 도구로써 신뢰로운 검사인지 등의 문제는 추가로 규명되어야 할 부분지만, 지필식 시험 점수에 대한 예측력을 갖추었다는 점에서 학생의 능력을 측정할 수 있는 도구로써의 가능성도 보여주었다고 할 수 있겠다.

결론 및 제언

본 연구는 초등학생용 컴퓨터화 개방형 수학 시험을 활용하여 풀이된 예제 효과에 기반해 수학 평가를 개선하고, 그 효과를 검증하였다. 이때 풀이된 예제는 선다형 답지의 형식으로 제시하여, 틀린 풀이 과정 사이에서 올바른 풀이 과정을 찾

도록 하였다. 그 이유는 선다형이 학생들에게 익숙하고 선호되는 시험 방식일뿐더러 한정된 답지 내에서의 탐색활동이기 때문에 학습자들의 작업기억에 과부하를 초래하기 보다는 적절한 인지부하를 초래할 것으로 예상하였기 때문이다. 또한, 선다형 답지의 형식을 올바른 풀이 과정과 틀린 풀이 과정으로 제시할 경우 학생들은 추측전략을 통해 적극적으로 풀이된 예제를 탐색할 것으로 예상하였다. 이와 함께, 풀이된 예제 효과와 시험 효과의 결합을 통한 학습 이득의 극대화를 기대했다.

실험 결과는 이상의 예상을 지지한다. 먼저, 실험집단과 비교집단 간에 사후검사 점수에서 유의미한 차이가 발견되었다. 사후검사에서 문제를 풀고서 올바른 예제와 틀린 예제로 구성된 선다형 답지에서 올바른 예제를 선택한 실험집단의 점수가 문제를 풀고서 정답만 선택한 비교집단보다 더 높았다(60.56 대 53.45). 이는 틀린 예제 사이에서 올바른 예제를 탐색하는 과정이 학습을 촉진하였음을 시사한다.

그러나 Großе와 Renkl(2007)의 연구와 마찬가지로, 올바른 예제와 틀린 예제로 구성된 선다형 답지의 학습 효과는 하위집단에서는 발견되지 않았고, 상위집단에서 두드러지게 나타났다. Großе와 Renkl(2007)의 연구와 차이점이 있다면, 상위집단에서의 사후검사 점수 차이가 전체 집단 간 유의미한 차이를 이끌 정도로 컸다는 점이다. 또한, 하위 집단에서 풀이 과정으로 구성된 답지가 처치검사 점수에 비해 사후검사 점수를 향상시키기는 하였으나 집단 간 사후검사에서 유의미한 차이는 발견되지 않았다. 따라서 문제 사이에 풀이된 예제만 제시하는 조건이 추가된 후속 연구가 필요하다.

다음으로, 처치검사를 보는 데 소요된 시간에서 실험집단과 비교집단 간에 차이가 발견되지 않았다(1365초 대 1346초). 이는 풀이된 예제를 탐색하는 과정이 학생들에게 많은 시간을 요구하지 않음을 보여준다. 그러나 전체 시험 시간에서는 차이가 없었지만, 단답형과 선다형에서 응답한 시간이 집단별로 다를 가능성이 있다. 선다형 답지가 정답으로만 구성된 비교집단의 경우, 단답형 문제를 해결하지 못하면 선다형 답지에서 아무런 단서 없이 정답을 선택해야 한다. 반면, 선다형 답지가 풀이된 예제로 구성되어 있는 실험집단의 경우, 단답형에서 문제를 해결하지 못하더라도 선다형 답지에서 추측 전략을 통해 정답을 선택할 수 있다. 따라서 비교집단의 학생들은 실험집단의 학생들보다 단답형에서 오랜 시간 동안 문제를 풀었을

수 있다. 반면, 실험집단의 학생들은 단답형에서 문제가 잘 풀리지 않을 경우 빨리 선다형 답지를 요청했을 가능성이 있다.

이러한 설명은 인지부하이론과도 일치한다. 즉 단답형에서 문제가 잘 풀리지 않을 경우, 선다형 답지로 넘어가 풀이된 예제로 구성된 답지 내에서 탐색활동을 하는 것이 작업기억의 과부하를 줄일 것이다. 또한, 단답형에서 문제를 푼 학생들도 틀린 예제 사이에서 올바른 예제를 찾는 탐색활동은 풀이된 예제를 적극적으로 탐구하도록 한다. 그러나 선다형 답지가 정답으로 제시되는 비교집단은 추측 전략이 불가능하기 때문에 단답형에서 계속 문제풀이를 시도해야 한다. 이는 작업기억에 과부하를 초래할 수 있다. 따라서 이 평가 시스템은 학생들에게 문제풀이 전략을 최소화하도록 하는 동시에 풀이된 예제를 적극적으로 탐색하도록 할 수 있다. 그러나 이는 단답형 응답 시간과 선다형 응답 시간을 각각 측정함으로써 경험적으로 검토될 필요가 있다.

