인지과학, 제29권 제2호
Korean Journal of Cognitive Science
2018, Vol. 29, No. 2, 85～104.
http://dx.doi.org/10.19066/cogsci.2018.29.2.001

동료평가의 현황과 전망*

박 주 용† 박 정 애
서울대학교 심리학과 & 심리과학 연구소

동료평가란 학생들이 다른 학생의 글, 발표, 집단 활동과 같은 평가 대상이 될 수 있는 모든 결과물이나 수행 과정에 대해 점수를 매기거나 의견을 제공하는 활동을 가리킨다. 동료평가는 채점부담으로 인해 잘 활용되지 않는 글쓰기 기회를 늘릴 수 있을 뿐만 아니라, 평가를 하는 과정을 통해 학습도 향상시킬 수 있다는 장점이 있다. 본 개관 논문에서는, 주요 동료평가 시스템과 동료평가를 사용한 경험적 연구 결과를 소개한 다음, 동료평가 확산의 장애 요인을 살펴보고, 앞으로의 연구에서 해결되어야 할 과제가 논의되었다. 이 논의를 통해 동료평가에 대한 후속 연구가 활성화되고 동료평가가 실제 교육 장면에서 활용되어, 우리나라의 평가와 수업 방법에 큰 변화를 가져오기를 기대해본다.

주제어 : 동료평가, 피드백, 글쓰기, 웹기반 시스템

---

* 본 연구는 서울대학교 사회과학대학 융복합 연구의 지원으로 수행되었음.
† 교신저자: 박주용, 서울대학교 심리학과, 서울특별시 관악로1 서울대학교
연구분야: 인지(학습과 기억, 문제 해결)
Tel: 02-880-9050, E-mail: jooypark@snu.ac.kr

‘구슬이 서 말이라도 꿰어야 보배’라는 속담은 평가에서 학생 개입의 현황을 보여주기에 안성맞춤이다. 구슬을 꿰듯 학생들 간에 서로 평가를 하게 하면 다양한 교육적 이점이 있는데, 아직은 꿰어지지 않은 구슬로 머물고 있기 때문이다. 이 글은 학생들로 하여금 평가의 대상이 되는 동시에 평가의 주체가 되도록 하는 동료평가의 필요성을 제기하고자 한다. 그리고 현재 개발되어 있는 동료평가의 주요 시스템이 왜 아직 널리 사용되지 않고 있는 지에 대한 이유에서 시작하여 지금까지 동료평가에 대해 이루어진 주요 연구 결과를 살펴보고, 앞으로 어떻게 활용될 수 있는지를 탐색해보고자 한다. 이 탐색을 통해 동료평가의 교육적 가치를 부각시켜 궁극적으로 교육 현장에서의 활용을 촉진하고자 한다.

1. 동료평가의 필요성

동료평가란 학생들로 하여금, 다른 학생의 글, 작품, 발표, 집단 활동, 치료나 상담과 같은 전문기술 등처럼, 평가 대상이 될 수 있는 모든 결과물이나 수행 과정에 대해 점수를 매기거나 의견을 제공하도록 하는 활동을 가리킨다. 동료평가는 통상 교육이나 훈련 장면에서 명시적이고 공적으로 이루어진다. 명시적으로 이루어지지 않더라도 여전히 사적인 동료평가는 일어난다. 예를 들면, 어느 한 학생이 교수자에게 질문을 하거나 혹은 교수자의 질문에 대해 답변할 때, 다른 학생들이 보이는 몸짓이나 표정에 평가가 담겨 있다. 우리나라 학생들이 질문을 하지 않는 이유는 이런 암묵적 동료평가로부터의 부담 때문이다. 분명한 점은 동료평가는 다양한 방식으로 자주 일어나는 사건이라는 것이다.

동료평가가 의도적으로 활용될 때, 상황에 따라 그 구현 방식이 여러 가지로 달라질 수 있다. 예를 들어, 서로 짝을 짓거나 아니면 무선적으로 배정한 다음 평가하게 할 수도 있다; 익명으로 할 수도 있고 기명으로 할 수도 있다; 동료평가 결과를 그대로 성적에 반영할 수도 있고, 아니면 참여만 하면 모두 같은 점수를 줄 수도 있다. 동료평가 절차의 이런 다양성과 복잡성 때문에, 동료평가 전반에 대한 종합적 판단을 내리기는 쉽지 않다. 그럼에도 Topping(1998)의 역사적 개관 논문 이래로 발표된 많은 연구 결과들은 (예, Van den Berg, Admiraal, & Pilot, 2006; Van Zundert, Sluijsmans, & van Merriënboer, 2010; Ashenafi, 2017) 수렴적으로 동료평가는 학생들의 추가적인 시간과 노력을 필요로 하지만, 그만큼 그로부터 얻어지는 여러 장점이 있다는 것을 보여준다.

동료평가의 장점은 크게 두 가지로, 하나는 교수자의 평가에 대한 부담을 덜어주는 것이고 다른 하나는 학생들이 평가에 참여함으로써 얻을 수 있는 교육적 효과이다. 우선, 교수자의 평가 부담 측면을 살펴보자. 시험은 많은 경우 선발이나 서열을 매기기 위해 사용되지만, 다른 어떤 방식보다도 뛰어난 학습 활동이기도 하다. 이에 대한 많은 실증적 연구가 있음에도 불구하고, 시험이 교수자의 부담을 가중시키기 때문에 학습도구로 충분히 활용되지 않고 있다.

형성 평가의 중요성과 방법에 대한 연구는 Black과 동료들에 의해 다양하게 이루어졌다(Black & Wiliam, 1998; Black, Harrison, Lee, Marshall, & Wiliam, 2003). 이들의 연구의 출발점은 교수 학습이 상호작용적이어야 한다는 것이다. 이들은 광범위한 문헌 개관을 통해, 수업 장면에서 형성 평가가 적절히 활용되면, 전통적인 교사 중심의 수업에 비해 학습 효과가 크다는 것을 확인하였다(효과크기 0.3~0.7). 이 발견을 바탕으로 형성 평가를 위한 네 가지 방법 즉, 보다 효과적으로 질문하기, 피드백 제공, 학습 목표와 기준을 학생들과 공유하기, 그리고 자기 평가와 동료평가, 각각의 효과를 확인하는 동시에 이들을 수업 장면에 적용할 수 있는 구체적인 방법들을 제시하였다. 예를 들면, 교수자가 던지는 질문이 어느 정도 어려워야 하고 학생들로 하여금 질문에 대한 답을 생각할 수 있도록 충분한 시간을 주어야 한다는 것이다. 피드백의 경우, 비교가 목적이 아니라 수행에 대해 잘 한 부분을 인정해주는 동시에 개선에 도움이 되는 조언을 제공해야 한다는 것이다.

