심리학 석사 학위논문

동료평가의 정확성 향상을 위한 하위 평가자 제외 방식들의 비교 연구

2017년 2 월

서울대학교 대학원

심리학과 인지심리 전공

김혜성

동료평가의 정확성 향상을 위한 하위 평가자 제외 방식들의 비교연구

지도교수 박 주 용

이 논문을 심리학 석사 학위논문으로 제출함.
2017년 2 월

서울대학교 대학원
심리학과 인지심리 전공
김 혜 성

김혜성의 심리학석사 학위논문을 인준함.
2017 년 2 월

위 원 장 고 성 룡 (인)

부위원장 김 청 택 (인)

위 원 박 주 용 (인)

초록

교육 현장에서 학습으로서 평가의 기능에 주안점을 둔 형성평가에 대한 관심이 점점 높아지고 있다. 형성평가의 의의를 효과적으로 살릴 수 있는 방법 중에 하나는 동료평가이다. 동료평가는 학생들끼리 서로의 과제물이나 수행 과정에 대해 점수를 부여하거나 의견을 제시하는 평가 활동이다. 특히 글쓰기 수업에서 동료평가를 활용하면 학생들이 서로의 글을 비교·검토하면서 글을 평가하는 안목을 기를 수 있고, 자신의 글에 대해서 반성적으로 점검해볼 수 있는 기회를 제공한다. 결과적으로 학생들은 자신뿐 아니라 다른 학생의 사고과정을 반추하게 되어 글쓰기 능력을 발전시킬 수 있다.

동료평가는 또한 교수자들이 다수 학생들의 글쓰기 과제를 평가하는 데 드는 시간과 노력을 경감시켜 주는 부수적 효과도 있다. 원칙적으로 교수자들이 학생들의 과제를 매번 평가하고 피드백을 주는 것이 이상적이지만, 현실적으로 상당한 부담이 될 수밖에 없다. 최근 동료평가를 더욱 효율적으로 시행하기 위해 웹 기반 동료평가 시스템도 개발되고 있어 앞으로 활용도가 더욱 높아질 것으로 예상되고 있다.

하지만 동료평가 결과를 성적에 반영하는 문제에는 논란의 여지가 있다. 전문가가 아닌 비전문가 학생들이 실행한 동료평가 결과는 과연 얼마나 정확할까? 만약 학생들의 동료평가 결과가 정확하다면 이를 성적에 반영함으로써 더욱 효과적으로 이용할 수 있다.

본 연구에서는 동료평가의 정확성을 향상시키는 방안들을 분석하였다. 특히 동료평가 사후에 통계적 절차에 따라 결과의 정확도를 개선하는 방법에 주목하였다, 구체적으로 우수 평가자의 평균을 적용하는 방안(select-crowd strategy)을 응용하여 상대적으로 정확도가 낮은 하위 평가자들을 제외시켜 나가는 방식들을 실행하고 비교하였다. 이를 통해 동료평가 결과의 정확성이 어떻게 개선되는지 살펴보았다.

분석대상은 한 학기 동안 서울대학교 심리학과 학부 수업 중 이루어진 동료평가를 대상으로 삼았다. 이 동료평가는 클래스프렙(ClassPrep) 시스템을 활용하여 이루어졌다. 이 시스템은 예습을 위해 고안된 것이다. 학생들에게 한 가지 주제에 대해 클래스프렙에서 제시된 자료를 읽게 한

후 각자 그에 대한 글을 쓰도록 한 다음 다른 4명이 쓴 글들을 평가하도록 하였다. 그런 뒤 준전문가 2명이 모든 학생들의 글들을 각각 평가하고 채점하게 하여, 동료평가의 정확도를 비교하는 기준으로 삼았다.

우선 보통 통계에서 많이 쓰이는 이상치(outlier)를 제외하는 방법의 하나로 최고값과 최소값을 제거한 후 동료평가 점수를 준전문가 점수와 비교해 보았으나 유의미한 결과를 얻지 못하였다. 그 다음으로 하위 평가자 그룹을 제외한 점수를 준전문가 점수와 비교하였다. 이 때 준전문가척도와의 편차, 준전문가척도와의 상관관계, 동료평가 점수 평균과의 편차, 동료평가 점수 평균과의 상관관계를 보는 4가지 방법으로 측정하였으며, 각각 하위 25%, 50%를 순차적으로 제외하고 얻어진 동료평가의 점수를 준전문가 점수와 비교하였다.

연구 결과 각각 25%, 50%의 하위 평가자 그룹을 제외했을 때 모두 동료평가의 정확도가 개선되었다. 단순히 이상치만 제거했을 때 동료평가의 정확성은 개선되지 않았으나, 하위 평가자를 제외했을 때 준전문가척도와 비교한 경우나 동료평가 내부에서 비교한 경우 정확성이 유의미하게 향상되었다. 이러한 연구 결과는 동료평가의 사후적 통계 처치 방법에 따라 동료평가 결과의 정확도를 향상시킬 수 있음을 보여준다. 후속 연구를 통해 이 방법의 효과가 반복적으로 확인된다면, 실제 수업에서 동료평가가 학생과 교수자 모두에게 유용한 학습 도구로 활용될 수 있는 가능성을 높여줄 수 있을 것으로 기대된다.

주요어 : 동료평가, 글쓰기, ClassPrep, 하위 평가자 제외, 정확성

학 번 : 2014-22316

목 차

표 목 차

그 림 목 차

1. 서 론

우리는 삶 속에서 다양한 평가와 마주하게 된다. 입학, 취직, 승진 등에서 시험이라는 평가를 통해 중요한 결정이 내려진다. 평가가 가장 빈번히 행해지는 곳 중의 하나는 학교 등의 교육현장이다. 보통 교육현장에서의 평가라고 하면 특정 인원을 선발하거나 시험을 치고 성적을 산출하는 일련의 과정을 떠올리게 마련이다. 물론 교육평가의 기능으로서 교육의 결과를 측정하고 서열을 정하는 기능을 무시할 수 없다. 그러나 최근 교육평가가 학생의 학습에 도움을 주는 목적으로 활용되어야 한다는 주장의 목소리가 높아지고 있다. 학습으로서 평가의 기능은 특히 형성평가(formative evaluation)로 불린다. 형성평가는 학습이 이루어지고 있는 도중에 실시하는 평가이기 때문에 궁극적으로 학생들의 성취를 향상시키는 것을 목표로 삼고 있다.

형성평가의 가장 큰 특징은 피드백과 교정에 있다. 수십 명 이상의 집단으로 이루어지는 수업에서 모든 학생들이 동일한 성취 과정을 보이기는 어렵다. 그러므로 수업의 중간에서 교수자가 형성평가를 통해 학생 개개인의 성취를 점검하고 피드백을 통해 문제행동이나 오류를 교정하여 학생들 모두가 성장할 수 있도록 하여야 한다.

형성평가는 수업을 이끌어가는 교수자가 일차적인 책임을 지고 시행하는 것이 원칙이다. 그러나 교수자가 수업마다 전체 학생들에게 일일이 피드백을 해주기란 현실적으로 매우 어려운 일이다. 특히 글쓰기 수업에서는 교수자의 피드백이 매번 이루어지기가 쉽지 않다. 학생이 쓴 글을 읽는 데만도 상당한 시간이 소요된다. 자신의 생각을 근거로 하여 논리적으로 쓴 글에 대한 평가에는 더욱 많은 노력이 요구된다. 따라서 글쓰기 수업에서 매번 피드백이 가능하도록 하는 다른 방안을 생각해볼 필요가 있다.

글쓰기 수업에서 교수자에 의한 형성평가의 대안으로 주목할 만한 방

법이 동료평가이다. 동료평가는 학생들 각자가 평가자와 피평가자의 입장이 되어 평가과정에 참여하게 된다. 동료평가의 장점은 다음과 같다. 첫째, 즉각적인 피드백이 가능하다. 교수자가 모든 학생의 과제에 대해 피드백을 하려면 상당한 수고가 요구되며 정해진 시간에 맞추어 피드백을 주기 쉽지 않은 경우가 많다. 동료평가는 비교적 빠른 시간 내에 수행할 수 있으므로 학생들이 바로 피드백을 받아 자신의 글을 검토해볼 수 있다. 둘째, 여러 학생들로부터 피드백을 받을 경우 다양한 각도에서 생각해 볼 수 있는 기회를 얻게 되므로 자신의 글을 객관적으로 다시 바라볼 수 있다. 셋째, 학생들 스스로 평가에 참여하기 때문에 학습내용을 확인하는 동시에 평가 과정에 대한 인식을 제고하는 학습의 기회가 된다.

이상과 같은 동료평가의 장점을 극대화하기 위한 방편의 하나로 웹 기반 동료평가 시스템이 만들어지고 있다. 웹 기반 동료평가 시스템은 평가에 걸리는 시간을 단축할 뿐 아니라 과제의 제출과 수합이 편리하며 통계 처리가 용이하다는 장점이 있다. 대표적인 웹 기반 동료평가 시스템으로는 타이완 대학의 Web-Based Peer Review system (WPR), 오클랜드 대학의 Aropa, UCLA의 Calibrated Peer Review (CPR)과 피츠버그 대학의 Peerceptive 등을 들 수 있다. 이 시스템들은 적용한 전공 분야나 적용 후 기대 효과에서는 각기 차이를 보이고 있으나, 모두 수업에서 학습한 내용을 복습하는 차원에서 이용되고 있다는 공통점이 있다.

그런데 최근 개발된 클래스프렙(ClassPrep)은 이전의 웹 기반 동료평가 시스템과는 달리 동료평가를 역전학습(flipped learning), 즉 예습으로써 활용한다는 점에서 새로운 시스템이다. 클래스프렙을 통해서 학생들은 심도 있는 예습을 할 수 있게 되고, 예습의 과제로 주어진 질문에 답하는 형식으로 글을 쓰며, 동료 학생들의 글도 평가하고 자신의 글에 대한 평가를 받게 된다. 이러한 과정에서 학생들은 수업에 대한 이해도를 높일 수 있고 수업 시간에 활발한 토론이 유발될 수 있다. 예습으로서 글쓰기를 활용한다는 점에서 클래스프렙은 형성평가로서도 새로운 기능

을 제시한 프로그램이라고 할 수 있다.

