인지과학, 제30권 제4호
Korean Journal of Cognitive Science
2019, Vol. 30, No. 4, 175~197
http://dx.doi.org/10.19066/cogsci.2019.30.4.1

동료평가 정확도 향상 방안의 비교: 평가 기준에 대한 학생들 간 토론 대 전문가 평가 사례 제시

박 정 애 박 주 용†
서울대학교 심리학과 & 심리과학 연구소

고차적 사고를 요구하는 글쓰기는 배운 지식을 활용하고 발전시키게 한다. 학습 장면에서 글쓰기를 더 많이 활용하기 위해서는 동료평가를 도입하는 동시에 동료평가의 정확도를 높일 필요가 있다. 본 연구에서는 동료평가의 정확도를 높이는 방안으로, 평가 기준에 대해 학생들끼리 토론하게 하거나 전문가의 평가를 참고하도록 하는 두 방안을 탐색하였다. 대학생 150명을 대상으로 한 실험 결과, 동료평가 후 전문가의 평가를 참고했던 집단이 평가했던 글을 다시 평가할 때는 평가 정확도가 향상되었지만, 새로운 글을 평가할 때는 향상되지 않았다. 반면에, 동료평가 후 평가 기준에 대하여 토론을 진행했던 집단은 평가했던 글을 다시 평가할 때는 평가 정확도가 향상되지 않았지만, 새로운 글을 평가할 때는 향상됨을 발견하였다. 또한 토론 집단의 경우 평가 기준에 대한 토론에서 총 발언 수가 많아질수록 평가 정확도가 높아졌다. 이상의 결과는 평가 기준에 대한 토론에서 적극적이고 자발적인 논의가 활발할수록 이후 동료평가의 정확도가 향상됨을 시사한다.

주제어 : 동료평가, 글쓰기, 토론, 동료평가 정확도

† 교신저자: 박주용, 서울대학교 심리학과, 서울특별시 관악로1 서울대학교
연구분야: 인지 심리
Tel: 02-880-9050, E-mail: jooypark@snu.ac.kr

글쓰기는 배운 지식을 활용하고 발전시키게 하여, 학습과 사고를 촉진한다. 복잡한 인지과정을 통해 글쓰기 활동은 고차적 사고를 활성화시킨다. Hayes와 Flower(1980)의 글쓰기 인지 모델에 따르면, 글쓰기 과정은 계획(planning), 변형(translating), 검증(reviewing)의 반복적인 과정으로 이루어진다. 즉, 글의 목적과 대상에 맞게 글의 주제와 내용을 계획하고, 자신의 아이디어를 쓴 글을 검증하면서 자신의 계획을 수정하게 되고, 다시 검증하는 반복적인 과정이 글쓰기에서 나타난다. Fisher와 Frey(2004)도 글쓰기를 통해 학생들이 알고 있는 것을 인출하여 활용하고, 새로운 문제제기를 하며, 통찰력 있는 해답을 제시할 수 있다고 글쓰기 활동의 중요성을 강조한 바 있다. 따라서 글쓰기는 자기주도적 학습자를 양성하는 교육 방법이 될 수 있다.

이러한 글쓰기의 효율을 높이기 위해서는 학생들이 글을 많이 쓰고, 좋은 글을 많이 접해 볼 필요가 있다. 또한 학생들이 쓴 글에 대해 효과적이고 즉각적인 피드백이 필요하다(Russell, Van Horne, Ward, Bettis III, & Gikonyo, 2017). 그러나 글을 읽고 평가하는 과정이 시간이 많이 걸려 교수자의 업무를 과중시키는 현실적인 한계점 때문에 글쓰기가 교육과정에서 널리 사용되지 않고 있다. 이러한 문제를 해결하기 위한 한 방법으로 동료평가를 고려할 수 있다(Tsai & Liang, 2009; Li, Xiong, Zang, Kornhaber, Lyu, Chung, & Suen, 2016; Cheung-Blunden & Khan, 2018).

동료평가란 글쓰기 과제물에 대하여 평가 기준에 따라 동료 학습자 간에 서로 점수를 부여하고 피드백을 제공하는 일련의 과정을 말한다. 잘 규정된 평가 기준에 따라 동료평가가 이루어지면, 교수자 평가와 유사한 결과를 얻을 수 있다(Falchikov & Goldfinch, 2000; Topping, 1998; Cho, Schunn, & Wilson, 2006). 동료평가는 전문가만큼의 피드백은 아니더라도 즉각적인 피드백을 제공할 수 있고, 평가의 횟수를 늘릴 수 있다는 큰 장점이 있다. 글쓰기를 평가하고 적절한 피드백을 제공하는 것은 동료평가자 본인의 글쓰기 실력 향상에도 큰 영향을 줄 수 있다. 따라서 동료평가는 실용적인 평가 도구로써 활용될만한 여지가 충분하다.

본 논문에서는 먼저 이러한 동료평가의 유용성 측면을 살펴보고, 동료평가 정확도 향상을 탐구한 연구들을 검토하였다. 이를 바탕으로 동료평가의 정확도를 향상시키기 위한 방안으로 동료평가 후 평가 기준에 대해 토론하는 방안과 전문가 평가 사례를 참고하는 두 방안을 제시하였다. 제시된 두 방안의 효과를 비교하기 위한 실험 결과가 논의되었다.

동료평가의 유용성

학생들이 평가에 직접 참여하도록 하는 동료평가는 다양한 교육적 효과가 있다(박주용 & 박정애, 2018). Topping(1998)은, 동료평가의 교육적인 효과를 인지적 측면, 감정적 측면과 사회적 측면으로 나누어 설명한다. 첫째, 인지적 측면에서 동료평가는 학생들이 학습에 초점을 맞추어 학습을 스스로 계획하고 장단점을 파악하여, 메타 인지를 발달시킬 수 있게 도와준다. 동료평가에는 설명, 단순화, 명료화, 요약, 재구성과 인지적 구성을 반영하는 과정들이 포함되어 있기 때

문이다. 동료평가를 통해 피드백을 받을 때보다 피드백을 제공하면, 자신의 글쓰기도 향상된다 (Cho & Cho, 2011). 동료평가는 또한 자기 평가 능력을 발달시키는데도 기여한다 (Reinholz, 2016), 두 번째, 감정적 측면에서, 동료평가는 동료평가 초기에 다소 긴장감과 불안감을 조성하지만, 책임감, 동기, 흥미, 상호작용, 정체성 확립, 유대감, 자기 확신, 동정심을 강화시키고, 부정적 피드백과 다양성에 대한 수용력 등을 증대시켜준다. 세 번째, 사회적 측면에서 동료평가는 팀워크와 적극적이고 능동적인 학습 태도를 향상시키며, 대화와 협상 능력, 비판을 주고받는 능력, 자기 입장을 정당화하고 제안을 거절하는 능력 등을 향상시켜 줄 수 있다.

동료평가는 이러한 교육적 효과 뿐 아니라, 평가과정에서 교사의 부담을 실질적으로 덜어줄 수 있다. Cho 등(2006)에 따르면, 교사는 학생들의 글 수준이 전반적으로 낮다고 보는 바닥 효과 (floor effect) 때문에 학생들을 글을 차별적으로 평가하지 못한다고 한다. 또한 교사는 많은 분량의 글을 혼자 오랫동안 평가하기 때문에 평가 기준이 중간에 달라질 수 있고, 시간적 여유가 없어 오류를 범할 수도 있다. 교사의 특정 학생에 대한 편향도 평가에 영향을 줄 수 있다. 반면, 동료평가자들은 서로의 글들을 상당히 다르다고 인식하고 있기 때문에 차별적인 점수를 부여할 수 있고, 여러 명이 평가한 평균 점수로 동료평가 점수를 부여하기 때문에 특정 학생에 대한 편향을 상쇄시킬 수 있다. 실제로 여러 명의 평가자로 구성될 때 동료 평가의 신뢰도가 더 높아진다는 사실은 잘 알려져 있다(Cho 등, 2006). 이와 더불어, Cho와 MacArthur(2010)은 다양한 동료 학생들에게서 받은 피드백이 한 교사에게 받은 것보다 글쓰기 개선에 도움이 된다는 결과를 제시하였다. 이 결과는 여러 명으로 진행되는 동료평가가 교사 한명의 평가보다도 객관성을 유지할 수 있으며, 여러 명의 평가 피드백이 글의 개선에도 도움을 줄 수 있다는 점을 시사한다.

Park(2017)은 이러한 동료평가 방법을 예습도구로 활용하여 글을 쓰고 평가하게 하는 것은 물론 질문을 하게 해, 토론 중심 수업을 가능하게 하였다. 이 시스템을 이용하여 15주간의 수업 이후 실시한 설문에서, 학생들은 글쓰기와 동료평가를 통해 교과 내용을 이해하고 창의적으로 생각하는데 도움이 되었다고 진술하였다. 이러한 웹 기반의 동료평가는 제출, 저장, 분배, 인출과 평가가 시공간의 제약이 없이 이루어질 수 있다는 것이 장점이다(Liu, Li, & Zhang, 2017). 익명으로 무선 배분이 가능하고 상호 피드백을 주고받고 쉽게 확인할 수 있을 뿐만 아니라, 결과 분석도 쉽게 해준다.