Roediger, Putnam 그리고 Smith(2011)도 형성평가를 포함하여 시험을 통한 학습의 이점을 다음과 같은 10가지로 정리하였다. 1) 시험은 나중의 인출을 향상시키며, 2) 무엇을 모르는지 알게 해주며, 3) 다음 번 학습 시 더 잘 배울 수 있게 하며, 4) 지식을 더 잘 구조화시키며, 5) 새로운 맥락에로의 전이를 촉진하며, 6) 시험을 보지 않은 내용에 대한 인출 가능성을 높이고, 7) 상위 인지적 점검을 향상시키고, 8) 새로운 내용을 학습할 때 이전에 배운 내용으로부터의 간섭을 차단하며, 9) 가르치는 사람에게 피드백을 제공하고, 10) 학생들로 하여금 공부하도록 한다는 것이다. 이상의 10가지와 더불어, 한 가지를 더 추가할 수 있다. 그것은 시험 문항이 단지 지식을 묻는 것이 아니라 문제해결과 같은 고차적 사고 능력과 관련되면 성취도를 향상시킬 뿐만 아니라 학습 동기를 높인다는 것이다(Wenglinsky, 2002). Wenglinsky(2002)는 8학년을 대상으로 실시된 미국 국가학업 성취평가(NAEP)결과를 분석하였다. 이 분석에서 그는 학생들의 성취도를 종속변인으로 하고, 교사의 수학 관련 학위여부와 교육 연한, 교수 경력, 교원 연수에 참여정도, 그리고 교실 수업 활동 각각의 여러 하위 요소들을 변인으로 하는 경로분석을 실시하였다. 그 결과 학생들의 성취 수준과 높게 연관된 5개의 변인이 발견되었는데, 이들은 교사의 수학 전공 학위, 고차 사고능력 개발 연수에의 참여, 다양한 교원 연수에의 참여, 교구 활용, 그리고 고차 사고 기술을 사용하게 하는 수업 활동이었다. 여기서 고차적 사고란, 배운 내용을 새로운 문제에 적용하도록 하거나 배운 개념을 구체화시키거나 예를 찾아내는 활동과 연관이 있다.

이처럼 평가가 다양한 방식으로 학습에 도움이 될 수 있다면, 평가가 교육 장면에서 더 많이 활성화되어야 할 것이다. 이럴 경우 바로 문제가 되는 것은 채점과 피드백으로 인한 교수자의 부담이다. 이 부담은 고차적 사고를 요하는 문항을 만들기에 더 적합한 논술 시험일 경우 급격하게 커진다. 이런 부담을 해결하는 방안으로 Black 등은 물론 많은 연구자들이 동료평가를 추천한다(예, Black et al., 2003; Sadler & Good, 2006). 형성 평가의 중요성이 대두되기 전부터 동료평가가 고려되어 온 영역은 글쓰기 혹은 작문 교육이다. 글쓰기 수업은 학생들로 하여금 글을 쓰

게 하고 이에 대해 첨삭을 하는 방식으로 진행된다. 대개 소수의 학생을 대상으로 수업이 이루어지는 만큼 비용이 많이 드는 수업이다. 이런 수업에서 교수자의 부담을 덜기 위한 방법으로 동료평가가 고려되어 왔다. 글쓰기에서 동료평가가 활용되면, 설사 다른 동료들의 글에 대한 학생들의 평가가 정확하지 않다하더라도, 일단 글을 쓰도록 하는 기회를 늘릴 수 있다는 장점이 있다. 글쓰기는 글을 써야 느는데, 동료평가는 그런 기회가 많아지게 할 수 있다. 다행히도 여러 명이 평가할 경우 전문가의 평가에 준한다는 연구 결과가 많다(Cho, Schunn, & Wilson, 2006; Jeffery, Yankulov, Crerar, & Ritchie, 2016). 우리가 시험이나 과제 수행을 잘 해내지 못하더라도 일반적으로 평가는 상대적으로 잘 한다는 사실을 고려하면 놀라운 일이 아니다. 여기에 평가가 신속하게 이루어져 평가 결과의 활용도를 높일 수 있다는 점도 동료평가의 매력 중 하나이다.

동료평가의 또 다른 장점은 이 활동이 학생들에게 여러 교육적 효과를 가져다 줄 수 있다는 것이다. Sadler와 Good(2006)은 이를 교수적(pedagogic), 상위인지적(metacognitive), 그리고 정서적(affective) 장점으로 세분하였다. 교수적 장점은 다른 동료들의 글을 보면서 다각도로 생각해 볼 기회를 갖게 된다는 점이다. 상위인지적 장점은, 채점 과정을 통해 자신의 장점과 단점을 파악할 수 있게 되며, 또한 타인은 물론 스스로에 대해 평가하는 능력을 향상시킨다는 점을 가리킨다. 정서적 장점은 서로 협조적이고 우호적인 학급 분위기를 만들고, 평가에 대해 긍정적인 태도를 갖도록 도울 가능성을 가리킨다.

동료평가는 또한 학생들의 학습과 사고를 다양한 방식으로 촉진할 수 있다. 즉, 동료평가 과정은 평가 기준을 이해하고 동료 글에서의 문제를 파악하는 사고의 과정이 포함되어 있기 때문에 학생들은 동료평가 과정을 통하여 전반적인 평가에 대한 통찰력을 향상시킬 수 있다. 그리고 점수만 부여할 때 보다 구체적인 피드백을 제공하도록 하면 이 과정에서 발생하는 갈등을 통해 비판적 사고를 함양할 수 있다(Russell, van Horne, Ward, Bettis III, & Gikonyo, 2017). Tsai와 Chuang (2013)도 동료평가를 통해, 학습자들의 비판적 사고력을 향상시킬 수 있고, 초고를 수정하여 글을 다듬으려는 동기가 높아진다는 결과를 보고하였다. 따라서 교사는 동료평가를 통해 학생들로 하여금 서로의 차이점들을 이해하고 비판적 사고를 발전시키도록 지도할 수 있다. 이런 결과는 다른 연구에서도 반복적으로 확인되었다(Joordens, Pare & Pruesse, 2009; Çiftçi & Koçoğlu, 2012; Xiao & Lucking, 2008).

동료평가의 또 다른 교육적 이점은 학생들의 고차적 사고를 향상시킨다는 점이다. 동료의 글을 읽고 장단점에 대해 코멘트를 하고 채점을 하는 활동 자체가 고차적 사고이다. 사고를 기본적 사고와 고차적 사고로 나누는 것 자체가 불가능하다는 연구자도 있지만(예, Resnick 1986), 이 글에서는 편의상 Bloom의 교육 목표학에서의 구분을 활용하고자 한다. Bloom, Engelhart, Furst, Hill, & Krathwohl (1956)에서는 사고를 지식, 이해, 적용, 분석, 통합, 평가로 구분하였는데, 이 중 적용, 분석, 통합, 평가를 고차적 사고라 할 수 있다. Bloom 등 (1956)은 이 구분을 위계적으로 설명하지 않았음에도 후속 연구자들은 이를 위계적으로 해석하여, 고차적 사고활동에 앞서 지식

과 이해가 선행되어져야 한다고 오도하였다. 그 결과 사고의 여러 하위 활동이 서로 균형을 이루며 발전해야 함에도 불구하고, 지식과 이해를 지나치게 강조하여 고차적 사고를 위한 교육이 충분히 이루어지지 않았다. 동료평가는 이런 잘못된 관행을 변화시킬 수 있는 한 방법으로 활용될 수 있다.

2. 주요 웹기반 동료평가 시스템

위에서 언급된 여러 동료평가의 장점을 검증하는 한편 실제 교육 장면에 적용하기 위해 다양한 웹 기반 동료평가 시스템이 개발되었다. 이러한 웹 기반 동료평가의 장점은 시, 공간의 제약이 없으며, 제출, 저장, 분배, 인출, 그리고 평가가 하나의 플랫폼에서 이루어지고 익명으로 무선 배분이 가능하다는 것이다. 또한 다양한 시청각 자료의 제공이 가능하고, 상호작용을 통한 학습 기능을 담당할 수 있다(Liu, Li, & Zhang, 2018). 이 절에서는 주요 시스템에 대한 간략한 개관을 통해 웹기반 동료평가의 적용 현황을 알아보고자 한다.

