韓國數學敎育學會誌 시리즈 A <數學敎育>
https://doi.org/10.7468/mathedu.2018.57.3.215
2018. 08. 제 57권, 제 3호, 215-229.

J. Korean Soc. Math. Ed. Ser. A:
The Mathematical Education
August. 2018, Vol. 57, No. 3, 215-229.

증명 동료평가의 신뢰도 및 타당도 분석: 대학 정수론 수업의 사례를 중심으로

오예린 (서울대학교 대학원)†
권오남 (서울대학교)
박주용 (서울대학교)

Ⅰ. 서론

대학수준의 수학교육에서 증명학습의 중요성이 계속 강조되고 있는데, 정작 학생들은 증명학습에 어려움을 겪는다는 보고가 반복적으로 이루어지고 있다(Moore, 1994, Weber, 2001; Harel & Sowder, 2007; Stylianou, Blanton & Knuth, 2009). 증명학습에 어려움을 겪는 학생들을 지원하기 위한 시도 중 하나는 탐구기반학습(inquiry-based learning: IBL)이다. Smith와 동료들(Smith, 2006; Smith et al., 2009)에 따르면 IBL 방법을 사용한 수업을 들은 학생들이 전통적인 방법의 수업을 들은 학생들에 비해 증명의 수학적 아이디어에 더 집중한다는 것이 관찰되었다. Smith가 사용한 IBL 방법의 수업에서는 수업 전에 학생들로 하여금 미리 증명을 해결하도록 하였고 수업 시간에는 이미 해결한 문제에 대해 발표하고 토론을 하도록 하였다. 이 같은 수학적 의사소통은 증명학습을 향상시킨다. 같은 맥락에서 Weber & Mejía-Ramos(2014)도 학부 학생들이 수학자들과 비슷한 증명 관념을 갖게 하려면 학생들로 하여금 수학적 의사소통에 참여시켜야 한다고 주장하였다. 실제로 학생들로 하여금 수학적 의사소통에 참여하게 하기 위해 수업시간에 발표와 토론을 활성화하는 방법이 가장 흔하게 사용되고 있다.

그렇지만 성공적인 발표 및 토론식 수업이 항상 가능한 것은 아니다. 의사소통하기에 적합한 강의실 환경이 갖추어져 있고, 발표와 토론을 위한 시간이 확보되어야 하며, 강의자가 발표나 토론의 촉진자 역할을 잘 해낼 수 있으며, 대다수의 학생들이 적극적으로 참여할 때에만 성공적으로 이루어질 수 있다. 학생들로 하여금 토론과 발표에 적극적으로 임하게 하는 한 방법은 웹 공간을 이용하여 학생들로 하여금 예습을 해오게 하는 것이다. 학생 개개인이 가진 컴퓨터나 스마트폰을 통해 수업 전에, 자신의 증명을 익명의 글로 올리도록 하고 이를 서로 평가하도록 하는 것이다. 이렇게 하면 발표나 토론에 소극적인 학습자도 온라인 동료평가에는 적극적으로 참여할 수 있다.

동료평가란 학생들이 평가자가 되어 서로의 과제를 평가하는 활동을 지칭한다. 동료평가는 주로 대학 수준에서 글쓰기나 실습과제의 평가에 활용되고 있다(배수정, 박주용, 2016). 또한 동료평가가 개별 교수의 수업 개선을 가능하게 하며 학과나 대학차원에서 다양한 긍정적 효과를 가진다는 것이 확인되었다(신종호, 2014). 그럼에도 불구하고 동료평가가 증명학습에 사용된 연구를 찾아보기 어렵다. 증명을 쓰고 읽는 활동은 수학적 의사소통 과정으로, 증명 쓰기는 일종의 글쓰기 활동이며 증명 읽기는 일종의 읽기 활동이라는 점에서, 증명 학습에서 동료평가의 활용을 고려해볼 만하다. 다양한 글을 읽으며 좋은 글에 대한 인식을 가져가듯이, 다양한 증명을 읽어보는 것은 좋은 증명에 대한 학생들의 관점을 확립시키는 역할을 할 수 있다. 전통적인 대학 수학 수업에서 학생들의 증명 읽기 활동은 교과서에 제시되거나 교

* 접수일(2018년 2월 2일), 수정일(1차: 2018년 4월 24일, 2차: 2018년 7월 5일), 게재확정일(2018년 8월 29일)
* ZDM분류 : D75
* MSC2000분류 : 97C40
* 주제어 : 증명, 동료평가, 온라인 평가, 신뢰도, 타당도
† 교신저자
* 본 연구는 서울대학교 사회과학대학 융복합 연구의 지원으로 수행되었다.
* 본 논문은 오예린(2018)의 서울대학교 대학원 석사학위논문. 의 일부입니다.

수자가 수업시간에 제시하는 증명을 읽는 식으로 이루어진다. 수학 수업에서 동료평가를 도입하면, 학생들이 수업을 듣는 다른 동료학생들에 의해 작성된 증명을 읽을 수 있는 기회를 제공할 수 있다. 동료평가 활동의 주요한 특성 중 하나는 학생들로 하여금 평가자가 되는 경험을 하게 한다는 것이다. 이러한 경험은 완성된 증명을 읽고 이해하고 기억하는 경험과 달리, 아직 확실하지 않은 논증을 읽고 그 증명이 완성된 것인지, 증명의 흐름이 잘 읽히는지 등을 평가하는 기회를 제공한다. 즉, 학생들이 비판적으로 증명을 읽게 할 수 있다. 물론 동료평가는 증명 문제에 국한될 필요가 없다. 글쓰기와 증명 과정의 유사성에 착안하여 일단 증명에서 시작하지만, 다른 수학교과로 확장될 개연성은 매우 높다.

