인지과학, 제27권 제2호 http://dx.doi.org/10.19066/cogsci.2016.27.2.002
Korean Journal of Cognitive Science
2016, Vol. 27, No. 2, 221～245.

대학 수업에서 누적 동료평가 점수를 활용한 성적 산출 방법의 타당성

배 수 정 박 주 용†
서울대학교 심리학과 & 심리과학연구소

동료평가 (peer assessment)란 통상 교수자가 주관하는 평가를 학생들로 하여금 서로에 대해 하도록 하는 활동을 가리킨다. 동료평가는 고등교육에서 글쓰기를 훈련시키기 위한 도구 혹은 배운 지식을 적용하거나 확장하는 학습 활동을 위한 도구로 사용된다. Park(2016)은 최근 동료평가를 예습 활동의 한 부분으로 확장하는 방안을 제안하였다. 학생들은 매주의 수업에 앞서 스스로 공부하고, 공부한 내용을 바탕으로 하여 대답할 수 있는 질문에 대해 한 페이지 글을 쓰고 다른 학생들의 글을 평가하였다. 본 연구에서는, 이 시스템을 S 대학의 학부 수업에 두 학기에 걸쳐 적용한 결과가 분석되었다. 학생들 간의 채점 신뢰도와, 학생들의 채점 결과와 교수자에 의한 최종 보고서 평가 간의 상관이 분석되었다. 매주의 동료평가에서 학생들이 부여한 점수간의 신뢰도는 그리 높지 않았지만, 이들이 누적된 점수와 교수가 평가한 기말 보고서 점수간의 상관은 두 수업 모두에서 유의미하게 높았다. 논의에서는 이 결과를 바탕으로 대학에서 예습을 위한 동료평가의 결과를 성적에 반영하는 방안의 확장 가능성이 다루어졌다.

키워드 : 동료평가, 글쓰기, 예습, 신뢰도, 타당도

† 교신저자: 박주용, 서울대학교 심리학과 & 심리과학연구소, 서울특별시 관악구 관악로 1 서울대학교
연구분야: 인지(학습과 기억, 문제 해결)
Tel: 02-880-9050, E-mail: jooypark@snu.ac.kr

서 론

평가는 전통적으로 여러 지원자 가운데 소수를 선발하거나 교육과정 이수 후 성취도에 따라 성적을 매기기 위해 사용되어져 왔다. 이러한 전통적 기능에 추가하여 최근에는 학습 기능이 강조되는데, 이는 형성평가라 불린다(Andrade & Cizek, 2010; Roediger et al., 2012; Sadler, 1989). 형성평가를 통해, 학생들은 최종 평가에 대비하여 자신의 학습 수준을 점검할 수 있고, 정보의 인출 훈련을 통해 후속 인출이 더 잘 일어나게 할 수 있다(Karpicke & Roediger, 2008). 따라서 적절한 간격을 두면서 평가를 하면, 학생들의 학습 효과를 높일 수 있다. 문제는 잦은 평가에 따른 채점 부담을 어떻게 해결할 것인가이다. 채점 문제는 특히 한 문단 이상의 긴 서술형 답안일 경우에 심각하다.

서술형 답안의 채점 문제를 해결하는 방법은 크게 두 가지이다. 하나는 컴퓨터를 이용한 자동채점이다. 이 방식은 자연어 처리에 대한 연구가 상당히 이루어진 영어의 경우 널리 활용되고 있다. 대표적으로 외국어로서의 영어 시험(Test Of English as a Foreign Language: TOEFL)과 경영대학원 입학시험(Graduate Management Admission Test: GMAT), 그리고 우리나라의 국가 영어시험(National English Ability Test: NEAT) 등에서 사용되고 있다. 그렇지만 자동 채점을 시행하기 위해서는, 동일한 주제에 대해 쓴 글의 편수가 충분히 많아야 하고, 사전에 각 글에 대한 세밀한 분석과 평가 작업이 선행되어야 한다. 이 때문에 특별히 데이터베이스가 구축되어 있지 않은 경우에는 컴퓨터 자동 채점을 활용하기 어렵다. 실제로 한글은 영어와 달리 어미와 어간의 동사 결합 방식이 다양하여, 작성된 글의 의미를 자동으로 분석해내는 과정이 상대적으로 복잡한 편이다. 이러한 이유로 한국어로 작성된 글쓰기, 특히 수업에서 제출된 서술형 답안에 대한 자동채점은 아직 갈 길이 멀다.

서술형 답안의 채점 문제를 해결하는 또 다른 방법이자 본 연구의 관심사는 동료평가를 활용하는 것이다. 동료평가는 같은 수업을 듣는 학생들이 보고서 등의 수행 결과를 서로 평가하도록 한다. 각 결과물은 한 명 이상의 다른 학생에 의해 평가되며, 평가 결과는 과제를 수행한 학생에게 피드백으로 보내진다. 결과적으로 동료평가는 교수자의 채점부담을 크게 경감시켜주면서도 학생들이 피드백을 즉각적으로 받을 수 있게 해주는 역할을 한다.

학생들로 하여금 평가에 참여하도록 하는 것은 단지 채점의 부담을 덜기 위해서만은 아니다. 학생들이 평가에 참여하는 것은 그 자체로 또 하나의 중요한 학습기회가 된다(Cho & McArthur, 2011; Cho & Cho, 2011). 평가자가 되어 봄으로써 학습 내용과 평가 과정에 대한 새로운 조망을 얻을 수 있기 때문이다. 실제로 학생들에게 서로 평가를 하게 하면 학습이 촉진된다는 것이 확인되었는데(Brown & Smith, 1997; Davies, 2000; Falchikov, 1986), 특히 대학생들에게서 두드러진다(Falchikov, 1995; Freeman, 1995; Rada, 1998; Strachan & Wilcox, 1996). 여러 명으로부터 피드백을 받을 경우 다양한 각도에서 조망을 할 수 있기 때문에 자신의 글을 수정하는 데 중요한 통찰을 제공한다.

동료평가는 종이와 펜을 이용하여 진행될 수 있다. 하지만 다수의 학생들이 참여하는 수업에서 ① 제출된 과제를 수합하고, ② 평가할 수 있게끔 분배한 다음, ③ 평가가 종료되면 다시 수합하여 ④ 제출자에게 돌려주는 일련의 과정이 복잡하고 각 과정마다 고려해야 할 사항이 많다. 또한 보통 평가자와 평가과제가 1:1로 분배되기 때문에, 동시에 하나의 과제를 여러 명이 평가하거나, 한 명의 평가자가 몇 개의 과제를 평가하도록 만들 수 없다. 더욱이 동료평가의 결과를 확인하기 위해서는 각 과정에서의 수행을 개인별로 일일이 추적하여 연결시켜야 하는 번거로움이 존재한다. 웹기반 동료평가 시스템(Web-based peer review system)은 이런 어려움을 쉽게 해결할 수 있다. 시스템 내의 알고리즘에 따라 과제 업로드와 분배, 수거가 신속하고 자동적으로 이루어지고, 평가 과정과 결과가 체계적으로 기록된다. 그리고 목적에 따라 평가자 수는 물론 평가요소, 평가절차 등이 유연하게 조정될 수 있다(예, Gielen, Peeters, Dochy, Onghena, & Struyven, 2010; Park, 2016).