컴퓨터 공학자에 의해 개발된 Aropä는 대규모 수업에서의 학생 활동을 보조하기 위한 목적으로 만들어졌다(Aropä 시스템의 특징과 사용 현황에 대한 상세한 자료는 개발자들이 직접 쓴 Purchase와 Hamer (2018)의 논문에서 볼 수 있다). 이 시스템은 2009년에 개발되어 사용되었고 사용자의 반응을 통해 그 기능이 지속적으로 확장되고 있다. 이미 사용하고 있는 사람들의 추천을 받으면, 누구나 이 시스템을 무료로 사용할 수 있다. 처음에는 컴퓨터 과학 수업 과제를 위해 개발되었지만 다양한 기능이 추가되어, 현재에는 영문학, 의학, 약학, 상법 등에서 폭넓게 사용되고 있다.

또 다른 시스템은 UCLA에서 개발된 조정된 동료평가 시스템(Calibrated Peer Review system: CPR)이다. CPR은 1990년대에 주로 과학 관련 과목을 중심으로 개발되어 사용되어졌지만 지금은 모든 과목에 적용될 수 있고, 현재 300개 이상의 고등학교, 대학교, 그리고 대학원 교육 장면에서 사용되고 있다. 주로 학생들에게 글쓰기 기회를 더 많이 제공함으로써, 교과내용에 대한 이해를 향상시키기 위해 사용된다. CPR의 가장 큰 특징은 학생들의 개인차가 채점 점수에 반영된다는 점이다. 이 시스템의 작동 방식은 다음과 같다. 학생들은 학습 내용을 공부한 다음 과제 글을 써서 제출한다. 그 다음에 평가 능력 측정을 위해, 3개의 샘플 글을 채점한다. 각 샘플 글에 대해, 예를 들면 첫 문장에서 전체 글의 주제가 정확히 소개되었는지, 글에 핵심 내용이 포함되어져 있는지 등에 대해 예/아니오로 답한 다음, 점수를 매기도록 한다. 예/아니오 답의 결과와 전문가가 채점한 샘플 점수의 편차를 이용하여 각 학생마다 채점의 정확도가 계산되고 이 정확도에 따라 채점에서의 가중치가 결정된다. 정확도와의 편차가 큰 학생들은 편차가 작은 학생들에 비해 상대적으로 낮은 가중치가 주어진다. 이 시스템은 현재 http://cpr.molsci.ucla.edu/Home.aspx 에서 상용화된 서비스를 제공하고 있다.

피츠버그 대학에서 개발되어 상용화된 Peerceptive 시스템도 빼놓을 수 없다. 이 시스템은 처음에 sword system 이라는 이름으로 개발되어 연구가 이루어졌는데, 교과내용에 대한 이해보다는 학생들의 글쓰기 능력 향상에 초점을 둔다. 제시된 주제에 대해 초고를 쓰고 다른 학생의 글을 평가한 후 이를 바탕으로 수정하는 절차가 추가되었다는 점이 차별적이다. 초고에 대한 동료평가 결과를 확인하고, 다른 학생들이 지적한 문제들을 반영하여 수정한 후 최종적으로 과제를 제출하도록 한다. 이 시스템은 현재 다음 사이트에서 상용화된 서비스를 제공하고 있다: http://www.peerceptiv.com/. 지금까지 소개된 시스템들은 상용화되거나 비교적 활발한 연구가 이루어져 학술지를 통해 보고된 것들이다. 이밖에도 개별 대학 혹은 연구자 수준에서 개발, 연구되는 시스템들은 훨씬 더 많을 것으로 추측된다.

위에서 본 시스템들은 일단 수업시간에 새로운 내용을 배운 다음, 이를 활용하여 상대적으로 오래 걸리는 글쓰기나 프로그래밍과 같은 과제를 수행하도록 하는데 사용된다. 동료평가를 이용하면, 배운 내용을 활용할 수 있는 기회가 많아지고, 일단 제출한 과제에 대해 빠른 피드백을 받을 수 있게 된다.

국내에서 개발된 Classprep 시스템은 기본적인 작동 방식은 Aropä나 Peerceptive와 같은 동료평가 시스템과 크게 다르지 않지만, 그 사용 방식에서 차별화된다. Classprep은 수업을 위한 예습 도구로 개발되었다(배수정 & 박주용, 2016; Park, 2017). 학생들은 수업에 앞서 주어진 학습 자료로 스스로 공부하고, 수업에서 다루어질 가장 중요한 내용을 다룬 한 개의 질문에 대해 한 페이지 글을 써야 한다. 글의 길이를 한 페이지 정도로 줄인 것은 매주 수업시간마다 글을 쓰는 동시에 동료평가를 해야 하는 부담을 줄이기 위해서였다. 글쓰기 과제와 함께 학생들은 스스로 공부하고 글을 쓰면서 생기게 된 질문들을 올리도록 하였다. 가능하면 많은 질문을 올릴 수 있도록, “내용 이해를 위한 질문”, “동료들과 토론을 위한 질문”, 그리고 “누구도 할 것 같지 않은 창의적 혹은 엉뚱한 질문” 이 세 가지 질문의 범주를 제시하였다. Classprep은 동료평가가 단지 평가도구에서 끝나는 것이 아니라 깊이 있는 예습을 하도록 하는 도구로 그 활용가능성을 확장한 점에서 중요하다.

구체적으로 이 시스템이 수업에 적용되는 절차는 다음과 같다. 학생들은 전용 사이트에서 회원가입 후 자신의 수업에 등록을 하면, 해당 수업 기간 동안 교수가 제시한 여러 과제를 수행할 수 있다. 학생들은 매주 수업 시간 전에 다음과 같은 일련의 예습 활동을 수행한다. 먼저 (그림 1)에 제시된 화면을 통해 수업 이틀 전까지 미리 제공된 학습 자료를 공부한 다음 과제에 대해 글을 쓰고 질문을 올려야 한다. 정해진 시간 내에 제출된 글은 무선적으로 뒤섞인 다음, 글을 쓴 학생 한 명 한 명에게 각각 4개의 글이 보내진다. (그림 2)에서처럼 학생들은 누가 쓴 글인지 모르는 상태에서 보내진 글을 하나씩 기준에 따라 채점하고 피드백을 제시한다. 학생들의 피드백과 채점 결과는 다시 글을 쓴 학생에게 보내지고, 글을 쓴 학생은 다른 학생들의 피드백이 얼마나 도움이 되었는지를 평가한다.