동료평가의 정확도 확보

동료평가의 정확도는 동료평가의 타당도와 신뢰도로 구분된다. 동료평가의 타당도는 통상 동료평가 점수와 전문가 점수와의 일치도로 측정되고, 신뢰도는 동료평가 점수들 간의 일치도로 각각 측정된다(Russell 등, 2017; Cheung-Blunden & Khan, 2018; Cho 등, 2006; Jeffery, Yankulov, Crerar, &

Ritchie, 2016). 그런데 이러한 동료평가 점수의 신뢰도와 타당도에 대한 의구심 때문에 그 교육적 효과에도 불구하고 여전히 널리 활용되지 않고 있다. 동료평가자는 전문가에 비해 배경 지식이 부족하고, 글쓰기 훈련도 덜 되어 있으며, 평가의 경험이 부족하고, 동료라는 인식 때문에 공정성을 잃을 수 있기 때문이다 (Topping, 1998; Cho 등, 2006; Liu 등, 2017). 무엇보다 평가할 자격이 없다는 생각에 학생 스스로가 평가 능력에 대해 확신이 없고, 동료평가를 신뢰하지 않는다 (Kaufman & Schunn, 2011). 그리고 동료평가는 대개 4~5개의 글을 읽고 점수를 부여하기 때문에, 모든 보고서들을 다 읽고 평가하는 교사의 평가보다 공정하지 않을 수 있다고 생각한다(Cho 등, 2006). 또한 학생들은 평가자 역할에 익숙지 않고 긴장과 경쟁의 관계 놓여있기 때문에 동료 점수가 불공정하며, 덜 정확하고, 덜 형식적이며, 덜 엄격하다고 생각하는 경향이 있다.

그러나 학생들의 생각과는 달리 동료평가는 전문가의 평가와 크게 다르지 않다(Rushton, Ramsey, & Rada, 1993). 동료평가자와 전문가 간의 신뢰도와 타당도에 관한 많은 연구들이 진행되어 왔으며, 그 결과 동료평가 점수가 전문가와의 점수간의 상당한 일치도가 보고되고 있다 (Toppong, 1998; Falchilov & Goldfinch, 2000; Li 등, 2016). 또한 동료 평가 훈련을 통해 동료평가의 정확도를 향상시킬 수 있다는 결과도 많다(Sluijsmans, Brand-Gruwel, & van Merrienboer, 2002; Liu 등, 2017).

동료평가의 타당도와 신뢰도

동료평가의 타당도는 동료평가 점수와 전문가 점수간의 높은 상관을 통해 입증되어 왔다 (Topping, 1998; Falchikov & Goldfinch, 2000; Jeffery 등, 2016; Li 등, 2016). Topping(1998)은 30개의 동료평가에 관한 연구 문헌들을 메타 분석한 결과 72%의 결과물에서 높은 타당도를 관찰하였다. Falchikov와 Goldfinch(2000)에서는 Topping(1998) 연구에서 다루지 않은 48개의 동료평가에 관한 연구들을 분석한 결과, 상관계수가 .14부터 .99까지였는데, 평균은 .69로 높은 상관관계가 있다는 것을 확인하였다. Li 등(2016)은, Falchikov와 Goldfinch(2000)에서 다루지 않은 컴퓨터 기반의 동료평가 연구들을 분석하여, 전문가와의 상관관계가 .63임을 보고하였다.

동료평가의 신뢰도는 동료평가자들 간의 일치도로 분석함으로써 입증하는데 계급내상관계수, 동료평가자들의 편차 점수의 차이 또는 크론바하 알파로 측정한다(Cho 등, 2006; Jeffery 등, 2016; Cheung-Bluden & Khan, 2018). Cho 등(2006)은 동료평가자들 간에 얼마나 차이가 있는지를 계급내상관계수의 단일측도 계수와 동료평가자들의 평균 점수와 개별 점수가 얼마나 차이가 나는지를 보는 평균측도 계수를 각 과제별과 동료평가자별로 계산하여 동료평가자가 얼마나 일관되게 평가하는지를 통해, 평가 결과와 동료평가자들 간의 일치성 정도를 파악하였다. 단일측도 계수는 학부생을 대상으로 했을 때 .20에서 .47이고 그 평균은 .34였고, 대학원생의 경우 .17에서 .56이고 그 평균은 .39였다. 평균측도 계수의 경우 학부생은 .45에서 .84이고 그 평균 .70이었고, 대학원생은 .45에서 .88이고 그 평균은 .71이었다. 학부생과 대학원생간 차이가 거의 없음을 보여주

는 결과이다. Cheung-Bluden과 Khan(2018)는 동료평가자들의 개별 평가 능력을 알기 위해서 모든 평가자들의 전체 그룹에 대해 크론바하 알파를 구한 다음, 각 특정 평가자의 점수를 삭제시키면서 그 평가자의 일치성의 기여도를 파악하는 방법으로 신뢰도를 살펴보았다. 크론바하 알파의 범위는 .60에서 .72이고 그 평균은 .71이었다. 일반적으로 .7 이상인 신뢰도는 높은 것으로 간주되므로, 동료평가의 신뢰도는 받아들일 만한 수준이 된다고 볼 수 있다. 이상의 연구는 동료평가 점수가 전문가 점수만큼이나 정확할 수 있음을 시사한다.

동료평가 점수 정확도에 미치는 요인

동료평가의 정확도에 미치는 중요 요인으로 동료평가 훈련(Sluijsmans 등, 2002; Liu와 Li, 2014), 동료평가물의 수(Jeffery et al., 2016)와 동료평가자의 수(Cho 등, 2006), 그리고 동료평가자의 인지적 능력으로 세 가지를 꼽을 수 있다(Patchan & Shunn, 2015; Russell 등, 2017). Sluijsmans 등(2002)의 연구에서는 동료평가 훈련이 학생들의 평가 기술을 향상시키고, 과제 수행률을 높이는데도 도움이 된다는 것을 확인하였다. 이들의 연구에서 동료 평가 훈련은 평가 기준에 대한 학습과 피드백에 대한 훈련이었다. Liu와 Li(2004)의 연구에서도 점수 기준에 대하여 이해시키고 동료와 교사의 평가를 비교하도록 하는 훈련이 학생들의 평가 능력과 과제 수행률을 향상시켜 준다는 것을 보여주었다.

동료평가물의 수와 동료평가자의 수도 동료평가의 정확도에 큰 영향을 미친다(Jeffery 등, 2016; Cho 등, 2006). Jeffery 등(2016)의 연구에 따르면, 동료평가물의 수가 6개일 경우 동료평가 점수와 전문가와의 점수가 가장 일치되었고, 동료평가를 한 횟수와 평균 점수 간의 신뢰도와 타당도와의 사이에 정적 상관이 있었다. 즉, 동료평가를 한 횟수가 많아질수록 전문가와의 점수와의 편차가 작아진다. 그리고 효과적인 동료평가의 훈련, 경험과 인지적 능력에 따라 이상적인 평가물의 수는 줄어들 수 있다고 결론 내리고 있다. Cho 등(2006)의 연구에서는 4개 대학, 16개 다른 강의의 708명의 학생들을 대상으로 동료평가점수의 신뢰도와 타당도를 분석한 결과 3명, 4명 그리고 6명의 동료평가자들의 수를 비교하여 6명의 점수를 합산하였을 때 가장 효과적인 신뢰도를 얻을 수 있음을 보였다. 동료평가자들의 수와 신뢰도간의 상관관계는 .78이었다. 이 연구에서는 최소 4명의 동료평가점수의 합산이 전문가만큼 신뢰도가 있고, 타당도가 있다고 보고하였다.

마지막으로, Patchan과 Shunn(2015)의 연구에 따르면 평가자의 인지적 능력이 동료평가 타당도에 영향을 준다. 학업 성취가 높은 상위 학생들은 비판이 많고 문제점을 많이 지적하면서 대안을 제시하는 경향이 있는 반면, 학업 성취가 낮은 하위 학생들은 칭찬 위주의 차등성 없는 피드백을 제공한다. 또한 Jeffery 등(2016)의 연구에서도 동료평가의 신뢰도를 높이는 요인으로 높은 학업 점수가 관찰되었다. 즉, 높은 학업 점수를 받은 집단일수록 동료평가자와 전문가와의 편차가 작고, 동료평가자 점수들 간의 차이도 작았다. 이상의 결과는 인지적 능력이 높은 학생들이 낮은 학생들보다 동료평가를 더 잘한다는 것을 시사한다.