이런 배경에서 정보통신 기술을 이용한 웹기반 동료평가 시스템들이 개발되었다(예, Fabos & Young, 1999; Kwok & Ma, 1999). 이중 몇 개를 간략히 살펴본 다음, 동료평가를 예습도구로 확장하는 방법을 소개하고자 한다. 본 연구에서는 소개된 방법을 실제 대학 수업에 적용했을 때 얻어진 결과가 분석되었고, 이를 바탕으로 동료평가를 이용한 예습 방법의 확장 가능성이 논의되었다.

웹기반 동료평가 시스템은 크게 두 유형으로 구분된다. 첫째는 이론을 학습한 후에 실습과제를 수행하고, 이에 대한 평가와 피드백 생성을 위해 동료평가 시스템을 활용하는 것이다. 둘째는 학생들이 교과 내용을 학습한 후 그 내용을 바탕으

로 이를 복습하는 차원에서 글쓰기를 하고, 동료평가 시스템을 통해 글에 대한 평가가 이루어지게 하는 것이다. 두 유형 모두 작성된 글을 다수의 동료들에게 평가하도록 한다는 점에서는 공통적이나, 주된 전공 분야나 기대효과 측면에서 차이를 보인다.

첫째 유형에 속하는 오클랜드 대학의 Aropa는 대규모 수업에서의 학생 활동을 보조하기 위한 목적으로 만들어진 웹기반 동료평가시스템으로서, 2002년에 개발되어 다양한 학과에서 사용되고 있다. 특히 수업 후의 실습 과제에 대한 피드백을 제공할 수 있어 공학과 컴퓨터 과학 전공 개론 수업에서 주로 사용된다. 학생들은 수업에서 컴퓨터 프로그래밍 언어에 대해 배운 후 이를 적용하는 과제를 하고, 각 과제는 다른 3-4명의 학생들에 의해 평가된다. Hamer, Purchase, Denny, 그리고 Luxton-Reilly(2009)는 2년에 걸쳐 5개의 수업에서 Aropa를 이용하여 동료평가를 실시하였고, 1500여명의 학생들의 평가 결과를 분석하였다. 이들은, 동료평가 시행 결과 학생들은 양질의 피드백을 즉각적으로 받을 수 있었고 교수자는 과제에 대한 채점 부담을 덜 수 있었다는 결과를 보고하였다. 타이완 대학에서는 WPR(Web-based Peer review system)이라는 동료평가 시스템을 개발하여 사용하였다(Liu, Lin, Chiu, & Yuan, 2001a). Aropa와 마찬가지로 컴퓨터 과학 전공 학생들을 대상으로 하였으며, 주어진 실습과제를 제출한 후 다른 6명의 학생이 제출한 과제를 서로 채점해주었다. Aropa와 달리 과제 제출자가 평가결과를 받은 후 자신의 과제를 수정할 수 있는 기회가 주어졌고, 그 다음에 최종적으로 과제를 제출할 수 있다. 이처럼 웹기반 동료평가 시스템은 필요에 따라 동료평가와 수정 기회를 여러 번 반복하여 과제의 완성도를 높일 수 있도록 할 수도 있다. Aropa와 WPR은 실습이 중요한 과목들에서 실습기회와 즉각적인 피드백을 제공하기 위한 목적으로 사용된 대표적인 예이다.

또 다른 유형의 동료평가는 내용이해와 더불어 글쓰기에 초점을 둔다. UCLA의 조정된 동료평가 시스템(Calibrated Peer Review system: CPR)과 피츠버그 대학의 Peerceptive가 여기에 해당된다. CPR은 1990년대에 개발된 후 300개 이상의 고등학교와 대학교에서 사용되고 있다. 주로 학생들이 글쓰기 기회를 더 많이 제공함으로써, 교과내용에 대한 이해를 향상시키기 위해 사용된다. CPR의 가장 큰 특징은 학생들의 개인차가 채점 점수에 반영된다는 것이다. 학생들은 학습 내용을 숙지한

다음 내용 전문가에 의해 만들어진 시험을 본다. 이 시험 점수로 각 학생의 이해 수준이 측정되는데, 이해 수준이 높은 학생의 동료평가 점수에 더 높은 비중이 주어진다. Peerceptive 시스템은 학생들의 글쓰기 능력 향상에 초점을 둔다. 제시된 주제에 대해 초고를 쓰고 다른 학생의 글을 평가한 후 퇴고하는 절차가 추가되었다는 점이 차별적이다. 초고에 대한 동료평가 결과를 확인하고, 다른 학생들이 지적한 문제들을 반영하여 수정한 후 최종적으로 과제를 제출하는 것이다. 동료평가 결과를 바탕으로 과제를 수정하고 다시 제출하게 한다는 점에서는 앞서 언급된 WPR과 유사한 면이 있다.

글쓰기를 향상시키기 위해서이든 과제 수행을 위해서이든 지금까지의 동료평가 시스템들은 주로 수업에서 배운 내용을 활용하는데 적용되어 왔다. 즉 전통적인 수업형태를 그대로 유지하면서, 수업과 관련된 별도의 과제에 대해 동료평가를 하도록 하였다. Coursera와 EdX와 같은 광역 개방강좌에서도 동료평가가 활발히 사용되고 있는데(Suen, 2014), 온라인으로 강의를 듣게 된다는 것을 제외하면, 위에서 소개된 시스템들과 크게 다르지 않다.

Park(2016)은 기존의 동료평가 시스템들과 달리, 동료평가를 예습도구로 확장하는 방안을 제안하였다. 사실 예습은 교수자에 의해 강조되기는 하지만, 학생들이 실제로 예습을 했는지를 확인하기가 어려워 그 중요성에 비해 널리 사용되지 않는 학습 활동 중 하나이다. Park이 구현한 Classprep 시스템에서는, 예습활동의 일환으로 동료평가를 이용한 글쓰기와 질문 만들기 활동을 하도록 한다. 학생들은 먼저 지정된 자료를 읽고 한 페이지 분량의 글을 쓰고 나서, 다른 학생들의 글을 평가한다. 평가 결과는 글쓴이에게 돌아가며, 마지막으로 글쓴이는 자신이 받은 평가들에 대해 재평가를 한다. 이렇듯 Classprep을 통해 수업 이전에 학습내용에 대해 자기 생각을 담아 글을 쓰고 서로의 글을 읽고 평가하게 되며, 일련의 과정은 모두 데이터로 기록된다. 이 과정에서 학생들은 깊이 있는 예습을 하고, 동시에 교수자는 학생들의 예습 상황을 확인할 수 있어 교수-학습도구로서의 활용가치가 매우 높다.

Classprep은 2014년부터 학부와 대학원의 여러 수업에서 활용되어 왔다. 이때 동료평가 점수를 단지 수집하는데 그치지 않고 성적에 직접 반영하였다. 대부분의 선행 연구에서는 동료평가 결과를 점수에 반영하더라도 과제 자체가 간헐적으로

주어졌기 때문에 그 비중이 그리 높지 않았다. 하지만 Park은 강의계획서에 성적 산출 방법이 '동료평가 50%, 기말 보고서 점수 50%'임을 명시하고, 매주 과제를 하게 한 후 그 기준대로 성적을 산출하였다. 본 연구는 이렇게 성적을 산출하는 것이 얼마나 타당성이 있는지를 알아보기 위해 수행되었다. 이를 위해 실제 이 방식으로 수업을 진행한 학부생 대상의 두 강좌의 결과가 분석되었는데, 그에 앞서 동료평가의 심리측정학적 특성에 대한 선행 연구를 살펴볼 필요가 있다.