하지만 클래스프렙에도 다른 글쓰기 동료평가에서 제기되고 있는 문제가 남아 있다. 동료평가의 결과를 얼마나 신뢰할 수 있는가 하는 것이다. 전문가들조차도 동일한 학생의 글에 대한 평가의 결과가 서로 일치하지 않는 경우가 많다. 그렇다면 학생들이 동료의 글에 대해 평가한 결과는 얼마나 신뢰할 만하고 정확한 것일까? 글쓰기 동료평가의 신뢰도를 높이기 위해 모색된 방법들로는 첫째, 동료평가를 시행할 때 그 절차를 정교하게 하는 것이다(이현정, 2016). 예를 들어 채점기준을 더 명확히 제시하거나, 채점에 참고가 될 비교 예시를 제시하여 평가 기준의 근거를 더욱 확실하게 보여주는 방법 등이 그것이다. 둘째, 동료평가 후 통계적 절차를 활용하여 점수를 보정하는 방법이 있다. 이 방법은 특히 동료평가의 결과물을 활용하기 때문에 다양한 시도가 가능하다는 점에서 활용도가 높다.

본고에서는 특히 두 번째 방법으로서 동료평가 후 통계적 절차에 따라 결과의 정확도를 개선하는 방법에 주목하고, 하위 평가자 집단을 제외하는 방식을 검토하였다. 최근 의사결정에 관련한 연구에서, 특정 집단이 내린 의사결정의 정확도를 개선하는 데에 있어서 집단 구성원 전체의 평균이 아닌 하위 그룹을 제외한 우수 평가자의 평균을 적용하는 방안(select-crowd strategy)이 더 효과적이라는 이론이 제기되었다(Mannes et al., 2014). 본고는 이 이론을 차용하여 하위 평가자 집단의 점수를 제외하는 것이 동료평가의 정확도를 어떻게 개선시키는가에 대해 연구를 진행하였다. 이와 같이 동료평가 결과의 통계적 처치를 통해 정확도를 향상시킬 수 있다면 향후 수업에서 동료평가의 활용도는 더욱 높아질 것으로 기대되었다.

2. 이론적 배경

2.1. 형성평가의 도구로서 동료평가

일반적으로 평가는 선발 또는 학습 성취 정도를 확인하기 위한 수단으로 사용되고 있다. 교육현장에서도 평가는 교수학습 활동을 통해 학습목표가 얼마나 달성되었는가를 측정하기 위해 활용되거나, 또는 차후의 교육활동과 관련된 의사결정을 내리는 데 필요한 정보를 수집하는 방편으로 여겨져 왔다.

그런데 최근에는 교수학습을 촉진하고 학생들의 성장을 돕기 위한 하나의 교육활동으로서 평가가 강조되고 있다. 평가를 수업 진행 과정 가운데 한 부분으로 인식하고 교수학습 개선을 위해서 활용해야 한다는 것이다. 즉, 과거의 평가 경향이 결과중심의 평가였다면, 최근에 등장한 새로운 평가 경향은 과정중심에 초점을 둔다. 학습자의 행동 변화 및 학습자 중심의 평가에 더 주목하고 있는 것이다(이명실, 2008).

이에 따라 평가에 대한 관심이 공식적·결과지향적인 총괄평가에서 비공식적·과정지향적인 형성평가로 전환되고 있다. 평가가 단지 결과나 성과의 측정 도구라기보다 학습 수준이 점검되고 교육 목표를 효과적으로 달성하기 위한 훈련의 한 방법으로 주목되게 된 것이다. 이제 평가는 학습의 성취도를 측정하고 점수를 매기는 것으로 끝나는 것이 아니라, 학습 효과를 높이기 위해 적절한 간격을 두고 제공되는 학습의 한 과정으로 여겨지게 되었다.

그 이론적 바탕에는 학습에 대한 구성주의적 시각이 전제되어 있다. 포스트모더니즘과 함께 등장한 구성주의는 지식이 발견되기보다 구성되는 것이라고 본다. 구성주의 교육에는 Piaget의 발달이론에 영향을 받은 인지적 구성주의와 Vygotsky의 사회적 상호작용에 기반 한 사회적 구성주의 등 세분화된 시각이 존재한다(서진원, 2005). 입장의 차이는 있으나 모든 구성주의 이론들은 학습과정에서 학생이 맥락 속에서 지식을 스스

로 구성해 가는 것을 중요한 요소로 들고 있는 점에서 유사하다.

구성주의 수업원리의 구체적 내용은 ①학습에 대한 책무성ㆍ자율성 강화, ②협동학습을 통한 사회적 상호작용 촉진, ③학습 과정과 성과에 대해 스스로 반성할 계기 제공, ④실제 수업의 맥락에서 학습을 평가할 것 등등 이다(김종문 외, 1998). 구성주의적 시각에서 본다면 평가 또한 학습의 과정이며 학습자가 적극적으로 개입해야 하는 과정이다. 형성평가를 통해 학생들은 자신의 학습수준을 점검할 수도 있고, 정보 인출 훈련을 함으로써 후속 인출이 더 잘 일어나도록 유도할 수도 있다(Karpicke & Roediger, 2008).

한편 형성평가는 교수자의 교수학습 지도 업무에도 도움이 된다. 교수자는 형성평가를 통해 학생의 학습 성취 정도를 확인하고 개선할 점에 대한 정보도 얻을 수 있다. 이러한 정보를 근거로 하여 학생들의 학업 성취를 격려하고 동기를 강화하거나 또는 교수학습방법을 교정하게 된다.

이와 같이 형성평가의 교육적 효과가 강조되면서 글쓰기 학습에서도 형성평가를 적극적으로 도입하려는 시도가 늘어나고 있다. 본 연구에서 평가의 대상으로 삼은 것도 학생들이 수업 전에 예습을 하고 제출하도록 한 글이었다. 매주 수업 시간마다 학생들이 제출하는 글쓰기 과제에 대해 평가하는 과정이 단순히 성적 매기기에 그치지 않고 학생들에게 의미 있는 학습활동이 되기 위해서는 적절한 피드백이 주어져야 한다.

그러나 현실적으로 한 명의 교수자가 매주 수업 시간마다 수십 명의 학생들이 제출한 글에 대해 일 대 일로 첨삭지도하고 피드백을 주기란 매우 어려운 일이다. 이러한 상황에서 학생들이 가진 글쓰기 능력의 성장 과정을 점검하고 촉진하는 형성평가로서 활용할 수 있는 구체적 방안 중 하나가 바로 동료평가이다.

동료평가(peer assessment)는 다른 학생의 과제를 읽고 이에 대해 평가하는 활동 전반을 가리키며, 나아가 교수자가 주관하는 평가에 학생들을 참여시켜 서로의 과제에 대하여 일정한 점수나 등급을 부여함으로써

실제 평가에 참여시키는 활동을 말한다. 이런 점에서 동료평가는 학습 목표에 얼마나 도달하였는가를 평가해보는 것을 주목적으로 하여 평가자가 되는 학습자가 특정 채점표 등을 기준으로 해서 상대방 학습자의 학습 결과물을 평점 하여 학습 목표에 달성한 정도를 알아보는 방식의 효과적인 형성평가 도구이다. 학생 간에 학습 성취를 진작시키고 그 과정에서 궁극적으로 학습을 성공적으로 이끄는 것을 목적으로 하고 있다(김민정, 2008).

글쓰기 수업에서 동료평가를 활용할 때 얻을 수 있는 장점이 있다. 첫째, 교수자의 채점 부담을 덜어주는 한편 학생들은 피드백을 즉각적으로 받아 볼 수 있다. 둘째, 다른 학생들로부터 피드백을 받을 경우 다양한 각도에서 생각해 볼 수 있는 기회를 얻게 된다. 이를 통해 자신의 글을 수정하고 보완하는데 필요한 통찰력을 얻을 수 있다(Brown & Smith, 1997; Falchikov, 1986; Davies, 2000). 셋째, 학생들 스스로 평가에 참여함으로써 학습내용을 확인하면서 평가 과정에 대한 인식을 제고하는 학습기회를 얻을 수 있다(Cho & Cho, 2011; Cho & McArthur, 2011). 이러한 효과는 중등학생들보다도 대학생들에게 더 두드러지게 나타나는데(Freeman, 1995; Strachan & Wilcox, 1996; Rada, 1998), 대학생 정도의 지적 수준이 되어야 타인의 글을 객관적인 준거에 의해서 평가할 수 있다고 여겨지기 때문일 것이다.

동료평가를 통해 학생들은 즉각적으로 다른 학생이 평가한 자신의 글에 대한 피드백을 받아보게 된다. 동료 피드백의 성과에 관한 많은 연구들에서는 대체로 동료 피드백에 대해 긍정적인 평가를 내리고 있다. 그 중에서도 글쓰기에 대한 동료 피드백은 학생들에게 좋은 효과를 주고 있음이 입증되었다. 피평가자의 자신감을 높여 줌으로써 글쓰기의 결과보다 글쓰기 과정에 더 많은 도움을 주고(Rizzolo, 1982; Bell, 1982; 이윤빈 & 정희모, 2014), 특히 대학 글쓰기 교육 과정에서는 피평가자로서의 글쓰기 성적 뿐 아니라 평가자로서의 쓰기 능력까지도 향상되었다고 한다(Holladay, 1990; 이윤빈 & 정희모, 2014).

많은 학생들이 글쓰기를 할 때 정작 자신의 글이 가지고 있는 문제점을 잘 파악하지 못하고 있는 경우가 많다. 동료평가를 통해 학생들은 다른 사람의 글을 읽고 평가하면서 그 글의 완성도를 판단할 수 있는 훈련을 쌓게 된다. 그리고 그 판단 기준을 그대로 자기 글에 적용해 봄으로써 자신의 글을 객관적으로 판단해 볼 수 있게 된다. 이런 의미에서 동료평가는 평가자의 평가능력과 학습능력 신장에 상당히 큰 도움이 된다.

이처럼 동료평가 훈련을 통해 평가자의 시선을 습득하게 된 학생들은 자신의 글에 대해서도 동일한 기준을 적용해 평가하는 것이 가능해진다. 즉 자신이 쓴 글에 대해 다른 사람의 시선으로 보고 객관적으로 평가할 수 있게 되어(Nicol & Macfarlane-Dick, 2006), 동료평가는 글쓰기를 개선시킬 뿐만 아니라 그 다른 학생들의 글의 논리와 입장을 이해하고 평가하는 과정에서 학생들 자신의 글을 객관적으로 평가하는 능력도 향상시킨다(Boud, 1995; Bostock, 2000; Brown et al., 1994; DeGrez et al., 2012; Dochy et al., 1999; Falchikov, 1986; Liu & Carless, 2006).