추가적으로 본 연구에서 제안된 새로운 시험 방식은 선다형 응답과 단답형 응답을 동시에 수집할 수 있는 구성적 선다형의 장점을 갖고 있다 (박주용 & 민경석, 2009). 이를 통한 평가 도구로써의 이점은 다음과 같다. 첫째, 학생들이 문제를 푸는 과정정보를 저장할 수 있다. 둘째, 컴퓨터에 저장된 과정정보는 지필식 시험지에 비해 그 보관이 용이하며, 언제든 편리하게 다시 볼 수 있고, 영구적으로 보관할 수 있다. 셋째, 교사는 선다형 응답과 과정정보를 대조함으로써 추측을 통한 정답 여부를 확인할 수 있다. 넷째, 즉각적인 피드백이 제공된다. 답지가 풀이된 예제로 구성되어 있기 때문에 문제의 정답만 제공하면, 학생들은 답지로 돌아가 쉽게 해결 방법을 확인할 수 있다. 다섯째, 자동화 채점이 가능하다.

마지막으로 본 연구의 한계를 지적하고, 이를 통해 후속 연구를 제안하면 다음과 같다. 먼저, 풀이 과정을 답지로 이용한 시험 방식에서 컴퓨터를 통해 수학 문제를 푸는 과정은 연습을 통해 학생들이 어렵지 않게 이용할 수 있지만, 여전히 지필식보다 익숙한 방식은 아니다. 또한, 본 연구에 국한된 문제는 아니지만, 새로운 문항 형식은 본래 측정하고자 하는 능력 이외에 외부적 기술을 평가할 수 있다 (Huff & Sireci, 2001). 다시 말해 어떤 학생들은 문제를 풀 수 있음에도 컴퓨터에서 문제를 푸는 과정이 어려워서 문제를 풀지 못할 수도 있다. 따라서 본 연구에서 제안된 선다형 답지 제시 방법을 지필식 환경에서 검증해볼 필요가 있다. 이와 함

께 평가 시스템 전반을 학생들이 더욱 편리하게 사용할 수 있도록 구현하는 것 또한 앞으로의 과제이다.

다음으로, 본 연구에서는 학생들이 자주 범하는 오답 유형을 분석하여 이를 답지의 틀린 풀이 과정으로 구성하였는데, 이는 사전 지식이 낮은 학습자들의 학습을 방해했을 수 있다. 이를 규명하기 위해 답지 구성의 곤란도에 따른 학습 효과를 검증해볼 필요가 있다. 또한, 본 연구에서 제안된 평가 시스템이 정확히 측정하는 구인이 무엇이며, 측정 도구로써 신뢰로운 검사인지 등이 더 구체적으로 규명할 필요가 있다. 마지막으로, 풀이된 예제를 이용한 선다형 답지의 구성은 초등학교 수학 범위를 벗어나 풀이 과정이 길고 복잡한 문제에서는 사용하기 어려울 수 있다. 이를 극복할 수 있는 대안 역시 후속 연구의 주제가 될 수 있을 것이다.

참고문헌

교육과학기술부 (2008). **초등학교 교육과정 해설(Ⅳ)**, 서울: 교육과학기술부.

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

박주용, 김용국 (2010). 초등학생용 컴퓨터화 개방형 수학 시험 방식의 사용가능성 검증. **인지과학, 21**(2), 283-307.

이종성 (역) (1985). **행동과학연구를 위한 측정이론의 기초.** *M. J. Allen. M. J., & W. M. Yen의 Introduction to Measurement Theory.* 서울: 중앙적성출판사.

Becker, W. E., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record, 75*, 348-357.

Betts, L. R., Elder, T. J., Hartley, J., & Trueman, M. (2009). Does correction for guessing reduce student' performance on multiple-choice examination? yes? no? sometime?. *Assessment & Evaluation in Higher Education, 34*(1), 1-15.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-73.

Bible, L., Simkin, M. G., & Kuechler, W. L. (2008). Using multiple-choice tests to

evaluate students' understanding of accounting. *Accounting Education: an international journal, 17*, 55-68.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education, 31*(4), 399-413.

Boud, D., & Falchikov, N. (2007). Assessment for the longer term. In. Boud, D., & Falchikov, N. (Eds.), *Rethinking assessment in higher education.* Routledge: London and New York.

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education, 25*(2), 157-163.

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*, 514-527.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23*(6), 760-771.

Clark, R. C., Nguyen, F., & Sweller, J. (2006). *Efficiency in learning: evidence-based guidelines to manage cognitive load.* San Francisco: Pfeiffer.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293-332.

Delgado, A. R., & Prieto, G. (2003). The effect of item feedback on multiple-choice test responses. *British Journal of Psychology, 94*(1), 73-85.

Earl, M. L. (2003). *Assessment as learning: Using classroom assessment to maximize student learning.* Thousand Oaks, CA: Corwin Press.