동료평가의 이런 장점에도 불구하고 실제 대학 수학 교육에서 잘 활용되지 않고 있다. 그 가장 큰 이유는 학생들의 평가를 신뢰하지 못하기 때문이다. 이러한 맥락에서 학생들의 평가 결과를 신뢰하고 사용해도 되는지를 검증하기 위한 노력들이 있어왔으나(Cho & Schunn, 2007; Stefani, 1994) 주로 글쓰기에 국한되어 있고, 증명 학습을 대상으로 한 연구는 찾아보기 어렵다. 이에 본 연구에서는 증명 장면에서 동료평가의 신뢰도와 타당도를 살펴보고자 한다. 구체적인 연구 문제는 다음과 같다.

1) 증명 과제에서 동료평가의 채점자간 신뢰도는 어떠한가?

2) 증명 과제에서 동료평가의 타당도는 어떠한가?

위 두 가지 연구 문제에 대답하기 위해서 대학 <정수론> 수업에서 한 학기 동안 온라인 동료평가를 수행하도록 하고 그 결과를 분석하였다. 정수론 과목은 수학 전공 학생들이 엄밀한 증명을 본격적으로 배우기 시작하는 과목이다. 첫 번째 연구 문제에 답하기 위해, 학생들이 주별로 수행한 동료평가 결과와 추가로 수집한 증명 평가과제의 결과를 이용하여 학생 채점자간 신뢰도를 조사하였다. 두 번째 연구 문제를 위해서는 학생들과 전문가가 수행한 평가 결과가 얼마나 유사한지를 조사하였다.

Ⅱ. 이론적 배경

1. 동료평가

동료평가는 학습자가 비슷한 지위를 가진 이들의 생산물이나 수행을 살펴보고 그것의 등급이나 가치 또는 질을 평가하는 활동이다(Topping, 2009). 객관성과 신뢰도를 확보하기 위해 대개 과제물 당 다수의 평가자가 교차 채점을 하도록 한다(한국교육평가학회, 2004). 동료평가는 오랜 역사를 갖고 있다. 1774년부터 1826년까지 글라스고 대학의 교수로 재직했던 George Jardine은 일찍이 글쓰기에서의 동료평가 방식과 그 이점을 설명하였고(Gaillet, 1992) 그 후 많은 실증적 연구가 이어졌다(O'Donell & Topping, 1998). 동료평가는 실제로 초·중·고등학교 모두에서 성공적으로 활용되어 왔다(Scruggs & Mastropieri, 1998).

동료평가가 학생들에게 미치는 긍정적인 효과는 학습적인 측면과 정의적인 측면으로 나누어 볼 수 있다. 먼저 학습적인 측면에서, 학습자는 동료의 과제를 분석하고 오류를 파악하는 등의 인지적, 메타인지적 활동들로 해당 교과 및 주제 영역에 대한 이해를 증진시킬 수 있다(Topping & Ehly, 1998). 또한 정의적인 측면에서, 학습자는 능동적인 학습참여를 통해 자기주도적 학습력을 신장시키며, 평가자로서의 주도권과 주인의식의 경험을 통하여 자아 존중감을 향상시키며, 부정적인 피드백에 대한 두려움을 덜 느끼게 되고, 교수자로부터 피드백을 받는 경우에 비해 피드백을 긍정적으로 받아들이게 된다(김정환, 조유미, 2006; Topping et al, 2000; Dochy, Segers, & Sluijsmans, 1999; Hunter & Russ, 1996).

동료평가가 제대로 이루어지면 학생들뿐만 아니라, 교수자도 수혜자가 된다. 채점 부담이 줄어들기 때문이다. 만약 학생들끼리 채점하고 피드백을 제공하는 것이 충분히 신뢰성 있고 타당하게 이루어진다면, 교수자는 채점에 할애하던 시간을 수업 준비 등의 다른 일에 사용할 수 있다. 뿐만 아니라 동료평가를 사용하면 학생들은 보다 즉각적인 피드백을 받을 수 있다. 교수자가 학생들의 모든 과제물을 채점하고 피드백을 제공하는 것에 비해, 학생들이 몇몇 동료의 과제물을 채점하는 데에 더 적은 시간이 소요될 것이기 때문이다.

동료평가는 현재 글쓰기 과제에 주로 활용되고 있는데 이는 수학에서의 증명으로 확장될 수 있다. 증명 쓰기를 일종의 글쓰기 활동으로 볼 수 있기 때문이다. 그

렇지만 이는 어디까지나 가능성이기에, 동료평가의 도입에 신중을 기하기 위해, 증명에 대한 동료평가의 신뢰도와 타당도 검증이 선행될 필요가 있다.

2. 평가의 신뢰도와 타당도

증명에서 동료평가의 신뢰도와 타당도를 조사하는 방법을 구체적으로 설정하기에 앞서 신뢰도와 타당도에 대해 간략히 소개하고자 한다. 중학교에서 동료평가를 활용하여 수행평가를 채점하고 그 결과를 통해 신뢰도와 타당도를 조사한 연구에서 강애남과 이규민(2006)은 수행평가 결과에 영향을 주는 주된 요인으로 채점자를 꼽으면서 채점자내 신뢰도나 채점자간 신뢰도에 대한 분석이 수행평가의 신뢰도 검토의 중요 요소라고 주장하였다. 채점자간 신뢰도는 서로 다른 채점자들의 채점 결과가 얼마나 유사한지를 의미한다. 특히 전체 학생이 모든 동료들의 과제물을 평가하는 경우가 아니고 하나의 과제물을 배정된 일부의 학생들만이 평가하는 경우에는 채점자간 신뢰도의 확보가 반드시 필요하다. 다시 말해 어떤 채점자가 배정되는지에 상관없이 유사한 평가결과가 나오는지에 대한 검증이 필요하다. 본 연구에서 시행한 동료평가에서는 시간적 제약으로 인해 모든 학생이 모든 동료학생의 과제를 평가하도록 하지 않고, 일부 학생들의 과제를 배정하였다. 이럴 경우 채점자간 신뢰도 확보가 더욱 중요하기 때문에 채점자간 신뢰도를 조사하였다.

