심리학석사학위논문

다중 평가를 활용한 확률 분배에서의 상위인지 측정과 보정 방안에 대한 연구

2016년 2월

서울대학교 대학원

심리학과 인지심리 전공

현 익 주

초 록

본 연구는 상위 인지적 점검 능력을 정확하게 측정하고 피드백을 통해 그것을 조정할 수 있는 방법을 알아보기 위해 수행되었다. 이를 위하여 다중 평가(Multiple Evaluation) 방식을 사용하였다. 다중 평가란 객관식 선다형 방식을 개선한 문제풀이 형태로서, 주어진 답지들로부터 하나의 정답을 찾는 것이 아니라, 문항 내의 모든 답지에 대해 각각이 정답이 될 가능성을 확률로 나타내되 합이 100%가 되도록 기입하는 방법이다(Dirkzwager, 2001; Homles, 2002). 다중 평가는 비교적 적은 문항을 통해서도 응시자들의 확신 정도(confidence)을 평가할 수 있으며, 이를 기반으로 피드백을 주어 향후 수행을 조정하는 것이 가능하다.

본 연구에서 수행된 실험에서는 대학생들을 대상으로 총 네 차례에 걸친 상식 문제로 이루어진 시험을 치게 하였다. 통제 집단의 참가자들은 네 차례의 시험 사이에 어떠한 피드백도 받지 않고 문제를 풀었고, 실험집단은 매 시험 후에 확신 정도에 대한 피드백을 제공받았다. 그 결과, 첫 번째 시험부터 세 번째 시험까지 두 집단의 확신 정도에 대한 유의미한 차이는 없었다. 그러나 네 번째 시험에서 두 집단 간에 확신 정도에서 유의미한 차이가 나타났으며, 피드백을 받은 집단이 그렇지 않은 집단에 비해 확신 정도가 더 정확하였다.

피드백을 받지 않은 통제 집단의 네 차례의 시험에 대한 응답을 재분석하여 피드백 제공방식 활용의 가능성을 탐색하였다. 다중 평가의 확신 수준을 나타내는 값인 A는 현실적인 확신 수준에서의 이상적인 확률 평가인 p와 응시자의 주관적인 확률 평가 r과의 선형적 관계($p=Ar+B$)를 전제로 도출되었다. A 값은 시험을 볼 때마다 도출되기 때문에 과거의 A값을 현재의 응답에 대한 가중치로 수정하는 것이 가능하다. 이 경우 t기에 (t-1)기를 반영하여 수정된 r 값은 $At_{-1}+(1-At_{-1})k$가 된다. 분석 결과, 이전 차례의 확신 정도를 반영하여 수정한 점수로부터 도출된 확신 점수가 기존의 확신 점수에 비해 유의미하게 과신 편향이 줄어든 것

으로 나타났다.

정리하면, 다중평가 방식을 통한 피드백은 확신 정도를 현실적으로 조정하는데 효과가 있었는데, 이 효과는 즉각적으로 나타나기 보다는 여러 차례 피드백을 받은 후 나타났다. 또 다른 결과는 이전 수행에서 나타난 확신 정도를 다음 확률 평가에 반영하여 수정한다면, 과신 편향이 줄어든다는 점이다. 이상의 연구 결과는 다중 평가 방식이 확신 정도를 현실적으로 조정하기 위한 피드백 도구로서 활용할 수 있으며, 또한 확률적으로 응답한 내용을 수정을 위한 도구로서 활용될 수 있을 가능성을 보여주었다.

주요어 : 상위인지, 점검 능력, 다중 평가, 조정, 확률 평가, 확률 수정
학 번 : 2013-22825

목 차

서 론 ····· 1
선행 연구 및 이론적 배경 ····· 3

실 험 ····· 14

종합 논의 ····· 21

참고문헌 ····· 24
부 록 ····· 28

표 목 차

[표 1] ····· 16
[표 2] ····· 17
[표 3] ····· 19
[표 4] ····· 19

그 림 목 차

[그림 1] ····· 4
[그림 2] ····· 9
[그림 3] ····· 11
[그림 4] ····· 12
[그림 5] ····· 15
[그림 6] ····· 17
[그림 7] ····· 18

서 론

학교나 대학 입학시험 등 거의 대부분의 교육 장면에서, 학업에 대한 수행은 시험의 형태로 드러나게 된다. 우리는 시험을 통해 자신의 지식 수준, 학업 상태 등을 확인하고, 이를 기반으로 향후 공부에 대한 계획을 세운다. 만약 어떤 학생이 시험에서 좋지 않은 성적을 받고 있다면 우리는 자연스럽게 그에 대한 해결책이 '조금 더 노력해서 공부를 하는 것'이라고 생각한다. 시험을 보기 위해 가장 중요한 것이 필요한 지식일 테니 이는 틀린 말은 아니지만, 이에 더하여 어떠한 공부 전략을 활용할 것인지, 무엇을 더 공부하고 무엇을 덜 공부할 것인지, 혹은 시험을 볼 때 정답인지 아닌지 확실치 않은 문항을 어떻게 응답할지에 대한 선택 혹은 지식은 공부한 내용에 대한 지식 그 자체에 못지않게 수행에 큰 영향을 줄 것이다. 즉, 지식이나 습득한 정보 등에 못지않게 지식 이상의 것들, 이를테면 "자신이 무엇을 알고 있는지"를 아는 것 역시 중요한 문제이다. Flavell(1979)에 의해 크게 조명되었던 이러한 자신의 지식, 혹은 학업이나 인지 수준에 대한 이해력은 상위인지, 혹은 초인지라 불리며 많은 학자들에 의해 연구되어왔다. 특히 이러한 연구들은 상위인지가 무엇인지부터 시작해서 그것을 어떻게 측정할 것인지, 상위인지의 역량을 어떻게 강화시킬 것인지에 대해 다루고자 하였다(Dunlosky & Lipko, 2007).

상위 인지적인 역량을 강화시키기에 앞서 가장 중요한 문제는 어떻게 상위 인지적인 역량을 측정할 수 있을지에 대한 것이다. 특히, 학습한 내용에 대한 인출 시의 확신 정도를 측정하는 것은 시험에서의 수행과 직접적인 연관이 있으며, 학업 계획을 수립하고 조정하는 등의 상위 인지적 통제를 위한 정보로 활용될 수 있다는 점에서 매우 중요하다고 할 수

있다. 확신 측정 방법은 인출시의 응답을 통해 도출되기 때문에 대부분 시험의 형식과 관련이 깊다. 그렇기 때문에 기존의 연구들에서는 시험을 본 전/후에 이에 대한 자신감을 묻거나(pre/post-diction), 문항 별로 응답에 대한 확신을 묻는(confidence weighting)이 많았다(Schraw & Dennison, 1994; Nelson & Narens,1990; Tobias & Everson, 2002). 그러나 이러한 측정 방식은 상위 인지적 점검 능력을 제대로 측정하고 있지 못할 가능성이 있는데, 사후 추정 방식의 경우 문항 수와 배열에 따라 정확도에 영향이 있을 수 있을 가능성이 있으며, 확신 가중치를 활용할 경우 상위 인지적 점검이 아닌 문항에 대한 위험 회피 성향 등에 대한 정보가 될 가능성 등이 있기 때문이다.

이러한 문제에 대한 대안으로 본 연구가 활용하고자 하는 것은 Dirkzwager에 의해 제안된 다중평가(Multiple Evaluation)이다(Dirkzwager, 1996;2001, Holmes, 2002. 다중 평가 방식은 확률 평가(probability assesment)를 기반으로 하고 있기 때문에, 선다형 문항에서 응시자의 지식수준에 대해 비교적 많은 정보를 추출할 수 있으며, 이를 기반으로 응시자에게 확신 정도에 대한 피드백을 주는 것이 가능하다. 또한 다중 평가는 문제 풀이 이후 복잡한 처리 과정 없이 확신 정도를 측정할 수 있기 때문에 다양한 방식으로 상위 인지적인 점검 능력을 훈련시킬 수 있는 도구로서 활용이 가능하다. 그러나 선행연구들에서는 상위 인지적 점검 능력에 대한 측정과 조정을 위해 다중 평가를 활용하는 방식에 대해서 연구가 미흡하였다. 따라서 본 연구에서는 다중 평가를 활용하여 확신 수준에 대한 피드백을 주었을 때 응시자들이 이를 바탕으로 확신 수준을 현실적으로 조정 가능한지 보고자 하였다.