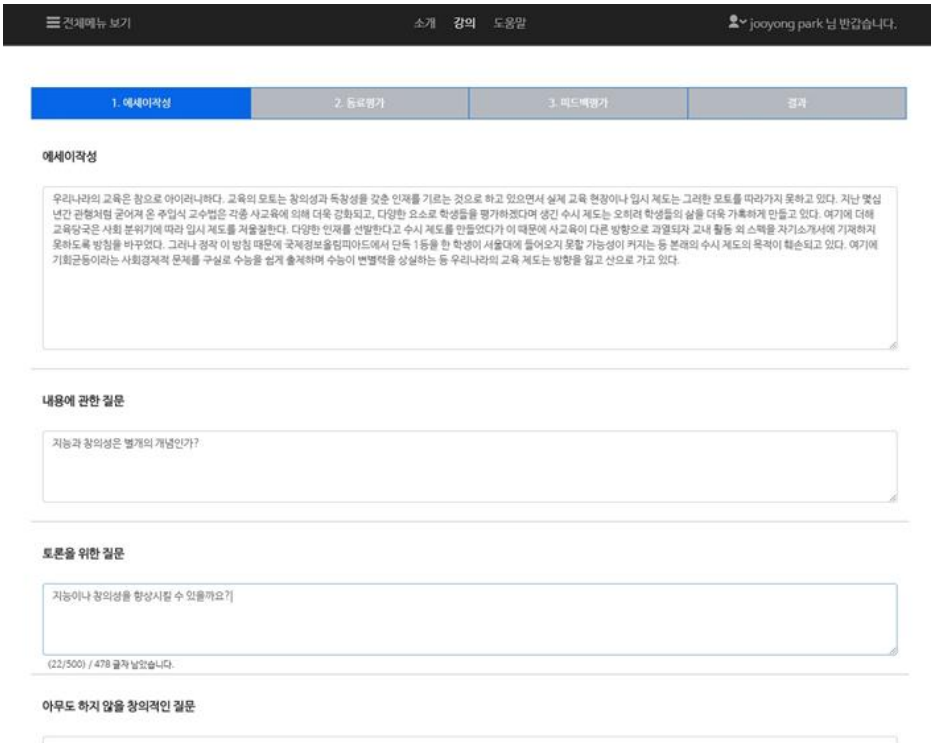

그림 1. classprep에서 글쓰기와 질문 과제 제출 화면

그림 2. Classprep에서 동료평가 화면

학생들이 글을 평가하는 동안 교수는 학생들이 올린 질문을 정리한다. 학생 수가 적어 질문이 많지 않을 때는 비슷한 내용끼리 묶은 다음 수업 자료로 활용할 수 있다. 하지만 30명 이상이 되면 질문의 수가 100개 이상이 되기 때문에, 질문들을 훑어보면서 수업시간에 다룰 수 있을 정도의 질문을 선별한다. 이렇게 정리된 질문은 수업시간에 교수가 짧은 강의를 제공할 때 참고하거나, 아니면 그 질문에 답을 할 수 있는 학생이 설명하게 하거나, 경우에 따라서는 학생들 간의 토론거리로 활용할 수도 있다. 교수는 Classprep 시스템을 통해 학생들의 글은 물론 일련의 활동을 실시간으로 확인할 수 있다.

동료평가를 예습 도구로 활용하는 방안은 최근 국내에서 확산되고 있는 역전학습(flipped learning)의 한계를 극복하는데 도움이 될 수 있다. 역전학습에서는 동영상을 미리 보고 오거나 수업자료를 미리 공부하고 수업에 들어오도록 한다. 수업시간은, 강의 대신 문제풀이와 토론으로 대체된다. 이렇게 하면 수업시간을 보다 생산적으로 사용할 수 있다는 점은 분명하다. 그런데 이런 방식의 한 문제는 미리 동영상이나 수업자료를 이용하여 예습을 해 오지 않은 학생들이 있다는 점이다. 예습을 했는지를 확인하려면 퀴즈를 보는 것과 같은 별도의 조치가 필요하다. Classprep을 사용하면 이런 문제가 발생하지 않는다. 과제를 수행했는지의 여부를 바로 확인할 수 있기 때문이다. 그리고 글쓰기 과제는 다른 어떤 과제 못지않게 학생들로 하여금 생각을 깊게 하도록 만들기 때문에 후속 수업에서 토론의 질을 높일 수 있다. 요컨대 Classprep은 단지 글쓰기나 과제 수행의 도구로서 뿐만 아니라 수업 방식과 내용을 바꾸도록 하는데 기여할 수 있다.

3. 동료평가에 대한 주요 경험적 연구 결과

위에서 동료평가의 이점을 교수의 채점 부담 감소와 학생들의 채점을 통한 다양한 교육적 기능으로 나누어 살펴보았다. 이중 교수의 채점 부담을 실질적으로 덜려면 반드시 충족되어야 하는 선결 요구사항이 있다. 그것은 측정이론에서 중요시되는 신뢰도와 타당도이다. 동료 평가의 많은 연구들에서 신뢰도와 타당도를 측정하고, 이를 향상시키는 방법들이 논의되고 있다. 이들 연구에서 통상 신뢰도는 동료 평가자들 간 일치도로 보고 있으며, 타당도는 동료 평가자들 간의 점수와 전문가 점수와의 일치도로 보고 있다 (Jeffery et al., 2016; Johnson, Penny, Gordon, Shumate & Fisher, 2005; Li, Xiong, Zang, Kornhaber, Lyu, Chung, & Suen, 2016). 따라서 신뢰도의 경우, 한 개의 글을 여러 명이 채점한 결과, 점수 차이가 적을수록 신뢰도가 높아진다. 타당도에 대한 연구는 동료 평가 점수와 전문가 점수와의 차이가 거의 없다는 것에서 큰 차이를 보인다는 것까지 다양한 결과를 보이고 있고, 이들을 통합하기 위한 메타분석이 꾸준히 이루어지고 있다 (Topping, 1998; Falchikov & Goldfinch, 2000; Van Zunder et al., 2010; Purchase & Hamer, 2018; Li et al., 2016). 메타분석 결과, 동료평가 점수와 교수의 채점 결과 간에는 0.6 이상의 비교적 높은 상

관이 있다는 것이 밝혀졌다. Jeffery 등 (2016)은 대학교 4학년생을 대상으로 한 실제 수업에서의 동료평가 결과를 이용하여, 동료평가의 정확성에 영향을 주는 요인들에 대한 회귀분석을 실시하였다. 분석 결과, 평가자 수가 큰 영향을 주고 평가자의 과제 수행 점수도 약간의 영향을 준다는 것을 발견하였다. 3명이상이면 차이가 있기는 하지만 전문가에 준하는 결과를 얻을 수 있고, 4, 5명이면 더 근접하는데, 이 회귀직선을 실제로는 그렇게 하지 않았지만 6명이 평가를 했다고 가정하고 확장시키면, 전문가와의 차이를 없앨 수 있을 것으로 추정하였다. 이 추정은 실증적으로 검증될 필요가 있지만, 4~6명 정도면 전문가의 점수와 크게 다르지 않다고 볼 수 있겠다.

그렇다면 신뢰도와 타당도를 높이기 위해서는 어떻게 해야 할까? 하나는 보다 상세한 채점기준 혹은 루브릭(rubric)을 만들어 제공하는 것이다. 루브릭이란, 어떤 과정이나 결과에 대한 수행수준을 체계적으로 기술하는 목록으로 각 수준에 대해 점수가 부여되는 채점기준을 가리킨다. 루브릭을 만들기 위해서는 학습 목표(learning targets)와 이 학습 목표에 도달한 상태, 그리고 시작점에서 목표에 도달한 상태에 이르는 일련의 단계 구분이 필요하다. 학습 목표를 하나로 묶는지 아니면 몇 개로 세분하는 지에 따라, 총체적 루브릭과 분석적 루브릭으로 나눌 수 있다. 전자는 사용이 쉽다는 장점이 있지만 미묘한 차이를 포착하기 어렵다는 문제가 있고 후자는 사용하기는 복잡하지만 보다 구체적인 평가 정보를 얻을 수 있다는 장점이 있다. 어떤 방식을 사용하든 루브릭은 교수자에게는 물론 학생에게 유용한 정보를 제공한다. 교수자에게는 가르치는 내용과 평가를 일치시키도록 하고, 학생에게는 무엇을 위해 공부하는 지를 이해하게 하고 그 구체적인 모습을 확인할 수 있게 한다.