동료평가 점수 정확도 향상 방안

앞서 소개한 것처럼, 동료평가는 학생들에게 다양한 교육적 도움을 줄 수 있고, 동료평가 자체가 전문가의 평가만큼이나 실질적으로 중요한 평가 도구가 될 수 있다. 따라서 동료평가 점수의 정확도를 향상시킬 수 있다면 그 사용가능성을 더 높일 수 있다. 이에 본 연구에서는 동료평가 점수의 정확도를 향상시킬 수 있는 방안으로 '평가 기준에 대해 토론하기'와 '전문가의 평가 사례를 참고하기'를 제안하고 그 효과를 비교하고자 한다.

Zheng, Cui, Li 그리고 Huang(2018)의 연구에 따르면, 글쓰기 후 동료평가에 대해 토론한 집단은 토론하지 않은 집단보다 글쓰기와 피드백의 질이 향상되었다. 이 결과는 직접적으로 평가 정확도가 향상된 것을 보여준 것은 아니지만, 토론이 글쓰기에 대한 이해와 자기 평가를 향상시킬 수 있음을 보여준다. 동료평가 후 점수 부여의 이유에 대한 토론은 학생들이 어떻게 점수를 부여하는 지를 비교할 수 있게 해준다. 토론은 평가에 대한 적절한 근거가 무엇인지 지식을 탐색하게 하고, 탐색과정 가운데 비판적이고 분석적인 사고를 활성화시킨다(Murphy, Firetto, Wei, Li, & Croninger, 2016). 또한 Van Loon, Dunlosky, Van Gog, Van Merrienboer 그리고 De Bruin(2015)의 연구에 따르면, 잘못된 지식에 대한 확신도가 높을수록, 토론을 통해서 그 지식이 틀렸다는 피드백을 받게 되면, 잘못된 지식을 더 잘 수정할 수 있게 된다고 한다. 즉, 토론은 자신의 잘못된 사고를 바로잡을 기회를 얻게 하는 것이다. 따라서 토론 과정 동안 학생들은 평가 준거에 대하여 깊이 생각할 수 있고, 자신의 평가 기준 체계를 새롭게 확립할 수 있는 기회를 가질 수 있다. 평가의 기준이 더욱 명료해진다면 그 기준을 적용할 수 있게 되고, 이에 평가 정확도의 향상이 일어날 것이라고 기대할 수 있다.

이상의 논의에 근거하여, 본 연구에서는 평가 기준에 대한 토론을 통해 동료평가의 정확도가 향상되는지 확인하면서, 학생들의 토론이 얼마나 잘 활발히 이루어지는지 관찰하고자 하였다. 토론이 잘 이루어지지 않는다면 평가 정확도의 향상을 기대하기 어렵기 때문이다. 토론 집단에서 평가의 정확도가 향상된다는 예측된 결과를 얻을 경우, 토론이 얼마나 잘 활발히 이루어지는지가 평가 정확도의 향상에 도움을 주게 되었는지 확인해 볼 필요가 있다. 본 연구에 있어서의 토론은 평가기준에 의해 점수를 부여한 이유에 대한 논의이므로 토론에서 근거가 있는 이유가 제시되었는지를 확인하는 것이 중요하다. 평가한 점수에 대하여 근거를 제시한 발언이 많은 토론 집단이 구체적인 근거 없이 막연하게 글이 좋고 나쁨을 이야기한 집단 보다 토론의 질이 좋을 것이라고 예상할 수 있다. 마찬가지 이유로 토론 집단에 따라 토론의 질이 다르다면 그 결과는 평가의 정확도에 영향을 줄 것이라고 예측할 수 있다.

그러나 학생들 간의 토론이 평가 기준을 정확히 확립하는데 있어서 얼마나 영향을 줄지는 알 수 없다. 학생들 간의 토론이 엉뚱한 결론에 이를 수 있기 때문이다. 이런 가능성을 알아보기 위해, 전문가의 평을 참고하는 집단과 평가 정확성 향상 정도를 비교하였다. 전문가의 의견을

참고하는 집단은 정확한 평가 준거에 대해서 습득할 수 있는 기회가 주어지기 때문에 정확한 평가 향상도가 일어날 것이라고 기대할 수 있다. Liu와 Li(2004)의 연구에서도 동료와 전문가의 평가 결과를 비교하여 보여주는 훈련을 실시한 결과, 동료평가 점수와 교사 점수의 일치도가 향상된다는 것을 확인하였다.

그렇다면, 평가의 기준에 대해서 토론하는 것과 전문가의 평을 참고하는 것 중 어떤 방법이 평가의 정확도 향상에 더 도움을 줄까? 이는 ICAP(interactive, constructive, active, passive) 학습 프레임워크 이론으로부터 추론해 볼 수가 있다(Chi, 2009; Chi & Wylie, 2014). ICAP 학습 프레임워크는 네 가지 기본적인 학습 모드가 있다고 본다. 수동적 모드(passive)는 수업에 참여하여, 강의에 집중하고 듣는 모드로 다른 활동은 전혀 하지 않는 상태이다. 능동적 모드(active)는 새로운 지식을 덧붙이는 것 없이 주어진 자료를 조작하고 문제 풀이를 쓰고 지식을 외우는 등의 활동이다. 구성적 모드(constructive)는 제시된 자료 외에 새로운 추리를 기반으로 다이어그램을 그리거나 설명을 덧붙이거나 질문을 하거나 해결책을 제시하는 등의 활동이다. 상호적 모드(interactive)는 협동적으로 학생들이 서로 다이어그램을 같이 그리거나 다른 학생의 질문에 답을 하는 활동 등을 말하며, 함께 지식을 구성하는 과정(coconstructing)이 일어나게 한다. 이 이론에 따르면 학습자의 몰입 정도가 제일 높은 순서인 I>C>A>P 순서로 학습 효과가 가장 크다고 한다.

ICAP 학습 프레임워크 이론에 따르면 토론하기는 상호작용이 가장 활발히 일어나는 학습 활동으로 학습효과가 크게 일어나고, 수동적인 활동인 전문가의 평가 사례 참고하기는 학습 효율성이 낮을 수 있다. 이에 본 실험에서는 토론한 집단이 전문가의 평가 사례를 참고하는 집단 보다 평가에 대한 이해가 더 높아져서 평가의 정확도가 높아질 것으로 예측하였다. 이와 함께 토론의 질이 평가 기준에 미치는 영향을 살펴보았다.

실 험

본 연구는 동료평가 점수의 정확도를 향상시킬 수 있는 방안으로써, 글을 평가한 후 토론을 진행하는 방식(실험집단1)과 글에 대한 전문가의 평가 사례를 참고하는 방식(실험집단2)을 고려하였다. 그리고 동료의 글을 평가한 후 순서를 바꾸어 한 번 더 평가하게 한 통제집단과 위의 두 실험집단의 평가 정확도 향상 정도를 비교하였다. 그리고 토론을 진행한 실험집단1에서 토론의 어떤 요소가 동료평가 점수 정확도에 영향을 주었는지 탐색하고자하였다. 토론이 활발하게 일어나지 않으면 토론을 통해 평가 기준에 대한 의견을 주고받기 어려워지기 때문에 동료평가 점수 정확도 향상이 일어나지 않을 수가 있다. 따라서 토론이 얼마나 잘 활발히 일어났는지, 그리고 어떤 발언들이 토론의 질에 영향을 미쳤는지를 분석하였다.

방 법

참가자

서울 소재 S대학교 학부생 164명이 실험에 참여하였다. 참여자는 심리학과에서 운영하는 연구참여시스템을 통하여 모집하였다. 참여자 중 응답지에 20% 이상 성실하게 반응하지 않은 14명의 자료는 결과 분석에서 제외되었다. 이 실험에 참가한 학생들은 심리학개론 수업 이수 시 받을 수 있는 실험 참가 점수를 인정받았다. 최종 분석 대상이 된 실험참가자 150명의 평균 연령은 21.82세(SD=1.88)였으며 남자는 84명, 여자는 66명이었다. 각 실험참가자들은 세 집단에 무선배정 되었다. 집단 간 연령이나 성별 등 인구통계학적 변인상의 차이는 존재하지 않았다. 실험집단1에는 62명이 배정되었으며, 네 명으로 구성된 조가 14개조, 세 명으로 구성된 조가 2개조로 총 16개 조로 구성되었다. 실험집단2와 통제집단은 44명씩 배정되어 개인별로 과제를 수행하고, 동료평가 점수를 산출하기 위하여 4명 단위로 명목상의 16개조를 구성하였다. 즉, 네 명의 평가점수를 한 조의 동료평가점수로 할당하기 위하여 실제 실험집단2와 통제집단은 조별 단위로 실험이 실시되지는 않았으나, 44명을 무작위로 4명씩 중복되지 않도록 하여 각각 16개씩 조로 구성하였다. 따라서 실험집단2와 통제집단은 실험집단1만큼의 실험참여자 수가 필요하지 않았으므로 집단별 실험참가자수는 다르나, 각 집단별로 16개 조로 동일하게 구성되도록 하였다. 동료평가 정확도는 조별 점수로 측정되므로 각 집단별 조의 수가 동일하도록 구성한 것이다.