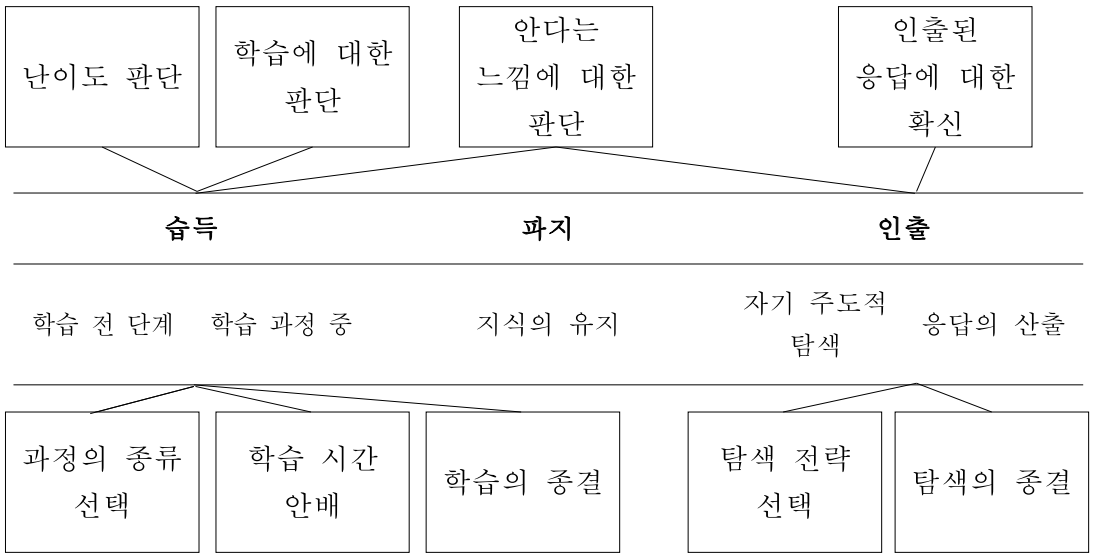
선행 연구 및 이론적 배경

상위 인지적 점검 능력

다중 평가(Multiple Evaluation)란 무엇이며, 어떻게 상위 인지적 점검 능력을 측정할 수 있는지를 밝히기 전에, 먼저 상위 인지적 점검 능력이 무엇인지에 대해 알아볼 필요가 있다. '지식에 대한 지식'이라고 불리는 상위 인지에 대한 연구는 Flavell(1979) 이후에 매우 활발히 이루어져 학습의 단계에 따라, 상위 인지가 개입하는 관점에 따라 서로 다른 체계가 있을 수 있다는 것이 주장되었다([그림 1] 참고). 즉, 지식의 습득에서 파지, 인출에 이르기까지의 학습 과정에서, 혹은 지식을 점검(Monitoring)하는 역할로서의 상위인지와 학습을 통제(Control)하는 역할 구분에서 상위인지의 기능과 구성은 나누어 질 수 있다(Nelson & Narens, 1990). Pintrich, Wolters, 와 Baxter(2000)는 이러한 체계를 다시 상위 인지적 지식과 상위 인지적 판단 및 점검, 자기 규제와 통제라는 세 가지 역할로 구분하였다. 이 외에도 서로 다른 방식으로 상위 인지를 구분한 시도들이 있었는데, 이렇게 상위인지를 구분하는 것이 중요한 이유는 각 분야들이 서로 어느 정도 독립적으로 기능하고, 때로는 같은 구분 하에 있는 하위 구성 요소들 역시 서로 독립적일 수 있다는 연구들이 있기 때문이다(Nelson & Narens, 1990; Pintrich, Wolters, & Baxter, 2000). 무엇보다 각 하위 요소들에 대한 사람들의 상위 인지적 역량은 서로 상이한 방법으로 측정되며, 이렇게 측정된 상위 인지 역량은 분야에 따라 의미하는 바가 다르기 때문에 상위 인지와 측정을 논할 때에는 정확히 어떤 요소를 살피고자 하는지 인식하는 것이 중요하다.

본 연구는 상위 인지의 다양한 영역 중에서도 상위 인지적 판단 및 점검에 대해 다루고 있으며, 그 중에서도 확률로 표현되는 확신 정도를 측정하고 조정할 수 있는 방법에 주목하고 있다. 즉, Nelson과 Narens(1990)의 구분에 따르면 인출된 응답에 대한 확신을 측정할 수 있는 방법과 그러한 응답의 정확도를 높이고 확신 수준을 현실적으로 조정

할 수 있는지의 여부를 밝히고자 한다. 상위 인지의 여러 영역 중에서 점검 능력, 그 중에서도 확신 정도에 주목하는 이유는 먼저 확신 정도는 응답에 직접적인 결과로 드러나기 때문이다.



[그림 1] 상위 인지의 연구에 대한 체계 (Nelson & Narens, 1990)

이는 상위 인지적 통제 능력이나 일반적인 상위 인지적 지식에 비해 양적으로 접근하기 수월하며, 그렇기 때문에 측정과 조정이 상대적으로 용이할 수 있도록 한다. 무엇보다 점검 능력의 향상은 학습 장면에서 학업 수행 향상과도 연관이 있음이 밝혀진 바 있다. 학업 수행이 저조한 사람들이 자신의 능력을 점검할 수 있는 역량 역시 부족하다는 점을 지적한 Kruger 와 Dunning, (1999)의 연구는 일명 더닝-크루거 효과로 널리 알려져 점검 능력과 학업 수행이 깊은 관계에 있음을 밝힌 바 있다. Isaacson와 Fujita(2006)는 더 나아가, 대학생들을 대상으로 한 연구에서 실제 10주간의 수업을 하는 동안 학생들에게 자신이 학습한 시간, 자신감의 정도, 그리고 시험 점수의 예측을 하게 하였고, 그 결과 수행이 좋았던 학생들은 자신의 시험 결과를 예측하는데 더 정확하였으며 시험 결과에 따라 자신의 자신감을 조정하는데도 더 유능하였음을 밝혔다. 이러한 연구는 상위 인지적 점검 능력이 단순히 어떠한 개별적인 지표가 아

닌, 학습의 본래 목표인 지식의 습득 및 학업 수행에도 기여할 수 있음을 시사한다. 또한 Thiede, Anderson, 과 Therriault(2003)의 연구는 상위 인지적 점검 능력이 상위 인지적 통제와 연관이 있어 학습 수행을 높일 수 있음을 주장하였다. Was, Beziat, 와 Isaacson(2013)의 연구 역시, 250명의 학생들을 대상으로 한 학기 동안의 수업이 이루어지는 동안 12번의 시험을 시행하였으며 매 시험마다 학생들이 스스로의 점수에 대한 예측을 하게끔 하였다. 한 학기의 과정이 끝나고 12번의 시험을 통한 점수와 그 예측을 분석한 결과, 점수에 대한 예측의 정확성과 실제 시험점수가 정적인 상관관계가 있음을 밝힐 수 있었다. 이를 통해 연구자들은 자신의 학업 수준에 대한 스스로의 측정이 정확해질수록 학업 역량 역시 강화될 수 있음을 보여주는 가능성을 발견하였다.

그러나 연구자들도 스스로 지적했듯 이 연구는 몇 가지 한계점들을 지니고 있었다. 먼저 응시자들이 학기 후반으로 갈수록 보여주었던 예측의 정확성은, 상위인지의 조정(calibration)이 정교해져서가 아니라 단순히 시험 점수가 전반적으로 후반으로 갈수록 좋아지기 때문이었을 가능성이 있다. 즉, 응시자들은 학기 초반의 예측과 학기 후반의 예측 모두 절대적으로 비슷한 수준의 점수에서 이루어졌으나 실제 시험점수는 점점 높아졌기 때문에 그 간극이 좁아져 상관관계가 높은 것으로 드러났을 수 있다. 또한 연구는 집단을 대상으로 분석하였기 때문에, 개개인의 수준에서 얼마나 차이가 드러났는지 살펴보는 데에는 무리가 있었다. 무엇보다 상위인지를 측정하는 데 있어서, 문항별로 세부적으로 살필 수 없다는 측정 방식에서의 한계점 또한 존재하였다.

상위 인지적 점검 능력 중, 인출에 대한 확신 정도에 주목하는 또 다른 이유는, 점검 능력에 대한 하위 요소인 난이도 평정(Ease-of-learning judgments)이나 학습에 대한 판단(Judgments of learning)과는 달리 확신 측정 방식은 상위 인지적 판단이 이루어지는 특정 영역에 대해 특수하게 접근해야만 하는 것이 아니라는 점이다. 확신에 대한 측정은 학업 과제나 상황 등에 대한 정보보다는 학습자, 혹은 확률 평가자가 응답하

는 내용을 기반으로 하기 때문에 영역과 별개로 비교적 보편적인 방법을 통해 도출해낼 수 있으며, 그렇기 적합한 측정 방식을 찾아 한 영역에 대해 점검 능력을 효과적으로 향상 시킬 수 있다면 이러한 효과가 다른 영역에서도 나타나길 기대해 볼 수 있다.

본 연구는 이러한 관점에서 상위 인지적인 점검 능력을 측정하고 조정할 수 있는 방법을 탐색하고자 하며, 그러한 가능성을 다중 평가방식(Multiple Evaluation)에서 찾고자 한다. 다중 평가 방식은 확률 평가(Probability Assessment)를 기반으로 학생들의 학업 수행을 평가하고자 하는 방식인데, 이러한 방법은 학생들의 수행을 전통적인 객관식 선다형 방식 등 보다 더 정확하게 측정할 수 있을 뿐만 아니라, 확신 정도를 평가할 수 있는 정교하고 다루기 쉽게끔 한다. 다중 평가 방식이 어떻게 확신 정도를 평가하는지를 이해하기에 앞서, 다중 평가방식이 무엇인지에 대해 객관식 선다형 방식과의 비교를 통해 이해해 볼 필요가 있다.