평가 도구의 신뢰성도 중요하지만, 타당성이 결여되면 평가의 의미가 없어진다. 평가의 타당도란 “검사도구가 측정하고자 하는 능력이나 특성을 제대로 측정하고 있는 정도로, 검사목적에 따른 검사 도구의 적합성”을 의미한다(한국교육평가학회, 2004). 평가의 타당도를 확보하는 방법으로 내용타당도, 구인타당도, 공인타당도, 예언타당도 등을 확보·검증하는 방법이 있다(Cohen, Manion & Morrison, 2011). 그렇지만 동료평가에서의 신뢰도와 타당도는 통상적인 검사 도구에서 사용되는 의미와 달리, 신뢰도는 동료 평가자들 간 일치도로, 타당도는 동료 평가자들의 점수와 전문가 점수와의 일치도로 보고 있다 (Jeffery, Yankulov, Crerar, & Ritchie, 2016; Johnson, Penny, Gordon, Shumate & Fisher, 2005; Li, Xiong, Zang, Kornhaber, Lyu, Chung, & Suen, 2016).

이런 이유에서 본 연구에서는 증명 동료평가의 신뢰도와 타당도를 검증할 때에 Cronbach 알파 방법(Cronbach, 1951)을 사용하였다. 이 방법은 Pearson 상관계수 측정과 함께, 동료평가의 채점자간 신뢰도와 타당도 검증에서 가장 많이 사용되는 방법(Panadero, Romero, & Strijbos, 2013)으로 내적일관성으로써 신뢰도와 타당도를 계산하는 방법 중 하나이다. 이때 계산되는 계수를 Cronbach 알파 계수(Cronbach’s alpha coefficient)라고 부른다. 이 계수는 문항들 간의 내적 일관성의 척도이며 다중항목 척도(multi-item scales)에서 사용된다(Santos, 1999). Cronbach 알파 계수는 보통 0에서 1 사이의 값을 가지는데 때때로 음의 값을 가지기도 한다. 용인할 수 있는(acceptable) 신뢰도/타당도 수준에 대한 기준은 연구자마다 약간씩 다른데 보통 신뢰도가 0.67 이상이면 용인할 수 있는 수준으로 취급한다. Nunnaly(1978)는 알파 계수가 0.7이 넘으면 용인할 수 있는 신뢰도/타당도로 보았고 Bryman & Cramer(1990)는 0.8 이상이면 용인할 수 있는 신뢰도/타당도라고 주장하였다. 최근에 와서는 알파 계수를 통한 신뢰도/타당도 판단 기준의 세분화가 이루어졌다. 대표적으로 Cohen, Manion & Morrison(2011)은 알파 계수 측정을 통한 신뢰도 검증의 가이드라인을 [표 1]과 같이 제시하였다.

[표 1] 알파 계수 측정을 통한 신뢰도 검증의 가이드라인(Cohen, Manion & Morrison, 2011, p. 640)
[Table 1] Guideline for reliability test by measuring alpha coefficient(Cohen, Manion & Morrison, 2011, p. 640)

Cronbach 알파 계수	신뢰도
0.90 이상	신뢰도 아주 높음 (very highly reliable)
0.80 이상 0.90 미만	신뢰도 높음 (highly reliable)
0.70 이상 0.80 미만	신뢰할 만함 (reliable)
0.60 이상 0.70 미만	미미하게 신뢰할 만함 (marginally/minimally reliable)
0.60 미만	용납할 수 없게 신뢰도 낮음 (unacceptably low reliability)

Cohen, Manion & Morrison은 신뢰도의 정도를 신뢰도 아주 높음, 신뢰도 높음, 신뢰할 만함, 미미하게 신뢰

할 만함 그리고 용납할 수 없게 신뢰도 낮음의 다섯 가지 단계로 분류하였다. Nunnaly나 Bryman & Cramer가 용납할 수 있는 신뢰도인지 아닌지를 결정하는 기준을 제시한 데에 그친 것과는 차이가 있다. 본 연구에서는 알파 계수 측정을 통해 학생들 간의 채점 일치도인 신뢰도와 학생과 전문가 간의 채점 일치도인 타당도를 계산할 때에는 Cohen, Manion & Morrison의 가이드라인을 따르도록 하겠다.

Ⅲ. 연구방법

1. 연구참여자

본 연구에는 대학생과 전문가가 참여하였다. 대학생 참여자는 수도권 소재 대학교 <정수론> 과목의 수강생 25명 중 연구 참여를 희망한 학생들로 구성되었다. 해당 과목의 수강 신청 전에 수강생들에게 해당 수업에서 주별로 동료평가가 이루어질 것임을 안내하였다. 25명의 학생들 중 18명은 수학교육 전공 학생이었고 5명은 과학교육 전공 학생이었으며 나머지 2인은 인문계열 전공 학생이었다. 본 과목은 학부 2학년 대상 과목이었고, 수학교육 전공 수강생 중 대부분이 2학년 학생들이었다. 25명의 수강생 중 여학생은 2명, 남학생은 23명이었다. 본 연구에서는 자료를 수집할 때마다 별도로 수강생들에게 연구 참여 의사를 확인하였는데, 이 때문에 수집된 자료별 참여자 수가 달라졌다. 수집된 자료별 참여자 수는 [표 2]에 제시되었다.