또한 학생들은 한걸음 더 나아가 고차원적 사고를 할 수 있게 된다. 학생들 스스로 다른 학생들의 글을 읽으며 사고과정의 오류나 문제들을 발견해 내고 이 과정을 자기의 글에 적용하여 글의 논점을 더 강화하거나 문제점을 교정할 수 있기 때문이다. 평가를 한다는 것 자체가 이미 다른 학생들이 이해한 내용을 보고 논리적으로 해석하는 작업인 것이다(Boud, 1990).

학생들은 피평가자 학습자의 글쓰기를 검토하고 피드백을 줄 때 점검, 명료화, 요약, 기대 수준과 실제 수준의 차이 파악 등과 같은 전략을 사용하게 되는데, 이 전략들은 인지적, 메타인지적인 활동과 관련된 것이다(김민정, 2008). 이렇게 볼 때 글쓰기에 대한 동료평가는 형성평가로서 학습에 도움을 줄 뿐 아니라 학생들의 인지 발달에도 도움이 되는 매우 유용한 방법이다.

그런데 동료평가는 수기 작업으로 진행되는 경우가 많다. 하지만 다수의 학생이 참여하는 수업에서 제출된 과제를 배분하고 평가 후 다시 수

합하여 작성자에게 돌려주는 일련의 과정은 너무 복잡하고 시간이 많이 걸린다. 동료평가의 효율성을 높이기 위한 방편의 하나로 개발된 것이 바로 웹기반 동료평가 시스템이다. 이 시스템을 활용하면, 과제를 업로드 하여 분배한 후 평가 결과를 수거하여 피드백 하는 일련의 과정이 신속하고 자동적으로 이루어지고, 체계적으로 기록되며, 평가 시 목적에 따라 변화하는 평가자 수와 평가 요소와 절차 등을 반영할 수 있다(Fabos & Young, 1999; Kwok & Ma, 1999: Gielen, Peeters, Dochy, Onghena & Struyven, 2010; Park, 2015). 웹 기반 동료평가 시스템은 번거로운 과정을 덜어주는 편리함과 통계 처리 분석의 용이함으로 인해 최근 활발하게 개발되고 있다.

기존의 웹기반 동료평가 시스템에는 크게 두 가지 유형이 있었다. 첫 번째 유형은, 이론을 학습한 후 실습과제를 수행하고 그에 대한 평가와 피드백을 생성하여 학생 활동을 보조하는 방식이다. 대표적인 사례로 2001년 보고된 타이완 대학의 WPR (Web-based Peer Review system)을 들 수 있다. 이것은 컴퓨터 과학 전공 학생들을 대상으로 실습과제를 제출케 하고 다른 6명의 학생 과제를 채점하게 한 후 돌려받은 자신의 평가결과를 보고 과제를 수정할 수 있는 기회를 준 다음 최종적으로 과제를 제출하게 하는 것이었다(Liu, Lin, Chiu, & Yuan, 2001).

또한 오클랜드 대학에서 개발된 Aropa도 들 수 있는데, 이것은 특히 대규모 수업에서 학생들의 학습 활동을 돕기 위해 고안되었다. 컴퓨터 과학 전공 학생들에게 이론 수업을 하고 이를 적용하는 과제를 하게 한 후 이것을 3~4명의 다른 학생들로 하여금 평가하도록 한 것이었다.

두 번째 유형은 학생들이 교과 내용을 학습한 후 그 내용을 바탕으로 복습하는 차원에서 글쓰기를 하고 그 글에 대한 평가를 하는 방식이다. 대표적인 시스템으로 1990년대 개발된 UCLA의 CPR (Calibrated Peer Review system) 과 피츠버그 대학의 Peerceptive를 들 수 있다. 그 중에서도 CPR이 다른 동료평가와 크게 차이를 보이는 특징은 채점점수에 학생의 개인차가 반영된다는 점이다. 즉 학생들은 먼저 학습내용을 공부하

고 난 다음에 내용 전문가들에 의해 출제된 시험을 치르게 된다. 이 시험에서 받은 점수를 근거로 각 학생의 이해수준이 측정되고, 이해 수준이 높은 학생들의 동료평가 점수에 더 높은 비중이 주어진다는 것이다.

그런데 이 두 가지 유형의 동료평가는 전공 분야나 기대 효과 측면에서 다소 차이는 있으나, 주로 수업에서 교수된 내용을 활용하는데 적용된다는 공통점이 있다. 즉, '선수업 후과제'의 전통적 수업형태를 유지하며 동료평가를 하고 있는 것이다.

이에 비하여 최근 개발된 Park(2016)의 클래스프렙은 새로운 유형의 웹기반 동료평가 시스템으로 주목할 만하다. 클래스프렙의 특징은 예습활동의 일환으로 동료평가를 이용한 글쓰기와 질문 만들기 활동을 하는 방식이라는 점에 있다(배수정 & 박주용, 2016). 클래스프렙을 통해 학생들은 깊이 있는 예습이 가능했으며 예습 후 과제인 글쓰기를 동료 간에 평가하고 피드백을 교환하게 되므로 형성평가로서도 활용가치가 큰 시스템이라고 할 수 있다.

본고는 형성적 동료평가를 활용하는 방안의 연장선상에서 Park(2016)의 클래스프렙을 활용하여 동료평가의 정확도를 높이기 위한 방안을 모색해보았다. 이를 통해 클래스프렙이 예습 도구로서 학생들의 학습능력 향상에 기여할 뿐 아니라, 신뢰할 만한 성적 산출 도구가 될 수 있다는 점을 살펴보고자 하였다.

2.2. 동료평가의 문제점과 한계

이상에서 살펴본 바와 같이 형성평가로서 동료평가를 활용할 때 학생들의 학습에 도움이 될 수 있다는 긍정적 측면이 있다. 그렇지만 동료평가가 가진 문제점과 한계도 존재한다. 우선 형성평가로서 동료평가가 유용하게 사용되려면 많은 준비가 필요하다는 점이다. 형성적 동료평가는 현실적으로 활용하려 할 때 계획 단계를 명확히 세우지 않고 지나가는 경우가 많다. 뚜렷한 학습목표를 세운 후 형성적 동료평가를 학습에 활용하는 명확한 근거와 계획이 필요하지만 실제로는 그다지 철저하지 못

한 채 시행되는데 그러한 경우에는 적절한 효과를 보기 어렵다.

게다가 동료평가에 대한 훈련이 제대로 이루어지지 않고 있다는 점도 문제점으로 들 수 있다. 동료평가 시 구체적인 수행 방법에 대한 훈련이 잘 이루어지지 않고 간단한 평가기준표만 제시한 후 학생들에게 채점을 하거나 피드백을 주라고 하는 경우가 많은데, 이런 경우 평가와 피드백은 명확하지 않을 수 있다. Orsmond 등(2000)의 연구에 따르면, 동료평가에서 주어지는 부정확한 피드백은 학습적인 측면 뿐 아니라 피평가자 학습자들에게 불편한 감정을 들게 하고, 더 나아가 동료 학습자의 평가 능력에 대한 불신을 불러일으키기도 한다. 그 뿐만 아니라 동료평가에서 평가자로서 학습자 자신이 가지는 능력에 대한 회의를 보인다고 한 연구도 있다(Ballantyne, R., Hughes, K. & Mylonas, A., 2002). 그러다보면 학생들이 동료평가에 소극적으로 참여하거나 포기하게 되는 경우도 생겨날 수 있다.

그러므로 학생들에게 동료평가에 적극적으로 참여하도록 유도하기 위해서는 동료평가의 결과를 성적에 반영하는 것이 효과적인 방법의 하나이다. 그런데 문제는 동료평가의 결과를 얼마나 신뢰할 수 있는가 하는 것이다. 동료평가의 정확도에 대한 논란은 항상 문제로 남아 있다. 심지어 전문가 간, 또는 전문가와 준전문가 간의 평가가 항상 일치하리라는 보장도 없는 상황에서, 비전문가인 학생들의 동료평가는 정확도가 더 떨어질 것이라고 의심할 수밖에 없다.

비전문가인 학생들은 어떤 주제에 대해 심도 있는 이해가 부족하기 때문에 다른 학생들의 글을 평가할 때에도 핵심을 잡아내지 못하거나, 심지어 잘못된 조언을 할 수도 있다. 예를 들어 전문가들이 글을 평가할 경우 논리적인 부분이나 내용적인 면에 비중을 크게 두는데 비하여 학생들은 상대적으로 글의 형식적인 부분이나 문체에 더 치중하는 경향이 있다(Flower, Hayes, Carey, Schriver, & Stratman, 1986). 그렇다면 글을 평가하는 데 있어 폭과 깊이 면에서 부족한 동료평가의 결과를 성적에 반영하는 것은 정당하지 않을 수 있다. 우리나라 중등교육 현장에서의

글쓰기 수행평가에서 동료평가는 실제적으로 큰 효과를 거두고 있지 못하며 점수에 반영되지도 못하고 있다(오택환, 2010).

또한 동료평가 능력에 있어서 학생들의 개인차가 크다는 점도 무시할 수 없다. 전문가 수준에 가깝게 평가를 하는 학생들이 있는가 하면 그렇지 않은 학생이 있을 수 있다. 학생들 간의 일치도와 평가 방식에 대한 문제는 꾸준히 제기되어 왔다. 동료평가에 대한 메타 분석 결과를 보면 그 결과는 안정적이지 못하다. 예를 들어 학생들 간 또는 학생들과 교수자의 평가 일치도를 분석한 결과 r=0.29로 보고된 바가 있다(Kovach, Resch, & Verhulst, 2009).

Topping 등(2000)은 12명의 교육심리학 대학원생들의 글을 동료평가를 수행하고 그 결과를 분석했는데, 그 결과를 보면 학생들은 글의 구조의 대해 평가하는 점에서는 높은 일치도를 보였으나, 글의 내용적인 측면(얼마나 새로운 아이디어가 있는지? 데이터의 질은 어떤지? 등등)을 평가할 경우에는 비교적 낮은 일치도를 보였다. 따라서 학생들의 동료평가 결과가 낮은 일치도를 보인다면 전체 학생들의 동료 평가 결과를 성적에 반영하는 것에 신중해야 한다.

그런데 이와는 달리 동료평가의 정확도를 긍정적으로 보고한 연구들이 있어 주목이 된다. Hamer 등(2009)의 연구에서는, 대학 수업에서 두 학기동안 제출된 5개의 과제에 대한 동료평가 결과들이 교수자 평가와 어떤 관계에 있는지를 분석하였다. 그 결과 평가 점수가 교수자의 점수와 높은 상관(r=0.71)이 있다고 보고되었으며, 동료평가가 어떤 방식으로 진행되는지에 따라 결과 값에 차이가 있다는 점이 밝혀졌다(Hamer et al, 2015; Topping et al., 1998). 이 결과는 동료평가의 결과가 그 사용 맥락에 따라 정확성이 달라질 수 있음을 시사한다.