객관식 선다형과 확신 가중, 다중 평가방식의 비교

초등학교 때부터 수능시험에 이르기 까지, 혹은 때로는 대학교에서도 우리가 제일 자주 접하는 시험 방식은 객관식 선다형 문제이다. 2016 학년도 대학수학능력시험에서도 언어 영역은 45문항의 객관식이었으며, 외국어 영역 역시 45문항 객관식이었다. 객관식 선다형 문제의 풀이방식은 간단하다. 두 개 혹은 그 이상의 선택지 중에서 응시자가 가장 정답에 가깝다고 생각하는 것을 보통 한 개를 고르는 것이다. 대부분의 시험들이 4지 선다형 혹은 5지 선다형의 객관식 문제로 구성되어있고, 문제에 대한 학생들의 지식수준이 어떠하든 학생들은 선택지를 고르는 것만으로 정답과 오답이 결정되기 때문에, 가끔은 재미있는 현상들이 발견되기도 한다. '찍기'가 그것이다. 4지 선다형 문제를 받았을 때, 문제에 대한 지식이 전혀 없더라도 정답을 고를 수 있는 확률을 25%나 된다. 그래서 학생들 사이에선 연필 굴리기와 일렬로 찍기 등 나름의 추측 전략들을 농담처럼 공유되곤 한다.

학생 개인의 입장에서 문제를 알지도 못하면서 정답을 맞힐 수 있다면 나쁠 게 없지만, 교수자의 입장에서 이것은 문제가 된다. 하나의 문제가 주어질 경우, 교수자는 정답을 맞힌 학생들이 과연 알고 문제를 푼 것인지 모르고 문제를 푼 것인지 알 수 있는 방법이 없다. 혹은 더 나아가서, 틀린 학생들 중에서도 전혀 모른 채로 문제를 틀린 학생과, 조금은 정답과 오답 사이에서 고민을 하다가 오답을 고른 학생을 구분해줄 방법이 없다. 시험이 끝날 때 마다, '긴가 민가 하다가 오답을 골랐는데…….' 라는 학생들이 탄식이 나오는 것도 이 때문이다. 객관식 선다형에서 오답과 정답의 경계는 명확하며, 시험은 학생들이 무엇을 알고 있는지 그 자체보다는, 정답인 한 개를 고르는 행위를 통해 수행을 측정한다.

다시 말해, 객관식 선다형의 가장 큰 문제는 학생들의 지식을 시험으로 전환하는 과정에 있어 추출하고자 하는 정보량을 매우 축소시킨다는 점이다. 이것이 객관식과 비교되곤 하는 주관식, 혹은 서술형 시험문제와의 가장 큰 차이이다. 장문 서술형 시험의 경우 객관식에 비해 지니는 정보량은 매우 크다. 같은 문제가 주어진다고 하더라도 객관식은 몇 번을 선택하였는지에 대한 정보만 가지고 있는 반면, 장문 서술형 시험은 그 대상에 대해서 얼마나 이해하고 있는지, 어느 부분에서 심화된 지식을 보이고 어떤 부분은 그렇지 않은 지에서 더 나아가 필요하다면 글쓰기 실력과 응시자가 가지고 있는 통찰, 창의력 등 발산적인 사고를 평가하는 것 역시 가능하다. 장문 서술형 시험 문제를 푸는데 있어서 '찍는' 행위란 불가능하며, '아는데 까지'라도 쓰는 것이 응시자의 최선의 전략이다. 또한 이는 응시자로서도, 교수자로서도 자신의 지식수준만큼 평가하고 평가받을 수 있다는 점에서 더 공정하고 엄밀한 시험 방식이 될 수 있다.

그럼에도 불구하고, 고등교육과정이 아니고는 여전히 객관식 시험방식이 선호되는 이유는, 장문 서술형 시험 방식의 장점을 모르고 있기 때문이 아니다. 2012년에 초등학교 교사들을 대상으로 이루어진 설문 연구에서 역시 무려 91.8%의 교사들이 수학 교과에서 서술형 평가가 조금 필

요하거나 혹은 꼭 필요하다고 응답을 한 반면, 학급당 학생 수가 과다한 점, 문항 개발이 어려운 점, 객관성 확보가 어려운 점 등을 서술형 평가 실시의 문제점으로 지적하였다(김민경, 조미경, & 주유리, 2012). 서술형 시험 방식을 시행하는 데에는 객관식 시험 방식에 비해 큰 비용이 소모되며, 교수자에게 주어지는 부담 역시 크다. 학급 당 학생 수는 점점 적어지고 있지만, 초등학교의 경우에도 여전히 최대 수용 인원은 30~35명이며, 이렇게 많은 인원을 대상으로 모든 과목을 서술형 등으로 실시하는 것은 거의 불가능하다. 또한 시험이 교내에서 이루어지는 것이 아니라 TEPS나 TOEIC 등의 영어시험, 혹은 수능 시험과 같은 대규모로 이루어지는 시험일 경우에는 더더욱 그러하다. 채점 방식 역시 문제가 된다. 서술형 평가 방식은 응답 내용에 대한 정보가 많지만 이것을 어떻게 채점하고, 무엇에 중점을 두느냐에 따라 학생들이 받게 되는 점수와 등차 역시 매우 달라진다. 서술형에서 채점의 어려움은 이미 많이 지적되었지만, 채점 방식뿐만 아니라 최근에는 더 나아가 서술형을 위한 문제를 만드는 것 자체가 매우 어렵다는 점 역시 제기되고 있다(백순근, 1998).

이러한 점으로 인해 객관식의 형태는 그대로 두되, 이를 보완할 수 있는 시험 방식들이 제안되었는데, 대표적인 것이 문제에 대해 확신하는 정도에 따라 가중치를 주는 방식(Confidence Weighting)이다([그림 2] 참고). 확신 가중치를 사용하는 가장 보편적인 방법은 객관식, 혹은 주관식의 문항에 대한 응답 후에, 자신이 응답한 내용에 대하여 얼마나 확신하는지를 물은 후 이것을 통해 점수에 가중치를 부여하여 계산하는 방법이다. 확신 가중치가 초기에 주목을 받았던 이유는 상위인지를 측정 도구로서보다는 객관식 문항이 가지고 있는 적은 정보를 보완할 수 있다는 점 때문이었다. 즉, 적절한 가중 체계를 활용한다면 응답자들은 우연히 문제를 맞힐(“찍어서” 맞히는 등)수 없게 된다. 만약 잘못된 응답에 대한 가중 처벌 점수가 있다면, 선택지 중 하나를 찍더라도 그것에 대해 100%의 점수를 줄 수 없기 때문이다. 응답한 내용이 확실하지 않고, 모

※객관식 선다형 지시사항: 5가지의 객관식 선택지 중 정답이라고 생각되는 1개를 고르시오. 다음 중 영어로 “개”를 의미하는 것은? 1 Cat 2. Deer 3. bear 4. Dog 5. Raccoon	※확신 가중 방식 지시사항: 5가지의 객관식 선택지 중 정답이라고 생각되는 1개를 고른 후 정답에 얼마나 확신하는지를 확률로 표기하시오. 다음 중 영어로 “개”를 의미하는 것은? 1 Cat 2. Deer 3. Bear 4. Dog 5. Raccoon 확신 정도 (%)

[그림 2] 객관식 선다형과 확신 가중 문항의 비교

험을 하려는 경우가 아니라면 응시자는 50% 이하의 확신 정도를 표기하는 것이 시험점수를 극대화 하는 방법이 되며, 교수자는 이러한 응답을 통해 실제 지식에 가까운 응시자의 지식수준을 평가할 수 있게 된다. 이러한 장점으로 인해 최근에도 컴퓨터를 이용하여 확신 가중치를 시험 방식으로 활용하고자 하는 연구들이 이루어지고 있다(Yen, Ho, Chen, Chou, & Chen, 2010).

그러나 확신 가중치를 활용 하는 것 역시 단점이 존재한다. 먼저 확신 가중치를 시험 방식으로 사용할 경우 시험에 대한 신뢰도는 높을 수 있지만 타당도는 그렇지 않다는 비판이 제기되어 왔으며, 오히려 객관식 선다형보다 유용하지 않을 수 있음을 시사한 연구들도 적지 않다(Ebel, 1965; Hopkins, Hakstian, & Hopkins, 1972; Holmes, 2002, Wang & Stanley, 1970). 또한 시험 방식이 지니는 문항에 대한 응답의 낮은 정보량 역시 여전히 문제가 될 수 있다. 즉, 확신 가중치를 활용하면 교수자

는 응시자가 고른 선택지에 대해 얼마나 확신하는지 알 수 있지만, 나머지 선택지들에 대해서는 어떻게 판단하고 있는지 확실히 알 수 있는 방법이 없다. 응시자의 입장에서 확신 정도를 확률로 표기하는 것에 대한 이해가 어려움 점도 부수적인 문제가 될 수 있다. 이를테면 문제가 묻는 문항에 대한 지식이 전혀 없는 응시자가 4지 선다형의 문제에서 하나를 골랐을 경우, 그는 확신 정도를 0으로 표기해야 할까, 아니면 25%라고 해야 할까. 응시자는 문제에 대해서 전혀 아는 것이 없으니, 0을 기입해야겠다고 생각할 수도 있고, 4지 선다형이니 25%를 기입해야한다고 생각할 수도 있다. 혹은, 그는 이 선택지가 정답인지 아닌지 전혀 알 수 없으니 50%을 기입해야 된다고 생각할 수도 있다. 특별한 지시사항이 없다면, 교수자는 응시자가 정확히 어떤 의미로 확률을 기입했는지 알 수 없다.

이에 대한 대안으로 제시된 것이 다중 평가(Multiple Evaluation)방식이다(Dirkzwager, 1996). 다중 평가란 객관식 선다형 방식을 개선한 문제풀이 형태로서, 쉽게 말해 객관식 선다형 문제에서 단순히 정답 하나를 찾는 것이 아니라, 각각의 문항에 대해 자신이 정답이라고 생각되는 만큼 확률을 기입하는 방법을 의미한다. [그림 3]은 다중 평가 방식을 통한 문제 풀이의 예시를 보여준다. 다중 평가 방식은 Shuford Jr, Albert, 와 Massengill(1966)이 제안하였던 확률 평가(probability assessment)를 기반으로 하고 있으며, 이 방식은 단순히 문제 하나를 고르는 객관식 선다형 문제에 비해 각 선택 사항에 확률을 기입한다는 점에서 응시자가 가지고 있는 지식수준에 대한 정보를 더 많이 담을 수 있다는 우수성이 있다.