[표 2] 자료별 대학생 연구 참여자 수
[Table 2] Number of student participants of each data collection

자료 분류	참여자 수
온라인 증명 동료평가	16명
증명 평가과제 1	20명
증명 평가과제 2	14명

연구자는 학기 말에 수강생들을 대상으로 온라인 설문을 통해 본인의 동료평가 활동 자료를 연구에 활용하는 것에 동의하는지를 확인하였다. 수강생 25명 중 총 16명의 학생이 동의한다고 응답하였다. 추후에 신뢰도를 구하는 과정에서는 본 설문에서 동의한다고 응답한 학생들의 동료평가 활동 자료만 활용되었다. 타당도 조사를 위해 실시한 '증명 평가과제 1'과 '증명 평가과제 2'에는 각각 20명, 14명이 참여하였다. 강제로 연구 참여하도록 하지 않기 위해 각 과제는 익명으로 진행되었다.

증명 동료평가의 타당도를 조사하기 위하여 전문가 2인에게 정수론 내용에 해당하는 10개의 증명에 대해 채점을 의뢰하였다. 동료평가의 타당도를 탐색한 연구에서 대개 전문가 평가 자료로 교수자 1인이 평가한 점수를 사용하였다(Cho, Shunn, & Wilson, 2006). 본 연구에서는 보다 정확한 전문가 평가 자료를 얻기 위해 전문가 2인이 매긴 점수의 평균을 전문가 점수로 사용하였다. 이들이 채점한 증명은 학생들이 참여한 '증명 평가과제 1'과 '증명 평가과제 2'의 내용과 같았다. 정수론을 큰 틀에서 대수학으로 보고 대수학을 전공하는 박사과정 학생들을 섭외하였다. 연구에 참여한 전문가 2인 중 1인은 대수학 박사 학위 소지자로 본 연구에 참여하였을 때에는 박사 학위를 취득한 직후였다. 나머지 1인은 대수학 전공 박사과정 학생으로 본 연구에 참여하였을 때에는 박사과정을 시작한 지 3년 정도 지난 시점이었다.

2. 주별 동료평가의 설계 및 절차

서울 소재의 한 대학교의 <정수론> 수업에서는 수강생들에게 예습을 위한 과제로써 서면과제와 온라인 동료평가를 수행하도록 하였다. 수학전공과목에서 학생들은 엄밀한 논리체계의 수학을 학습하게 되는데 이러한 논리체계에서 정리(theorem)와 증명이 핵심이 된다고 할 수 있다. 어떤 수학적 주장은 증명되었을 때에야 비로소 정리라고 불린다. 하나의 정리를 증명하는 방법은 여러 가지일 수도 있는데, 정리를 증명하는 방법을 배우고 익히는 것이 수학과 전공수업의 주된 목표 중 하나로 여겨진다. 해당 과목에서도 수학적 정리들의 증명 작성을 배우는 것이 주된 목표였다. 해당 과목의 수업은 학생들의 발표와 토론으로 진행되었는데 학생들은 수업에 앞서 미리 배정된 정리들의 증명을 스스로 작성하도록 요청받았다. 학생들에게 배정된 정리 중 주별로 4개의 정리가 동료평가에 활용되었다. 동료평가 시에 학생들은 정리별로 3명의 동료가 작성한 증명을 채점하였다. 채점할 때 학생들은 교수자가 미리 제공한 채점기준에 따라 증명의

논리성, 명료성, 참신성에 대한 점수를 부여하였다. 작성한 증명을 제출하고, 채점자에게 채점할 증명이 분배되고, 평가자가 평가를 수행하고, 평가된 결과가 피평가자에게 전달되는 등 증명 동료평가의 모든 과정은 Park(2017)이 개발한 온라인 동료평가 도구인 Classprep을 통해 이루어졌다.

1) 증명 채점기준 설계

연구자들은 학생들이 증명 동료평가 시에 활용할 수 있는 채점기준을 작성하였다. 글쓰기 과제를 위한 증명 평가 기준은 제시된 바 있으나(배수정, 박주용, 2016 등) 수학적 증명은 특수한 장르의 글이라는 점에서 증명 동료평가를 위한 별도의 평가 기준이 고안되었다. 이때 매주 서로 다른 4개의 정리에 대한 증명들을 채점해야 한다는 점을 감안하여 일반적인 채점기준을 추구하였는데, Brwon & Michel(2010), Moore(2016), Lavy & Shriki(2014) 등의 연구가 참조되었다. 해당 연구들에서 제시한 증명 채점 기준은 다음 [표 3]과 같다.

[표 3] 선행연구 및 본 연구의 증명 채점 항목
[Table 3] Proof evaluation criteria in literature and this research

Brown과 Michel (2010)	Moore (2016)	Lavy와 Shriki (2014)	본 연구
타당성	논리적 정확성	관례적인 증명 구조, 옳은 해답, 옳은 결론의 형성, 수학적 영역 사이의 연결 지식의 입증	논리성
가독성, 유창성	명료성, 유창성	쉽고 이해되는 해답, 문제에 대한 명확한 개요, 명확한 수학적 기호의 사용, 합리적인 조직과 명료성	명료성
		지적 도전성	참신성

먼저 Brown & Michel(2010)은 좋은 수학적 증명의 특징을 가독성, 타당성, 그리고 유창성 세 가지로 제시하였다. Moore(2016)는 이러한 기준에 다른 수학자들도 동의할 수 있는지, 다른 수학자들도 학생들의 증명을 채점할 때에 위 세 가지 기준으로 채점할지에 대해 의문을 가졌다. 그는 네 명의 대학 교수들을 인터뷰하여 학부 학생들이 작성한 증명을 채점해보도록 한 뒤에, 채점 기준을 묻고 좋은 증명의 특징이 무엇인지를 물어보았다. 네 명의 교수 모두가 좋은 증명은 논리적 정확성, 명료성, 유창성, 그리고 증명의 이해에 대한 입증을 지닌다고 대답하였다. Lavy & Shriki(2014)는 예비수학교사들이 기하 증명을 위해 만들어 낸 평가 요소들이 어떻게 변화하였는지에 대해 연구하였다. 이 연구에서 예비교사들이 만들어 낸 9가지 평가 요소는 위 [표 3]에 수록하였다.