2.3. 동료평가의 정확성 향상 방안

동료평가에 대한 문제 제기는 대체로 학생들의 동료평가의 결과적 측면에 집중한 결과 일어난다고 할 수 있다. 동료평가를 형성평가로 활용하여 학생들의 학습능력 신장에 도움이 되도록 하면서, 동시에 동료평가의 결과를 성적에 반영할 수 있는 방법은 없을까? 이것이 가능하다면 형성평가의 본래의 목적이 달성될 뿐 아니라 교수자도 시간을 절약할 수 있다. 이를 위하여 본고에서는 동료평가의 정확성을 높일 수 있는 사후적 통계 처치 방안을 탐색해 보고자 한다.

사회심리학자들은 오랫동안 통계화된 그룹의 힘을 인정해 왔다. 어떤 사실에 대한 개인적 평가들을 합쳐서 평균을 냈을 때 그 평균적 견해가 대부분의 개인적인 평가들보다 더 정확하다고 믿는다. 이러한 것을 우리는 종종 대중의 지혜(wisdom of crowds)라고 말한다. 대중의 지혜가 발휘되려면 여러 가지 조건이 있는데, 그 중 하나는 집단의 다양성이 보장되어야 한다는 것이다. 다양성은 다른 사람들이 미처 알지 못했을 수 있는 관점을 추가해 주며, 집단 의사 결정이 일어날 수 있는 집단적 사고(group think) 및 편향적인 판단을 약화시키는데 기여한다. 이 다양성의 효과는 실험으로도 확인되었다. 예를 들어 구성원 10~12명이 서로 다른 종류의 기술을 보유한 집단을 여러 개 만들어 상대적으로 복잡한 문제를 풀게 했다. 그 결과, 현명한 사람과 그렇지 않는 사람이 섞여 있는 집단이, 현명한 사람들로만 구성된 집단보다 좋은 결과를 낸 것으로 밝혀졌다(Surowiecki, 2004). 이 연구 결과에 따르면 구성원이 다양한 집단일수록 문제 해결 능력이 전반적으로 높아진다고 볼 수 있다.

두 번째는 집단 구성원들이 서로 독립적으로 작동해야 한다는 것이다. 구성원들이 서로 독립적이라면 집단의 지혜가 정확한 예상치를 구해 내거나 좋은 결정을 낼 가능성이 훨씬 높다. 독립성은 사람들이 저지른 실수가 서로 연관되는 것을 막아 주며, 개개인의 판단에서 생긴 오류가 집단 전체의 판단을 손상시키지 않는다. 이 독립성이 사람들의 다양한 관점을 가지게 유지 시켜주는 방법이기도 하다(Surowiecki, 2004).

이런 대중의 지혜를 사용하여 실제 의사결정 현장에 적용한 연구들이 있다. Mannes et al. (2014)은 다수로 이루어진 그룹들을 대상으로 몇 가지 실험을 하여 다수 평가의 정확도가 가장 높아질 수 있는 방안을 모색하였다. 이를 위해 다음과 같은 세 가지 전략이 사용되었다. 첫째, 가장 우수한 구성원의 평가를 채택하는 전략(best-member strategy), 둘째, 구성원 전체의 평가의 평균을 채택하는 전략(whole crowd strategy), 마지막으로 구성원 가운데 소수의 우수 평가자들의 평균을 채택하는 전략(select-crowd strategy)이다. 이 세 가지 전략의 효과를 측정하는 시뮬레이션 실험이 진행되었는데, 그 결과 오차가 가장 적은 전략은 우수 평가자의 평균을 내는 방안이었다. 또한 우수 평가자의 수는 모집단의 크기와 관계없이 상위 4명에서 8명까지 선별하여 평가하는 것이 가장 최적이라는 사실이 확인되었다 (Mannes et al., 2014).

이밖에 학생들에게 경제적 수치를 예측하게 한 과제 실험에서도 우수 평가자들의 평균이 가장 정확도가 높다는 연구 결과가 있다. 80명의 학생들을 대상으로 향후 반년 간의 경제 전망 특집 기사 속 11명 경제학자들의 예측 값들을 참고하여 6개월 후 국채의 이자율을 예측하는 과제가 주어졌다. 이때 학생들에게 위의 세 가지 전략 중 하나를 선택하여 과제를 수행하도록 하였고, 학생들이 매번 선택을 할 때 마다 예측된 값의 오차를 알려주었다. 총 24번의 예측을 실시한 결과 다시 선택할 기회가 주어졌을 때 우수 평가자의 평균을 내는 전략으로 바꾸는 학생들이 많았다. 그리고 실제로도 우수 평가자 평균을 사용하여 예측된 이자율의 정확도가 가장 높았다(Larrick et al., 2011). 이 연구 결과에 따르면 평가의 정확도를 높이는 방법으로 우수 평가자들을 선택하는 방안이, 가장 우수한 구성원의 평가를 채택하는 전략이나 구성원 전체의 평가의 평균을 채택하는 전략에 비해 여러 변수의 환경에 효과적이고 더 정확하다고 판단되었다.

위의 실험은 예측에 관련된 것이므로 그 결과를 그대로 동료평가에 적용할 수는 없지만 동료평가의 정확도를 높이기 위한 방법을 모색하고자

할 때 여기서 유용한 아이디어들을 많이 얻을 수 있었다. 특히 동료평가를 예습도구로 확장시킨 클래스프렙에서 평가의 측면에서 정확도를 개선하는 방안으로 우수 평가자 전략(select-crowd strategy)을 활용할 수 있는지 모색해보도록 하겠다.

평가의 정확도는 참값과 평가에서 나타난 오류의 차이를 줄여나갈수록 높아진다. 그러면 클래스프렙에서의 참값을 생각해 보도록 하자. 클래스프렙에서의 동료평가는 앞으로 있을 강의 내용을 예습하고 그와 관련된 주제에 대하여 글쓰기를 한 것을 서로 평가하는 것이다. 총 4번의 평가를 하게 되는데, 이때 평가를 상대적으로 우수하게 하는 학생들이 있을 것이고, 반면에 그렇지 않은 학생들이 있을 것이다. 그렇다면 평가를 제대로 하지 못한 학생의 결과를 최소화함으로써 동료평가에서의 오류를 줄여가야 할 것이다. 이미 클래스프렙에서의 동료평가의 정확성에 대한 연구가 진행된 바 있다. 그 연구에서 예습활동의 일환으로 쓴 글에 대한 동료평가가 성적에 반영될 수 있는 방안의 확장 가능성이 긍정적인 입장에서 논의되었다(배수정 & 박주용, 2016). 매주 이루어지는 동료에 의한 평가 신뢰도는 그리 높지 않았지만 이들 점수를 합산한 점수는 교수자의 최종평가점수와 상관관계가 높았기 때문이다.

본 연구에서는 동료평가의 정확도를 향상시키기 위해서 동료평가 실행 후 통계적 절차를 활용하여 점수를 보정하는 방법에 대한 연구를 실행하였다. 앞서 언급한 우수 평가자들을 뽑아 그들이 내린 평가의 평균을 선택하는 것이 평가의 정확도를 가장 높일 수 있는 방안임을 이야기한 바 있다. 하지만 선행 연구의 실험들이 전문가를 대상으로 하여 실행된 것에 비해, 본 연구는 학생들의 동료평가를 분석하였다. 따라서 기존의 우수 평가자 전략과는 다르게 평가에서 정확도가 떨어지는 학생들의 평가결과를 전체평가에서 제외함으로써 상대적으로 평가자의 정확성을 높이는 방법을 채택하였다. 구체적으로 어떤 기준으로 정확도가 떨어지는 학생들의 평가결과를 전체평가에서 제외할 것인지, 그리고 그에 따라 평가의 정확도가 어떻게 달라지는지 비교하는 것에 중점을 두었다.

3. 연구

본 연구에서는 클래스프렙을 사용한 서울대학교 학부 심리학과 글쓰기 과제에 대한 동료평가의 결과가 분석되었다. 학생들은 한 학기 동안 매주 클래스프렙으로 제시된 소주제에 대한 텍스트나 논문을 예습하고, 주어진 질문에 대해 한 페이지 가량의 글을 작성하였다. 글의 내용은 소주제 자료 요약과 함께 비판, 적용, 혹은 후속 연구 제안 등으로 구성되었다. 학생들의 글은 정해진 시간까지 클래스프렙에 업로드 하도록 하였다. 이와 함께 내용 이해에 곤란을 느낀 부분이나 함께 토론하고 싶은 사항이 있는 경우 이를 클래스프렙에 쓰도록 하였다. 이후 클래스프렙에 수합된 전체 학생들의 글은 다른 학생들 4명에게 익명으로 배부되었다. 학생들은 동료들의 글 4편에 대해 미리 공지된 채점기준에 따라 평가하고 코멘트를 달았다. 그리고 자신의 글을 평가받은 학생은 평가자 4명으로부터 받은 코멘트가 얼마나 도움이 되었는지 평가하였다.

본 연구에서는 학생들이 각자 다른 평가자 4명에게 평가를 받은 결과까지만 분석대상으로 삼았으며, 평가자의 코멘트에 대한 피평가자의 평가는 다루지 않았다. 또한 매주 실행된 총 12주의 동료평가 결과 중에서 7주차 데이터를 분석하였다. 7주차 데이터를 대상으로 삼은 것은 학기 중간인 7주차에는 학생들이 어느 정도 동료평가 시스템에 익숙해져 있을 것으로 기대되었기 때문이다.

방법

참가자

연구 1은 2016년도 1학기 서울대학교 심리학과 학부 수업 중 7주차 동료평가를 분석대상으로 삼았다. 동료평가를 수행한 인원은 58명이었다.

수업을 수강한 학생 외에 별도의 평가자로 준전문가를 두었다. 대학원생 1명과 경험이 풍부한 심리학 전공 학부생 1명에게 모든 학생들의 글을 각각 평가하고 채점하게 하였다.

절차

클래스프렙은 다음과 같은 절차로 수업에서 활용되었다. 먼저 교수자는 매주 학습할 자료를 클래스프렙에 탑재해두고, 수업 1주일 전에 학생들에게 학습 자료를 안내해 주었다. 학습 자료에는 논문이나 단행본, 교과서 등 다양한 자료들이 포함되었다. 그리고 학생들이 클래스프렙에서 학습 자료를 볼 때 질문과 채점기준이 함께 제시되었다.