동료평가 결과를 성적에 반영하고자 할 때, 배경지식이나 평가 경험이 별로 없는 학생들의 평가 결과를 신뢰하고 사용해도 될 것인지에 대해 의문이 제기되어 왔다 (Boud, 1989; Cho & Schunn, 2008; Lynch & Golen, 1992; Magin, 2001; Rushton, 1993; Stefani, 1994; Swanson, Case, & van der Vlueten, 1991). 이에 많은 연구자들이 동료평가와 교수자 평가 간의 상관을 통해 그 타당성을 알아보고자 하였다. Topping(1998)은 대학 수업에서 동료평가를 시행하고 그 점수를 집계한 연구 30개를 종합적으로 살펴본 다음, 동료평가의 메커니즘과 효과, 가치를 긍정적으로 평가하였다. Falchikov와 Goldfinch(2000)는 Topping(1998)에 의해 다루어지지 않은 48개의 동료평가 연구들에 대한 메타분석을 실시하였는데, 동료평가 결과는 교수자 평가와 높은 상관(M=0.69)을 보인다는 것을 확인하였다. 보다 최근에 이루어진 Hamer와 그 동료들(2009)의 연구에서는, 대학 수업에서 두 학기동안 제출된 5개의 과제에 대한 동료평가 결과들이 교수자 평가와 어떤 관계에 있는지를 분석하였다. 연구 결과, 학생들의 코멘트는 교수자보다 세밀하지 않지만, 평가 점수는 교수자의 점수와 높은 상관(r=0.71)이 있음을 확인하였다. 후속 연구에서 Hamer와 동료들(2015)은 59개의 에세이에 대해 학생과 교수자의 점수를 대응표본 t 검증으로 비교분석한 결과 통계적인 차이가 관찰되지 않았다. 이상의 연구결과는 전반적으로 동료평가 결과가 교수평가와 크게 차이가 없음을 보여주지만, 동료평가가 어떤 방식으로 진행되는지에 따라 결과 값에 차이가 있다는 점을 명시하고 있다(Falchikov, 2005; Hamer et al, 2015; Topping et al., 1998). 다시 말해 동료평가를 사용하면 다 좋은 결과를 얻는 것은 아니고, 어떤 조건에서 어떻게 사용되었는지에 따라 그 결과가 달라질 수 있다는 것이다.

Park(2016)의 Classprep 시스템은, 기존의 동료평가들과는 달리, 수업 전에 글을 쓰고 동료평가를 하도록 한다. 이 때문에 내용에 대한 숙지도가 상대적으로 떨어질

수 있고, 이 점이 동료평가의 타당성을 더 떨어뜨릴 가능성이 높다. 따라서 동료들 간의 평가의 일치성은 물론 동료평가의 결과가 교수자 평가와 어느 정도 일치되는지에 대한 경험적 확인이 요구된다. 본 연구는 이들을 확인하기 위해 실시되었다.

동료평가의 신뢰도는 크게 세 측면에서 접근할 수 있다. Davis와 Rose(2000), Cohen 등(2000)은 동료평가의 신뢰도를 평가의 일치도(equivalence), 일관성(consistency), 그리고 안정성 (stability)으로 구분하고, 이들에 대한 확인이 필요함을 강조하였다. 먼저 일치도는 한 주의 동료평가 내에서 각 과제에 대해 4명의 평가자가 얼마나 비슷하게 점수를 매기는지와 관련되어 있다. 평가해야 할 과제가 받아야 할 진점수(true score)가 있다고 가정하고, 서로 다른 평가자가 부여한 점수들이 얼마나 수렴되는지를 확인하는 것이다. 두 번째로 일관성은 한 학기에 걸친 전체 동료평가 결과들이 각 개인의 실력을 일관적으로 평가하고 있는지를 확인하는 것이다. 차원별로 학생들이 매주 받은 평균 점수들이 누적되면, 점수들 간의 분산이나 상관계수를 통해 평가 결과의 일관성을 판단하게 된다. 마지막으로 안정성은 동일한 동료평가 도구를 서로 다른 집단을 대상으로 사용해 보았을 때 비슷한 결과가 도출되는지를 통해 관찰될 수 있다.

학생들의 평가는 배경지식이나 학습정도, 평가실력 등의 차이로 인해 그 점수들이 서로 일치하지 않을 수 있다. 하지만 다수의 평가자들이 준 점수를 여러 주차에 걸쳐 수렴적으로 반영했을 때 그 결과가 학생들의 실력을 일관성 있게 반영할 수 있어야 한다. Rosenthal과 Rosnow(1991)는 다수의 평가자를 둠으로써 비전문가인 학생들의 평가 신뢰도(agreement) 문제를 극복할 수 있다고 주장하였다. 이러한 관점의 연장선상에서, Classprep 시스템을 활용한 동료평가에서는 평가자간 일치도는 어느 정도의 수준이며, 한 학기 동안의 4명의 평가결과를 수렴했을 때 전체 동료평가의 내적 일관성은 어떠한지를 실증적으로 확인할 필요가 있다. 또한 서로 다른 집단에서 시행되었을 때 그 결과가 안정적으로 반복되는지 확인하기 위해 2 개의 다른 수업에서 얻어진 자료가 분석되었다.

연 구 1

방 법

참가자

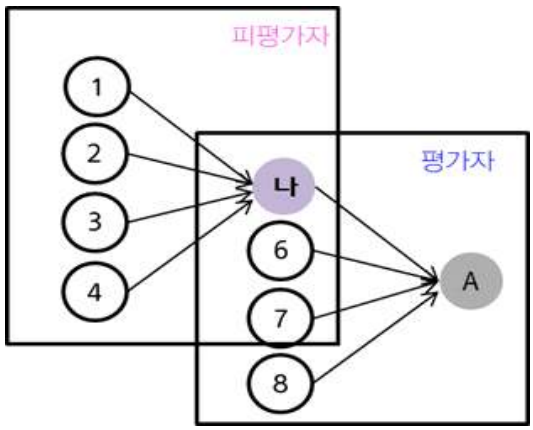
연구 1에서는 서울대학교 2014학년도 2학기 심리학 학부 전공 수업(N=11)이 분석되었다. 초기 수강신청자는 20명이었고 도중에 수강철회를 한 9명의 데이터는 제외되었다. 전체 참여자 가운데 36%가 여성이었고, 평균 나이는 25.8세(SD=2.0)이었다.

학생들은 한 학기동안 전공분야 내에서 다양한 소주제에 대해 자료를 학습하였고, 각 소주제에 대한 하나의 질문에 대해 한 페이지 분량의 글을 쓰도록 하였다. 학생들은 동료들의 글 4편을 평가하였다. 글을 쓴 학생은 다른 학생들의 코멘트와 점수를 본 다음 이에 대한 피드백으로서 코멘트와 함께 점수를 매겼다. 이 점수가 평가자에게 보내지는 것으로 한 주 과제가 종료되었다. 첫 주의 연습기간과 중간고사 기간을 제외한 매주 수업에서 동일한 과정을 반복하였고, 학기 말에는 5 ~ 10페이지 분량의 최종 보고서를 작성하게 하였다. 최종 보고서는 교수자 1인에 의해 일괄적으로 평가되었다.