루브릭은 선다형 시험을 중심으로 한 평가방식의 대안으로 수행평가가 널리 사용되면서 이에 따른 채점부담을 완화하기 위해 도입되었다. 그렇지만 형성평가의 한 방법으로 자기평가가 중요해지면서, 중등 교육에서는 물론 대학 교육에서도 루브릭이 널리 쓰이고 있다. 에세이 평가의 경우 아이디어와 내용, 글의 구성, 논조 (voice), 단어 선정, 문장의 가독성, 그리고 형식 준수 등이 고려된다. 수학의 경우 수학적 지식, 문제 해결, 그리고 수학적 의사소통 등이 고려될 수 있는데, 과목에 따라서는 논리성 (logical reasoning), 명료성 (clarity), 그리고 참신성 (novelty)으로 구분되기도 한다.

과학의 경우 세부 학문 분야에 따라 다를 수밖에 없는데, 물리학의 경우 Rutgers 대학에서 Etkina와 동료들이 공동으로 개발한 루브릭이 있다 (Etkina, Karelina, Ruibal-Villasenor, Rosengrant, Jordan, & Hmelo-Silver, 2010). 이 루브릭은 정보를 다양한 방식으로 표상하는 능력, 관찰실험을 설계하고 수행하는 능력, 가설/설명 혹은 수학적 관계를 검증하기 위한 실험을 설계하고 수행하는 능력, 응용 실험을 설계하고 수행하는 능력, 과학적 개념을 전달하는 능력, 실험 결과를 모으고 분석할 수 있는 능력, 그리고 모델, 등식, 해법 그리고 주장을 평가할 수 있는 능력을 포함한다. 각 능력은 다시 4~8개의 하위 능력으로 구분되고 각 하위 능력은 결측, 미흡함, 개선요망, 그리고 충분함의 4점 척도 상에서 평가된다. 평가되는 하위 능력의 개수가 많기 때문에 한꺼번

에 이들 모두를 사용하면 채점 부담이 커지므로, 과제의 속성에 따라 필요한 부분을 2~3개만 사용해야 한다.

이런 루브릭을 사용하여 평가하게 하였을 때 전문가와의 일치도가 높다는 실증적 연구들은 쉽게 찾아볼 수 있다. 대학생의 구두 발표에 대한 동료평가를 분석한 Hafner와 Hafner (2003)의 연구와 고등학생을 대상으로 한 고급 영어 수업에서 동료의 글을 평가하도록 하고 그 정확성을 확인한 Schunn, Godley 그리고 De Martino (2016) 연구가 대표적인 예이다. 경우에 따라서는 학생들과 함께 평가 전에 채점기준을 만드는 것도 가능하고, 연구 목적에 따라 새로운 루브릭이 필요할 경우 고려해야 하는 변인들은 Dawson (2017)에서 찾아볼 수 있다.

동료평가의 정확도를 높이는데 또 다른 방법은 동료평가를 연습시키는 것이다. 이 때 채점된 사례를 제시하는 것이 좋은데, 사람들은 추상적 원리보다 구체적인 사례가 제시될 때 이해를 더 잘하기 때문이다. 실제로 이현정 (2017)은 사례를 주고 채점을 했을 때 그렇지 않았을 때보다 전문가가 채점한 결과와 더 높은 상관을 보인다는 것을 확인하였다. Liu와 Li (2014) 또한 78명의 학부생들을 대상으로 한 연구에서 동료평가 기술에 있어서 평가 훈련의 영향이 있음을 보고하였다. 훈련 동안에는 학습 개념, 점수 기준과 점수 기준에 의한 프로젝트 예시에 대한 동료와 교사의 평가를 비교한 자료를 제시하였다. 그리고 동료평가 훈련 전과 후의 각각의 학생들의 프로젝트 결과물과 예시로 제공한 프로젝트에 대한 점수와 동료평가자들의 피드백을 분석하였다. 분석 결과, 훈련을 통해 학생과 교수 간의 예시로 제공된 프로젝트에 대한 평가 점수가 일치하게 되었고, 동료평가의 피드백에 기반하여 학생들의 프로젝트 개선에도 큰 도움이 된다는 것을 확인하였다. Sluijsmans, Brand-Gruwel, & van Merriënboer (2002)의 연구는 동료평가 훈련을 실시한 실험 집단과 훈련을 하지 않은 통제 집단 간의 비교 분석을 통하여 실험 집단이 통제 집단 보다 동료평가 기술에 있어서 능가함을 보여주었고, 마지막 과제물에서도 더 우수한 점수를 받았음을 보고하였다. 동료평가 훈련에는 평가 기준과 피드백 주기, 글쓰기 평가 보고에 관련된 내용이 포함되었다. 따라서 루브릭과 함께 몇 개의 샘플을 제시를 통한 동료평가의 훈련은 동료평가의 정확성을 높일 수 있다.

동료평가 훈련은 가르치는 것보다 직접 해보고 의견을 나누는 방식으로 진행될 때 더 효과적이라는 결과도 있다. Zheng, Cui, Li 그리고 Huang (2018)는 토론을 통한 정확성 향상 방안에 대한 글을 쓰고 동료평가를 한 다음 서로의 평가에 대해 토론을 하게 하면 그렇지 않을 때보다, 글쓰기와 피드백의 질이 개선됨을 발견하였다. 토론은 또한 글쓰기에 대한 상위 인지적 자각과 자기효능감도 유의미하게 향상되도록 하였다. 만일 평가가 한 번으로 끝나는 경우가 아니라 여러 번 지속되는 경우에는, 학생들의 평가에 대해 교수가 피드백을 제공함으로써 평가 능력을 향상시킬 수 있다는 점이 추가될 필요가 있다. 모든 학생에게 피드백을 제공하는 것이 바람직하지만 이렇게 하려면 교수의 부담이 너무 커질 수 있다. 이 부담을 줄이는 한 방법은 다른 학생들과의 편차가 큰 학생들부터 선별적으로 몇 명만에게라도 피드백을 제공하는 것이다. 피드백을

제공하는 방식도 여러 가지인데, 잘 된 부분을 두 가지 언급하고 개선할 부분을 한 가지만 언급하는 방법을 고려해 볼 수 있다. 피드백 제공 방식의 기본 방향은 변화를 일으킬 수 있는 일종의 대화여야 한다는 점인데, 보다 구체적인 피드백 제시 방법은 후속 연구를 통해 밝혀져야 겠다.

4. 동료평가의 확산을 막는 요인들

이미 쉽게 사용할 수 있는 웹 기반 동료평가 시스템이 있고 그 장점도 어느 정도 확인되었음에도 불구하고, 실제 교육 장면에서 그 활용도는 낮다. 가장 큰 이유는 학생들은 물론 교수자들도 이 방법에 대해 그리 호의적이지 않기 때문이다. 학생들은 이미 기존의 시험과 과제 수행에 익숙한 상태인데, 여기에 추가적으로 동료 평가를 하게 하면 그리 좋아하지 않는다. 국내에서는 이와 관련된 연구가 없어 미국의 경우를 살펴보면, 현재 대학생들은 과도한 부모의 보살핌 속에서 많은 것을 누리면서도 예전보다 훨씬 더 불안해하는 세대로 인식되고 있다. 이에 동료 평가는 높아진 '고객으로서의 대학생이 누릴 수 있는 특권'을 침해하는 활동으로 받아들여질 수 있다(예, Singleton-Jackson, Jackson, & Reinhardt, 2010). 이는 한국 대학생들의 실정과도 크게 다르지 않은 것 같다. 시험이나 과제만으로도 버거운데, 교수자들이 당연히 해야 할 평가를 자신에게 떠넘겨지는 것으로 여길 수 있기 때문이다. 실제로 대학을 중심으로 하여 동료평가가 점차 활발히 도입되고 있는 유럽과 호주 등의 여러 연구 결과에서, 동료평가에 대한 부정적 태도는 반복적으로 확인되고 있다. Yucel, Bird, Young 그리고 Blanksby (2014)은 대학 1학년 학생들을 대상으로, 채점기준을 제공하고 예시 글을 평가하도록 한 다음, 서로 토론을 하게 하였다. 그 다음에 실제 과제 글에 대해 서로 채점을 하도록 하고, 이 경험에 대한 학생들의 반응을 수집하였다. 그 결과 65% 정도만이 이런 훈련이 도움이 되었다고 긍정적으로 반응하였다. 나머지 학생들은 동료들의 피드백에 대해 불만스러운 반응을 보였다.