질문은 “산업사회에서 제품을 디자인 할 때에 사용자를 염두에 두지 않고 디자인을 한 적은 없었다. 그렇다면 디자인 사고에서 말하는 인간중심적 디자인이나 감성적 디자인 또는 디자인 혁신이 전통적인 디자인과 다른 점이 무엇인지 (혹은 근본적으로 다른 점이 있는지) 논의하라.” 였다.

채점기준은 Cho 등(2006)에서 사용된 채점기준을 활용하여 변형한 것이다. 본디 채점 기준은 ‘통찰(insight)’과 ‘흐름(flow)’ 두 가지 측면에서 제시되었으나, 이 연구에서는 ‘통찰’의 점수만 활용하였다. 선행연구에 따르면 ‘통찰’과 ‘흐름’ 차원 간 상관은 비교적 높은 수준으로 크게 차이가 없었으므로(배수정·박주용, 2015), 결과를 보다 단순하고 명확하게 도출하여 분석하기 위해 ‘통찰’ 차원만 다루기로 하였다. ‘통찰’의 채점기준에는 1~7점까지 단계별 점수와 그 기준 내용이 서술되었다(그림 1).

<채점 기준>

글을 쓸 때와 동료 평가를 할 때 참고할 기준은 다음과 같습니다.

A. 통찰 (비판, 활용, 혹은 후속 연구 제안)

점수	기준
7	더 이상 좋은 통찰을 해낼 수 없다.
6	높은 수준의 비판, 활용 혹은 후속 연구 제안이다.
5	통찰을 찾아보기 어렵지만 새로운 비판, 활용 혹은 후속 연구 제안이다.
4	고민한 흔적이 있지만 누구나 제시할 수 있는 비판, 활용 혹은 후속 연구 제안이다.
3	자료에 대한 숙고 없이도 쓸 수 있는 통념이나 상식이다.
2	엉뚱한 답이거나 아이디어나 관점을 찾아볼 수 없다.
1	통찰이 전혀 없다.

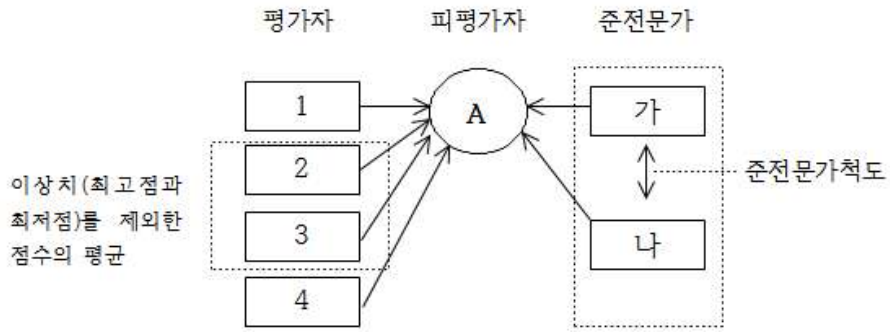
(그림 1) 7주차 채점 기준 중 ‘통찰’ 내용

분석 방법

산출된 통찰 차원의 동료평가 점수 데이터를 다양한 통계 처리 방법을 통해 검토하였다. 본 연구에서는 통계에서 일반적으로 사용되는 상관관계 분석을 포함하여 편차 검증을 준전문가척도와의 비교 및 동료평가자 내부에서 각각 실시하였다. 편차 검증은 동료평가 데이터가 많지 않을 때 활용 가능하며 상대적으로 용이하게 접근할 수 있기 때문에, 교육현장에서 쉽게 활용 가능하다고 판단하였다. 동료평가 점수 데이터는 다음과 같은 단계를 거쳐 분석하였다.

우선 동료평가의 정확성을 판단하기 위한 기준으로 준전문가 2명이 채점한 점수를 척도로 삼았다. 그런데 준전문가 2명의 점수가 어떤 상관관계를 가지고 있는가에 따라 정확도가 달라질 수밖에 없다. 준전문가척도의 정확성을 담보하기 위하여 준전문가 2명이 독립적으로 각자 채점한 점수의 상관관계를 상관계수(product-moment correlation coefficient)인 r을 이용하여 계산하고, 이 상관관계가 높으면 하나의 척도로 삼기로 하였다.

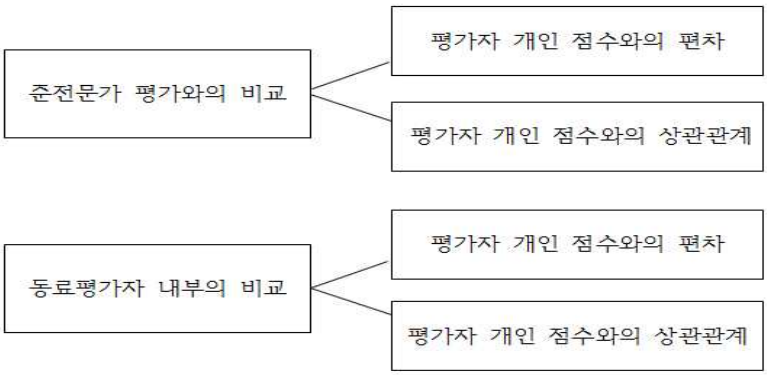
그런데 동료평가 결과의 정확도를 개선하기 위한 방법으로서 우선 고려해 본 것은 이상치(outlier) 제외 방식이었다. 통계에서는 평균 점수를 낼 때 이상치를 찾아서 그것을 제외한 점수를 활용하는 방법이 가장 일반적으로 사용되는 방법 중 하나이다. 이상치란 관측된 데이터의 범위에서 많이 벗어난 아주 작은 값이나 아주 큰 값을 말하는데, 이 연구의 데이터에는 4개의 값밖에 없으므로 그 중 최고값과 최소값을 제외한 평균을 이상치 제거 평균으로 설정할 수 있다. (그림2)와 같이, 한 명의 피평가자에게 4명의 평가자가 점수를 부여하므로 그 점수 중 최고점과 최저점을 제외한 두 점수의 평균을 내었다. 그리고 그 평균점수를 준전문가척도와 비교하여 도출된 수치를 계산하였다.



(그림 2) 이상치를 제외한 점수의 평균과 준전문가척도의 상관관계

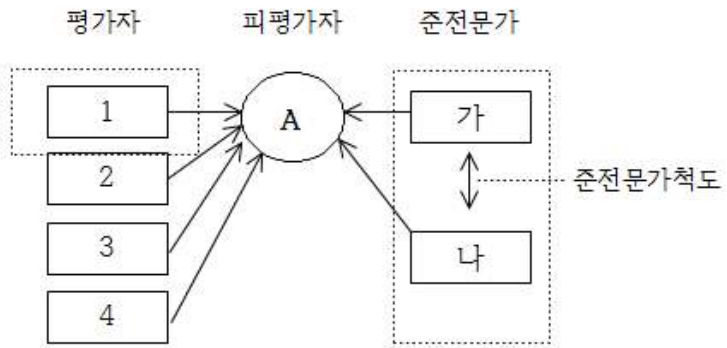
이상치를 제외한 점수의 평균을 사용하는 것은 일반적으로 많이 쓰이는 통계 처리 방법이므로 만약 이 방법이 동료평가의 정확도를 개선하는데 효율적이라면 충분히 활용할 수 있을 것이다. 그런데 이상치를 제외하는 방식 외에도 본 연구에서는 하위 평가자 그룹을 제외하는 방식을 주요하게 다루고자 하였다. 이상치를 제외하는 방식과 하위 평가자 그룹을 제외하는 방식을 각각 활용했을 때 동료평가 결과가 어느 정도로 개선되는지 비교하여 하위 평가자 그룹의 제외 방식이 사후적 통계 처치로서 얼마나 효율적인가 검증하고자 하였다.

본 연구에서는 '평가의 정확도 면에서 하위에 속하는 평가자 그룹'을 제외하는 방식으로, 준전문가척도와 비교하거나 동료평가자 내부에서 비교하는 두 유형에 각각 편차와 상관 계수를 활용하였다. 이를 그림으로 표현하면 다음과 같다.



(그림 3) 동료평가에서 학생 개인 점수의 정확도 산출 방법

첫 번째는 평가자의 점수를 준전문가척도와 비교하는 것이다. 클래스프렙에서는 학생 1명이 동료 4명의 글을 평가하게 된다. 즉 모든 학생들은 자신의 글에 대해 동료 4명으로부터 평가를 받게 되는 것이다. 이 때 한 학생이 동료 4명에게 부여한 각각의 점수를 준전문가척도와 비교할 수 있다. 편차를 이용할 경우 4개의 부여한 점수와 준전문가의 점수와의 차이를 계산하여, 평균값을 계산하는 것이다. 이것은 (그림 4)와 같이 피평자 A에 대해 평가자 1이 부여한 점수와 준전문가척도의 편차를 비교하였다. 이 때 평가자 점수가 준전문가척도와 차이가 크면 클수록 평가의 정확도가 떨어진다고 보았고, 그와 반대로 준전문가척도와 차이가 작을수록 평가의 정확도는 올라간다고 보았다. 이에 따라 준전문가척도와 편차가 큰 평가자를 순차적으로 제외시켜 나가면서 정확도의 차이를 비교할 수 있다. 준전문가와 비슷한 평가를 내릴수록 정확도가 높다고 판단할 수 있었다.

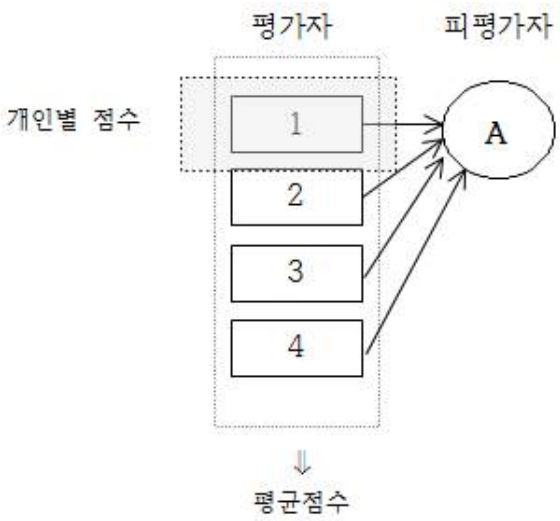


(그림 4) 평가자 점수와 준전문가척도의 비교

두 번째로 준전문가 척도와의 상관계수를 비교해 보았다. 상관계수를 이용할 경우, 한 학생이 동료 4명에게 준 각각의 점수와 준전문가의 점수와의 상관계수 r을 계산하였고 이때 r의 값이 0에 가까울수록 평가의 정확도가 떨어지고, 1에 가까울수록 높은 것으로 설정하였다.