확률 평가를 활용하는 시험방식은 여러 가지 장점에도 불구하고 지필시험에서 이용하기 용이치 않은 점 등 때문에 실제 시험에서 이용하기에는 어려움이 있었다. 그러던 것이 컴퓨터가 보급화 된 이후 필요한 소프트웨어의 개발이 가능해지며, Dirkzwager (1996)에 의해 개선되어 채점

※다중 평가 방식
지시사항: 5가지의 선택지 각각에 대해 정답이라고 생각하는 만큼의 확률을 기입하되, 그 합이 총 100%가 되도록 하시오.

다음 중 영어로 “개”를 의미하는 것은?
1 Cat %
2. Deer %
3. Bear %
4. Dog %
5. Raccoon %

[그림 3] 다중 평가 문항의 예시

방식 등이 정교화 됨과 함께, 이를 구현할 수 있는 TestBet이라는 프로그램이 개발되었다. 또한 Holmes(2002)는 이에 대해 객관식 선다형과 Confidence weighting, Multiple Evaluation 등 다양한 평가 기법들을 다각적으로 정리하며, Dirkzwager에 의해 개선된 Multiple Evaluation의 방식이 다른 객관식 평가기법들보다 장점들을 많이 가지고 있음을 밝혔다.

상위 인지적 점검 능력 측정 도구로서의 다중평가

특히 다중평가는 응시자가 기입한 확률을 바탕으로 자신의 지식수준에 비해 어느 정도의 자신감을 가지고 응답을 하고 있는지를 판단할 수 있으며, 이에 따라 피드백을 주어 응답자의 확신 수준을 높일 수 있는 가능성이 있다. [그림 4]는 다중평가에서 응시자의 확신 수준을 측정할 수 있는 공식이 유도이다(Holmes, 2002).

위 공식에서 A란 확신 수준(Realism)을 의미하며, r(i)는 정답인 선택지에 기입한 확률, r(j)는 정답이 아닌 선택지에 기입한 확률, k는 선택지

- 완벽히 현실적인 경우의 확률평가: $p(i) = r(i) \quad \frac{R_t}{R} = r(i)$
- $p = A \times r + B$
- $B = \frac{1-A}{k}$
- $f(A,r) = \sum\sum R(r) \times (\frac{R_t(r)}{R(r)} - \left(A \times r + \frac{1-A}{k}\right))^2$
- $\frac{\delta f(A)}{\delta A} = 0$
- $A = \frac{\sum\sum R_t \times r - R_t(r)/k - (R(r)\times r)/k + R(r)/k^2}{\sum\sum R(r)\times r^2 - (2R(r)\times r)/k + R(r)/k^2}$

- $\sum R(r) = m \times k \ , \ \sum r = 1$
- $\sum R(r) \times r = m$
- $\sum R_t(r) = m$
- $\sum R_t(r) \times r = \sum r_c$
- $\sum R_t(r) \times r^2 = \sum\sum r^2$
- $A = \frac{k \times \sum_{i=1}^{m} r_c(i) - m}{k \times \sum_{i=1}^{m} \sum_{j=1}^{k} r(i,j)^2 - m}$

[그림 4] 다중 평가에서의 응시자 확신 수준(A) 유도(Homles, 2002)

의 수, m은 문제의 수이다. 완벽히 현실적으로 다중 평가 문항에 확률을 기입할 경우, 주관적인 확률 기입인 r(i)는 p(i)와 일치하게 된다. 응시자의 확률 평가와 완벽히 현실적인 확률 평가의 관계를 선형적이라고 보았을 때, 이는 p = A×r+B로 표현할 수 있는데, 다중 평가 문항의 선지 수는 정해져있으므로 B=(1-A)/k가 된다. 위의 공식은 이러한 관계를 통해 A의 값을 최소제곱법으로 유도한 것으로, 10 문항 정도 이상이라는 비교적 적은 표본이 있을 경우에도 경향성을 계산해 낼 수 있다(Holmes, 2002). 또한 응답 그 자체 외의 부가적인 질문들을 필요로 하는 기존의 사후 평가(Postdict) 방식과는 달리, 응답 내용만을 가지고 계산한다는 점에서 응답자에게 불필요한 개입 없이 계산해 낼 수 있다는 차별성이 있다. 이 점은 기존의 확신 정도를 평가하는 방법들에 비해 큰 의의를 지닌다. 먼저 객관식 선다형과 이에 대한 사후 평가를 활용할 경우, 응시자는 전면적(global)인 점검 과정을 통해 자신의 확신 정도를 추측하고 사후 평가를 내리게 된다(Schraw, 1994). 이러한 점검의 경우 최신효과로 인해 문항의 배치나 문항의 개수에 따라 같은 확신을 가지고 있어야 할 시험에서도 서로 다른 응답이 가능할 수 있다. 확신 가중의 경우 전면적 점검 과정의 문제는 없지만, 시험에 대한 응답과 별개로 매번 문제에 대하여 추가적인 응답(확신에 대한)을 해야 하는 과정이 부자연스러우며 부담스러운 단계가 될 수 있다. 또한 각 선지들을 기준으로 확률을 배분하는 다중 평가방식과 달리 명확한 기준점이 없기 때문에 응답한 확신 수준에 대한 해석에 문제가 있을 수 있다.

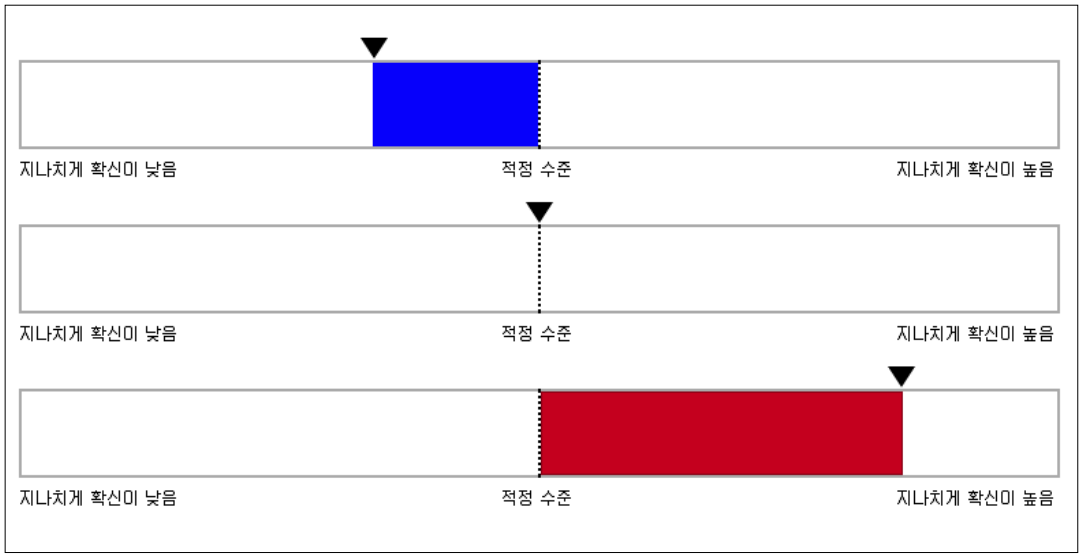
본 연구는 이러한 점을 바탕으로 선행연구들이 지녔던 한계점들을 보완함과 동시에, 효과적인 피드백을 줄 수 있는 방법으로 다중 평가 방식의 피드백을 활용하였다. 본 연구의 목적은 실제로 다중 평가 방식을 통해 시험을 볼 때 확신 수준에 대한 피드백을 받을 경우 다음 시험에서 자신의 확신 수준을 보다 현실적으로 조정 가능한지 알아보고자 하였다. 이를 위하여 20 문제씩 4 세션에 걸친 총 80개의 다중 평가 문항을 구성하였으며, 시험을 푸는 동안 세션 사이에 피드백을 받은 집단이 그렇지 않은 집단에 비해 현실적으로 조정된 확신 수준을 갖게 되는지 확인하고자 한다. 본 연구는 다중 평가 방식을 통한 피드백이 확신 수준을 현실적으로 조정할 수 있을 것이라고 예상한다.

실 험

실험 참여자 : 본 연구의 참여자는 S 대학교에 재학 중인 학부생으로, 2014년도에 심리학개론 혹은 인간의 이해 수업을 듣는 학생들이며, 연구 참여자 모집 시스템인 R-Point를 통하여 60명의 학생이 모집되었으며, 참여자들은 통제 조건과 확신 정도에 대한 피드백을 받는 조건으로 나누어진 두 집단으로 구성되어, 무선 배분되었다.