학생보다 더 큰 거부감을 보이는 것은 교수들이다. 만일 교수들이 거부감을 보이지 않았다면, 설사 학생들이 싫어하더라도, 교육 효과를 위해 이 방법을 썼을 가능성이 높다. 교수들은 왜 거부감을 보일까? 사실 교수들은 동료평가라 거부감을 보이기보다는 거의 모든 새로운 교수법에 대해 기본적으로 거부감을 보인다. 가장 오래된 방법인 강의법이, 학습에 크게 도움이 되지 않는다는 많은 실증적 연구에도 불구하고 (Bligh, 1972), 여전히 가장 널리 사용되는 이유가 바로 이 거부감 때문일 수 있다. 여기에 동료평가가 얼마나 정확한 지에 대해 충분히 신뢰하지 못한다.

동료평가에 대한 이 같은 거부감은 서울대학교 자연과학 대학 교수(N=102)와 학생(N=276)을 대상으로 한 한 설문결과를 통해서 간접적으로나마 확인할 수 있다 (노정혜, 전헌수, 김상현, 이철범, 이준호, 허영숙, 홍성욱, 박주용, 곽지훈, & 윤창규, 2016). 이 설문에는 학생과 교수로 하여

〈표 1〉 자연대 교수들과 학생들의 설문 조사 결과

	질문하기		상호교수		발표코멘트		보고서코멘트		과제평가	
	중요성	실현 가능성	중요성	실현 가능성	중요성	실현 가능성	중요성	실현 가능성	중요성	실현 가능성
교수	75	61	56	40	65	57	51	51	40	34
학생	78	74*	65**	45	67	58	63**	55	52**	52**

금 교육적 역할을 담당하게 하는 것이 얼마나 중요하다고 생각하고 실제로 그런 역할을 담당하게 하는 것이 실현 가능한지를 각각 100점 척도에서 평정하는 항목이 포함되었다. 각 학생들과 교수들의 평정 점수는 <표 1>과 같다. 중요성과 실현가능성의 교수 점수는 서로에게 질문하기와 발표에 대해 코멘트하기에 대해서는 각각 75점, 61점 그리고 65점, 57점으로 비교적 높았지만, 상호 교수(서로 가르치기), 보고서 코멘트, 그리고 과제평가는 각각 56점과 40점, 51점과 51점, 그리고 40점과 34점으로 상대적으로 낮은 점수를 주었다. 이 결과는 질문하기나 발표 코멘트에 대한 중요성에 대해서는 인식하고 있으나, 상호 교수, 보고서코멘트, 과제 평가에 대한 중요성 인식이 상대적으로 떨어진다는 것을 보여준다. 학생들의 반응은 교수들의 반응과 전반적으로는 비슷한 패턴이지만 평점 점수는 교수들보다 학생들이 전반적으로 높았다. 상호 교수, 보고서 코멘트, 그리고 과제 평가의 중요성과 질문하기와 과제평가의 실현 가능성 항목에서는 학생들이 교수보다 유의미하게 높게 반응하였다. 즉, 학생들은 이와 같은 항목에서 교수보다 더 중요하고 실현 가능성이 있다고 판단하고 있다는 것이다.

비록 한 단과대학에서의 설문결과이지만 다른 대학의 경우에도 크게 다르지 않을 것이라고 추측해 볼 수 있다. 실제 수업 장면에서 상호 교수나 질문하기 그리고 과제평가가 거의 일어나지 않고 있기 때문이다. 그리고 교수는 물론 학생들도 동료평가에 대해 그리 중요하다고 그리고 가능하다고 생각하지 않고 있어, 이러한 상황이 변할 가능성은 희박해 보인다.

5. 동료평가에 대한 거부감을 극복해야 하는 이유

이런 무관심이나 거부감에도 불구하고 학생들로 하여금 평가하게 하는 기회를 늘리고 그 정확성을 향상시키기 위한 노력이 필요한데, 그 이유 중 세 가지만 언급하면 다음과 같다. 먼저, 교육의 한 목적이 자기주도적 학습자가 되도록 하는 것인데, 자기주도성을 발달시키는데 평가가 절대적으로 중요한 역할을 담당하기 때문이다. 평가에 참여하는 것을 통해 학생들이 무엇이 왜 중요한지를 배우게 되고, 평가 방법을 점차 내면화하여 어느 수준에 이르면 다른 사람이 아니라 스스로에 대해 평가할 수 있게 된다. 이 평가를 바탕으로 장점은 살리고 부족한 점을 보완해 갈 때 숙달 수준을 높일 수 있다. 전문 기능인이 훈련되는 과정은 이를 잘 보여준다. 이들은 도제

로서 시작하여 점차 기술을 연마하고 연마한 기술을 사용하여 물건을 만드는데, 그 최종 산물에 대해 처음에는 스승의 도움으로 평가를 내리지만 점차 스스로 평가를 내려야 한다. 요컨대 정확한 평가능력이 없이는 고도의 기술을 연마하기 어렵다. 따라서 동료 평가를 시행하면, 평가의 기회를 학생들에게 많이 제공할 수 있고, 이는 학생들이 자기주도성을 발달시키도록 도와줄 수 있다.

두 번째는 학생들 그리고 교수자들의 거부감에도 불구하고 동료평가는 위에서 이미 언급된 것처럼 효과적인 교육 방법이기 때문이다. 그 효과를 확인하는데 시간이 걸릴 수 있고 처음에는 거부감이 있을 수 있지만 그렇다고 쓰지 않기에는 장점이 너무 많다. Yucel 등 (2014)도 언급한 것처럼 학생들은 동료들로부터 받은 피드백의 질이 높지 않다는 것에만 주목하지, 피드백을 제공하면서 얻게 되는 교육적 이점을 인식하지 못한다. 충분히 여러 번 경험하여 그 효과를 경험하게 되면, 상당수의 학생들은 동료평가에 대해 긍정적으로 반응한다(예, Park, 2017). 따라서 거부감 때문에 동료평가를 사용하지 않는 것은, 마치 아이들이 싫어하기 때문에 몸에 좋은 음식을 먹도록 하는 노력을 게을리 하는 것과 다를 바 없어 보인다. 교육 효과를 높이기 위한 한 방법은 홍보와 훈련을 통해 동료평가의 장점을 경험하게 하고 확산시키는 것이다.