세 번째로 평가자 점수를 동료평가 데이터 내부에서 비교해 보았다. 클래스프렙에서 한 학생이 동료 4명으로부터 평가 점수를 받았을 때, 평

가자 4명이 각각 부여한 점수의 편차를 살펴보는 것이다. 동료평가자 내부의 비교 방법은 (그림 5)와 같다. 이때에도 평가한 4개의 점수와, 예를 들어 평가자 1이 피평가자 A에게 준 점수를, A가 평가자 1~4에게서 받은 전체 점수의 평균과 비교하여 그 편차를 계산하였다.



(그림 5) 평가자 내부(동료평가평균)의 비교

이와 같은 동료평가 데이터 내부의 비교는 통계의 사후 처리 방법에 있어서 중요한 함의를 가진다. 준전문가척도 등과 같이 외부에서 주어지는 기준이 없을 경우 동료평가자 내부에서 평가의 정확도를 높일 수 있는 가능성을 모색해볼 수 있기 때문이다. 준전문가척도를 사용한 경우와 마찬가지로 이 경우에도 그 편차가 클수록 정확도가 떨어지는 평가자로 판단하고 순차적으로 제외시켜 나감으로써 오류를 줄일 수 있다.

마지막으로, 동료평가평균 점수와의 상관계수를 비교하는 것이다. 앞서와 마찬가지로 학생이 동료 4명에게 부여한 점수와 동료평가평균 점수와의 상관계수 r을 계산하였다. 이때 r의 값이 0에 가까울수록 평가의 정확도가 떨어지고, 1에 가까울수록 높은 것으로 설정하였다.

위와 같이 준전문가척도나 동료평가자 내부 평균이라는 두 유형을 사용하여 그 편차와 상관관계를 비교함으로써 제외할 수 있는 정확도 하위

평가자들의 선별이 가능하다. 통상적으로 상관관계를 많이 사용하나, 편차를 이용한 우수 평가자 그룹의 선별 방식은 비교적 소수의 데이터를 가지고도 분석이 가능하다는 점에서 유리하다.

이를 바탕으로 해서 제외 기준에 따라 동료평가 정확도가 어떻게 변하는지 살펴보기 위해 정확도 하위 평가자 그룹을 처음에는 하위 25%로, 그 다음에는 하위 50%로 순차적으로 확대해 나갔다. 즉 ①평가자 전체, ②정확도가 낮은 평가자 25%를 제외한 평가자 집단, ③정확도가 낮은 평가자 50%를 제외한 평가자 집단의 채점점수를 준전문가척도와의 상관계수를 이용하여 분석하였다.

결과

준전문가척도의 설정

전체 수업 중 7주차 수업에서 이루어진 동료평가를 분석하기 위한 기준으로 먼저 준전문가척도를 설정하였다. 준전문가 2명이 수강생 58명 전체의 글을 각각 평가하고 그 상관관계를 분석한 결과, 준전문가들이 채점한 점수의 상관계수는 r=.80로 상당히 높은 상관관계가 나타났다. 이는 준전문가들이 서로 유사한 채점 기준을 가지고 채점을 했다는 사실을 보여준다. 후속 작업으로 2명의 준전문가는 각자의 채점 점수에서 1점 이상 차이가 발생한 글들에 대해 서로 토의 과정을 거쳐 점수를 조정하였다. 그 결과 최종 상관관계는 r=.81로 비슷하였다. 준전문가 2명의 평가는 상관관계가 상당히 높은 것으로 검증되었으므로 준전문가척도로 활용할 수 있음이 판명되었다.

Outlier(이상치)를 제외한 평균과 준전문가척도와의 상관관계

평가자 4명이 부여한 점수 중 최고점과 최저점을 이상치로 보아 제거하고 나머지 2명의 점수의 평균을 내고, 그 점수를 준전문가척도와 상관관계를 계산한 결과 r=.55로 도출되었다. 이 수치는 전체학생들의 동료평가 점수를 포함하여 계산하여 도출된 r=.46과 비교해보았을 때, 통계적으

로 크게 유의미한 차이를 보이지 않았다.

하위 평가자 그룹을 제외했을 때 준전문가척도와의 상관관계

앞에서 제시한 네 가지 방법에 따라 준전문가척도와의 편차 및 상관관계, 동료평가 평균과의 편차 및 상관관계를 사용해서 준전문가 척도와 상관계수를 계산하였다. 그 결과는 다음 [표 1]과 같다.

[표 1] 연구 1에서 준전문가 척도와의 상관관계

	전체	하위 25% 제외	하위 50% 제외
준전문가점수			
편차	0.46	0.68*	0.79*
상관관계	0.46	0.67*	0.69*
동료평가평균점수			
편차	0.46	.57	.73*
상관관계	0.46	.69*	.7*

*p< 0.05, N=58

준전문가척도와의 편차

동료평가 점수와 준전문가척도와의 상관관계는 동료평가자 전체의 점수로 계산해보았을 때 .4 가 도출되어 그다지 높지 않았다. 그러나 준전문가척도와 편차가 컸던 하위 평가자 25%를 제외한 나머지 평가자의 채점 점수와 준전문가척도의 상관관계를 분석해 본 결과 그 상관은 .68(p= 0.03)로 향상되어서 모든 학생들의 포함했을 때의 상관계수와 통계적으로 유의하게 다른지 상관계수 차이의 검정을 실시하였다. 다음으로 하위 평가자 50%를 제외하고 분석했을 경우 상관이 .79(p<0.005)로 모든 학생을 포함했을 때보다 더 높아졌다. 우수 평가자 집단의 수를 상위 그룹 중심으로 더욱 좁혔을 때 준전문가척도와의 상관관계가 더욱 높아졌다.

준전문가척도와의 상관관계

동료평가 점수와 준전문가척도와의 상관계수에 대한 상관을 살펴보면, 역시 전체 동료평가자의 점수를 계산했을 때는 .4로 높지 않은 경향을 보였다. 그렇지만 하위 평가자 그룹 25%를 제외하고 분석했을 때는 .67로 상관이 처음보다는 유의미하게 증가하였다 (p=0.04). 하위 평가자 그룹 50%를 제외한 결과는 .69로 처음에 비해서 통계적으로 유의미하게 증가하였다 (p=0.03). 그러나 25%를 제외시켰을 때와 50%를 제외시켰을 때 각각 도출된 결과는 유의미한 차이를 보이지 않았다. 즉 우수 평가자 그룹의 범위를 더 좁혔으나 상관관계가 크게 향상되지는 않았다.

동료평가 점수 평균과의 편차

동료평가점수 평균과 편차가 컸던 하위 평가자 25%를 제외하고 다시 계산해 본 결과 상관관계는 .57로 약간 높아졌으나 통계적으로 유의미한 차이를 보이지는 않았다. 하위 평가자 그룹 50%를 제외했을 경우 상관은 .73(p=0.01)로 통계적으로 유의미하게 개선되었다. 적어도 동료평가 점수 평균과의 편차와의 상관관계에서 우수 평가자 그룹의 범위를 좁혔을 때 상관이 유의미하게 향상되었음을 알 수 있다.

동료평가 점수 평균과의 상관관계

앞에서와 마찬가지로 동료평가 점수의 전체 평균과 준전문가척도와의 상관관계는 .46라는 낮은 상관관계가 도출되었다. 이번에는 동료평가점수 평균과 상관관계가 작았던 하위 평가자 25%를 제외하고 계산해 본 결과 상관은 .69(p=0.02)가 나와서 통계적으로 유의미한 향상을 보였다. 그런데 하위 평가자 그룹 50%를 제외했을 경우 상관은 .7(p=0.02)로 25%를 제외했을 때와 비슷한 수준을 유지하였다.

논의

이 연구에서는 동료평가 결과의 정확도를 높이기 위한 사후 통계 처치 방법들을 비교 검토하여 어떤 방법이 가장 유용하게 사용될 수 있는지 살펴보았다. 먼저 동료평가의 정확도를 측정할 수 있는 기준으로 준전문가척도를 설정하였다. 그리고 통계에서 평균점수를 내기 위해 통상적으로 사용하는 방법으로 최고값과 최소값의 이상치를 제외한 평균과 준전문가척도와의 상관관계를 비교해 보았으나, 그 결과 동료평가의 정확도가 유의미하게 개선되지는 않았다.

다음으로 학생들의 동료평가 점수 정확도를 순위대로 정렬하고 그 중에서 하위 평가자 점수를 제외했을 때 동료평가의 정확도가 얼마나 개선되는지 비교해보았다. 준전문가척도와의 편차 및 상관관계, 동료평가 내부의 편차 및 상관관계를 각각 계산하고, 그 중 정확도가 낮은 하위 평가자 그룹 25%를 먼저 제외하고 다음으로 50%를 제외하는 방식으로 선별하였다.

먼저 준전문가척도와의 편차를 기준으로 하여 하위 25%의 평가자들을 제외했을 경우 .68(p= 0.03)이라는 유의미한 향상을 보였으며, 하위 50%를 제외했을 경우 .79(p<0.005)로 개선되었다. 즉 동료평가의 정확도는 준전문가척도와의 편차로 하위평가자들을 제외할 때 정확도가 전반적으로 개선되었음을 확인할 수 있다.

준전문가점수와의 상관관계도 하위 25%를 제외했을 때 .67(p<0.05)로 개선되었다. 그러나 우수 평가자의 범위를 50%로 더 좁혔을 때는 여전히 높은 상관을 보이기는 하지만, 하위 25%를 제외했을 경우와 큰 차이를 보이지 않았다. 두 경우 거의 비슷한 수준을 유지했다고 평가할 수는 있으나 우수 평가자 그룹의 범위가 좁아졌을 때 미세하게 상관관계가 낮아진 것이다. 그 원인에 대해서 좀 더 분석이 요구된다.

이처럼 준전문가점수와 비교할 때 동료평가의 정확도가 높을 것이라는 사실은 일반적으로 예상된다. 준전문가라고 하는 전문가에 가까운 일종의 비교 기준이 있기 때문이다. 그렇다면 준전문가의 평가 없이 동료

평가 자체만을 분석해서 정확도를 높일 수 있다면 여기에 더 주목할 필요가 있을 것이다.