실험설계

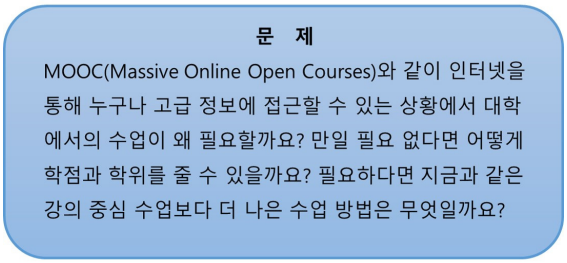
본 실험은 반복측정설계로 동료평가 정확도 향상 방식에 따라 실험집단1, 실험집단2와 통제집단의 1차, 2차, 3차 평가의 정확도를 비교하였다. 실험집단1에서는 동료평가 후 평가 기준에 대한 토론을 실시하고, 실험집단2에서는 동료평가 후 전문가 평가 사례를 읽도록 하였다. 통제집단은 같은 글이 다른 순서로 제시되어 개인이 재평가하는 시간을 갖도록 했다. 종속변인은 1차, 2차, 3차 단계에서의 동료평가 점수와 전문가 점수의 상관관계 점수(타당도)와 동료평가점수들 간의 편차점수(신뢰도)이다.

실험자극

이 연구를 위해 수집 된 자료는 2015년 2학기에, S대학에서 진행된 “심리학: 인간의 이해” 강의를 수강한 수강생이 ClassPrep이라는 온라인 동료평가 시스템에 올린 글들이다. 이 강의에 참여한 학생들은 매주 수업 전에 지정된 텍스트를 읽고 관련 내용의 요약과 함께 비판, 적용, 또

는 후속 연구를 제안한 글을 A4 1장 이내로 작성하여 ClassPrep에 업로드 하였다. ClassPrep 서버에 기록된 자료들 중 2주차에 실시되었던 주제에 대한 8개 글을 선정하였다. 글의 주제는 대학에서의 수업 필요성에 대한 것이었다, 그림 1. 글의 평가 기준은 글의 통찰과 흐름 두 가지로 구분되었다. 글의 통찰은 글의 아이디어, 비판, 활용, 후속 연구 제안에 관한 것이고, 글의 흐름은 글의 구성, 문장의 가독성에 초점을 둔 것이다.

평가 기준에 따라 전문가가 채점한 글들 가운데 상위 점수의 2개 글, 중위 점수의 4개 글, 하위 점수의 2개 글을 선정하였다. 1차 평가 단계에서 상위글 1개, 중위글 2개, 하위글 1개로 구성된 4개 글이 무작위 순서로 제시되고, 2차 평가 단계에서는 1차 평가 단계에서의 4개 글이 다른 순서로 다시 한 번 더 제시된다. 마지막 3차 평가 단계에서는 1차, 2차 평가 때와는 다른 상위글 1개, 중위글 2개, 하위글 1개가 제시된다.



(그림 1) 실험 자극에 사용된 문제

(그림 2) 실험 절차

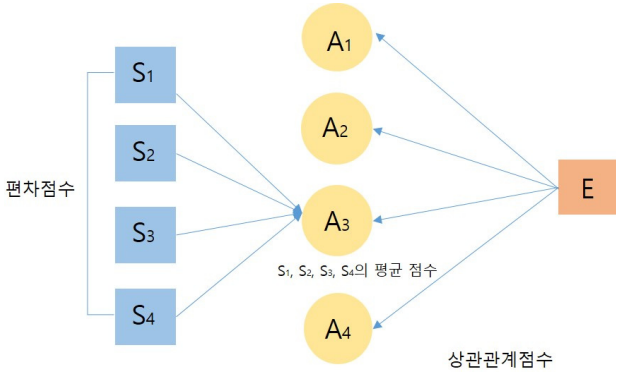
실험절차

실험참가자들은 외부와 차단된 실험실에 방문하여, 실험 진행 방법 및 내용에 대한 설명서를 읽은 후 실험 참가 동의서에 서명한 이후에 실험에 참가하였다. 실험참가자는 1차 평가 단계에서 20분 동안 4개 글을 읽고 통찰과 흐름에 대하여 7점 척도 상에서 점수를 부여한다. 실험참가자들이 글에 대하여 평가 점수를 줄 때, 평가 이유를 함께 적도록 하여, 평가에 충실히 할 수 있도록 하였고, 실험 분석에서 불성실하게 답변한 경우는 분석에서 제외시켰다. 글의 제시와 평가는 온라인상에서 이루어졌다. 2차 평가 단계에서 실험집단1에서는 4명이 한 그룹이 되어 서로 점수를 부여한 이유와 근거에 대해서 10분 동안 토론을 하고, 다른 순서로 제시된 같은 글에 대해서 5분 동안 다시 평가를 하도록 한다. 실험집단2에서는 전문가의 평가 내용을 10분 동안 읽고, 다른 순서로 제시된 같은 글에 대해서 5분 동안 다시 평가를 하도록 한다. 통제집단에서는 다른 순서로 제시된 같은 글에 대해서 15분 동안 다시 평가를 하도록 한다. 3차 평가 단계에서는 20분 동안 1차 평가 단계에서 제시되지 않았던 4개 글을 읽고 통찰과 흐름에 대하여 7점 척도 상에서 점수를 부여한다. 그림 2에 실험절차를 제시하였다.

결과 1

동료평가의 타당도는 동료평가 점수와 전문가 점수의 일치도로, 동료평가의 신뢰도는 동료평가자들 간의 일치도로 측정하였다. 동료평가의 타당도는 Cho 등(2006)의 연구에서와 같이 동료평가자들의 점수와 전문가 점수와의 적률상관계수로 구하였다. 동료평가의 신뢰도는 Jeffery 등(2016)에서처럼 동료평가자들 점수의 편차점수로 구하였다.

평가 1단계에서는 A_1, A_2, A_3, A_4 네 개의 글이 제시되었다. 그림 3과 같이 각 글의 동료평가 점수는 한 조의 네 명의 동료평가자들의 평균 점수가 된다(즉, A_3의 동료평가 점수는 S_1, S_2, S_3, S_4이 A_3에 부여한 점수들의 평균 점수). 각각 네 개의 글에 대한 동료평가 점수와 전문가(E) 점수와의 상관관계 점수가 그 조의 동료평가 점수의 타당도가 된다. 동료평가 타당도 향상 정도를 관찰하기 위하여, 각 집단별 16개 조의 상관관계 점수에 대한 반복측정 분산분석을 실시했다. Mauchly의 구형성 검정 결과, Mauchly의 W=0.60(p<.01)로 구형성 가정이 만족되지 않아 개체 내 효과 검정에서 Greenhouse-Geisser로 보정한 값을 사용하여 분석하였다. Greenhouse-Geisser 검정 결과, 상관관계 점수 변화에 대한 유의미한 차이가 없었고(F(1.43, 64.41)=2.47, MSE=.17, p=.11, η=.05), 상관관계 점수의 집단 간 상호작용은 유의미하였다(F(2.86, 64.41)=4.45, MSE=.30, p=.007, η=.17). 개체 간 효과 검정 결과, 집단별 차이는 발견되지 않았다(F(2, 45)=.37, MSE=.04, p>.05, η=.02).



(그림 3) 동료평가 점수 산출 방식

〈표 1〉 집단 간 조별 전문가 점수와의 상관관계 점수의 평균과 표준편차

집단	1차 평가	2차 평가	3차 평가
실험집단1(n=16)	.48(.31)	.49(.33)	.69(.17)
실험집단2(n=16)	.42(.26)	.68(.17)	.46(.23)
통제집단(n=16)	.50(.27)	.50(.21)	.49(.28)
전체(N=48)	.47(.28)	.56(.26)	.54(.25)

각 집단 간 조별 상관관계 점수의 평균과 표준편차를 표 1에 제시하였다. 각 평가 시기별 집단 간 상관을 일원분산분석으로 검증한 결과, 1차, 2차 평가에서는 차이가 발견되지 않았고(F(2, 45)=.34, MSE=.03, p=.71, η^2=.02; F(2, 45)=2.29, MSE=.19, p=.06, η^2=.12), 3차 평가에서만 차이가 관찰되었다(F(2, 45)=4.84, MSE=.26, p=.01, η^2=.18). Tukey HSD 사후 분석 결과, 실험집단1의 상관관계 점수가 실험집단2(p=.02)와 통제집단(p=.04) 보다 통계적으로 유의미하게 높았다.