절차

Classprep은 다음과 같은 절차로 수업에서 활용되었다. 제일 먼저 교수자는 매주 학습할 자료를 정해두고, 수업 1주일 이전에 학생들에게 학습 자료를 안내해 주었다. 학습 자료에는 논문이나 단행본, 교과서 등 다양한 자료들이 포함되었다. Classprep을 통해 글쓰기를 할 수 있는 문제와 채점기준이 함께 주어졌는데, 이는 Cho 등(2006)에서 사용된 채점기준을 내용에 맞게 변형한 것이었다. 채점기준은 통찰(insight)과 논리(logic), 흐름(flow)의 세 차원에서 5점 만점(0.5점 단위)으로 구분되었고, 각 점수에 대한 의미가 서술되었다(1점 단위).

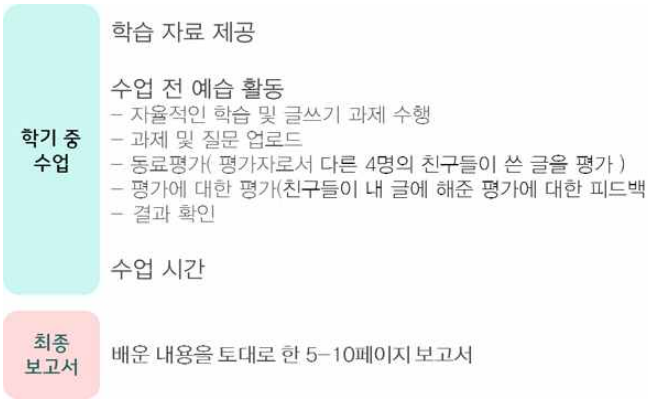
수업자료를 학습한 후에는 A4 한 페이지 분량으로 글쓰기 과제를 한다. 그리고 지정된 시간까지 Classprep에 과제와 질문을 함께 업로드한다. 각 과제는 과제 작성자를 제외한 4명의 동료들에게 익명으로 무선분배 되고, 모든 학생들은 자기에게 분배된 4명의 글을 평가한다. 아래(그림 1)에서와 같이 '나'는 1~4까지의 평가자로부터 작성한 과제를 평가받는 피평가자이면서, 'A'라는 학생의 과제를 평가하는 4명의 평가자 중 한 명이기도 하다.



(그림 1) Classprep에서의 과제 분배 알고리즘

초기의 과제 평가가 종료되면 4명의 점수와 코멘트들이 과제 작성자에게 보내진다. 과제 작성자는 4명의 평가자로부터 받은 평가를 확인한 후 '점수에 얼마나 수긍할 수 있는지', '코멘트가 얼마나 도움이 되는지'에 대해 다시 평가한다. 마지막으로 피평가자의 피드백이 다시 각 평가자들에게 보내지고, 평가자의 점수와 코멘트에 대한 피드백을 확인한다. 이상의 절차는 주 1회 수업에서 매주 진행되었다. 과제 업로드와 동료평가, 그리고 평가에 대한 평가는 각각 24시간이 주어졌다.

동료평가 과정을 통해 각 개인별로 서로 다른 4명의 평가자가 준 점수들이 기록되었으며, 차원 내에서의 평균점수와 차원을 통합한 전체 차원의 점수들이 도출되었다. 이 점수들은 각 주별 점수와 전체 시행 주차에 걸쳐 누적된 점수로 각각 정리되었다. Classprep은 각 과제가 4명의 평가자에 의해 평가되기는 하지만, 평가



〈그림 2〉 Classprep 을 도입한 수업의 한 학기 진행과정

자가 고정되어 있지 않았으며 과제를 제출한 학생들 내에서 무선적으로 4명씩 평가자가 선택되는 방식이었다. 따라서 모든 학생이 4명의 평가자에게 점수를 받더라도, 각자 서로 다른 4명의 평가자에 의해 점수를 받게 되었다.

결 과

전체 13주 가운데 오리엔테이션이 이루어지는 첫 주와 중간고사, 기말고사 기간을 제외하고 10주에 걸친 수업이 이루어졌다. 수업 내용과 Classprep의 소개를 하면서 10주 동안 과제를 하고 그 과제를 동료들에게 평가받게 되는데, 그 중 상위 7개의 과제 점수가 성적에 반영됨을 고지하였다. 매주 이루어지는 평가로 인한 부담감을 줄이고, 과제나 수업 주제에 따른 저조한 수행 결과 등의 부정적인 영향을 줄이기 위한 의도를 가지고 있었다. 7주에 걸친 세 차원에 대한 학생들의 평균 점

〈표 1〉 연구 1에서 동료평가의 세 차원에 대한 평균과 표준편차

	통찰	논리	흐름
7주 평균(표준편차)	3.58(.24)	3.77(.19)	3.98(.22)

수와 표준편차는 <표 1>에 제시되었다. 차원들 간의 상관은 비교적 높았는데, 통찰과 논리는 0.84, 논리와 흐름은 0.75, 그리고 통찰과 흐름은 0.62였다.

동료평가의 신뢰도

먼저 동료평가 점수의 신뢰도를 알아보기 위해 계급내상관계수(Intraclass Correlation Coefficient: ICC)와 크론바하 알파 (Cronbach's alpha)를 구하였다. ICC는 각 유형에 따라 서로 다른 모델과 분석방법이 사용되므로, 적절한 유형을 선택하는 것이 중요하다(Shrout & Fleiss, 1979). Classprep의 경우 각 과제마다 과제 제출 학생들 전체 집단에서 4명의 평가자가 무선적으로 선발되어 평가를 하므로, 각 과제가 서로 다른 k명의 평가자에게 평가되며, k명은 전체 평가자 집단(population of judges)에서 무선적으로 선택될 때 사용하는 유형으로 분석되었다. ICC값은 일반적으로 0에서 1사이에 분포하며(Taylor, 2010), 추정 오차가 클 경우 음의 값으로 보고되기도 한다. 그리고 학생들의 평가가 서로 일치할수록 높은 값을 나타낸다.

전체 데이터 (평가 과제 수, N=105)를 분석한 결과, 각 평가자가 세 차원에 대한 ICC의 평균과 표준편차는 각각 0.18, 0.22 였다. 범위는 -0.17에서 0.49로 비교적 넓었다. 세 차원 각각에 대한 평균과 표준편차는 통찰 0.21(0.17), 논리 0.10 (0.23), 그리고 흐름 0.11 (0.1) 이었다. 각 과제에 4명의 평가자가 참여하도록 되어 있었지만, 소규모 수업 상황에서 수집되는 데이터의 특성상 결측치(missing data)들이 많았다[1]. 결측치가 있는 경우를 모두 배제하고 4명의 평가자에 의해 평가된 결과(N=75)를 이용하여 분석하였을 때에도 결과에서 큰 차이가 없었다. ICC 결과만을 보면 선행 연구에서 얻어진 값과 비슷한데, 동료평가의 신뢰도가 그리 높다고 할 수 없다.