셋째로 글에 대한 평가는 학생이 아니라 전문가가 하더라도 신뢰도가 높지 않다는 점을 지적할 필요가 있다. 우리는 앞에서 동료평가의 정확성을 학생들의 점수와 전문가 점수 간의 일치도로 보았다. 평가는 많은 경우 전문가에 의해 이루어지기 때문이다. 그런데 이 때 우리가 짚고 넘어가야 할 부분은 과연 전문가의 평가는 제대로 이루어지고 있는가이다. 이를 알아보는 한 방법은 같은 글을 여러 명의 전문가로 하여금 채점하도록 하고 그 일치도를 알아보는 것이다. 실제로 이런 연구가 수행되었는데, 전문가들 간의 점수도 편차가 크다는 사실이 반복적으로 확인되어져 왔다 (예, Finlayson, 1951; Hunter & Docherty, 2011; Bloxham, den-Outer, Hudson & Price, 2016). Bloxhom 등 (2016)은 평가는 채점기준이나 전문가 협력 작업 등으로 보완할 필요가 있지만 그렇게 하더라도 “너무 복잡하고, 직관적이며, 암묵적이기 때문에 편차는 불가피하다”고 결론짓는다 (479쪽). 요컨대 글을 정확히 평가하는 것은 학생들은 물론 전문가들에게 조차 쉬운 일이 아니라는 것이다. 우리가 할 수 있는 일은 좀 더 많은 사람들이 개입하고 논의를 통해 합의를 만들어가는 것이겠다. 따라서 채점의 정확성을 문제 삼아 동료평가를 폄하하는 것은 재고의 여지가 있다. 보다 적극적으로 동료평가를 사용함으로써, 여러 교육적 효과를 높이는 한편, 평가의 정확성을 향상시킬 수 있는 방법들을 모색하는 것은 한 대안이 될 수 있다.

6. 표절 문제

동료평가가 실제 교육 장면에서 쓰이게 될 경우 예상되는 문제점 중 하나는 표절이다. 표절 문제는 사실 동료평가와 직접 관련이 없어 보이지만, 글쓰기의 빈도가 많아지고 또 같은 내용에

대해 다른 사람의 글을 평가하기 때문에 그 발생가능성이 높다. 이 문제는 학문적 정직성과 관련된다. 우리나라는 산업 경쟁력을 높이기 위해 빠른 추적자 전략을 사용해왔다. 이 과정에서 우리는 다른 사람들의 노력과 성취에 대한 충분히 인정하고 그에 맞게 대우하지 않았던 부분이 있었다. 그 결과 도덕성에 대한 우리의 기준이 상대적으로 낮고 이런 상황은 학문 영역에서도 마찬가지라 할 수 있다. 그 구체적인 예로, 대학 내에서 이루어지는 연구에서는 물론 학생들이 과제나 시험에서 발생하는 표절에 대한 기준을 찾아보기 쉽지 않다. 물론 연구나 교육 장면에서의 표절은 우리나라만의 문제는 아니다. 다른 나라에서도 이 문제에 대해 고민을 하고 있는데, 표절 문제를 보다 체계적이고 학문적으로 다루기 위해 학문적 정직성을 위한 국제센터(International Center for Academic Integrity: ICAI)가 설립되기도 하였다. 이 센터의 홈페이지에는 2002년에서 2015년까지 McCabe 교수가 축적한 설문 결과가 소개되고 있다 (www.academicintegrity.org). 학부생 약 71,300명으로부터 얻어진 결과에 따르면 시험 부정행위는 39% 보고서에서 표절은 62%였는데, 둘 중의 하나라도 한 적이 있다고 대답한 학생은 전체의 68%나 되었다. 1,700여명 정도의 대학원생으로부터의 결과는 학부생들보다는 낮았지만, 시험부정은 17%, 보고서 표절은 40%, 둘 중의 하나라도 한 적이 있는 경우는 42%였다. 이 결과는 미국이 시험 부정행위는 물론 표절에 대해 우리나라보다 엄격한 것으로 알려진 점을 고려하면 충격적이라고 하지 않을 수 없다. 이 결과는 우리의 학문적 정직성에 대한 실태를 파악하고 필요한 조치, 예를 들어 학생들에게 학문적 윤리에 대해 지속적으로 교육해야 할 필요가 있다는 것을 시사한다.

앞서 소개된 서울대학교 자연대학에서 실시된 연구에서도, 표절과 관련된 설문 문항이 포함되어 있다. 학부생 (N=276)과 대학원생 (N=616)을 대상으로 한 설문 결과, 과제나 연구 등을 진행할 때 '표절 문제를 매번 생각'하거나 혹은 '평소에도 정직성에 대해 생각'하는 학생들은 학부생의 경우 45%였으며, 대학원생의 경우는 이 보다 높은 69%였다 (노정혜 등, 2016). 학부생의 경우는 "부정과 관련된 어떤 사건들이 일어나면 그 때 잠시 생각한다"고 응답한 학생들의 비율이 상대적으로 가장 많았으며 (41%), 별 생각이 없는 학생들의 비율도 14%나 된다. 얼마만큼 자주 생각해야 하는지에 대한 이상적인 기준이 있을 수는 없지만, 글을 쓰는 상황에서는 표절에 대해 학부나 대학원생일 때 그 기준을 자주 생각하고 적용하는 훈련이 필요해 보인다. 따라서 더 많은 학생들이 학문적 정직성에 대해 자주 생각할 수 있도록 교수가 이에 대해 수시로 강조할 필요가 있다. 이런 상황은 설문조사를 통해 나타난, 표절에 대한 교수들의 인식에서 확인할 수 있다. 교수를 대상으로 한 설문 문항 중 하나는 "표절 등 부정행위가 일어난 적이 있다"인데 이 항목에 대한 평정 점수는 5점 만점에 2.65점이었다. 이 결과는 표절이나 부정행위가 실제로 일어나고 있음을 보여준다. 이 결과는 확인된 사실이기 때문에 확인되지 않은 부정행위나 표절까지 포함시키면 실제 발생 가능성은 높아질 것이다. 그런데도 교수들은 학내 망을 통해 무료로 사용할 수 있는 표절 확인 프로그램 사용법을 잘 모르고 (2.59 점), 부정행위의 기준과 부정행위

에 따른 불이익을 알려주는 것은 물론 실제 부정행위가 일어났을 때 규정에 맞게 불이익이나 제제를 가하는 지에 대한 항목 모두에서 4점 이하로 반응하였다. 그만큼 표절과 부정행위에 대한 예방은 물론 사후 관리가 느슨함을 보여준다. 따라서 학생들의 학문적 정직성에 대한 인식의 고취와 표절에 대한 시스템 관리에 대한 필요성은 계속 논의되어야한다.

7. 동료평가의 활성화를 기대하며

동료평가는 단지 하나의 평가 기법이 아니다. 그 이상으로 학습과 교육에 대한 새로운 시각을 갖도록 한다. 동료평가는 학생들을 단지 평가의 대상으로 보는 대신 평가의 주체가 되도록 하는 활동이다. 평가의 주체가 된다는 것은 다른 학생들의 생각을 알게 되고 배우는 내용을 반성적으로 검토할 기회를 갖게 됨을 의미한다. 그리고 평가의 기회가 많아질수록 어떤 글이 좋은 글인지 더 잘 알 수 있게 되고, 자신의 글에도 적용할 수 있는 힘이 생기게 된다. 다른 사람들의 글을 평가함으로써, 자신의 글에 대해서도 비판적으로 생각할 수 있고, 지식의 적용과 확장의 능력을 고취시킬 수 있다. 따라서 동료평가는 교수의 채점 부담을 감소시키는 것 외에도, 통상 가르치는 사람이 담당했던 역할을 하면서, 지식, 이해의 수준을 넘어서 적용, 분석, 통합, 평가의 수준까지 이르게 할 수 있다. 그리고 동료평가의 정확도에 대한 우려와는 달리, 동료평가의 정확도에 대한 많은 연구들이 수행되었고, 실제 정확도 향상을 위한 훈련 프로그램들도 개발되어 그 효과도 확인되고 있다.