연구자들은 논리성, 명료성과 참신성을 채점 항목으로 정한 뒤에 채점기준을 작성하였다. 작성된 채점기준은 다음 [표 4]와 같다.

[표 4] 증명 동료평가를 위한 채점기준
[Table 4] Rubric for peer assessment on proofs

영역	점수	설명
논리성	5점	논리적인 비약이나 오류가 없으며 증명이 완성되었다.
	3점	증명의 군데군데에 논리적인 비약 또는 논리적인 오류가 있다. 증명이 절반 정도 밖에 완성되지 않았다.
	1점	논리적인 비약이나 오류가 심하다. 또는 증명이 거의 완성되지 않았다.
명료성	3점	증명의 구성이 깔끔하다. 수학적 표현이 매끄럽다.
	2점	증명에 필요 없는 내용이 조금 들어있다. 수학적 표현이 약간 매끄럽지 않다.
	1점	증명에 필요 없는 내용이 많다. 수학적 표현이 매끄럽지 않다.
참신성	2점	참신하다.
	1점	그저 그렇다.

연구자와 교수자는 앞서 살펴본 선행연구들에서 논리성과 명료성이 참신성에 비해 많이 강조되고 있기 때문에 논리성과 명료성의 배점을 참신성보다 높게 책정하였고, 논리적으로 결함이 없어야 옳은 증명이 된다는 점에서 논리성 점수를 명료성 점수에 비해 더 높게 책정하였다. 그 결과 논리성, 명료성, 참신성의 배점을 각각 5점, 3점, 2점으로 설정하였으며, 보다 명확한 채점 기준을 제시하기 위해 영역별 선택지는 3개 이하로 설정하였다.

2) 온라인 동료평가 시스템

동료평가는 여러 효과와 장점이 있지만 동시에 여러

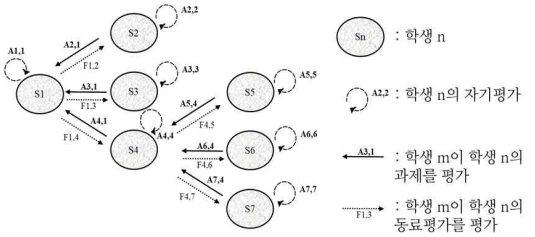
문제점과 우려사항이 지적되어 왔다. 무엇보다 학습자들의 채점에 대한 신뢰성이 확보되지 않았을 때에는 학습자들의 학습에 혼란을 일으킬 수 있다. Orsmond, Merry & Reilin(2000)의 연구에서는 부정확한 피드백을 받았을 때 평가를 받은 학습자는 불편한 감정을 느끼고 동료의 평가 능력을 불신하게 되었음을 확인하였다. 또한 평가자와 피평가자가 누구인지 공개되는 경우에 학생들은 환경적, 심리적으로 안전하지 않다고 느꼈다. 또한 평가자가 자신이 평가하는 과제를 작성한 사람이 누구인지 아는 경우 후광 효과로 인한 평가 결과 오염이 발생할 수 있고 이는 평가의 신뢰도를 떨어뜨리는 원인이 될 수 있다. 뿐만 아니라 수업 시간에 동료평가를 실시할 때에는 수업 시간 할애의 문제가 있다. 쪽지 시험이나 과제물을 교사가 채점하는 경우에는 수업 이외의 시간을 할애하여 채점할 수 있으나, 학생들이 수업 시간에 동료평가를 실시하는 경우에는 수업 시간을 할애할 수밖에 없다. 특히나 정기적인 동료평가를 실시하고자 하는 경우에는, 교수자가 이러한 수업 시간 할애에 부담을 느낄 수 있다.

온라인 동료평가의 도입은 이러한 여러 어려움을 해결하는데 효과적이다. 본 연구에서 사용된 온라인 동료평가 시스템인 Classprep은 채점할 과제물을 무작위로 평가자에게 배정해주고, 채점 결과를 다시 수집하여 피평가자에게 전달해준다. 모든 과정이 익명으로 진행되기 때문에 심리적 안정감을 높이고 후광효과를 쉽게 방지할 수 있다. 이뿐만 아니라 과제를 제출하고, 동료채점을 하는 모든 과정이 웹상에서 이루어지므로 수업 시간을 할애하지 않고 학생들이 개별적으로 수업 이외의 시간에 수행하도록 할 수 있다. 더불어 동료평가가 이루어진 이후에 평가자로부터 받은 점수와 피드백에 대한 평가 단계를 추가함으로써, 부정확한 평가가 이루어지는 것을 방지하고자 하였다.

Classprep에서 강좌를 개설하고 과제를 설정할 수 있다. 과제를 설정 시 과제 하나 당 평가자 수, 과제 일정, 평가 항목 및 채점점수 척도 등을 지정한다. 학생들은 개설된 강좌의 코드를 등록하면 글쓰기와 동료평가 활동이 가능해진다. 이 시스템에는 또한 학생들에게 과제를 수행하는 동안 생기는 질문을 입력하도록 하는 기능도 있다. 강의자가 과제를 등록하면 다음 세 가지 단계로 과제를 수행하게 된다. 먼저 과제 제출 단계로 학생들은 자신이 작성한 증명을 업로드한다. 그 다음은 동료평가와 자기평가 단계로, 자신에게 배정된 세 명의 동료학생들의 증명을 평가한 다음 자신의 증명을 평가한다. 마지막으로 피드백 평가 단계에서는 동료들이 자신의 증명에 대해 작성한 피드백을 보고 납득할 만한 평가인지, 피드백이 도움이 되었는지 등을 기준으로 동료의 피드백에 대한 평가를 한다.