Entwistle, A., & Entwistle, N. (1992). Experiences of understanding in revising for degree examinations. *Learning and Instruction, 2*, 1-22.

Frary, R. B. (1988). NCME instructional module on formula scoring of multiple-choice tests(correction for guessing). *Educational Measurement: Issues and Practice, 7*, 33-38.

Glover, J. (1989). The "testing" phenomenon: Not gone, but nearly forgotten. *Journal of Education Psychology, 81*(3), 392-399.

Großе, C. S., & Renkl, A. (2006). Effects of multiple solution methods in mathematics learning. *Learning and Instruction, 16*, 122-138.

Großе, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes?. *Learning and Instruction, 17*, 612-634.

Holder, W. W., & Mills, C. N. (2001). Pencils down, computers up: The new CPA exam. *Journal of Accountancy, 96*(3), 57-60.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*, 16-25.

Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors, 40*, 1-17.

Kalyuga, S., Chandler, P., & Sweller, J. (2001). Learner experience and efficiency of instructional guidance. *Educational Psychology, 21*, 5-23.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23-31.

Kvale, S. (2007). Contradictions of assessment for learning in institutions of higher learning. In. Boud, D., & Falchikov, N. (Eds.), *Rethinking assessment in higher education*. Routledge: London and New York.

Lee, I. (2007). Feedback in Hong Kong secondary writing classrooms: Assessment for learning or assessment of learning?. *Assessing Writing, 12*, 180-198.

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14*(2), 194-199.

Moreno, R. (2006). When worked examples don't work: Is cognitive load theory at an impasses?. *Learning and Instruction, 16*, 170-181.

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education, 31*, 53-64.

Park, J. (2005). Learning in a new computerized testing system. *Journal of Educational Psychology, 97*(3), 436-443.

Renkl, A. (1999). Learning mathematics from worked-out examples: Analyzing and fostering self-explanations. *European Journal of Psychology of Education, 14*, 477-488.

Renkl, A., & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational*

Psychologist, 38(1), 15-22.

Reisslein, J., Atkinson, R. K., Seeling, P., & Reisslein, M. (2006). Encountering the expertise reversal effect with a computer-based environment on electrical circuit analysis. *Learning and Instruction, 16*(2), 92-103.

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experiment Psychology: Learning, Memory, and Cognition. 31*(5), 1155-1159.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Psychological Science, 1*(3), 181-210.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology, 20*, 1209-1224.

Salden, R., Aleven, V., Renkl A., & Schwonke, R. (2009). Worked examples and tutored problem solving: Redundant or synergistic forms of support?. *Topics in Cognitive Science, 1*, 203-213.

Santos-Trigo, M. (2007). Mathematical problem solving: An evolving research and practice domain. *ZDM - The International Journal on Mathematics Education, 39*, 523-536.

Schoenfeld, A. (1992). Learning to think mathematically. In D. Grouws (Ed.), *Handbook for research on mathematics teaching and learning.* New York: Macmillan.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection?. *Decision Sciences Journal of Innovative Education, 3*, 73-97.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59-89.

Sweller, J. (2006). The worked example effect and human cognition. *Learning and Instruction, 16*, 165-169.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 196-251.

Tillema, H. H. (2009). Assessment for learning to teach: Appraisal of practice teaching lessons by mentors, supervisors, and student teachers. *Journal of Teacher Education, 60*, 155-167.

Williams, R. L., & Clark, L. (2004). College students' ratings of student effort, student ability and teacher input as correlates of student performance on multiple-choice exams. *Educational Research, 46*, 229-239.

1 차원고접수 : 2010. 9. 27
2 차원고접수 : 2010. 12. 6
최종게재승인 : 2010. 12. 20

(*Abstract*)

The Learning Effect of Test that Worked Examples Used as Options

Jeongman Lim
Department of Education
Sejong University

Jooyong Park
Department of Psychology
Seoul National University

The present study proposes and examines a new type of multiple-choice math test. In this format, the options are the intermediate derivatives of the math problem solution process rather than the final answers. This idea originates from the studies on the effect of worked-out examples. In these studies, it has been shown that students learn better when they were presented with worked-out examples than when presented with final answers by themselves. In line with these findings, we introduced the intermediate derivatives of the solution process as the options of multiple-choice items so that the test-taker will have a chance to examine the solution process. The test was implemented as a computerized test in which students can solve problems in a short answer format, and then pick a multiple-choice option which appears when requested. The experimental group had multiple-choice options which were intermediate derivatives of the solution process, and the control group had the final answers as the options as in most multiple-choice tests. The participants were 6th graders in elementary school. The posttest results revealed that the average score of the experimental group was higher than that of the control group. The results suggest that tests that use intermediate derivatives of the problem solution process as options can be used as learning tools in the classrooms. Finally, directions for further studies were discussed.

Keywords : Computerized Testing System, Assessment for Learning, Worked Examples