각 주마다 학생들이 받은 차원별 평균점수를 바탕으로, 전체 동료평가 시행에 걸쳐 해당 차원에 대한 일관성있는 평가가 이루어졌는지 살펴보기 위해 크론바하

1) 즉, 4명에게서 평가받아야 할 과제가 2명 혹은 3명의 평가자에게만 평가되는 경우가 있었고, 그것이 평가자간 상관계수(ICC)를 도출하는데 크고 작은 영향을 미칠 수 있었다. 분석 결과 결측치를 제거했을 때 일치도가 소폭 상승하고 표준편차 역시 조금 더 커지는 경향을 보일 뿐이었다.

알파를 구하였다. 과제를 제출하지 않거나 평가를 빠뜨리는 등의 이유로 빠진 자료는 각 개인의 차원별 평균 점수로 채워 넣었다. 이를 통해 4명의 평가자에 의한 평균점수들이 여러 주차에 걸쳐 축적되었을 때 전체 동료평가 결과가 해당 차원에 대한 학생들의 실력을 제대로 반영해낼 수 있는지를 확인할 수 있다. 분석 결과, 전체 세 차원을 통합했을 때는 .677이었고, 5주차의 결과를 제외했을 때에는 .715가 되었다. 세부 차원으로 나누어서 살펴보면, 통찰 차원은 .690이었고 2주차의 결과를 제외했을 때 .714가 되었다. 논리 차원은 .599로 통찰 차원보다는 낮은 수치였고, 10주차의 결과를 제외했을 때 .610이 되었다. 마지막으로 흐름 차원에서는 .682였고, 2주차를 제외했을 때 .738이 되었다. 이처럼 지정된 주차별로 그리고 차원별로 큰 차이를 보이는 것은 학습 자료와 그에 근거한 질문으로 주어지는 과제 특성에 기인하는 것으로 추정된다. 후속 연구에서는 어떤 특성이 이런 차이를 일으키는 지 탐색될 필요가 있다.

지금까지 ICC와 크론바하 알파를 통해 동료평가의 신뢰도를 알아보았다. 주차별로 또 차원별로 큰 차이를 보였고 전체적으로 그리 높지 않았는데, 이런 결과는 선행 연구에서도 마찬가지였다.

동료평가의 타당도

동료평가의 타당도를 알아보기 위해, 각 개인이 동료들에게 받은 점수와 교수자에게 받은 점수 간의 상관을 도출하여 동료평가 결과 받은 점수와 교수자 점수를 비교하였다. 각 주별로 세 차원의 점수, 즉 통찰과 논리 흐름 차원의 평균 점수와 세 차원의 점수를 평균낸 전체 점수 등 총 4개의 점수가 사용되었다. 7주의 평균 점수들을 합하여 최종적인 '통찰 점수', '논리 점수', '흐름 점수', '전체 점수' 도출되었으며 이 점수들이 분석에 사용되었다. 한편, 교수자평가 점수는 최종보고서에 대한 교수자 1인의 평가 점수(1회)였다. 두 점수간의 관계를 분석하고자 피어슨 적률 상관계수(pearson product moment correlation coefficient)를 사용하였다. 이는 점수의 분포에 크게 영향을 받지 않아 타당도 평가 등 여러 장면에 걸친 일반화에 더 적합하다고 알려져 있다(Hunter, 1983). 높은 상관 계수가 도출된다면 학생들의 동료평가 점수가 교수자평가 점수에 견주어 타당한 점수로 간주될 수 있다.

분석 결과. 세 차원의 동료평가 점수를 평균하여 도출된 점수와 교수자 평가 점수 간의 상관은 .51(p<.05)으로 중간 정도의 상관을 보였다. 각 차원의 점수를 세분하여 살펴보면, 통찰에 대한 동료평가 점수와 교수자평가 점수간 상관은 .70로 매우 높았다(p=.01). 논리에 대한 동료평가 점수와 교수자평가 점수도 .59(p<.05)로 비교적 높았다. 그러나 흐름에 대한 동료평가 점수와 교수자평가 점수는 유의한 정적상관이 관찰되지 않았다.

논 의

형성 평가의 학습적 기능이 최근 강조됨에 따라 학습 도구로서 평가의 중요성이 부각되고 있다. 평가가 학습에 분명 도움이 되기는 하지만, 평가에 대한 채점과 피드백이 빠지게 되면 그 효과가 줄어들 수밖에 없다. 이런 문제를 해결하는 한 방법은 동료평가이고 실제로 이를 활용하는 여러 시스템들이 개발, 활용되고 있다. 본 연구는 기존의 동료평가 시스템과 기술적인 면에서는 큰 차이가 없지만, 실제 수업에서 예습 도구로 활용하는 방안과 관련하여 동료평가의 신뢰도와 타당도를 탐색하였다.

예습 상황에서의 동료평가 결과는 전반적으로 복습을 목적으로 시행되었던 선행 연구 결과와 크게 다르지 않은 결과를 얻었다. 보다 세부적으로 4명의 평가 결과에 대한 신뢰도 분석 결과 각 과제에 대한 4명의 평가자간 일치도는 그리 높지 않은 정도로 보고되었다. Cho 등(2006)의 연구에서는 Classprep과는 다른 방식의 ICC 값을 도출하기는 하였으나 평균적으로 .2에서 .3정도의 낮은 신뢰도가 보고되었다. ICC 분석에 음수의 값이 보고된 주도 있었는데, 이는 매우 낮은 수준의 결과값에 대한 추정오차 때문이다 (Taylor, 2010). 신뢰도가 높지 않은 것은 평가 차원과 채점 기준 상의 문제 때문일 수 있다. 연구 1에서 제시된 3개의 차원이 학생들에게 제대로 이해되지 않았거나, 혹은 평가 장면에서 적절히 변별되지 않았을 가능성이 있다. 또한 평가 단위는 0.5점이었으나 각 점수에 대한 설명은 1점 단위로 제시되었다는 점이 학생들의 평가에 영향을 미쳤을 가능성이 있다. 크론바하 알파 값을 살펴보면 학생들에 대한 차원별 점수들이 전체 학기에 걸쳐 수용가능한 수준의 일

관성을 가짐을 알 수 있다. 이는 한 학기에 걸친 Classprep을 통한 평가가 학생들의 수행 결과를 일관적으로 평가할 수 있음을 시사한다.

교수자의 평가 점수와의 상관을 통해 살펴본 동료평가의 타당도는, 흐름 차원의 점수를 제외하고는, 높은 편이었다(통찰 차원 r=0.70, 논리 차원 r=0.59, 평균 r=0.51). 흐름 차원에서 상관이 높지 않은 이유는 여러 가지인데, 그 중 하나는 흐름 차원의 점수가 다른 차원의 점수들에 비해 높았기 때문인 것으로 보인다(M=3.98, SD=.22). 한 과제에 대해 3개의 차원에 걸쳐 평가를 하다 보니, 마지막 차원에 대해서는 크게 신경을 쓰지 않고 높은 점수를 주었을 가능성이 있다. 흐름 차원의 독립성은 물론 채점기준의 명확성 등도 문제가 될 수 있다. 하지만 일단 이런 가능성에 대한 탐구는 후속 연구로 미루도록 하고, 본 연구에서는 통찰 차원의 점수가 비교적 높은 신뢰도와 타당도를 보이는 점에 주목하고자 한다. 이 결과는 글에서 표현된 학생의 아이디어나 관점에 대해서는 학생들 간 그리고 학생과 교수 간에 어느 정도 수렴되는 평가가 이루어짐을 시사한다.