실험 설계 : 실험에 소요되는 시간은 약 한 시간 정도로, 연구 참여자들은 통제조건과 피드백 조건에 무선적으로 배분된다. 두 집단 모두 네 세션에 걸쳐 시사·문화 상식에 관련된 문제를 각 세션마다 20개씩 풀게 되며, 총 80개의 상식 문제를 풀게 된다. 각 세션과 그것을 구성하는 문제의 순서는 정해져 있으며, 각 문제를 푸는데 걸리는 시간은 40초로 고정되어있다. 통제 집단은 4 세션의 상식 문제 세트를 다중 평가 방식으로 응답하고 나면 실험이 종료된다. 피드백 집단은 4 세션의 상식 문제 세트 사이에 세 번의 확신 정도에 대한 피드백을 받게 되는데, 확신 정도는 각 세션 사이에 바로 이전 세션에 응답한 문항 20개를 바탕으로 계산된다. 따라서 세 번의 피드백은 각각 첫 번째와 두 번째, 세 번째 세션에 대한 것이며, 마지막 세션에 대한 피드백은 받지 않는다.

실험 자극 : 실험에 사용된 문항은 시사 상식 문제였으며, 실험 참여자들이 대학생들인 만큼 영역적인 특수성의 영향을 최소화 할 수 있도록 하기 위해 문화·예술, 자연과학, 역사 등 다양한 영역의 문제를 포함하였다. 선행 연구에서 다중 평가는 Dirkzwager에 의해 개발된 *TestBet*이라는 다중평가를 실제로 구현할 수 있는 소프트웨어가 사용되었으나, 개발 시점이 90년대 후반에 제작되었던 프로그램인 만큼 현재의 실험에 사용



[그림 5] 피드백의 예시

되기에는 열악한 점이 많았으며, 다중평가의 다양한 요소들을 조작하며 실험을 하기에는 더욱 제한되어있는 프로그램이었다. Holmes(2002)는 제언을 통하여 향후 등장할 프로그램이 갖추어야 할 요건을 나열하였는데, 본 연구에서는 실험에서 사용 가능한 수준으로 이러한 요건들을 충족시킬 수 있는 프로그램을 개발하고자 하였으며, 이는 웹기반 서베이 플랫폼인 Qualtrics를 응용하여 *Testbet*을 보완하여 구현하였다. 또한 Dirzwager(1996;2003)의 Holmes(2002)의 선행연구에서는 확률적인 표기나 직접적인 지시("조금 확신 높습니다" 등)으로 피드백을 주는 것을 제안하였으나, 본 연구에서는 응시자들이 피드백의 내용을 쉽게 이해할 수 있도록 하기 위하여 그래픽을 활용한 피드백을 제시하였다. 또한 A의 값이 0.5보다 작거나 1.5보다 이상일 경우 각각 "지나치게 확신이 낮음"과 "지나치게 확신이 높음"의 양 극단으로 표기되었다. [그림 5]는 피드백의 예시를 보여주는데, 응시자의 확신 수준에 따라 양 측으로 게이지가 차오르는 형식으로 구성되어있으며, 그림에서의 맨 위 그래프는 확신 수준이 낮을 경우, 아래는 과신하고 있을 경우, 가운데는 완전히 현실적인 확신 수준을 지닐 경우의 각 예시들을 나타낸다.

	인원	세션 1		세션 2		세션 3		세션 4	
		평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차
통제 집단	29	.48	.16	.49	.25	.57	.21	.39*	.21
피드백 집단	31	.55	.23	.53	.24	.67	.30	.52*	.29

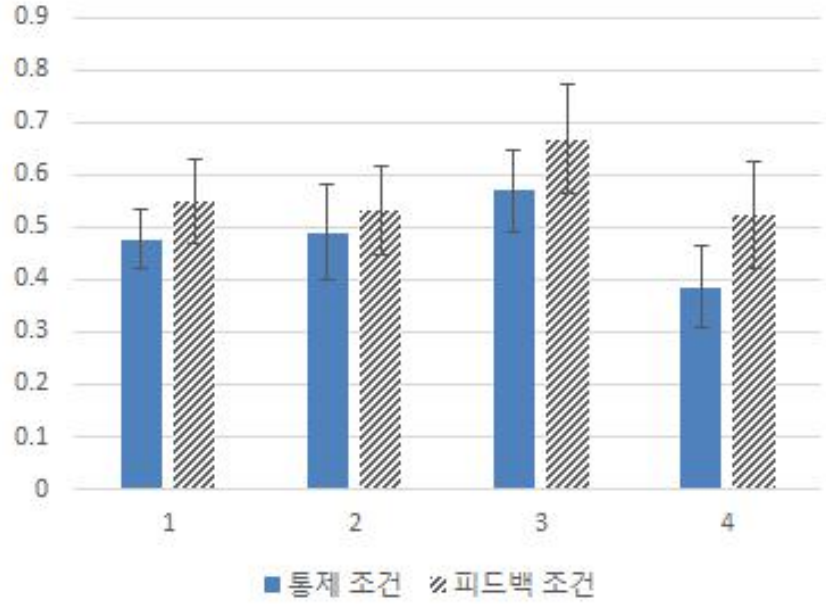
* : $p < .05$

[표 1] 통제 집단과 피드백 집단의 확신 수준

실험 결과 : 총 네 세션에 걸친 두 집단의 확신 정도에 대한 평균과 표준편차는 [표 1]에 제시되어있다. 집단을 비교하기에 앞서, 놀라운 점은 두 집단 모두 네 세션에 걸쳐 평균값이 전부 과신 경향을 보이고 있다는 점이다. 그 중에서도 통제집단의 경우 세션 3을 제외한 세 세션에서 평균값이 “지나치게 확신이 높음”의 영역에 존재함을 알 수 있다. 이러한 결과를 통해 우리는 다중 평가방식을 사용하여 측정한 평가 방식에서도 과신 편향 현상이 존재하고 있음을 알 수 있으며, 이는 과신 현상이 보편적이고 빈번하게 일어난다는 선행연구들의 결과와도 일맥상통한다(Lichtenstein, Fischhoff, & Phillips, 1977; Moore & Healy, 2008).

[그림 6]는 두 집단의 확신 정도를 비교한 그래프이다. 앞서 말하였듯 확신 정도를 나타내는 A 값은 작을수록 과신의 정도가 강함을 나타내며, 1에 가까울수록 현실적인 확신을 하고 있음을 의미한다. 본 연구에서 사용한 자극은 서로 다른 상식 문제 세트를 순차적으로 풀게 하였기 때문에, 집단 내의 세션 간 비교는 실시하지 않았다. 그러나 피드백의 효과를 살펴보기 위하여 각 세션별로 확신 정도를 비교하였는데, 세션1과 세션2, 세션3에는 통계적으로 유의미한 차이가 없었던 반면, 마지막 세션인 세션4에서는 피드백 집단이 유의하게 더 현실적인 응답을 보였다 ($t(58)=2.11, p<.05; d=.55$). 이는 참여자들이 피드백을 통해 바로 확신 수준을 조종하기 보다는, 여러 차례 피드백을 받았을 경우 조정을 할 수 있게 됨을 의미한다.

다음은 각 조건에 따른 실제 시험 점수의 비교이다. 시험 점수는



오차막대 : 95% 신뢰구간

[그림 6] 통제 집단과 피드백 집단의 확신 수준 비교

	인원	세션 1		세션 2		세션 3		세션 4	
		평균	표준편차	평균	표준편차	평균	표준편차	평균	표준편차
통제 집단	29	-.61	.40	-.50	.64	-.39	.55	-.76**	.55
피드백 집단	31	-.41	.47	-.31	.47	-.16	.48	-.36**	.49

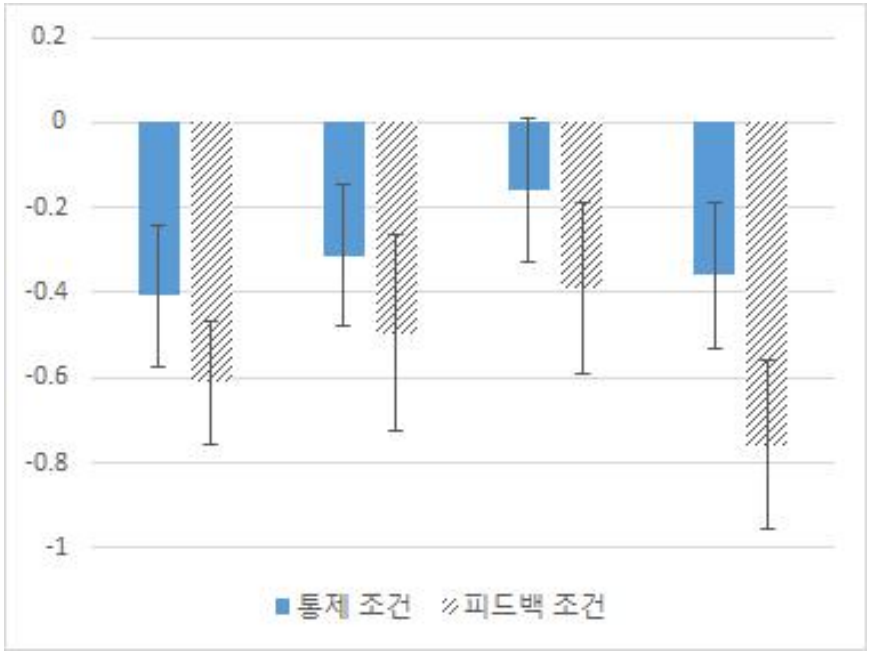
** $p < .01$

[표 2] 통제 집단과 피드백 집단의 시험 점수

Homles(2002)의 제안에 따라 다음과 같은 공식에 따라 채점되었다.

$$\text{시험 점수} = 1/n\sum\{1n(0.996\times rc(i) + 0.001) + 1.386\}/1.383$$

이러한 공식을 따른다면 응시자가 모든 정답인 선택지에 100%를 기입할 경우 점수는 최고점인 1점에 근사하게 되며, 반대로 오답 선택지에 모든 확률을 부여할 경우 최저점인 -4점에 근사하게 된다. [표 2]는 이렇게 얻어진 점수의 평균과 표준편차를 보여준다. 두 집단의 시험 점수는 확신 수준과 마찬가지로 세션 1~3에서는 유의한 차이가 없었으나, 마지막 세션에서는 피드백 집단의 시험 수행이 유의하게 더 높았다



오차막대 : 95% 신뢰구간

[그림 7] 통제 집단과 피드백 집단의 시험 점수 비교

($t(58)=-2.99, p<.01; d=.55$). 이는 다중 평가를 통한 확신 수준에 대한 피드백만으로도 응시자들이 더 정답에 가까운 응답을 할 수 있도록 기여할 수 있음을 보여준다.