다음으로 동료평가 평균 점수와 편차의 상관은 하위 25%를 제거했을 때 .4에서 .57로 향상되었다. 특히 하위 50%를 제거하자 .73(p<0.05)으로 큰 폭으로 상승하였다. 동료평가 내부의 편차를 살펴보는 방법에서 우수 평가자 집단을 활용했을 때 정확도가 향상되었다는 사실은 중요한 시사점을 제공한다. 교수자나 준전문가가 없을 경우에 동료평가에서 우수 평가자 집단의 평가 결과도 유의미한 평가도구가 될 수 있다는 사실을 보여주기 때문이다.

마지막으로 동료평가 평균 점수와 준전문가척도와의 상관관계 또한 25%를 제거했을 때 .69(p<0.05)로 향상되었다. 그러나 50%를 제거했을 때는 .7(p<0.05)로 여전히 유의미한 상관을 보이기는 했지만, 편차에서 드러났던 것만큼 큰 차이는 나타나지 않았다.

이상과 같이 전체적으로 동료평가 점수의 정확도는 하위 평가자 그룹을 제외하고 계산되었을 때 준전문가점수와의 비교, 동료평가 내부 비교에서 모두 상관관계가 높아져 개선되는 것을 확인할 수 있었다. 그러나 동료평가 평균 점수와 편차를 낸 경우를 제외하고는 전반적으로 하위 25%를 제거하든 하위 50%를 제거하든 큰 차이가 나타나지 않았다. 그 이유는 제외되는 학생 수가 많아지면서 3~4명이 아니라 1~2명의 평가 결과가 사용되었기 때문일 수 있다.

본 연구를 통해서 동료평가의 정확도를 향상시키려는 시도가 사후 통계 처리로 가능함을 알 수 있었다. 준전문가척도와 비교한 경우나 동료평가 내부에서 비교한 경우에 모두 편차와 상관관계 모두 유의미한 개선이 있었으며, 따라서 동료평가 자체만 가지고도 통계 분석을 통해 평가 결과의 정확도를 향상시킬 수 있다는 점에 주목할 수 있다.

4. 종합논의

최근 교육현장에서 학습의 한 방안으로 동료평가가 강조되고 있다. 특히 글쓰기에서 동료평가를 활용하면 학생들이 서로의 글을 비교, 검토, 평가함으로써 글을 평가하는 안목을 기를 수 있고, 글에 대하여 스스로 평가하는 능력을 신장시킬 수 있도록 도와준다. 이를 통해 학생들은 다른 학생의 사고과정을 반추하게 되어 글쓰기 능력을 발전시킬 수 있다.

동료평가가 성공하려면 학생들이 얼마나 적극적으로 평가에 참여하도록 하고 동료평가의 결과를 신뢰할 수 있게 해야 한다. 우선 동료평가에서 학생들의 참여를 더욱 적극적으로 이끌어내기 위해서는 동료평가의 결과를 성적에 반영할 필요가 있다. 이는 학생들의 학습 능력 향상에도 긍정적인 효과를 미친다. 동료평가를 통해 학습과 평가, 성적 간의 관계를 쉽게 연결시킴으로써 학생들이 학습목표를 설정하고 성취해나가는 과정을 내면화하는 데 도움을 얻을 수 있기 때문이다. 학생들의 적극적인 참여로 동료평가가 이루어지고 그로 인해 학습 능력이 향상된다면 동료평가의 결과를 성적에 반영하는 것은 타당하다고 여겨진다.

동료평가의 결과를 성적에 반영할 수 있다면 교수자의 수고를 덜어주는 부수적 효과도 발생된다. 교수자가 매주 학생들의 모든 과제를 점검해 줄 수 있다면 매우 이상적이지만 현실적으로는 시간적·물리적 제약이 많을 수밖에 없다. 따라서 동료평가 결과가 교수자 평가 결과와 어느 정도 비슷한 경향을 보인다면 동료평가의 중요도는 교수자의 수업 부담을 경감시켜준다는 측면에서 더욱 커질 것이다.

본 연구는 실제 대학 수업에서 모든 학생들이 매주 글을 쓰고 그에 대해 동료평가를 하도록 한 상황에서 얻어진 점수가 실제로 성적에 반영될 수 있는지 타당성을 알아보려는 의도가 전제되어 있다. 선행 연구에 따르면 여러 주 동안 실시된 결과를 누적했을 때 전공과목에서 교수자의 기말 보고서 평가와 동료평가의 상관관계가 비교적 높다는 것이 확인된

바 있다. 본 연구에서는 여러 주 동안 이루어진 평가의 누적이 아니라, 한 번의 평가 결과를 가지고도 동료평가의 정확성을 향상시킬 수 있는지 탐색해 보고자 하였다. 이것이 가능하다면 동료평가 결과를 더 적극적으로 성적에 반영할 수 있을 것이기 때문이다.

본 연구의 대상이 되었던 수업은 전공과목이 아니라 전공 탐색 과목이었기 때문에 심리학에 대한 사전 지식이나 관심 정도에 따라 학생들의 수준에서 상당히 큰 차이를 보였다. 더욱이 수강과제를 제출하는 학생들의 수도 매주 일정하지 않았다. 학생들의 동료평가 수준 또한 고르지 않고 큰 편차를 보였다. 그러므로 동료평가 결과 전체의 정확도가 높지 않은 경우 어떻게 성적에 반영할 수 있는지에 초점을 맞추었다. 문제는 다양한 수준을 가진 평가자 집단이 행한 동료평가에서 어떻게 하면 평가의 정확도를 향상시키는가 하는 것이었다.

이를 위해 동료평가의 결과를 비교할 대상으로 우선 준전문가 척도를 설정하였다. 준전문가 2명은 교수자를 대신하여 교수자와 비슷한 수준에서 평가를 할 수 있다고 보았고, 준전문가 내부의 높은 상관관계가 검증되어 준전문가 척도를 동료평가의 결과와 비교하여 정확도를 판단하는 기준으로 하였다. 평가의 정확도를 개선하기 위한 방법의 모색으로 우선적으로 평가 점수에서 최소값과 최대값을 이상치로 보고 제외한 결과를 준전문가척도와 비교하여 상관관계를 살펴보았다. 그러나 이상치 제외 기법은 정확도 개선에 있어서 하위 평가자 제외 방식에 비해 유의미한 결과를 도출하지 못하였다.

따라서 또 다른 방안으로 우수 평가자 전략(select-crowd strategy)에서 아이디어를 얻어, 학생들의 동료평가 능력 정확도를 순위로 정렬하고 그 중에서 상대적으로 정확도가 낮은 하위 평가자들을 제외시켜 나가는 방법에 주목하였다. 그리고 이 때 도출된 평가 결과를 준전문가 그룹의 평가 결과와 비교하여 얼마나 정확도가 높아지는가를 확인해 보았다.

하위 그룹을 제외하는 동료평가 결과의 통계 처리 방법은 동료평가의 문제점을 해소하는 장치이기도 하다. 동료평가가 가진 문제점으로 여

러 가지가 지적되고 있는데, 하위 학업 성취자가 주는 부정확한 피드백, 평가자로서 역할 수행에 대한 책임감 부족, 그로 인한 동료 평가자들의 능력에 대한 회의 등이 그것이다. 이러한 문제점들은 동료평가의 신뢰성과 타당성을 떨어뜨리는 결과마저 낳게 할 우려가 있다. 그 중에서 특히 하위 학업 성취자가 주는 부정확한 피드백은 동료평가의 정확도를 저해하는 요인이 된다. 다만 본 연구에서는 형성평가로서 동료평가의 긍정적 측면을 최대한 살리기 위하여 '하위 학업 성취자'가 아니라, 준전문가척도 또는 동료평가 점수 평균에서 차이가 큰 '평가의 정확도 면에서 하위에 속하는 평가자 그룹'을 제외하는 방식을 사용하였다.

이 연구에서는 하위 평가자 그룹을 제외했을 때 준전문가척도와의 편차 및 상관관계, 동료평가 평균과의 편차 및 상관관계에 있어 모두 동료평가 결과의 정확도가 개선되는 유의미한 효과가 확인되었다. 그런데 전반적으로 하위 25%를 제거하든 하위 50%를 제거하든 큰 차이가 나타나지 않았다. 그 이유 중 하나는 제외되는 학생 수가 많아지면서 3~4명이 아니라 1~2명의 평가 결과가 사용되었기 때문일 수도 있다. 이에 대하여 차후 더 면밀한 검증이 요구되어 진다.

또 하나 주목할 만 한 점은 동료평가 평균을 활용할 때 동료평가 정확도가 유의미하게 개선되었다는 점이다. 이와 같은 경향은 편차와 상관관계에서 모두 나타났으며, 특히 편차의 경우 우수 평가자 집단의 수를 줄일수록 정확도가 더욱 증가하였다. 이러한 결과는 교수자나 준전문가 평가가 이루어지기 어려운 경우에 동료평가가 그 자체로도 상당히 효과적인 통계적 처리 방법이 될 수 있음을 시사한다.

이상의 연구 결과는 동료평가를 실제로 성적에 활용할 수 있다는 점에서 중요한 시사점을 제공한다. 더욱이 가장 중요한 점은 동료평가가 이루어진 후에 통계적 처치를 어떻게 하는가에 따라 동료평가의 타당도, 정확도가 향상될 수 있다는 것이다. 본 연구에서 실행한 것처럼 정확도 하위 그룹을 제외하는 방법은 현재 동료평가 시스템에서 활용이 가능하며, 앞에서 언급한 여타 웹 기반 동료평가 시스템과도 차별성을 지닌다.

예를 들어 CPR 시스템에서는 학생들은 학습내용을 공부한 후 내용 전문가들에 의해 출제된 시험을 치르게 된다. 여기에서 얻은 점수로 각 학생의 이해 수준이 측정되어 수준이 높은 학생들의 동료평가 점수에는 더 높은 비중이 주어진다. 반대로 낮은 학생들의 동료평가 점수에는 상대적으로 낮은 가중치가 부여된다. 이러한 방식은 동료평가 정확도가 낮은 학생들의 동료평가 점수의 가중치를 0으로 설정하는 통계 처리 방식과 비슷한 측면이 있다.

그러나 클래스프렙을 활용하여 동료평가 정확도 하위 그룹을 제외시키는 방식은 별도의 시험을 볼 필요 없이 사후적인 통계 절차를 통해 정확도를 향상시킬 수 있으므로 학생과 교수자 모두의 부담을 덜어준다는 점에서 유용하며 활용도가 높다. 이 연구에서 하위 평가자를 제외하는 통계 처리는 연구자의 수작업으로 진행되었는데, 앞으로 클래스프렙 시스템 내에서 이러한 통계 처리가 이루어질 수 있도록 프로그램을 개발하여 적용한다면 클래스프렙의 활용도는 더욱 높아질 것으로 여겨진다.