연구 1에서는 유의한 결과를 얻기는 하였으나, 몇 가지 한계점을 가진다. 첫째로 수강생의 수가 너무 작았다는 점이다(N=11). 적은 수의 데이터 임에도 불구하고 유의한 결과가 도출되었다는 점에서 주목할 만하지만, 더 많은 학생들을 대상으로 하여 반복 검증이 필요하다. 둘째로, 평가 차원의 적절성 문제이다. 연구 1에서는 통찰, 논리, 흐름의 세 차원에서 평가를 하도록 하였는데, 흐름 차원의 경우 변별력이 떨어졌다. 동료평가는 구체적(specific)으로 접근할 때보다 전반적(global)으로 접근할 때 더 좋은 결과를 만들어 낼 수 있다는 주장을 고려할 때(Falchikov & Goldfinch, 2000), 차원수를 더 줄이는 방안을 생각해 볼 수 있다. 연구 2는 이런 문제점을 해결한 상황에서 연구 1의 결과를 반복 검증하기 위해 수행되었다.

연 구 2

연구 2에서는 연구 1에서와 마찬가지로 학부생을 대상으로 한 전공 수업이었다. 수강생 수는 24명이었고, 한 학기동안 동료평가를 하도록 하였다. 연구 1에서의 차이점은 동료평가의 평가 차원이 3개에서 통찰과 글쓰기의 두 개로 줄였고, 0.5점

단위로 5점 만점 체계 대신, 1점 단위로 1점에서 7점 만점으로 변경하였다. 각 채점 단위에 따라 점수별 설명을 제시하여, 채점 기준을 보다 명료히 하였다.

방 법

참가자

연구 2는 2015학년도 2학기 서울대학교 심리학 전공 수업의 학부생(N=24)을 대상으로 하였다. 초기 수강신청자는 39명이었고, 첫 2주 동안 12명이 수강 신청을 취소하였고, 추가로 수강변경마감까지 3명이 더 취소하였다. 24명중 46%가 여학생이었고, 전체 평균 나이는 25.8세(SD=2)였다.

절차

모든 절차는 연구 1과 동일하며, 평가 차원과 점수 체계만이 변화되었다. 연구 1에서는 세 개의 평가차원과 5점 만점 체계로 평가하게 하였으나, 연구 2에서는 한계점을 반영하여 두 개의 평가차원과 7점 만점으로 평가하도록 하였다.

결 과

동료평가의 신뢰도와 타당도가 연구 1과 동일한 방식으로 분석되었다. 전체 13주 가운데 첫 주와 중간고사, 기말고사 기간을 제외한 10주의 수업이 이루어졌고, 그 가운데 상위 7개의 과제 점수가 분석되었다. 7주에 걸친 두 차원에 대한 학생들의 평균 점수와 표준편차는 <표 2>에 제시되었다. 통찰과 글쓰기의 두 차원간 상관은 연구 1의 차원간 상관의 평균과 비슷한 수준인 0.71로 비교적 높은 수준이었다.

〈표 2〉 연구 2에서 동료평가의 세 차원에 대한 평균과 표준편차

	통찰	글쓰기
7주 평균(표준 편차)	4.03(.50)	4.12(.43)

동료평가의 신뢰도

동료평가 점수의 신뢰도를 알아보기 위해 연구 1에서처럼, ICC와 크론바하 알파가 각각 분석되었다. 전체 데이터 (평가 과제 수, *N*=203)의 ICC를 분석한 결과, 평균은 0.33 표준 편차는 0.15였고, 점수의 범위는 0.10에서 0.62였다. 통찰과 글쓰기를 나누어 분석했을 때, ICC의 평균과 표준편차는 통찰 0.28, 0.16이고, 글쓰기 0.28, 0.11이었다.

크론바하 알파를 도출한 결과는 <표 3>과 같다. 두 차원을 통합했을 때는 0.873이었고, 3주차의 결과를 제외했을 때는 0.871이었다. 세부 차원으로 나누어서 살펴보면, 통찰은 0.832이었고 3주차의 결과를 제외하면 0.841이었다. 글쓰기 차원에서는 0.871이었고, 3주차를 제외했을 때 0.868이 되었다. 연구 2에서도, 연구 1에서와 마찬가지로, 주차별로 그리고 차원별로 ICC 점수에서 큰 차이가 관찰되었다.

〈표 3〉 차원별 내적 일관성 계수 (연구 2, *N*=24)

	통찰 차원 (표준화 계수)	글쓰기 차원 (표준화 계수)	전체 차원 (표준화 계수)
크론바하 알파 계수	.832(.837)	.871(.878)	.873(.881)

연구 2에서의 ICC 점수는 연구 1에서 보다는 높았다. 그리고 두 차원의 ICC 점수는 비슷하였다. 크론바하 알파도 연구 1에 비해 높았는데, 이 결과는 연구 참여자의 수가 증가한데 기인한 것으로 보인다.

동료평가의 타당도

학생 개인이 동료평가를 통해 받은 점수(7주의 평균 점수 합산)와 교수자에게

받은 점수(1회의 평가 점수) 간의 상관을 도출하여 동료평가의 타당도를 알아보았다. 연구 1에서와 마찬가지로 피어슨 적률 상관 계수를 사용하였고, 두 차원(통찰과 글쓰기 점수) 각각의 점수와 두 차원을 평균한 전체 점수 등 총 3개의 점수가 사용되었다.

분석 결과는 <표 4>와 같다. 두 차원을 통합하여 도출한 점수와 교수자 평가 점수의 상관은 .72(p<.001)로, 연구 1에서의 0.51보다 높아졌다. 각 차원의 점수를 세분하여 살펴보면, 통찰에 대한 동료평가 점수와 교수자평가 점수간 상관은 .73 (p<.001), 글쓰기에 대한 동료평가 점수와 교수자 평가 점수도 .70(p<.001)로 비교적 높았다.

〈표 4〉 동료평가 점수와 교수자평가 점수간 상관 (연구 2)

	동료평가 점수(상위 7주 합산)		
	통찰 점수	글쓰기 점수	전체 점수
상관계수	.73	.70	.72
N	24	24	24
유의확률	p<.001	p<.001	p<.001

논 의

연구 2는 더 많은 수의 학생을 대상으로 동료평가를 시행하되, 평가 차원과 점수체계를 변화시켜 연구 1의 한계점을 보완하고자 시행되었다. 그 결과 연구 1에서보다 더 높은 신뢰도가 관찰되었고 타당도는 연구 1에서와 비슷한 정도로 높았다.

4명의 평가자간 일치도는 여전히 높지는 않았지만, Cho 등(2006)에서와 유사한 정도로 상향 조정되었다. 일부 주차에서 음수의 ICC값이 보고되었던 연구 1과 달리, 연구 2에서는 모두 0에서 1사이의 양수의 값이 나왔다. 그리고 크론바하 알파를 통해 살펴본 평가의 일관성은 수용가능한 수준이었던 연구 1에서보다 더 높아

졌다. 이런 변화의 한 원인은 연구 2에서는 두 차원으로 더 단순해진 기준을 사용하고, 각 점수 단위별로 그 점수에 해당되는 설명을 제시하여 동료평가의 어려움을 경감시켰기 때문일 수 있다.