해당 <정수론> 수업의 교수자는 학생들에게 주별로 15개가량의 정리의 증명을 작성하는 예습과제를 내주었다. 이에 따라 학생들은 매주 초에 그 주에 다룰 정리에 대한 증명을 손으로 작성하여 제출하였다. 서면으로 제출하는 예습과제와 더불어 온라인 동료평가가 이루어졌다. 주별로 배정된 15개가량의 증명들 중에 4개의 정리가 동료평가 과제로 지정되었으며, 학생들은 수업에서 해당 정리들을 다루기 직전 주에 Classprep을 통해 동료평가를 수행하였다. 학생들은 각 정리별로 배정된 3개의 증명을 평가하였다. 주별 동료평가는 학기 전체 15주 가운데 중간고사와 기말고사 기간을 제외한 11주에 걸쳐 이루어졌다.

한 개의 과제에 해당하는 동료평가 과정을 한 학생을 중심으로 도식화하면 [그림 1]과 같다.



[그림 1] Classprep 활동의 구조적 설명
[Fig. 1] Schematic structure of activity on Classprep

[그림 1]은 학생 4를 기준으로 동료평가의 구조를 나타낸 것이다. 학생 4는 학생 1의 증명을 평가하고($A_{4,1}$; 2단계 - 동료평가 및 자기평가) 학생 1은 학생 4의 동료평가에 대한 평가를 한다($F_{1,4}$; 3단계 - 피드백 평가). 또한 학생 4는 동료 자신의 증명을 평가한다($A_{4,4}$; 2단계 - 동료평가 및 자기평가). 한편, 학생 5, 학생 6, 학생 7은 학생 4의 과제를 평가하고($A_{5,4}$, $A_{6,4}$, $A_{7,4}$; 2

단계- 동료평가 및 자기평가) 학생 4는 학생 5, 학생 6, 학생 7의 동료평가에 대한 평가를 한다($F_{4,5}$, $F_{4,6}$, $F_{4,7}$; 3단계 - 피드백 평가). [그림 1]에서는 동료평가 활동의 모든 구조를 담기보다는 한 학생(학생 4)이 채점자인 동시에 피평가자인 것을 설명하고자 하였다.

3. 증명 평가과제의 설계 및 절차

1) 설계

학생들이 전문가와 유사한 평가를 하는지를 통해 학부 수업에서 실시하는 증명 동료평가의 타당도를 확인할 수 있다. 이를 위해, 연구자는 타당도 조사를 위하여 별도로 '증명 평가과제' 두 세트를 설계하여 학생들과 전문가에게 각각 평가하도록 하였다. 증명 평가과제를 하나의 정리에 대한 서로 다른 5가지의 증명으로 구성하기 위해 먼저 각각의 증명 평가과제에서 사용할 정리(theorem)를 선정하였는데 그 기준은 다음과 같다: (1) 수업시간에 비중 있게 다루어진 정리, (2) 다양한 증명이 가능한 정리. 먼저, 학생들의 지식 부족이 평가 결과에 영향을 미치지 않도록 하기 위해 수업 시간에 비중 있게 다룬 정리들 중 하나를 선정하였다. 또한 증명 평가과제에서 다양한 증명을 제시할 수 있도록 다양한 방법으로 증명이 가능한 정리를 선정하고자 하였다. 이러한 기준에 따라 '증명 평가과제 1'로 '소수의 무한성'에 관한 정리가, '증명 평가과제 2'로 '페르마의 소정리'가 선정되었다. 증명 평가과제에서 학생들에게 주어진 '소수의 무한성 정리(Infinitude of Prime Theorem)'와 '페르마의 소정리(Fermat's Little Theorem)'의 구체적인 서술은 [그림 2]에 제시되었다.

증명 평가과제 1: (Infinitude of Prime Theorem)

There are infinitely many prime numbers.

증명 평가과제 2: (Fermat's Little Theorem)

If p is a prime and a is an integer relatively prime to p, then $a^{p-1} \equiv 1 (\mathrm{mod}\ p)$.

[그림 2] 증명 평가과제에서 다룬 정리
[Fig. 2] The theorems used in Proof Assessing Tasks

소수의 무한성에 관한 정리와 페르마의 소정리는 각각 학기의 전반부와 후반부에 다루어진 정리들 중 핵심 내용에 해당하며 여러 가지 방법으로 증명이 가능하다. 따라서 이 정리들에 대한 증명을 평가하게 함으로써 학생들에게 <정수론> 과목의 핵심 내용에 관해 물을 수 있었으며, 다양한 증명을 제시하게 할 수 있었다.

연구자들은 증명 평가 과제 1, 2에서 다룰 정리를 선정한 다음, 각 증명 과제에 대해 가능한 여러 증명들을 작성하였다. 이때 학생들이 한 것과 비슷한 수행처럼 보일 수 있도록, 학생들이 실제로 동료평가 시에 제출하였던 증명들을 변형시켜 평가할 증명들을 구성하였다. 학생들의 증명을 변형시킨 이유는, 완전히 동일한 증명을 평가하게 되는 경우, 평가 시점에 상관없이 이전에 했던 것과 동일한 평가를 할 가능성을 배제하기 위해서였다. 이런 변형과 함께 평가할 증명들의 다양성을 확보하기 위해서 연구자가 의도적으로 작성한 증명도 있었고, 논리성과 명료성, 그리고 참신성 측면에서 다양한 증명을 다루기 위하여 증명들의 논리성, 명료성 그리고 참신성의 정도를 조정하고자 일부 답안을 수정하였다.