한편 본 연구는 다음과 같은 한계점도 가지고 있다. 첫째, 이 연구의 대상이 된 수업은 심리학과 학부 전공탐색 수업이었다. 따라서 학생들의 수준 차이가 크기 때문에 하위 평가자를 제외하는 방식이 효과적이었다. 그렇다면 반대로 학생들의 수준 차이가 고른 전공수업의 경우 하위 평가자를 골라내기가 쉽지 않다는 문제가 있다. 게다가 하위 평가자들의 점수가 크게 변별되지 않는다면 통계 처치에서도 유의미한 결과를 얻지 못할 가능성이 있다. 따라서 전공수업에서는 하위 평가자 그룹을 제외하는 기법 외에 또 다른 통계적 처리 방안을 고민해볼 필요가 있다.

둘째, 본 연구에서는 한 차례의 수업만을 분석하였기 때문에 전체 수업에서 이루어진 동료평가의 결과를 반복 분석하는 작업이 요구된다. 7주차 수업 후반부에 해당되기 때문에 학생들이 동료평가의 기준과 절차, 내용에 어느 정도 익숙해져 있는 상태였다. 수업 초반부의 주차에 이루어진 동료평가는 학생들의 적응 정도가 높지 않을 것으로 예상되어 7주차와는 다른 양상을 보일 가능성이 크다. 만약 그렇다면 매주 산출되

는 동료평가 결과를 실제 성적에 반영할지 여부는 고려해 보아야 한다.

셋째로 매주 제시되는 과제 내용의 종류와 수준 정도에 따라 동료평가의 정확성 여부도 달라질 수 있다. 7주차에 제시된 주제인 '디자인'에 대해서 학생들은 비교적 친숙함을 느끼며 글쓰기를 하였다. 하지만 다른 주제도 비슷한 양상을 보이는지 다시 살펴보아야 할 것이다.

그리고 근본적으로는 효과적인 동료평가를 위해 평가자 훈련을 더욱 정확히 실시할 필요가 있다. 평가자 훈련은 실제 평가에 앞서 교수자가 학생들을 대상으로 평가 전반에 대한 안내와 연습을 하는 것이며, 이를 통해 평가자 간의 신뢰도를 향상시킬 수 있다. 훈련 과정은 평가 기준의 숙지, 평가 기준의 상세화, 실제 평가 연습, 불일치 채점에 대한 토의와 피드백 등으로 진행된다. 이 때 교수자가 직접 실제 자료를 보여주고 평가하는 모습을 보여줌으로써 더욱 효과적인 훈련을 할 수 있다. 반드시 교수자가 아니더라도 본 연구에 참여했던 대학원생이나 수업 참여 경험이 많은 학부생 등 준전문가들이 학생들의 동료평가를 지도하는 것도 좋은 방법이 될 수 있다.

이상의 한계점을 보완한다면 하위 평가자 그룹을 제외한 동료평가 결과를 활용하는 방법은 교양 수업과 같이 다양한 수준의 학습자가 다수 참여하는 교육 현장에서 통계적인 사후 처치 방법으로 활용될 수 있는 여지가 크다. 교수자나 준전문가가 직접 채점을 하는 것이 가장 이상적인 평가이겠으나, 그것이 불가능하거나 곤란한 상황에서도 정확한 평가를 내릴 수 있는 가능성을 열어주기 때문이다. 또한 이러한 통계 처리 프로그램을 클래스프렙 시스템 내에 적용한다면, 클래스프렙을 활용한 동료평가는 학습과 평가를 함께 구현할 수 있는 유용한 교육적 도구로서 그 지평이 더욱 확장될 것으로 기대된다.

참고문헌

김민정 (2005). 학습방법으로서의 동료평가 : 평가자 및 피평가자의 역할이 학습자의 초인지, 학업성취, 학습동기에 미치는 영향, **교육공학연구,** 21-4

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

서진원 (2005). 학교교육방법의 핵심장치로서의 학교도서관에 관한 연구, 구성주의 교수 학습이론을 중심으로. **한국문헌정보학회지**, 39(4), 163-175.

이명실. (2008). 대학 글쓰기 교육에서 '평가'방법 재고. **작문연구**, 6, 67-97.

이윤빈 & 정희모 (2014). 대학생 글쓰기에서 동료 피드백의 양상 및 타당도 연구. **작문연구.** 20 (3), 299-334.

오택환 (2010). 쓰기 수행평가에서 동료평가자간 신뢰도 분석, **국어교육연구.** 47, 91-116.

지은림 (2010). 교사의 형성평가 피드백 수행 척도 개발 및 타당화, **교육평가연구.** 23-1

홍경선 (2007). 학습자 중심 사회에서의 수업평가 모델 개발 연구, **학습자중심교과교육연구.** 7-2

Ballantyne, R., Hughes, K. & Mylonas, A. (2002). Peer assessment in large classes using an action research process: developing procedures for implementing. *Assessment & Evaluation in Higher Education*, Vol. 27(5).

Bostock, S. (2000). Student peer assessment. *Learning Technology.*

Boud, D. (2007). Reframing assessment as if learning was important. *Rethinking Assessment for Higher Education: Learning for the Longer Term*, London: Routledge.

Boud, D., & Dochy, F. (2010). Assessment 2020. *Seven propositions for assessment reform in higher education.*

Brown, S., Race, P. & Smith, B. (1997). *500 tips on assessment*. London: Kogan Page.

Brown, S., Rust, C., & Gibbs, G. (1994). Strategies for Diversifying Assessment in Higher Education. *Oxford: Oxford Centre for Staff Development.*

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology* 103(1): 73 – 84.

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

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

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 – 26.

Davies, P. (2006). Peer assessment: judging the quality of students work by comments rather than marks. *Innovations in Education and Teaching International*, 43/1: 69 – 82.

DeGrez, L., Valcke, M., & Roozen, I. (2012). How effective are self- and peer

assessment of oral presentation skills with teachers' assessments? *Active Learning in Higher Education,* 13(2): 129–42.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer, and co-assessment in higher education: A review. *Studies in Higher Education,* 24(3): 331–50.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. *Assessment & Evaluation in Higher Education*, 11: 146 165.

Fabos, B. & Young, M. (1999). Telecommunications in the classroom: Rhetoric versus reality. *Review of Educational Research*. 69:3 217–259.

Flower,L. Hayes, J.R. ,Carey, L. ,Schriver, K., & Stratman, J. (1986). Detection, diagnosis, and the strategies of revision. *College Composition and Communication*, 37, 16-55.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment and Evaluation in Higher Education*, 20, 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 peerassessment in CS1. *In Proceedings of the fifth international workshop on Computing educationresearch 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.

Holladay, J. M. (1990). Writing across the curriculum: annual report 1989-90. Michigan: Monroe County Community College (E.D. 326260).

Karpicke, J. D., &Roediger, H. L., III. (2008). The critical importance of retrieval for learning. The *Science*, 319, 966–968.

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

Larrick RP, Mannes AE, Soll JB (2011). The social psychology of the wisdom of crowds. Krueger JI, ed. *Frontiers of Social Psychology: Social Judgment and Decision Making* (Psychology Press, New York), 227–242.

Liu, NF. & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education,* 11(3): 279–90.

Mannes AE, Soll JB, & Larrick RP (2014). The wisdom of select crowds. *Journal of Personality Social Psychchology*. 107(2):276–299.

Orsmond, P., Merry, S. & Reiling, K. (2000). The use of student derived marking criteria in peer and self assessment, *Assessment and Evaluation in Higher Education*, 25(1), pp. 23–38.

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

Strachan, I.B. &Wilcox, S. (1996). Peer and self assessment of group work: developing an effective response to increased enrollment in a third year

course in microclimatology. *Journal of Geography in Higher Education*, 20, 3, 343 – 353.

Surowiecki, J. (2004). The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations. *New York: Doubleday.*

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

Rada, R. (1998). Efficiency and effectiveness in computer-supported peer-peer learning. *Computers and Education*, 30, 3/4, 137 – 146.

Abstract

Comparison of methods for increasing accuracy of peer review by eliminating lower ranking peer reviewers

Haesung Kim

Cognitive Psychology

The Graduate School

Seoul National University

Peer review is defined as reviewing other students' work and giving feedback on their work. This process is beneficial for both teachers and students. In addition to reducing the workload of instructors, peer review can help students to develop evaluation skills, improve their own writing, and actively participate in their own learning process. However, there concerns using peer generated scores. Both instructors and students worry about the possibility of low accuracy. These concerns are valid since students are novices in their disciplines with respect to both content knowledge and writing genre of the discipline. On the other hand, there are reasons to believe that peer review scores are just as valid as instructors' scores. Students are typically given a much smaller set to evaluate and thus can spend

more time on the evaluation of a given paper. To solve the accuracy problem, this research applied the idea of wisdom of select-crowds (Mannes et al., 2014). Rather than using all students' peer review scores for grading, scores that eliminated lower ranking students based on their review skills were used after peer review was conducted.

This research analyzed the peer review scores from a undergraduate psychology course assignment at Seoul National University. For the assignment, students read a psychology research paper and wrote a one page essay on a question given by a professor. After submitting the assignment, each student then reviewed four other students' paper. This process was all done using ClassPrep, an online peer view system. In order to measure the accuracy of peer review scores, two semi-experts graded every students' paper. Semi-experts' scores were compared to students' peer review scores. Five methods were compared to see if the accuracy improved. First method eliminated extreme scores out of four scores from the peers; the highest and the lowest score were eliminated then averaged. The remaining methods eliminated lower ranking student according to students' review skills. Second and third methods calculated the review skill from the difference and correlation score from semi-experts' grades. Fourth and fifth methods used averaged peer view scores instead of semi-experts' score to assess students' review skill. Using four different methods, students were ranked according to their review skills, then peer review score were calculated after eliminating lower 25 % and 50% student reviewers.

The result of eliminating outlier (maximum and minimun peer review scores excluded) did not improve the accuracy. However, eliminating lower ranking students according to their review skills methods did significantly improved the accuracy. Both the semi-expert score and the averaged peer review score methods improved the

accuracy. Using the average peer review score method is especially useful if an expert is unavailable to grade every students' work. However, there was no statistically significant difference between eliminating the lower 25% or 50%. Further replication studies are needed but this research shows that peer review accuracy can be improved eliminating lower ranking students after the peer review is performed. Therefore, this method can help using peer review in education settings by increasing the accuracy.

keywords : peer review, ClassPrep, accuracy, eliminating lower ranking

Student Number : 2014-22316