다음으로, 동료평가 신뢰도 향상을 관찰하기 위하여 편차점수를 분석했다. 편차점수가 작을수록 평가 신뢰도가 높다고 볼 수 있다. 동료평가 신뢰도 점수는 그림 3에 제시된 바와 같이 한 글에 대한 네 명의 점수들 간의 편차 점수가 된다. 각 집단별 16개 조의 편차 점수에 대한 반복측정 분산분석을 실시했다. Mauchly의 구형성 검정 결과, Mauchly의 W=.96(p=.38)로 구형성 가정이 만족되었다. 개체 내 효과 검정 결과, 편차점수 변화에 대한 유의미한 차이가 발견되었고(F(2, 90)=8.28, MSE=.86, p=.001, η^2=.15), 편차점수의 집단 간 상호작용도 유의미하였다(F(4, 90)=3.19, MSE=.33, p=.02, η^2=.12). 개체 간 효과 검정 결과, 집단별 차이가 발견되었다(F(2, 45)=8.51, MSE=4.08, p=.001, η^2=.27). Tukey HSD 사후분석 결과, 실험집단1이 통제집단과 비교하여 편차점수가 통계적으로 유의미하게 작았다(p<.001). 각 평가 시기별 집단 간 편차 점수를 일원분산분석으로 검증한 결과, 1차 평가에서 차이가 없었고(F(2,45)=1.66, MSE=.32, p=.20, η^2=.07), 2

〈표 2〉 집단 간 조별 동료평가 편차점수의 평균과 표준편차

집단	1차 평가	2차 평가	3차 평가
실험집단1(n=16)	1.93(.45)	1.51(.58)	1.50(.58)
실험집단2(n=16)	2.07(.46)	1.77(.33)	1.94(.42)
통제집단(n=16)	2.21(.41)	2.12(.57)	2.36(.44)
전체(N=48)	2.07(.46)	1.80(.56)	1.93(.59)

차 평가와 3차 평가에서 점수의 차이가 유의미하였다(F(2,45)=604.23, MSE=1.52, p<.001, η^2=.21; F(2,45)=763.11, MSE=2.90, p<.001, η^2=.35). 2차 평가에서 Tukey 사후 검증 결과, 통제집단의 편차점수가 실험집단1보다 유의미하게 높았고(p=.004), 3차 평가에서는 통제집단의 편차점수가 실험집단2 (p=.048)와 실험집단1(p<.01) 보다 유의미하게 높았다. 각 집단 간 조별 동료평가 편차점수의 평균과 표준편차를 표 2에 제시하였다.

논의 1

통제집단의 경우 동료평가 점수의 타당도 결과를 보면, 1차, 2차, 3차 평가시기에 걸쳐 동료평가 점수와 전문가 점수와의 상관관계 점수가 달라지지 않았다. 글의 제시 순서를 달리하고, 읽었던 글을 다시 보면서 새롭게 정립된 평가 기준이 있는지 살펴보고, 실험참가자 개인이 혼자 다시 재평가하도록 하였을 때 동료평가 점수의 타당도가 향상되지 않은 것이다. 또한 통제집단의 신뢰도 결과를 살펴보면, 개인이 재평가할 때 동료평가자들 간의 편차 점수가 더 커지고, 실험집단1에 비해서 통계적으로 유의미하게 컸다. 이는 통제집단의 동료평가 신뢰도가 더 떨어졌음을 의미한다. 이 결과는 동료평가를 반복적으로 실시할 때, 다른 사람들과 공통된 평가기준으로 평가하기보다 개인적인 기준에 더 몰입했기 때문인 것으로 보인다.

전문가 의견을 참고했던 실험집단2는 2차 평가 시기에서 동료평가 점수와 전문가 점수와의 상관관계 점수가 향상되었는데, 이는 평가 타당도가 향상되었음을 의미한다. 그런데, 이 향상은 3차 평가에까지 유지되지 않았다. 2차 평가에서는 실험참가자들이 1차 평가 시에 읽었던 글을 재평가하는 과정에서 전문가의 평가 기준을 수용하여, 전문가 점수와의 일치도가 높아진 것으로 보인다. 실제 실험참가자들이 전문가의 의견을 참고한 것이 동료평가를 하는데 있어 어떤 영향을 미쳤는지를 설문 조사 결과를 통해 살펴보았다. 26%의 실험참가자가 전문가의 의견에 영향을 받지 않았다고 대답하였고, 74%가 다음과 같이 전문가의 의견이 평가 기준에 대한 이해도를 높이고 평가에 영향을 미쳤다고 반응하였다.

- “평가의 기준과 평가의 활용을 더욱 명확하게 해주어서 이전과 비교해 조금 더 객관적으

로 평가할 수 있었다."

- "글을 평가할 때 통찰과 형식 부분에서 어떤 점에 주목하는 게 효과적인지 알 수 있었다."
- "평가할 수 있는 기준에 대한 가이드라인이 되어서 조금 더 점수를 까다롭게 줄 수 있게 한 것 같다."
- "글의 기준을 다시 한 번 체크하게 한 것 같다. 특히 내용적인 부분에 조금 더 신경을 쓰게 되었다."

그러나 결과적으로 이러한 전문가 의견의 영향이 전혀 다른 글을 평가하는 3차 평가 시기에서는 미치지 못해 상관관계 점수는 다시 1차 평가 시기 때와 같은 수준으로 떨어졌다. 이는 전문가의 의견을 참고한 것은 해당 글을 다시 읽고 평가할 때에는 영향을 미치지만, 다른 글을 평가 할 때에는 영향을 주지 못한 것을 보여준다. 또한 실험집단2의 평가 신뢰도 측면에서도 전문가 의견을 참고한 직후에는 조별 점수간의 편차가 줄어들다가 다른 글을 평가하는 3차시기에 다시 커지는 것을 발견할 수 있다. 이 역시 전문가의 평가 기준의 내면화가 일어나지 않아서 편차점수는 계속 크게 발생한 것으로 보인다. 평가의 신뢰도 또한 향상되지 않은 것이다. 따라서 전문가 평가 사례를 참고했던 실험집단2에서 동료평가 점수의 타당도와 신뢰도는 향상되지 않았다.

1차 평가 이후, 평가 기준에 대해 조별 토론을 실시했던 실험집단1의 타당도 결과를 보면, 토론 후 다시 재평가하는 2차 평가에서 전문가 점수와의 상관관계 점수의 향상이 일어나지 않았다. 그러나 전혀 다른 글을 읽는 3차 평가 시기에서는 다른 두 집단에 비해서 통계적으로도 유의미하게 상관관계 점수가 향상되었는데, 이는 평가의 타당도가 향상되었음을 의미한다. 토론의 결과가 평가에 어떤 영향을 미쳤는가에 대한 설문에 대해서 25%가 점수에 별 영향을 미치지 않았다고 응답했고, 75%는 아래와 같이 평가에 영향을 미쳤다고 보고했다.

- "다른 사람들의 의견과 생각을 들을 수 있는 것만으로도 내가 잘못 생각하고 있는 것이 무엇인지 또는 논리가 얼마큼 타당하고 명확한지를 알 수 있었다고 생각한다. 예리한 비판과 깊은 생각이 드러나는 타인의 논리를 듣고, 글을 읽을 때 어떤 면에서 글을 읽고 파악해야하는지 알 수 있었다."
- "내가 전체적으로 점수를 후하게 주고 있다는 것과, 점수를 주는 방법이 조금 잘못되었음을 알게 되었다. 원래는 흥미위주의 흐름을 보려고 했지만 토론을 한 후 더욱 이성적으로 글을 평가할 수 있었다. 내 자신이 굉장히 자유로운 생각을 가진 사람이라는 것도 알게 되었다."
- "내가 원래 생각했던 글과 평과의 의도를 다른 의미에서도 바라 볼 수 있게 하였고, 내가

미처 발견하지 못했던 내용과 글의 의미 등을 파악하는 데에 도움이 되었다.”

- “나의 주장이 분명했던 글과는 달리 명료하게 읽지 못했던 글에 대한 평가가 달라졌다. 상대적으로 점수가 후한 사람이 있고 박한 사람이 있는 것 같아 내 기준이 너무 높거나 낮지는 않은지 생각해보게 되었다.”
- “글의 전체적 구조에 대해서는 신경을 못 썼는데 토론 중간에 이에 대해 언급해준 사람이 있어서 뒤에 평가에서는 흐름을 평가할 때 글의 전체적 구조가 어떤지를 고려해보게 되었다.”