다중 평가에서 수정(Correction)의 활용에 대한 탐색

다중 평가의 확신 정도 평가 공식은 위에서 본대로 A = {k×∑rc(i)-m}/{k×∑∑r(i,j)^2-m}와 같다. 이때의 A 값의 도출은 참된 값 p와 편향된 값 r의 선형적인 관계를 가정하고 있다. 즉, 현실적인 확신 수준에서의 확률 평가를 p라고 하고, 응시자의 주관적인 확률 평가를 r이라고 했을 때, p와 r의 관계는 p=Ar+b로 표현될 수 있으며, 이 때 A의 값을 도출한 것이 다중 평가에서 활용하는 확신 수준이 된다. 또한 이 때 B의 값은 (1-A)/k가 되기 때문에, p와 r의 관계는 다시 p=Ar+(1-A)/k 라고 표현될 수 있다. 그렇기 때문에 A의 값이 주어진다면, 참여자의 응답 r을 현실적인 수준인 p로 수정할 수 있는 가능성이 생긴다. Holmes(2002)는 이를 과잉 확신의 정도가 심한 학생들의 시험 값을 조정해주기 위해 활용하였다. 그러나 이런 활용은 문제가 있을 수 있는데, 먼저 이러한 변환은 응답자의 기입을 정답에 가깝게 조정해주는 작업으로 형평성에 문제가 있

으며, 응시자의 응답을 임의로 왜곡한다는 문제 역시 존재한다. 특히 이미 정답을 알고 있어야 활용할 수 있는 요소로 정답이 주어지지 않은 상황에서나 혹인 시험 결과를 알 수 없는 응시자의 입장에서는 활용이 불가능하다.

그러나 만약 이전의 응답 경향성이 현재에도 존재한다면, 과거의 A값을 현재의 응답 기입에 하나의 정보의 형태로 활용하는 것은 가능하다. 이 경우 t기에 (t-1)기를 반영하여 수정된 r 값은 $At_{-1}+(1-At_{-1})/k$가 된다. 이렇게 변형된 수정된 r값은 참여자들에게 피드백의 형태로 지급될 수 있으며, 참여자들은 이를 통해 각 예측을 실제로 어떻게 조정하면 되는지에 대한 구체적인 피드백을 받게 된다. 이러한 피드백의 지급은 다중 평가 방식의 활용으로는 시도된 바 없다. 본 연구에서는 이러한 수정의 유용성을 알아보기 위해, 피드백을 받지 않았던 통제 집단의 응답을 바탕으로, 기존의 응답과 수정된 응답의 확신 수준과 시험 점수를 비교해보고자 하였다. 이 때 응답의 수정은 과거 세션에서의 A값에 기반하고 있기 때문에($At_{-1}+(1-At_{-1})/k$), 과거의 정보가 없는 세션 1을 제외하고, 세션 1을 기반으로 수정된 세션 2와 원래의 응답, 세션 2를 기반으로 수정된 세션 3과 원래의 응답, 마지막으로 마찬가지 과정을 거친 세션 4를 비교해

	인원	세션 2		세션 3		세션 4	
		평균	표준편차	평균	표준편차	평균	표준편차
수정 전	29	.49***	.25	.57***	.21	.39*	.21
수정 후	29	1.10***	.60	1.43***	.89	.77*	.85

*** $p < .001$, * $p < .05$

[표 3] 수정 전과 후의 확신 수준

	인원	세션 2		세션 3		세션 4	
		평균	표준편차	평균	표준편차	평균	표준편차
수정 전	29	-.50***	.64	-.39***	.55	-.76***	.54
수정 후	29	.12***	.13	.13***	.10	.05***	.11

*** $p < .001$

[표 4] 수정 전과 후의 시험 점수

보고자 하였다. 이렇게 도출된 수정 전과 후의 A 값과 시험 점수에 대한 평균과 표준편차는 [표 3]과 [표 4]에 제시되어있다.

비교 결과, 놀랍게도 수정 된 r을 바탕으로 도출된 확신 수준은 수정되기 전보다 더 현실적이었으며(세션2:$t(56)=-5.04$, $p<.05$, $d=1.32$, 세션3:$t(56)=-5.079$, $p<.001$, $d=1.33$, 세션4:$t(56)=-2.34$, $p<.05$, $d=.61$), 시험 점수 역시 유의미하게 향상되었음을 확인할 수 있었다(세션2: $t(28)=-6.08$, $p<.001$, $d=1.34$, 세션3: $t(56)=-5.08$, $p<.001$, $d=1.31$, 세션4: $t(56)=-9.36$, $p<.001$, $d=2.07$). 이러한 결과는 응시자들이 자신의 과거의 응답 기록을 바탕으로 한 확신 수준을 미래의 응답에 반영하였을 경우, 서로 다른 문제 상황임에도 확신 수준과 정답률이 더 향상될 수 있음을 보여준다. 또한 이는 수정된 r값이 피드백을 위한 정보의 형태로 활용될 수 있을 가능성을 시사한다.

종합 논의

본 연구는 다중 평가를 통하여 상위 인지적인 점검 능력, 특히 확신 수준을 측정하고 이를 피드백을 통하여 조정할 수 있는지를 검토하였다. 이를 위하여 연구 참여자들에게 실제로 상식 문제들로 이루어진 시험 문제를 풀게 하였고, 그 결과 피드백을 받은 집단이 마지막 시험 세션에서 통제 집단에 비해 더 현실적인 확신 수준과 향상된 시험 점수를 받을 수 있었음이 확인되었다. 또한 과거의 확신 수준을 이후의 응답에 반영하여 수정한 새로운 응답을 수정 전과 비교한 결과 확신 수준과 응답 정확성 모두 전 세션에서 유의미하게 향상되었음을 알 수 있었다. 이러한 결과는 다중 평가를 통한 수정이 새로운 형태의 피드백으로 활용될 수 있을 가능성을 시사한다.

이러한 연구 결과를 더욱 상세히 논의해보자. 먼저 피드백이 확신 수준 조정에 미치는 영향을 알아보기 위한 본 실험에서 주목할 만 한 점은, 피드백의 효과가 즉각적으로 드러나지 않고 세 번째 피드백이 지급된 후의 네 번째 세션에서만 차이가 드러났다는 것이다. 이는 피드백에 대한 응시자의 반응이 즉각적으로 드러나기는 어려울 수 있음을 시사한다. 그럼에도 불구하고, 피드백은 약 한 시간이라는 짧은 시간동안 단 세 번만 주어졌다는 점에서 다중 평가를 통한 피드백이 효과적으로 적용되었음을 알 수 있다. 그러나 후속 연구가 진행된다면, 피드백 효과를 보다 더 빠르게 보일 수 있는 방안에 대한 탐색이 필요하다.

한 가지 가능성은 수정(correction)의 활용이다. 앞서 분석된 결과에서 과거의 확신 수준을 반영하여 수정된 응답은 기존의 응답에 비해 더 높은 정확성을 보임을 알 수 있었다. 이러한 결과를 바탕으로, 새로운 형태의 피드백을 제안해볼 수 있는데, 응시자가 다중 평가로 이루어진 어떠한 시험 문제를 다 풀고 나면 그것을 바로 제출하는 것이 아니라, 그가

기존에 보였던 확신 수준을 바탕으로 수정된 응답을 비교하여 확인한 후 자신의 응답을 조정하여 다시 제출하도록 유도하는 것이다. 이는 단순히 응시자에게 확신 수준에 대한 정보만을 제시하는 것이 아닌, 직접적으로 어떻게 응답을 조정해야하는 지 알려주는 적극적인 피드백이 될 수 있다.

본 연구의 함의를 확장을 위해 제안하는 연구는 다음과 같다. 먼저, 본 연구에서는 상식 문제를 활용하여 피드백의 효과를 확인하였다. 후속 연구에서는 보다 더 다양한 영역에서 다중 평가의 피드백이 효과가 있는지 확인해볼 필요가 있으며, 나아가 이러한 효과가 영역에 따라 특수하게 적용되는지 아니면 보편적으로 유지될 수 있는지 연구할 필요가 있다. 또한 본 연구에서는 피드백으로 인한 조정의 효과가 장기적으로 지속될 수 있는지 살펴보지 못하였다. 후속 연구를 통해 조정된 확신 수준이 얼마나 지속될 수 있는지 확인하는 것 역시 중요한 연구가 될 것이다. 본 연구가 지니는 또 다른 제한점은 확신 수준에 대한 피드백을 받았을 경우 응시자가 어떤 문항에서 조정을 해야 하는지를 스스로 판단할 수 있는지 확인하지 못하였다는 점이다. 다중 평가 방식은 앞서 말하였든 문항 내의 모든 선택지들을 기반으로 확신 수준을 평가하지만, 이 정보만으로는 평가된 문항과 선택지 내에서 정확히 어떤 부분을 조정해야 하는지 알려주지 않는다. 그렇기 때문에 확신 수준에 대한 피드백이 주어졌을 때 사람들이 자신의 응답을 바꾸어야 할 부분과 그렇지 않은 부분을 효과적으로 분별할 수 있는지, 나아가 개인에 따라서 어떤 사람들이 조정을 더 잘하는지 탐색해 볼 필요가 있다.