동료평가의 타당성을 확인하고자, 교수자의 평가와 상관관계를 살펴보았을 때, 두 차원 모두에서 높은 상관을 확인하였다. 특히 두 번째 차원인 글쓰기에 대한 상관이 통찰 차원과 유사한 정도로 높아졌다. 이 결과는 통찰 차원을 제외하고 논리와 흐름에서 낮거나 유의하지 않은 결과를 얻었던 연구 1과 대비된다. 두 연구 간의 주요 차이는, 차원의 개수를 줄이고 평가 점수의 의미를 보다 분명히 한 점임을 고려할 때, 이들이 동료평가의 신뢰도와 타당도를 높이는 한 방법일 수 있음을 시사한다. 어쨌든. 연구 2는 연구 1의 중요 결과인 여러 주에 걸친 동료평가 점수를 합한 점수와 교수자의 평가 점수 간의 높은 상관을 성공적으로 반복 검증하였다.

종합논의

서열을 매기기 위한 평가에서 학습을 위한 평가가 강조되면서 평가가 교육 장면에서 더 자주 활용될 가능성이 높아졌다. 그렇지만 무조건 평가를 많이 한다고 좋은 것은 아니다. 학습 효과를 높이려면 양질의 문항이 사용되어야 하고 평가에 대한 적절한 피드백이 주어져야 한다. 문제는 양질의 문항을 만들기 어렵고, 채점과 평가가 쉽지 않다는 것이다. 이 중 채점과 평가 문제를 해결하는 한 방법은 동료평가를 활용하는 것이다. 동료평가는 학생들의 학습을 촉진하고, 과제에 대한 피드백을 제공하며, 학습과 평가 과정에서 부수적인 학습 효과를 얻게 해준다. 그런데 그 결과를 성적에 무게감 있게 반영할 때, 동료평가의 효과는 훨씬 더 높아질 수 있다. 거꾸로 동료평가가 최종 성적에 반영되지 않거나 아주 낮은 비율로 반영된다면 학생들의 지속적인 참여를 이끌어내기 어렵다. 학생들은, 안타깝기는 하지만, 교수자가 무엇을 중요하게 여기고 무엇을 배우기 원하는지와 상관없이 좋은 학점을 받는데 일차적인 관심이 있기 때문이다.

동료평가는 또한 학습과 평가, 성적간의 관계를 쉽게 연결시킬(aligning) 수 있게

해주어 학생들로 하여금 학습목표를 설정하고 성취해나가는 일련의 과정을 내면화하는데 도움을 줄 수 있다(Stiggins, 2002). 수업에서 강조된 방식대로 학습과 평가가 이루어지고, 그 결과에 따라 성적이 산출되면, 학생들이 그에 따라 구체적인 학습 목표를 세우고 학습을 지속해나가는데 도움이 된다. 따라서 동료평가를 통해 학생들의 능동적인 학습과 평가 활동을 강조한 만큼 그 과정과 결과를 일정 비율 이상으로 성적에 반영하는 것은 자연스러워 보인다.

본 연구는 실제 대학 수업에서 예습의 일환으로 글을 쓰고 동료평가를 하도록 한 상황에서 얻어진 점수를 실제 성적에 반영하는 것의 타당성을 알아보기 위해 수행되었다. 동료평가와 관련된 선행 연구는 예습에 적용된 예가 없었고 동료평가 점수가 실제 성적에 반영되더라도 그 비중이 그리 높지 않았다. 본 연구에서는 학생들의 평가 결과를 합한 점수와 교수의 최종 보고서에 대한 평가를 합산하여 성적을 산출하였다. 그 결과 매 주별 이루어지는 4명의 동료에 의한 평가의 신뢰도는, 이전 연구에서와 마찬가지로 그리 높지 않았다. 그렇지만 이들 점수를 합산한 점수는 교수자의 최종 평가 점수와의 상관이 높다는 것을 2개의 연구를 통해 반복적으로 확인하였다. 본 연구는 동료 평가를 여러 번 실시하였을 때 얻어진 결과는 어느 정도 신뢰로운 평가도구가 될 수 있음을 시사한다.

두 연구 모두에서 나타난 한 가지 문제점은 중도 포기자가 많았다는 점이다. 매주 글을 쓰고 또 동료평가를 하는데서 오는 부담감이 가장 큰 이유였다. 중도 포기자가 많아진 것이 연구 결과에 영향을 줄 수 있는데[2], 이 영향은 중도 포기자의 특성에 따라 달라질 수 있다. 예를 들어 중도 포기자가 무선적으로 분포되었다면 표본의 크기만 달라지지만 전체 결과 패턴에 큰 영향을 주지 않을 수 있다. 현재의 판단으로는 상대적으로 우수하지 않은 학생들이 더 많을 것으로 추측된다. 이 경우 전체 분포에서 상대적으로 낮은 점수대에 위치한 학생들이 줄어드는 셈이다. 따라서 만일 이들이 그만두지 않았다면 점수 분포는 지금보다 더 넓은 분포를 보일 가능성이 높고 결과적으로 교수자의 평가 점수와의 .상관은 더 높아지게 된다. 요컨대 중도 포기자가 본 연구의 결론을 더 약화시킬 근거를 찾기는 어렵다.

동료평가는 수업을 같이 듣는 친구들끼리 서로의 글을 평가하게 되는 만큼, 학

2) 이러한 가능성을 지적해준 익명의 심사위원에게 감사드린다.

생들끼리 서로 좋은 점수를 주거나 나쁜 점수를 주게 되는 등 평가의 객관성이 떨어질 수 있다는 우려가 제기될 수 있다. 하지만 본 연구에서는 각 개인이 받는 점수를 세 가지 서로 다른 요소로 구분하여 이 문제를 방지하고자 하였다. 각 개인이 받는 점수는 1) 과제점수 : 4명의 평가자로부터 받은 점수의 평균, 2) 평가의 정확성 점수 : 4명의 피평가자에게 준 점수와 피평가자의 1) 점수간 편차, 3) 평가의 유용성 점수 : 4명의 피평가자가 평가에 대해 평가한 점수 등 세 가지였다. 특히 2) 평가의 정확성 점수는 학생들이 서로의 친분을 근거로 적절하지 않은 점수를 주었을 때 실제 그 과제가 받아야 할 점수와 비교하여 학생에게 패널티를 줄 수 있도록 하는 것이었다. 실제 성적에 반영된 것은 1)의 점수였지만, 학생들에게 세 가지 점수를 모두 소개함으로써 동료평가의 객관성과 일관성이 유지될 수 있도록 하였다.