동료평가 신뢰도 측면에서 보면, 평가 2차시기에 실험집단1의 편차점수는 확연히 감소하였다. 토론을 하면서 각자의 의견에 동조한 부분이 있어서 어느 정도 점수 조정이 일어났기 때문인 것으로 추측된다. 이때의 점수 조정은 평가 준거에 대한 명확한 이해를 바탕으로 한 조정이라기보다는 극단적 점수를 준 경우에 조별 평균 점수로 모였기 때문인 것으로 보인다. 따라서 평가 직후, 동료평가 점수에 대해서 서로 편차가 발생했을 때 점수를 조정하게 하는 것은 단순히 평균 점수로 모이게 할 뿐, 실제 평가 타당도에서는 변화가 없었다. 따라서 평가의 신뢰도가 높아진다고 해서 평가의 타당도가 확보되는 것은 아니며, 이는 궁극적인 평가의 정확도 향상으로 이어지게 하지 않았음을 알 수 있다. 그런데 3차 평가에서 편차 점수가 계속 낮게 유지 되고, 전문가와의 상관관계 점수도 높아졌다. 이는 토론 동안에 평가에 대한 준거의 내면화가 일어나고, 전혀 다른 글을 평가하게 될 때 그 준거들을 적용할 수 있도록 전이가 일어난 것으로 보인다. 실험집단1의 3차 평가 시기 결과와 같이 평가의 타당도와 신뢰도 모두 향상되었을 때 진정으로 평가의 정확도가 향상되었다고 볼 수 있다.

이상의 분석 결과는 통제집단과 같이 단순히 평가의 기회를 한 번 더 갖는다고 해서 평가 기준에 대한 학습이 일어나지 않으며, 이로 인해 평가의 정확도가 향상되지 않음을 시사한다. 또한 전문가의 평가 사례를 참고하는 것은 단기적으로 평가의 정확도가 향상될 수 있으나, 평가 준거의 전이가 쉽게 일어나지 않음을 보여준다. 반면에, 토론은 단기적으로는 그 영향이 나타나지 않지만, 평가 준거에 대한 내면화가 일어나고 다른 글에 적용 가능하게 해 주었다. 따라서 본 실험 결과는 동료평가 후 평가에 대한 토론이 전문가의 평가 사례를 참고하는 것보다 평가 기준을 더 명확하게 알게 해 줄 것이라는 첫 번째 가설을 지지한다. ICAP 이론에 따르면, 동료평가 후 평가 기준에 대한 토론은 상호적 모드(I)에 해당되며, 전문가 평가 사례를 참고한 것은 수동적 모드(P)에 해당된다. 이에 학습적 효과가 큰 상호적 모드인 토론이 동료평가 점수의 정확도 향상에 있어서 유용했던 것임을 알 수 있다. 그렇다면, 토론의 활성화 정도 및 토론의 질이 동료평가 점수의 정확도 향상에 도움을 주게 된 것인지 살펴볼 필요가 있다.

결과 2

토론의 활성화 정도에 따른 동료평가 정확도 향상에 있어서의 차이를 분석하기 위해, 녹음된 토론 내용을 전사하여 부호화 체계(coding scheme)에 따라 구분하였다. 토론 내용 분석을 위해 주로 많이 사용되는 부호화 체계로는 상호 분석 모델(Interaction analysis model, IAM; Gunawardena, Lowe, & Anderson, 1997)과 Newman, Webb, 그리고 Cochrane(1995)의 부호화 체계가 있다. IAM은 토론 동안의 정보의 공유와 분석, 불일치함의 발견 및 협상과 확증과 수정 단계, 동의와 적용 등으로 구성되어 있다. 그리고 이 부호화 체계는 그룹 토론 내에서 어떻게 지식 조직이 구성화되고 평가되어 활용되는지 알아보는 연구(Marra, Moore, & Klimczak, 2004; Moore & Marra, 2005)와 프로젝트 중심 학습 시 온라인 토론을 분석한 연구(Hou, Chang, & Sung, 2007)에서도 사용되었다. Marra 등(2004) 연구에서는 토론 시 비판적 사고를 분석하기 위하여 Newman 등(1995; Newman, Johnson, Webb, & Cochrane, 1997)의 부호화 체계를 사용하였다. 이 부호화 체계는 적절성, 중요성, 참신성, 지식, 명료성, 이해, 정당화, 비판적 평가, 실용성, 적용성의 지표로 구성되어 있다.

본 연구에 있어서의 토론은 각 평가 기준에 대하여 실험참가자 본인이 점수 부여에 대한 이유를 제시하면서 진행되었다. 이는 보통의 토론에서 문제를 제기하고, 새로운 아이디어를 제시하여 평가하여 문제를 해결해가는 과정을 담거나 갑론을박하며 하나의 결론을 도출하는 토론들과는 다르다. 본 연구에서의 토론은 IAM의 첫 번째 단계에 해당하는 것으로 정보의 공유와 비교 단계에 해당된다(Gunawardena 등, 1997). 이 단계에서는 각 토론 참여자들의 평가기준에 대한 의견들을 듣고, 각자가 부여한 점수와 그 이유에 대해서 비교하고 차이점을 관찰할 수 있는 단계이다. 그러므로 이 단계에서는 Newman 등(1995)의 부호화 체계 가운데 정당화 지표로 분석될 수 있다. 토론 시 동료평가 때 본인이 점수를 부여한 이유에 대하여 정당한 근거를 가지고 발언해야하기 때문이다. Newman 등(1995)의 부호화 체계의 정당화 지표는 적절한 근거와 예시를 통해 해결책 또는 판단을 제시하면 플러스 점수를 부여하고, 부적절하거나 애매한 질문이나 설명 없이 제시한 해결책 또는 판단의 경우에 대해서 마이너스 점수를 부여한다. 이에 본 연구에서 진행한 토론에서는 해결책을 제시하는 과정은 없었기 때문에 Newman 등(1995)의 부호화 체계에서 해결책과 관련된 부분을 제외하고, 근거가 있는 발언을 제시했는지 여부를 부호화 체계로 삼았다.

그리고 Chi(1997)에서 프로토콜 자료 분석의 지침에 따라 코딩할 때의 단위는 하나의 아이디어가 포함되어 있는 단위로 한 두 구절이 포함되는 것으로 구분하였다. Chi(1997)에 따르면 분석 단위가 꼭 완벽한 문장일 필요는 없다. 부호화 체계의 예시는 표 3에 제시하였다. 부호화 체계에 따라 전체 전사 자료의 약 20%에 해당하는 44개의 자료에 대한 코딩을 두 평가자가 실시하고, 각 코딩 점수 간의 일치도인 ICC 상관계수가 높게 관찰되어, 나머지 80%에 대해서는 한명의 연구자가 코딩하였다. 이는 Chi, Kang, 그리고 Yaghmourian(2017)에서도 방대한 양의 프로토콜 분

〈표 3〉 부호화 체계(coding scheme)의 예시

	토론 내용	근거 있음	근거 없음
A	근거를 제시하면서 대학수업과 무크를 비교했는데,	1	
	무크의 단점 설명하고, 대학교육 필요성을 제시하는 것이 더 설득력이 있었을 것이라고 생각한다.	1	
	대학교육이 이런 점이 좋다 이런 점이 좋다를 두 문단에 나눠서 쓰고 있고, 무크와의 비교는 없었다.	1	
	이런 주장에 대한 근거도 객관적인 것이 아니고, 주관적인 것이어서 설득력이 부족했다.	1	
	이상적인 주장이지, 현실적인 주장이 아니어서 좋은 점수를 주기 어려웠다.	1	
	총점	5	0
B	가장 기본적인 글쓰기 형식을 잘 따른 글이라고 생각한다.	1	
	“첫 번째”, “두 번째”라고 제시하여, 내용을 따라가기 쉽게 하여 글의 형식을 잘 갖추었다.	1	
	대학수업들의 장점을 나열하고 있지만	1	
	무크와의 비교언급이 없었다.	1	
	총점	4	0
C	이 글의 형식이 제일 익숙하고, 이해하기 좋았다.		1
	이 글은 읽기에 좋아서 통찰에 높은 점수를 주었다.		1
	총점	0	2
D	근거가 다소 뻔했다.		1
	두 가지 대안을 비교하지 않고	1	
	일방적인 대학수업만 얘기해서 설득적이지 않았다.	1	
	병렬식 구조를 쓰고 글의 형식을 잘 따라서 글을 이해하기 좋았다.	1	
	총점	3	1

석에 서 사용하였던 방법이다. 각 코딩에 대한 ICC 계수는 근거 있는 발언에서 ICC=.92(p<.001)과 근거 없는 발언에서는 ICC=.91(p<.001)이었다.