다중 평가는 확률 평가를 기반으로 하며, 그렇기 때문에 불확실성에 대한 판단을 다루는 학습과 시험, 예측, 선택 등 다양한 분야에 적용될 수 있는 도구이다. 또한 기존의 객관식이나 확신 가중 방식 등의 평가에 비해 응시자의 지식이나 선호 수준에 대한 정보를 더 많이 담을 수 있으며, 이를 기반으로 확신 수준에 대한 경향성을 용이하게 평가할 수 있다. 그럼에도 불구하고 현재까지 다중 평가 방식의 활용에 대한 연구는 충분히 이루어지지 못하였다. 본 연구는 실험을 통하여 다중 평가의 확신 수

준에 대한 피드백이 조정에 미치는 효과를 확인하였다는 점에서 그 의의가 있다. 무엇보다, 다중 평가를 통해 응답의 수정이라는 가능성을 제안하였다는 점에서 후속 연구를 위한 중요한 기반을 마련하였다고 할 수 있다.

참 고 문 헌

김민경, 조미경, & 주유리 (2012). 서술형 평가에 대한 인식 및 실태에 관한 조사연구-서울시 소재 초등교사를 중심으로. 한국초등수학교육학회지, 16(1), 63-95.

김병윤 (2009). 초등학교 학급, 학교 규모와 학생수용지표 관계 연구. 교육연구논총, 30(2): 1-19

백순근 (1998). 서술형 평가, 채점보다 출제가 어렵다. 중등우리교육, 149-152.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, 8(4), 559-573.

Dirkzwager, A. (1996). Testing with personal probabilities: 11-year-olds can correctly estimate their personal probabilities. *Educational and psychological measurement*, 56(6), 957-971.

Dirkzwager, A. (2001). Consensus measurement in multi-participant conversations. Kybernetes, 30(5/6), 573-588.

Dirkzwager, A. (2003). Multiple evaluation: A new testing paradigm that exorcizes guessing. *International Journal of Testing*, 3(4), 333-352.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension A Brief History and How to Improve Its Accuracy. *Current Directions in Psychological Science*, 16(4), 228-232.

Ebel, R. L.. (1965). Confidence Weighting and Test Reliability. *Journal of Educational Measurement*, 2(1), 49 - 57.

Everson, H. T., & Tobias, S. (1998). The ability to estimate knowledge and performance in college: A metacognitive analysis.

Instructional Science, 26(1-2), 65-79.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive - developmental inquiry. *American psychologist*, 34(10), 906.

Holmes, P. (2002). Multiple evaluation versus multiple choice as testing paradigm. Published master's thesis, Twente University, Ensehede, the Netherlands, ISBN 90 - 365 17 57 - 5.

Hopkins, K. D., Hakstian, A. R., & Hopkins, B. R. (1973). Validity and reliability consequences of confidence weighting. *Educational and Psychological Measurement.*

Isaacson, R. M., & Fujita, F. (2006). Metacognitive Knowledge Monitoring and Self-Regulated Learning: Academic Success and Reflections on Learning. J*ournal of Scholarship of Teaching and Learning*, 6(1), 39-55.

Isaacson, R. M., & Was, C. A. (2010). Believing you're correct vs. knowing you're correct: A significant difference. *The Researcher*, 23(1), 1-12.

Kelemen, W. L., Winningham, R. G., & Weaver III, C. A. (2007). Repeated testing sessions and scholastic aptitude in college students' metacognitive accuracy. *European Journal of Cognitive Psychology*, 19(4-5), 689-717.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology,* 77(6), 1121.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). *Calibration of probabilities: The state of the art* (pp. 275-324). Springer Netherlands.

Merkle, C., & Weber, M. (2011). True overconfidence: The inability

of rational information processing to account for apparent overconfidence. *Organizational Behavior and Human Decision Processes*, 116(2), 262-271.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, 115(2), 502.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The psychology of learning and motivation*, 26, 125-141.

Pintrich, P. R., Wolters, C., & Baxter, G. (2000). Assessing metacognition and selfregulated learning. In G. Schraw & J. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 43 - 97). Lincoln, NE: Buros Institute of Mental Measurement.

Schraw, G. (1994). The effect of metacognitive knowledge on local and global monitoring. *Contemporary educational psychology*, 19(2), 143-154.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary educational psychology*, 19(4), 460-475.

Shuford Jr, E. H., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, 31(2), 125-145.

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of educational psychology,* 95(1), 66.

Tobias, S., & Everson, H. T. (2002). Knowing what you know and what you don't: Further research on metacognitive knowledge monitoring. New York: College Board

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 663-705.

Was, C. A., Beziat, T. L. R., & Isaacson, R. M. (2013). Improving metacognition in a college Classroom: Does Enough Practice Work? *Journal of Research in Education* 23(1)

Yen, Y. C., Ho, R. G., Chen, L. J., Chou, K. Y., & Chen, Y. L. (2010). Development and evaluation of a confidence-weighting computerized adaptive testing. *Journal of Educational Technology & Society*, 13(3), 163-176.

부 록 (실험 자극)

세션 1. 20 문항

1)선진국형 인구구성 형태는?
종형, 표주박형, 방추형, 피라미드형

2)채플린이 출연한 영화가 아닌 것은?
쥐덫, 거리의 등불, 모던 타임즈, 위대한독재자

3)스탕달의 작품은?
이방인, 대지, 좁은 문, 적과 흑

4)병인양요를 일으킨 나라는?
영국, 미국, 프랑스, 러시아

5)사법(私法)상의 의무는?
근로의 의무, 부양의 의무, 납세의 의무, 교육의 의무

6)이타이아티아병의 원인은?
납중독, 카드뮴중독, 수은중독, 우라늄

7)대통령이 임명하지 않는 사람은?
대법원장, 국무총리, 감사원장, 국회의장

8)자본주의의 경제발전 원동력을 혁신(innovation)으로 보고, 혁신을 수행하는 기업가의 역할을 강조한 경제학자는?
슘페터, 밀, 해롤드, 마셜

9)우리나라 환율변동 주기는?
매일, 일주일, 3개월, 6개월

10)산성비의 기준이 되는 근거는?
pH5.6이하, pH5.6이상, pH7.0이하, pH7.0이상

11)헌법에 규정된 대통령의 임무는?
삼권의 조정, 개헌안의 제안, 대법원장의 임명, 조국의 평화적 통일

12)최초의 상업용 컴퓨터는?
ENIAC, IBM360, UNIVAC, FACOM230

13) 그리스의 철학자 플라톤의 핵심사상은?
이데아 사상, 공리 사상, 자유주의 사상, 변증법 사상

14) 성선설(性善說)을 주장한 사람은?
순자, 공자, 맹자, 노자

15)우리나라가 코리아(Korea)란 이름으로 서양에 알려진 것은 누구에 의해서 이루어졌는가?
하멜 표류기에 의해서, 당나라 상인들에 의해서, 마르코 폴로의 동방견문록에 의해서, 벽란도에서 무역을 행한 아라비아 상인에 의해서

16)프랑스 언론 매체가 아닌 것은?
AP, LeMonde, AFP, L'Express

17) 오라토리오에 대한 올바른 설명은?
오페라의 별칭, 성가극이나 그 음악, 오스트리아 민속음악, 12음 기법으로 작곡된 음악

18)A재화 가격이 상승할 때 B재화의 수요가 감소하는 경우, 두 재화의 관계는?
대체재, 보완재, 독립재, 경제재

19) 청조(淸朝)를 멸망시키고 아시아에서 처음으로 공화제 국가를 수립하게 된 사건은?
양무운동, 메이지 유신, 신해혁명, 의화단 사건

20)소선거구란?
유권자가 적은 선거구, 지역이 협소한 선거구, 한 선거구에서 한 사람의 의원만을 뽑는 선거구, 유권자 수가 10만 이하인 선거구

세션 2. 20 문항

1)UN총회의 의결 방법은?
만장일치제, 회원국과반수제,
원칙적으로2/3다수제,
주요문제2/3다수제 그 외 문제는
과반수제

2)인상주의 화풍에 속하지 않는
사람은?
모네, 마티스, 피사로, 로트레크

3)고려시대의 연중행사로서
불교제전을 의미하는 것은?
단오, 팔관회, 연등회, 상사일

4)잔 다르크가 활약하여 승리로 이끈
전쟁은?
십자군전쟁, 장미전쟁, 포에니전쟁,
백년전쟁

5)모라토리엄이란?
통화개혁, 지불유예, 채무청산,
약정이율

6)형사미성년자는?
13세미만자, 14세미만자, 16세미만자,
20세미만자

7)발트3국이 아닌 것은?
에스토니아, 벨로루시, 라트비아,
리투아니아

8)한국의 표준시는 세계의
표준시보다 어떠한가?
9시간 늦다, 9시간 빠르다, 7시간
늦다, 7시간 빠르다

9)민주정치의 3대원리로 이루어진
것은?
-국민자치-입헌주의-권력분립,
-정당정치-국민자치-입헌주의,
-권력분립-정당정치-국민자치,
-입헌주의-권력분립-정당정치