지금까지 예습활동의 일환으로 쓴 글에 대한 동료평가 결과를 성적에 반영하는 방안의 확장가능성에 대해 긍정적인 입장에서 논의하였다. 그렇지만 이 낙관적 견해는 다음과 같은 한계점을 고려하여 조심스럽게 받아들여져야겠다. 첫째로 본 연구의 대상이 된 수업은 학생 수가 30명 미만으로 비교적 작은 규모의 수업이었다는 점이다. 따라서 교양 강의나 대형 강의에서 동료평가를 시행했을 때의 결과에 대해서는 별도의 분석과 확인이 요구된다. 둘째로 이 수업을 끝까지 이수한 학생들의 특성 때문에 얻어진 결과일 수 있다. 본 연구는 과제도 많고 내용도 어려워 많은 노력을 필요로 하는 전공 수업에서 얻어진 자료에 기반하였다. 이들은 상대적으로 학습에 대한 동기나 열정이 높은 특징을 가지고 있을 수 있다. 그렇지 않은 학생들일 경우 예습 혹은 동료평가를 성실하게 하지 않을 수 있고, 이로 인해 수업 자체가 원래 의도한 대로 이루어지지 않을 수 있다. 따라서 다양한 수준의 더 많은 학생들을 대상으로 그 결과들을 분석하는 후속 시도가 더 필요해 보인다. 셋째로 평가의 신뢰도 문제이다. 연구 2에서는 조금 나아졌지만 개별 평가의 신뢰도는 다른 선행 연구에서와 마찬가지로 낮은 편이었다. 연구 2에서 논의된 것처럼 채점 기준을 명료히 하는 것과, 필요할 경우 채점 훈련을 시키는 것을 포함하여, 동료평가의 신뢰도를 향상시킬 수 있는 방안은 여전히 숙제로 남아 있다. 이와 관련하여, 평가 차원과 점수 체계에 대한 고민도 필요하다. 연구 1에서와 달리 연구 2에서 채택한 두 개의 차원, 그리고 1점 단위(7점 만점)의 평가 점수는 학생들의

평가 결과를 훨씬 안정되게 만들어 주었다. 하지만 여전히 학생들이 두 차원을 제대로 변별하여 평가를 시행했는지에 대해서는 단정적으로 결론짓기 어렵기 때문에, 후속 연구를 통해 탐구되어져야 하겠다.

이런 한계에도 불구하고 Classprep 시스템은, 그 효과나 중요성에 비해 대학 수업에서 크게 활용되지 않고 있는 예습과 동료평가가 효과적으로 이루어지도록 하는데 큰 도움을 줄 수 있다. 후속 연구를 통해 언급된 한계점들이 보완되는 동시에 다양한 전공분야와 규모의 수업에서 활용가능성이 확인되면, 대학 수업이 획기적인 변화를 일으킬 수 있을 것으로 기대된다.

참고문헌

Boud, D. (1989). The role of self assessment in student grading. *Assessment in Higher Education*, *14*(1), 20-30.

Brown, S., & Smith, B. M. (1997). *Getting to grips with assessment*. Birmingham: SEDA.

Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment & Evaluation in Higher Education*, *24*(3), 301-314.

Cho, K., Chung, T. R., King, W. R., & Schunn, C. (2008). Peer-based computer-supported knowledge refinement: An empirical investigation. *Communications of the ACM*, *51*(3), 83-88.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology*, *103*(1), 73.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, *48*(3), 409-426.

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.

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

Cohen, L., Manion, L., & Morrison, K. K. (2000). *Research methods in education*. London and New York: Falmer.

Davies, P. (2000). Computerized peer assessment. *Innovations in Education and Teaching International*, *37*(4), 346-355.

Davis, A., & Rose, D. (2000). The experimental method in psychology. *Research methods in psychology*, *2*, 42-58.

Fabos, B., & Young, M. D. (1999). Telecommunication in the classroom: Rhetoric versus reality. *Review of educational research*, *69*(3), 217-259.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. *Assessment and Evaluation in Higher Education*, *11*(2), 146-166.

Falchikov, N. (1995). Improving feedback to and from students. *Assessment for learning in higher education*, *1*.

Falchikov, N. (2005). *Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education*. Routledge.

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.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education*, *20*(3), 289-300.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, *20*(4), 304-315.

Hamer, J., Purchase, H. C., Denny, P., & Luxton-Reilly, A. (2009). Quality of peer assessment in CS1. In *Proceedings of the fifth international workshop on Computing education research workshop* (pp. 27-36). ACM.

Hamer, J., Purchase, H., Luxton-Reilly, A., & Denny, P. (2015). A comparison of peer and tutor feedback. *Assessment & Evaluation in Higher Education*, *40*(1), 151-164.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. *Performance measurement and theory*, *257*, 266.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning.

science, *319*(5865), 966-968.

Kwok, R. C., & Ma, J. (1999). Use of a group support system for collaborative assessment. *Computers & Education*, *32*(2), 109-125.

Liu, E. Z. F., Lin, S. S., Chiu, C. H., & Yuan, S. M. (2001). Web-based peer review: the learner as both adapter and reviewer. *Education, IEEE Transactions on*, *44*(3), 246-251.

Lynch, D. H., & Golen, S. (1992). Peer evaluation of writing in business communication classes. *Journal of Education for Business*, *68*(1), 44-48.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education*, *26*(1), 53-63.

Moskal, B. M., Leydens, J. A., & Pavelich, M. J. (2002). Validity, reliability and the assessment of engineering education. *Journal of Engineering Education*, *91*(3), 351-354.

Nunnally, J. C., Bernstein, I. H., & Berge, J. M. T. (1967). *Psychometric theory* (Vol. 226). New York: McGraw-Hill.

Park, J. (2016). ClassPrep: A peer review system for class preparation. *British Journal of Educational Technology*.

Rada, R. (1998). Efficiency and effectiveness in computer-supported peer-peer learning. *Computers and Education*, *30*(3), 137-146.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill Book Company.

Rushton, C. (1993). Peer Assessment in a Collaborative Hypermedia Environment: A Case Study. *Journal of Computer-Based Instruction*, *20*(3), 75-80.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, *18*(2), 119-144.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Stefani, L. A. (1994). Peer, self and tutor assessment: relative reliabilities. *Studies in Higher Education*, *19*(1), 69-75.

Stiggins, R. J. (2001). *Student-involved classroom assessment*. Prentice Hall.

Strachan, I. B., & Wilcox, S. (1996). Peer and self assessment of group work: developing

an effective response to increased enrolment in a third year course in microclimatology. *Journal of Geography in Higher Education*, *20*(3), 343-353.

Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning*, *15*(3).

Swanson, D. B., Case, S. M., & van der Vleuten, C. P. (1991). Strategies for student assessment. *The challenge of problem based learning*, 260-273.

Taylor, P. J. (2010). An introduction to intraclass correlation that resolves some common confusions. Unpublished manuscript, University of Massachusetts, Boston, USA. Retrieved from http://www.faculty.umb.edu/peter_taylor/09b.pdf.

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

1차원고접수 : 2016. 01. 28
1차심사완료 : 2016. 05. 19
2차원고접수 : 2016. 05. 28
최종게재확정 : 2016. 05. 30

(*Abstract*)

The validity of using cumulative peer assessed scores for final grades in college courses

Soo Jung Bae Joo Yong Park

Department of Psychology & Institute of Psychological Sciences

Seoul National University

Peer assessment refers to having students, rather than the instructor, make assessments of one another's work. Peer assessment is often used as a tool to train writing skills or a tool to apply or extend learning in higher education. Park(2016) recently proposed a system which utilizes peer assessment as a part of preparatory activity for college courses. Before weekly class, students studied given material on their own, wrote a one page essay on a given question based on their reading, and assessed the essays of other students. In this study, the system was implemented in undergraduate courses at S University over 2 semesters and the results were analyzed. The reliability of weekly scores given by students was not very high, but the correlation was high between the cumulative scores given by students across weeks and the scores of the end of the term paper assessed by the instructor. Based on these findings, the possibility of utilizing the results of the peer assessments as part of the final grades was discussed.

Key words : peer assessment, essay writing, preparation, reliability, validity