학습과 교육에 대해 새로운 시각이 필요한 이유는 분명하다. 우리의 교육 현황이 들이는 노력에 비해 결코 만족스럽지 않기 때문이다. 여러 국제간 성취도 비교 연구 결과를 종합해 보면, 중고등학교 때만 상대적으로 높은 성취를 보이다가 나이가 들어가면서 상대적으로 뒤쳐진다 (박주용, 2017). 아동기의 전반적인 삶의 만족도는 OECD 최하위권인데, 특히 학교에 들어가면서 급격히 떨어진다 (연합뉴스, 2014.11.4.). 지금의 교육 방식을 유지하면서 더 열심히 한다고 해서 이런 상황이 달라질 것 같지도 않다. 다른 나라에 비해 이미 엄청나게 많은 시간을 교육에 투자하고 있기 때문이다.

이제는 변화할 때이다. 한 때 우리나라 사람들의 희망이었던 교육이 고통이 되었다는 것은 단지 선거용 문구가 아니라 대다수의 사람들에게 현실이다. 교육이 다시 희망이 되기 위해서는 새로운 시도가 필요하다. 기존 교육의 문제점은 충분히 확인되었기 때문에 적극적으로 대안을 모색해야 한다 (예, 박주용, 2017). 동료평가는 그런 대안의 하나일 뿐이다. 다만 이미 많은 장점들이 확인되었기 때문에 막연한 거부감을 극복하고 우선적으로 시도해 볼 만하다는 것이다. 시도도 해보지 않은 상태에서 문제점만을 지적하는 것은 편하고 친숙한 방식에 안주하기 위한 변명일 수 있다.

동료평가를 사용하면, 지금까지 해왔던 것처럼 기계적 암기 학습을 유도하는 선다형 방식을

이용한 평가를 최소화할 수 있다. 선다형 대신 고차적인 사고와 창의력을 증진시킬 수 있는 글쓰기 과제로 평가를 대체할 수 있다. 아울러 수업도 강의를 최소화하면서 토론을 활성화할 수 있다. 그 과정에서 동료평가의 한계도 분명히 드러날 것이다. 요컨대 동료평가는 답답한 우리 교육의 새로운 활로를 찾으려는 유망한 대안이라는 것이다.

참고문헌

노정혜, 전헌수, 김상현, 이철범, 이준호, 허영숙, 홍성욱, 박주용, 곽지훈, 윤창규 (2016). **학문의 정직성에 관한 명예서약 제도 도입을 위한 정책연구 보고서**. 서울대학교 자연과학대학.

박주용 (2017). 대학 수업에서부터 시작하는 대한민국 교육개혁. **철학과 현실**, 113, 126-150.

배수정, 박주용 (2016). 대학 수업에서 누적 동료평가 점수를 활용한 성적 산출 방법의 타당성. **인지과학**, 27(2), 221-245.

연합뉴스 2014.11.4. **한국 아동 '삶의 만족도' OECD 최하위**

이현정 (2017). **채점예시답안이 동료평가의 정확성에 미치는 영향**, 서울대학교 석사학위논문.

Ashenafi, M.M. (2017). Peer-assessment in higher education-twenty-first century practices, challenges and the way forward. *Assessment & Evaluation in Higher Education*, *42*(2), 226-251.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning*. Maidenhead, Berkshire: Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

Bligh, D. (1972). *What's the use of lectures?* Harmondsworth, Middlesex: Penguin.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: David McKay Company.

Bloxham, S., den-Outer, B., Hudson J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education*, *41*(3), 466-481.

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.

Çiftçi, H., & Koçoğlu, Z. (2012). Effects of peer e-feedback on Turkish EFL Students' writing performance. *Journal of Educational Computing Research*, *46*(1), 61-84.

Dawson, P. (2017). Assessment rubrics: Towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education, 42*(3), 347-360.

Etkina, E., Karelina, A., Ruibal-Villasenor, M., Rosengrant, D., Jordan, R., & Hmelo-Silver, C.E. (2010). Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories. *The Journal of the learning sciences*, *19*(1), 54-98.

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.

Finlayson, D.S. (1951). The reliability of the marking of essays. *British Journal of Educational Psychology, 21*(2), 126-134.

Hafner, J., & Hafner, P. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education*, *25*(12), 1509-1528.

Hunter, K., & Docherty, P. (2011). Reducing variation in the assessment of student writing. *Assessment & Evaluation in Higher Education*, *36*(1), 109-124.

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.

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly*, *2*(2), 117-146.

Joordens, S., Pare, D.E., & Pruesse, K. (2009). peerScholar: An evidence-based online peer-assessment tool supporting critical thinking and clear communication. In *Proceedings of the 2009 International Conference on e-Learning* (pp. 236-240).

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, 43*(2), 207-222.

Park, J. (2017). ClassPrep: A peer review system for class preparation. *British Journal of Educational Technology, 48*(2), 511-523.

Purchase, H., & Hamer, J. (2018). Perspectives on peer-review: Eight years of Aropä. *Assessment & Evaluation in Higher Education, 43*(3), 473-487.

Resnick, L. (1986). Education and learning to think. Washington, DC: National Academic Press.

Roediger, H. L, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In B. Ross (Ed.), *Psychology of Learning and Motivation, 55*. Academic Press.

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.

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1-31.

Schunn, C., Godley, A., & De Martino, S. (2016). The reliability and validity of peer review of writing in high school AP English classes. *Journal of Adolescent & Adult Literacy*, *60*(1). 13-23.

Singleton-Jackson, J. A., Jackson, D. L., & Reinhardt, J. (2010). Students as consumers of knowledge: Are they buying what we're selling? *Innovation in Higher Education, 35*(5), 343-358.

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

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

Tsai, Y. C., & Chuang, M. T. (2013). Fostering revision of argumentative writing through structured peer assessment. *Perceptual and Motor Skills*, *116*(1) 210-221.

Van den Berg, I., Admiraal, W., & Pilot, A. (2006). Peer assessment in university teaching: Evaluating seven course designs. *Assessment & Evaluation in Higher Education*, *31*(1), 19-36.

Van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*(4), 270-279.

Wenglinsky, H. (2002). The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10,* 12.

Xiao, Y., & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. *The Internet and Higher Education*. *11*(3-4), 186-193.

Yucel, R., Bird, F. L., Young, J., & Blanksby, T. (2014). The road to self-assessment: Exemplar marking before peer review develops first-year students' capacity to judge the quality of a scientific report. *Assessment & Evaluation in Higher Education*, *39*(8), 971-986.

Zheng, L. Cui, P., Li, X., & 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*, *43*(3), 500-514.

1차 원고 접수: 2017. 10. 17
1차 심사 완료: 2017. 12. 12
2차 원고 접수: 2017. 12. 15
2차 심사 완료: 2017. 12. 15
3차 원고 접수: 2018. 06. 18
최종 게재확정: 2018. 06. 18

(*Abstract*)

The Current State and Prospects of Peer Assessment

Jooyong Park Jung Ae Park
Department of Psychology & Institute of Psychological Science

Peer assessment in educational setting, refers to activities in which students provide grades or comments on other students' accomplishments such as writing, presentation, or performances. In case of writing, peer assessment can allow instructors to give out more writing assignments, which were often avoided because of the burden of grading. Moreover, grading other students' writing can enhance learning by having students participate in the assessment process. This review will introduce major peer assessment systems and results from empirical studies on peer assessment, examine obstacles to its more wide-spread use, and discuss topics for further research. We hope this paper will facilitate further studies and use of peer assessment in actual educational settings, and eventually lead to many changes in assessment and teaching in Korea.

Key words : peer assessment, feedback, writing, web-based system