2) 절차

본 연구에서는 증명 동료평가의 타당도를 조사하기 위하여 정수론 수강생들과 전문가들에게 각각 증명 5개를 평가하도록 요청하였다.

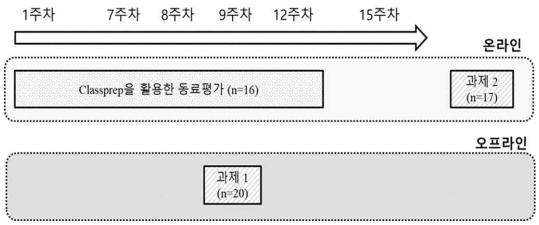
학생들은 학기 중반(8주차)과 학기가 끝난 이후에 각각 증명 평가과제 1과 증명 평가과제 2를 수행하였다. 연구자는 학생들에게 주어진 5가지의 증명이 동료가 작성한 증명이라고 가정한 상태에서 주어진 증명을 주별 동료평가에서 사용한 채점기준과 동일한 기준으로 평가할 것을 요청하였다. 또한 연구자는 학생들에게 해당 점수를 준 이유를 적도록 요청하였다. 증명 평가과제 1에는 총 20명의 수강생이 참여하였고 증명 평가과제 2에는 총 14명의 수강생이 참여하였다. '증명 평가과제 1'은 학생들이 소수의 무한성에 관한 증명을 배운지 얼마 지나지 않은 학기 중반에 이루어졌고, '증명 평가과제 2'는 학기가 종료된 뒤에 이루어졌다. 증명 평가과제 1은 오프라인에서 이루어졌으며 증명 평가과제 2는 온라인으로 진행되었다. 증명 평가과제 2는 학기가 끝난 이후에 이루어졌기 때문에 오프라인으로 자료를 수집하는 데에 어려움이 있어 온라인 설문조사의 형태로 자료를 수집하였다.

전문가 2인에게도 증명 평가과제 1, 2를 수행하도록 하였다. 수강생들에게는 “제시된 증명이 동료가 작성한 것이라 가정하고, 각 증명을 현재 Classprep 과제에서 사용하고 있는 채점기준을 적용하여 채점해주세요. 그리고 그렇게 채점한 이유를 써주세요.”라고 요청한 반면, 전문가들에게는 “제시된 증명을 학부생이 작성한 것이라 가정하고, 각 증명을 주어진 채점기준을 적용하여 채점하고, 그렇게 채점한 이유를 써주시길 바랍니다.”라고 요청하였다. 전문가들은 학생들이 동료평가 시에 사용한 채점기준에 익숙하지 않았기 때문에 연구자가 과제 수행 전에 이들에게 채점기준에 대해 설명해주었다. 전문가들의 답변은 전자우편을 통해 한글 문서파일로 수집되었다. 전문가들에게도 평가 시에 해당 점수를 준 이유도 함께 적도록 요청하였다.

4. 자료의 수집 및 분석 방법

1) 자료 수집의 절차

본 연구의 전체적인 자료수집 절차는 [그림 3]과 같다.



[그림 3] 자료 수집의 절차
[Fig. 3] The process of data collection

학생들은 1주차부터 12주차까지 7주차(중간고사 기간)를 제외한 11주에 매주 주별 온라인 동료평가 활동을 하였다. 7주차와 15주차에 중간고사와 기말고사가 실시되었으며 중간고사 이후와 기말고사 이후에 1, 2차 증명 평가과제를 실시하였다. 주별 증명 동료평가 자료와 증명 평가과제 1, 2의 결과를 사용하여 채점자간 신뢰도 검사를 실시하였고, 학생들이 작성한 증명 평가과제 1, 2 자료와 전문가가 작성한 증명 평가과제 1, 2 결과를 통해 타당도 검증을 실시하였다.

2) 신뢰도 검증

본 연구에서는 해당 수업의 학생들의 증명 동료평가 결과를 바탕으로 대학생들의 증명 동료평가의 채점자간 신뢰도를 살펴보았다. 먼저, 한 학기 동안 수집된 증명 동료평가 자료를 통해 3명의 채점자 간 신뢰도를 측정하였다. 더불어 추가로 수집한 ‘증명 평가과제 1’과 ‘증명 평가과제 2’의 결과 자료를 통하여 각각 20명과 14명의 채점자 간에 점수가 얼마나 일치하는지를 검사하였다. 모든 신뢰도 검사에서는 Cronbach 알파 테스트를 사용하여 급내 상관계수를 구하였다. 급내 상관계수 값은 보통 0에서 1 사이에 분포하며(Taylor, 2010) 추정 오차가 클 경우에는 음의 값이 측정되기도 한다. 자료들 간의 일치도가 높을수록 급내 상관계수 값은 1에 가깝다.

3) 타당도 검증

증명 동료평가의 타당도를 조사하기 위해 <정수론> 수강생들과 전문가를 대상으로 실시한 ‘증명 평가과제 1’과 ‘증명 평가과제 2’의 결과를 활용하여 수강생 집단과 전문가 집단 간의 증명 평가과제 점수의 상관을 조사하였다. 타당도를 보다 여러 가지 각도에서 살펴보기 위해 전체 점수뿐만 아니라 논리성, 명료성 그리고 참신성 점수들의 전문가 집단-학생 집단 상관도 조사하였다.