본 연구에서 1개조의 토론 녹음 자료가 누락되어 15개 조에 대하여 실시되었다. 그리고 각 토론 요소와 평가 정확도 점수 간에 특이한 관측치가 있는지를 검토하기 위하여 외면 스튜던트화 잔차(externally studentized residual)로 비교한 결과, 1개 조의 토론 분석점수가 이상치로 발견되었다 ($rstudent$=2.54, p<.05). 이에 총 14개 조에 대한 토론 분석 결과와 평가 정확도의 분석이 이루어졌

〈표 4〉 토론 요소와 평가 정확도와의 상관관계

	근거 있는 발언 수	근거 없는 발언 수	총 발언 수	근거 발언 수의 비율
평가 타당도	r=.49(p=.08)	r=.19(p=.52)	r=.53*(p=.045)	r=.25(p=.25)
평가 신뢰도	r=.31(p=.44)	r=-.37(p=.19)	r=.06(p=.84)	r=.42(p=.13)

다. 근거 있는 발언 수와 근거 없는 발언 수, 그리고 근거 있는 발언 수와 근거 없는 발언 수를 합한 총 발언 수와 근거 있는 발언 수의 비율에 대하여 상관관계 점수(타당도)와 편차점수(신뢰도)와의 적률 상관 계수를 구하였다. 평가의 신뢰도는 편차점수가 작아질수록 커지는 관계이기 때문에 토론 요소와 편차점수와의 상관관계에서 부적관계로 제시된 것은 신뢰도와의 정적관계로, 정적관계는 신뢰도와의 부적관계로 제시하였다.

그 결과는 표 4에 제시하였다. 분석 결과, 총 발언 수와 평가 타당도에서 통계적으로 유의미한 상관관계를 관찰하였다. 따라서 토론 요소 가운데 총 발언 수는 평가 타당도에 영향을 주었고, 평가 신뢰도에 영향을 준 토론 요소는 발견되지 않았다. 토론에서 총 발언 수가 높은 집단과 낮은 집단으로 구분하여 각 평가 시기별로 평가 타당도 향상에 있어서의 차이가 있는지 살펴보았다. 타당도 결과로 각 집단 간 상관관계 점수의 평균과 표준편차는 표 5에 제시되었다. 총 발언 수가 높은 7개 조와 총 발언 수가 낮은 7개 조에 대한 동료평가의 상관관계 점수를 반복측정 분산분석을 실시했다. Mauchly의 구형성 검정 결과, Mauchly의 W=0.39(p<.01)로 구형성 가정이 만족되지 않아 개체 내 효과 검정에서 Greenhouse-Geisser로 보정한 값을 사용하여 분석하였다. Greenhouse-Geisser 검정 결과, 평가 타당도 점수에 대한 유의미한 차이(F(1.24, 14.90)=5.19, MSE=.32, p=.03, η^2=.30)가 발견되었고, 평가 타당도의 집단 간 상호작용은 유의미하지 않았다(F(1.24, 14.90)=.13, MSE=.13, p=.18, η^2=.14). 3차 평가에서 총 발언 수가 높은 집단이 낮은 집단보다 평가 타당도가 통계적으로 유의미하게 높았다(t(12)=2.42, p=.03, d=1.29).

〈표 5〉 집단 간 전문가 점수와의 상관관계 점수의 평균과 표준편차

집단	1차 평가	2차 평가	3차 평가
총 발언 수 낮은 집단(n=7)	.51(.36)	.43(.38)	.58(.18)
총 발언 수 높은 집단(n=7)	.41(.27)	.49(.33)	.77(.10)
전체(N=14)	.46(.31)	.46(.35)	.67(.17)

논의 2

토론의 활성화 정도 및 토론의 질이 동료평가 점수의 정확도에 영향을 미치는지 알아보기 위해 Newman 등(1995)의 정당화 부호화 체계에 따라 부여한 점수를 바탕으로 14개조의 토론을 분석하였다. 이와 더불어 총 발언 수에 대한 근거 제시 발언 수의 비율도 동료평가 타당도와 신뢰도와의 상관관계를 함께 구하였다. 근거를 많이 제시할수록 토론의 질이 높아지고, 이는 동료평가 정확도에 더 많은 영향을 미칠 것이라고 예측하였다. 따라서 근거 제시 발언 수 또는 총 발언 수에 대한 근거 제시 발언 수의 비율이 높을수록 평가 정확도의 향상이 많이 일어날 것이라고 기대하였다. 그러나 결과는 이와 다르게 근거 제시가 있었던 발언이든, 근거 제시가 없었던 발언이든 발언을 많이 한 조일 높을수록 평가 타당도가 높아지는 것을 확인할 수 있었다.

이는 그룹 토론의 발언수가 높아질수록 그룹의 예측(forecasting)의 정확도가 높아졌던 Mellers, Ungar, Baron, Ramos, Gurcay, Fincher, Moore, Atanasov, Swift, Murray, Stone, 그리고 Tetlock(2014)의 연구 결과와 유사하다. Mellers 등(2014)은 토론에서의 발언 수가 높은 것은 의사소통이 활발했다는 것을 의미하며, 그 만큼 많은 정보를 공유할 수 있었기 때문에 팀의 예측도가 정확해질 수 있었다고 설명한다. 이에 평가나 예측 정확도에 대한 토론에서는 발언 수가 정확도 향상에 큰 영향을 준다고 볼 수 있다. 그리고 후속연구를 통해 토론의 종류에 따라 발언의 수, 발언의 종류 등이 토론의 질에 대해 미치는 영향에 대해 더 논의될 수 있다고 생각한다. 본 연구 결과는 토론의 질은 평가 기준을 더 명확하게 해 주는데 영향을 줄 것이라는 두 번째 가설에 대해서, 토론에서의 발언 수가 평가 기준 정확도 향상에 영향을 주었음을 보여준다.

종합 논의

본 연구는 글에 대한 동료평가의 정확성 향상 방안을 모색하기 위해 수행되었다. 정확성 향상 방안으로써, 동료평가 후 평가 기준에 대해 토론을 진행하는 방안과 평가에 대한 전문가의 의견을 참고하는 방안 두 가지를 탐색하였다. 연구 결과, 동료평가의 정확도는 평가 기준에 대한 토론을 통해 향상될 수 있음을 확인하였다.

토론은 학생들의 비판적, 분석적 사고력과 이해력 증진을 위해서 장려되는 학습 방식 중에 하나이다(Murphy 등, 2016). 학생들은 자기 사고와 추리력에 기반하여 토론을 펼치는데, 이런 토론은 학습을 촉진할 수 있다. 토론과 같은 사회적 상호작용은 또한 개인 추론을 증진시킬 수 있는 주요한 수단이다(De Lisi & Golbeck, 1999). 토론을 통해 학생들은 자신의 관점을 만들도록 격려받게 되고, 동료들의 의견을 통해 다른 관점을 고려할 수 있기 때문이다. 상반된 관점에 대해서는 갈등을 조화롭게 하도록 노력함으로써 활발한 인지과정을 촉발시키기도 한다(Soter, Wilkinson,

Murphy, Rudge, Reninger, & Edwards, 2008).

동료평가 후 평가 기준에 대한 토론 과정을 통해, 학생들은 글쓰기의 평가체계를 깨닫고, 본인의 잘못된 점을 바로잡을 수 있었던 것으로 보인다. 당장 그 평가 준거를 적용시키지는 못했지만, 다른 글을 평가하게 되었을 때는 그 평가 준거들을 내면화시켜 적용할 수 있었다. ICAP 이론에 따라 토론은 상호활동이 활발하게 일어나는 과정(I)으로써, 학습 효과가 큰 활동이다(Chi, 2009; Chi & Wylie, 2014). Chi 등(2017)도 학습 토론은 학습에 대한 동기와 몰입을 높이고, 토론을 통한 개념에 대한 공동구성(coconstructing)이 문제 풀이의 적용력을 향상시켜 준다는 것을 밝힌 바 있다. 토론을 통해 학생들은 자신의 생각을 표현하고 이에 대한 즉각적인 피드백을 받을 수 있기 때문에 자신의 오류를 수정할 수 있는 기회를 가질 수 있다(Reinholz, 2016). 본 연구에서도 3차 평가에서 동료평가와 전문가 점수의 상관관계 점수가 높아지고 동료평가자들 간의 편차점수가 감소했던 것은 토론의 효과로 보인다. 결국 동료평가 후 평가 기준에 대한 토론이 동료평가의 타당도와 동료평가의 신뢰도를 향상시킨 것이라 할 수 있다.

토론하는 과정에서 상대방에게 자신의 의사를 전달하기 위해 자신의 사고과정을 말로 표현하다보면 자신의 잘못된 확증 편향을 깨닫게 도와준다. 이런 언어적 표현은 특정 지식을 습득하는데 효과적이다(Murphy 등, 2016). 본 연구에서의 토론분석에서도 발언 수가 높을수록 동료평가점수와 전문가 점수의 일치도 향상이 일어나서 동료평가 타당도가 높아짐을 확인할 수 있었다. 발언 자체가 자신의 지식과 타인의 지식을 점검하게 하는 기제로 작용하여, 사고 과정을 더 견고하게 한다.