10)우리나라 최초의 한문소설은?
청구영언, 혈의 누, 금오신화,
홍길동전

11)음의3요소가 아닌 것은?
강약, 장단, 음색, 진동수

12)북한에서 인정되지 않고 있는
소유형태는?
국가소유, 법인소유, 개인소유,
협동단체소유

13)NATO설립의근본목적은?
유럽의 정치적인 통합, 공산 침략의
공동방위, 전후 유럽 경제의 부흥,
국제간의우호관계촉진

14)"형사소송법상 공소제기가 없는
한 심판은 없다"는 원칙을 가리키는
말은?
불고불리의 원칙, 일사부재리의 원칙,
형벌부소급의 원칙, 일사부재의의
원칙

15)형사재판의 원고는 누구인가?
경찰, 검사, 고발자, 피해자

16)"인간은 생각하는 갈대다"라는
유명한 말을 한 사람의 작품은?
팡세, 화폐론, 국부론, 역사의선국

17) 외교상의 중립정책, 즉 일종의
고립주의를 무엇이라 하는가?
먼로주의, 패권주의, 티토이즘,
삼민주의

18) 집안에서 은행거래를 할 수
있도록 한 시스템은?
텔리쇼핑, 인터뱅킹, 텔리뱅킹,
가정자동화

19) "보상 이면에 숨겨진 손해
(hidden cost of reward)"는 어떤
현상을 가리키는 말인가?
보상에의 중독, 보상에 대한
한계효용 체감, 보상으로 인한 내적
동기의 저하, 보상이 훈련됨으로써
장기적인 비용이 상승

20)
게리맨더링(Gerrymandering)이란?
선거비용과 선거운동의 일부를
정부가 부담하는 것, 특정 개인이나
정당에 유리하도록 선거구를
획정하는 것, 사표(死票)를 방지하기
위해 비례대표제를 채택하는 것,
군소정당의 난립을 방지하기 위해
소선거구제를 채택하는 것

세션 3. 20 문항

1)각국의 수도이다. 틀린 것은?
캐나다-오타와, 스위스-취리히,
레바논-베이루트,
사우디아라비아-리아드

2)섀도캐비닛(Shadow Cabinet)이란?
야당내각, 여당내각, 각외대신,
각내대신

3)실존주의 작가는?
말로, 위고, 카뮈, 플로베르

4)제6공화국 헌법에서 대통령의 임기는?
4년 중임, 4년 단임, 5년 단임, 7년 단임

5)부조리란?
카뮈에 의해 처음 쓰인 말,
데카르트에 의해 처음 쓰인 말,
하이데거에 의해 처음 쓰인 말,
야스퍼스에 의해 처음 쓰인 말

6)데탕트(Détente)란?
양극화, 냉전, 다극화, 긴장완화

7)대한민국 국민이 되는 요건을 규정하고 있는 법률은?
민법, 헌법, 호적법, 국적법

8)KS마크란?
선하증권, 무결점운동, 품질관리,
한국공업규격표시

9)후손들에게 훈요십조라는 교훈을 내린 왕은?
신라 문무왕, 고려 태조 왕건, 고려 광종, 조선 태조 이성계

10)우리나라에서 농사를 짓기 시작한 시기는?
구석기시대, 신석기시대, 청동기시대,
철기시대

11)우리나라 금융정책을 결정하는 기관은?
재경부, 경제기획원, 경제과학심의회,
금융통화운영위원회

12)광신적 애국주의를 가리키는 것은?
나치즘, 전체주의, 파시즘, 쇼비니즘

13)우리나라에서 한류와 난류가 교류되는 대표적인 어장은?
동해어장, 남해어장, 서해어장, 서해와 남해 어장

14)백서(白書)란 무엇인가?
형사피고인의 자백서이다,
국정감사보고서의 별칭이다,
백지위임장(委任狀)을 말한다, 일반적으로 정부가 발표하는 행정현황 조사보고

15)핵력이란?
원자폭탄이나 수소폭탄의 폭발력이다,
중성자가 원자핵에 충돌해 생기는 힘이다,
화학에너지를 열에너지로 바꾸는 힘이다,
원자핵의 구성입자인 양성자와 중성자를 결합시키고 있는 힘이다

16)주한미국대사는 다음 중 어느 것의 적용을 받는가?
자기 본국인 미국의 법이 적용된다,
거주하고 있는 곳인 대한민국의 법이 적용된다, 대한민국의 법도 미국의 법도 적용되지 아니한다, 미국의 법과 대한민국의 법 중 본인이 원하는 나라의 법이 적용된다

17) 사막지대에서 우기에만 물이 흐르는 일시적이 하천을 무엇이라고 하는가?
카르(Kar), 와디(Wadi),
바르한(Barchan),
페디먼트(Pediment)

18)다음의 사건 중 가장 오래된 것은?
명예혁명, 러시아혁명, 프랑스대혁명,
산업혁명

19) 엥겔 계수란?
물가지수, 생계비지수, 가격지수,
도매물가지수

20) 환경영향평가제란 무엇인가
환경 보존 운동의 효과를 평가하는 것,
환경보전법, 해상오염방지법,
공해방비법 등을 총칭하는 것, 건설이나 개발이 주변 환경과 인간에게 미치는 영향을 미리 측정하여 대책을 세우는 것,
공해지역 주변에 특별 감시반을 설치하여 환경보전에 만전을 기하는 것

세션 4. 20 문항

1)지방자치단체의 장이 제정 하는 것은?
조례, 규칙, 명령, 법률,

2)현대사회의 대중을 '고독한 군중'이라고 표현한 사람은?
로크, 쿨리, 마르크스, 리스먼

3)직접민주정치와 관계가 먼 것은?
국민대표, 국민투표, 국민발안, 국민소환

4)필리버스터(Filibuster)란?
다수당의 횡포, 야당의 최고 지도부, 국회의장의 의결권, 소수당의 의사진행 방해 행위

5)우리나라에 최초로 들어온 유럽인은?
하멜, 벨테브레, 오페르트, 마르코 폴로

6)전파와 빛의 차이점은?
진폭, 속도, 파장, 전달되는 방법

7)'구토'의 작가는?
보들레르, 지드, 사르트르, 세르반테스

8)다음은 음악의 빠르기이다. 가장 빠른 것은?
Adagio, Largo, Vivace, Presto

9)1932년 행해졌던 윤봉길 의사의 의거장소는?
경성 식산은행, 만주 하얼빈역, 동경 시의연병장, 상해 홍코우공원

10)"대한민국은 민주공화국이다"라는 말은 우리나라 헌법상 무엇을 밝힌 것인가?
정체(政體), 국체(國體), 국호(國號), 국호와 국가형태

11) 법의 근본이념은?
정의 실현, 선(善)의 실현, 질서 유지, 성(誠)의 실현

12)실질임금(실질소득)이란?
화폐로 지급 받는 임금, 생산고에 따라 실질적으로 지급되는 임금, 세금 등을 공제하고 실질적으로 받는 임금, 화폐로 지급된 임금이 가지는 소비재의 구매력

13) 사면권은 누구의 권한인가?
대통령, 대법원장, 국회의장, 검찰청장

14)국제 관례상 외교사절을 파견하기 전에 상대국의 동의를 구하는 것을 무엇이라 하는가?
비토, 신임장, 엠바고, 아그레망

15)팔만대장경인각대사업을 하게 된 동기는?
경전의 보급, 불교의 보급, 호국정신의발로, 고려 문화의 과히

16)법의 단계로 맞는 것은?
헌법-법률-조례-규칙-명령,
헌법-명령-규칙-법률-조례,
헌법-규칙-조례-명령-법령,
헌법-법률-명령-조례-규칙

17)법의 효력에 대한 설명 중 옳은 것은?
구법은 신법에 우선하여 적용된다, 특별법은 일반법에 우선하여서 적용된다, 법은 제정일부터 폐지일까지 효력을 갖는다, 속인주의를 원칙적으로 속지주의를 보충적으로 적용한다, 속인주의를 원칙적으로 속지주의를 보충적으로 적용한다.

18)화폐 단위가 잘못 연결된 것은?
Rouble-구소련, Dollar-캐나다,
Mark-독일, Lira-프랑스

19) 미국의 독립이 선언되었던 조약은?
베를린 조약, 파리 조약, 워싱턴 조약, 런던 조약

20) 영화와 관계없는 것은?
대종상, 그래미상, 아카데미상, 골든글로브상

Abstract

Enhancing Metacogitive Judgment Using Multiple Evaluation

Ik Joo Hyun
Psychology
The Graduate School
Seoul National University

Multiple Evaluation is a testing paradigm that uses probability assessment instead of multiple choice. Multiple Evaluation was developed to provide more accurate and reliable information compared to traditional testing paradigm such as multiple choice or confidence weighting. Since Multiple Evaluation uses examinee's subjective probability assessment as data, it is possible to derive degree of overconfidence. In this paper, an experiment was conducted to explore whether metacognitive feedback using Multiple Evaluation can enhance metacognitive judgment. Participants(N=60) were asked to solve 4 sections of common knowledge question, each containing 20 quizzes. Feedback group(N=31) received metacogitive feedback based on their personal degree of realism after each section where as control group(N=29) did not receive any feedback. Result showed that after three sections, level of overconfidence was significantly lower for feedback group than control group, and performed better on their test.

keywords : **Metacognition, Multiple Evaluation, Calibration, Probability Assessment**
Student Number : **2013-22825**