이에 반해 학생들이 전문가의 평을 읽게 한 조건에서는, 글에 대한 평가 기준에 대하여 비판적이고 분석적인 사고가 작용하기 보다는 오히려 그 의견을 수용하는 것으로 보인다. 이 때문에 전문가의 평을 읽은 직후에 같은 글에 대해서 다시 평가를 했을 때는 전문가의 점수와의 일치도가 향상되었으나 전혀 다른 글을 읽을 때는 그 평가 준거의 전이가 전혀 일어나지 않아 동료평가 점수의 정확도가 향상되지 않았다. 이 실험 결과를 ICAP 학습 프레임워크에 비추어 보면, 전문가의 평을 참조했던 집단은 수동적인 모드(P)로 학습했기 때문에 평가 기준에 대한 전이가 일어나지 않았던 것으로 해석할 수 있다. 전문가의 글에 대한 평가 의견을 참고하면, 자신이 평가했던 내용과 어떻게 다른지를 직접 비교할 수 있고, 자신이 놓친 부분을 점검할 수 있을 것이다. 그 결과 본인의 부족했던 평가 준거들에 대해서 전문가의 의견을 통해 재점검하고 체계화시킬 수 있다. 이로 인해 2차 평가에서는 평가의 정확도가 향상되었다. 하지만, 다른 글에 대한 평가로는 전이되지 않았다. 이상의 결과는 전문가의 의견을 정확한 평가의 지표로 삼는 데는 도움이 되지만, 자신의 것으로 내면화하지는 못한다는 것을 시사한다.

본 연구 결과는 글에 대한 평가가 동료평가를 통해 이루어질 수 있는 가능성을 보여준다. 평가 기준에 대한 토론은 학생들이 평가 기준을 이해하고 학습하도록 하고, 그 기준을 실제 평가에 응용할 수 있도록 도울 수 있다. 후속 연구에서는 토론 이외에 다른 방법으로 동료평가의 정

확성을 높이는 방법과 함께, 실제 글쓰기 평가에 대한 토론이 본인의 글쓰기를 향상시킬 수 있는지를 탐색할 필요가 있다. 이러한 동료 평가 연구 결과들은 그 동안 교사의 평가 부담 때문에 자주 사용되지 못하는 글쓰기 수업을 실제 교육 현장에서 확산하는데 기여할 수 있다. 본 연구에서 밝혀진 것처럼, 평가에 대한 토론이 추가되면 그 정확도를 높일 수 있기에 동료평가의 활용이 가속될 것을 기대해본다.

참고문헌

박주용, & 박정애 (2018). 동료평가의 현황과 전망. **인지과학, 29**(2), 85-104.

Cheung-Blunden, V., & Khan, S. R. (2018). A modified peer rating system to recognise rating skill as a learning outcome. *Assessment & Evaluation in Higher Education*, *43*(1), 58-67.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences*, *6*, 271-315.

Chi, M. T. (2009). Active constructive interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*(1), 73-105.

Chi, M. T., Kang, S., & Yaghmourian, D. L. (2017). Why students learn more from dialogue-than monologue-videos: Analyses of peer interactions. *Journal of the Learning Sciences*, *26*(1), 10-50.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219-243.

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science*, *39*(5), 629-643.

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 328-338.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891-901.

De Lisi, R., & Golbeck, S. (1999). *Implications of Piagetian theory for peer learning*. Mahway, NJ: Lawrence Erlbaum.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, *70*(3), 287-322.

Fisher, D., & Frey, N. (2004). *Improving adolescent literacy: Strategies at work.* Upper Saddle River, NJ: Pearson.

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in

computer conferencing. *Journal of Educational Computing Research*, *17*(4), 397-431.

Hayes, J. R., & Flower, L. (1980). Identifying the Organization of Writing Processes. In L. W. Gregg, & E. R. Steinberg (Eds.), *Cognitive Processes in Writing: An Interdisciplinary Approach* (pp. 3-30). Hillsdale, NJ: Lawrence Erlbaum.

Hou, H. T., Chang, K. E., & Sung, Y. T. (2007). An analysis of peer assessment online discussions within a course that uses project-based learning. *Interactive Learning Environments*, *15*(3), 237-251.

Jeffery, D., Yankulov, K., Crerar, A., & Ritchie, K. (2016). How to achieve accurate peer assessment for high value written assignments in a senior undergraduate course. *Assessment & Evaluation in Higher Education*, *41*(1), 127-140.

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: their origin and impact on revision work. *Instructional Science*, *39*(3), 387-406.

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung K. S., & Suen, H. K. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education*, *41*(2), 245-264.

Liu, X., & Li, L. (2014). Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment. *Assessment & Evaluation in Higher Education*, *39*(3), 275-292.

Liu, X., li, L., & Zhang, Z. (2018). Small group discussion as a key component in online assessment training for enhanced student learning in web-based peer assessment. *Assessment & Evaluation in Higher Education*, *3*(2), 207-222.

Marra, R. M., Moore, J. L., & Klimczak, A. K. (2004). Content analysis of online discussion forums: A comparative analysis of protocols. *Educational Technology Research and Development*, *52*(2), 23.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Moore, D,, Atanasov, P,, Swift, S. A., Murray, T., Stone, E., and Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106-1115.

Moore, J. L., & Marra, R. M. (2005). A comparative analysis of online discussion participation protocols. *Journal of Research on Technology in Education*, *38*(2), 191-212.

Murphy, P. K., Firetto, C. M., Wei, L., Li, M., & Croninger, R. M. (2016). What really works: Optimizing classroom discussions to promote comprehension and critical-analytic thinking. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 27-35.

Newman, D. R., Johnson, C., Webb, B., & Cochrane, C. (1997). Evaluating the quality of learning in computer supported co operative learning. *Journal of the American Society for Information science*, *48*(6), 484-495.

Newman, D. R., Webb, B., & Cochrane, C. (1995). A content analysis method to measure critical thinking in face-to-face and computer supported group learning. *Interpersonal Computing and Technology*,

3(2), 56-77.

Park, J. (2017). ClassPrep: A peer review system for class preparation. *British Journal of Educational Technology*, *48*, 511-523.

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: how students respond to peers' texts of varying quality. *Instructional Science*, *43*(5), 591-614.

Reinholz, D. (2016). The assessment cycle: a model for learning through peer assessment. *Assessment & Evaluation in Higher Education*, *41*(2), 301-315.

Rushton, C., Ramsey, P., & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: A case study. *Journal of Computer-Based Instruction*, *20*, 75-80.

Russell J., van Horne, S. V., Ward, A. S., Bettis III, E. A., & Gikonyo, J. (2017). Variability in students' evaluating processes in peer assessment with calibrated peer review. *Journal of Computer Assisted Learning*, *33*, 178-190.

Sluijsmans, D. M. A., Brand-Gruwel, S., & van Merrienboer, J. J. G. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education*, *27*(5), 443-454.

Soter, A. O., Wilkinson, I. A., Murphy, P. K., Rudge, L., Reninger, K., & Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research*, *47*(6), 372-391.

Topping, K. (1998). Peer assessment between students in college and universities. *Review of Educational Research*, *68*(3), 249-276.

Tsai, C. C., & Liang, J. C. (2009). The development of science activities via on-line peer assessment: The role of scientific epistemological views. *Instructional Science*, *37*(3), 293-310.

Van Loon, M. H., Dunlosky, J., Van Gog, T., Van Merrienboer, J. J., & De Bruin, A. B. (2015). Refutations in science texts lead to hypercorrection of misconceptions held with high confidence. *Contemporary Educational Psychology*, *42*, 39-48.

Zheng, L. Cui, P., Li, W., & Huang, R. (2018). Synchronous discussion between assessors and assessees in web-based peer assessment: Impact on writing performance, feedback quality, meta-cognitive awareness and self-efficacy. *Assessment & Evaluation in Higher Education*, 1-15.

1차 원고 접수: 2018. 11. 07
1차 심사 완료: 2019. 07. 12
2차 원고 접수: 2019. 09. 11
2차 심사 완료: 2019. 10. 10
3차 원고 접수: 2019. 10. 11
최종 게재 확정: 2019. 10. 11

(*Abstract*)

Student Discussion or Expert Example? How to Enhance Peer Assessment Accuracy

Jung Ae Park Jooyong Park

Department of Psychology & Institute of Psychological Science

Seoul National University

Writing is an activity known to enhance higher level thinking. It allows the writer to utilize, apply, and actively expand the acquired knowledge. One way to increase writing activity in classroom setting is to use peer assessment. In this study, we sought to increase the accuracy of peer assessment by having students discuss about the scoring rubric or by referring to an expert's assessment. One hundred and fifty college students participated in the experiment. In the group that referred to the expert's assessment, the accuracy of peer assessment increased when the same piece of writing was evaluated; however, no such increase was observed when another piece of writing was assessed. On the other hand, in the group that discussed about the scoring rubric, the accuracy of peer assessment remained the same when the same piece of writing was evaluated, but increased when another piece of writing was assessed. Also, in the discussion group, the accuracy increased in proportion to the number of comments during the discussion. The results suggest that active and voluntary participation of students increase the accuracy of peer assessment.

Key words : peer assessment, writing, discussion, accuracy of peer assessment