Ⅳ. 결과 분석 및 논의

1. 신뢰도 - 채점자간 신뢰도

1) 주별 동료평가에서의 신뢰도

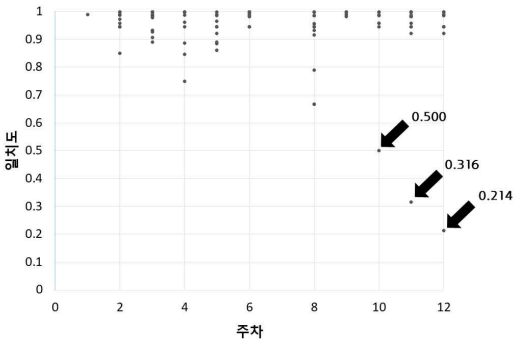
‘증명 과제에서 동료평가의 채점자간 신뢰도는 어떠한가?’라는 연구문제 1에 따라 채점자간 신뢰도를 분석하였다. <정수론> 수강생 25명 중 자신의 동료평가 활동 자료를 연구에 활용하는 것에 동의한 학생은 총 16명이었다. 이에 따라 연구자는 25명 학생 간에 이루어진 증명 동료평가 자료 중에서 16명의 학생들끼리 동료평가가 이루어진 자료만을 분석하였다. 즉, 자신의 동료평가 활동 자료의 사용을 동의하지 않은 9명의 학생이 평가한 결과나 평가받은 결과는 분석 시에 제외하였다. 9명의 학생의 자료를 제외하고 난 뒤 16명의 학생들이 받은 평가 중에 3명의 평가 결과가 모두 남은 경우만을 가지고 채점자간 신뢰도 조사를 실시하였다. 예를 들어 학생 A

가 학생 B, 학생 C, 학생 D로부터 평가를 받았는데 학생 B가 자료의 활용에 동의하지 않았다면 학생 A가 받은 점수는 채점자간 신뢰도 조사에서 제외하였다. 동료평가를 수행하지 않은 학생으로 인해 자료에 결손이 생겨, 한 학생이 2인 이하의 평가자로부터 평가를 받은 경우도 분석에서 제외되었다. 이렇게 조건에 맞는 자료들만 추려내자 과제별로 많지 않은 학생들이 받은 점수만이 남게 되었다. 그렇게 남은 점수들을 대상으로 알파 분석을 통해 급내 상관 계수를 측정하는 방법으로 채점자간 신뢰도 분석을 진행하였다. 다음 [표 5]는 총 143개의 표본에 대해 채점자간 신뢰도를 측정한 결과를 정리한 것이다.

[표 5] 주별 동료평가의 채점자간 신뢰도
[Table 5] The inter-rater reliabilities of weekly peer assessment

Cronbach 알파 계수	횟수	신뢰도
0.90 이상	129회 (90.2%)	신뢰도 아주 높음 (very highly reliable)
0.80 이상 0.90 미만	7회 (4.9%)	신뢰도 높음 (highly reliable)
0.70 이상 0.80 미만	2회 (1.4%)	신뢰할 만함 (reliable)
0.60 이상 0.70 미만	2회 (1.4%)	미미하게 신뢰할 만함 (marginally/minimally reliable)
0.60 미만	3회 (2.1%)	용납할 수 없게 신뢰도 낮음 (unacceptably low reliability)

0.6 미만의 값이 3회(0.214, 0.316, 0.500), 0.6 이상 0.7 미만의 값이 2회(0.667, 0.667), 0.7 이상 0.8 미만이 2회(0.750, 0.789) 0.8 이상 0.9 미만이 7회(0.846, 0.851, 0.861, 0.885, 0.889, 0.892, 0.892) 나온 것을 제외한 129회의 검사에서 0.9 이상의 값이 도출되었다. 이러한 결과는 하나의 증명에 대한 세 명의 평가자들이 준 점수들이 서로 매우 유사함을 의미한다. 다음 [그림 4]는 주차별 동료평가 표본들의 채점자간 신뢰도의 분포를 그래프로 나타낸 것이다.



[그림 4] 주차별 채점자간 신뢰도 분포
[Fig. 4] The scatter plot of inter-rater reliabilities

위의 그림을 통해 채점자간 신뢰도가 아주 낮은 표본들(0.500, 0.316, 0.214)이 10주차, 11주차, 12주차에 분포하고 있음을 확인할 수 있다. 동료평가를 통해 평가 훈련이 이루어지면, 학기말이 될수록 학생들이 보다 신뢰도 있는 평가를 해야 하는데, 학기 초에는 나타나지 않았던 아주 낮은 신뢰도 계수가 학기말에 소수의 표본에서 관찰된 것이다. 이에 대한 한 설명은 학기 후반에 이르면서 다루는 수학적 내용이 어렵고 복잡해졌고, 이에 따라 한 정리에 대한 다양한 증명이 가능해져 평가자가 자신이 평가하는 증명의 내용을 제대로 이해하지 못했기 때문이라는 것이다. 실제로 다른 영역에 비해 논리성의 평가 결과가 큰 차이를 보이는 것은 이를 방증한다.

2) 증명 평가과제 1, 2에서의 신뢰도

주별 동료평가 과제를 토대로 채점자간 신뢰도를 검사했을 때에 일치도가 대부분 0.9 이상으로 매우 높게 나왔으나 수집된 표본들의 크기가 모두 3으로 매우 작았다. 이처럼 작은 표본의 크기가 채점자간 신뢰도 측정에 영향을 주었을 수도 있기 때문에 보다 큰 표본에서 일치도 검사를 실시하였다. 수집된 '증명 평가과제 1'과 '증명 평가과제 2'의 결과를 통해 표본의 크기가 각각 20과 17일 때의 채점자간 신뢰도를 조사하였다. 주별 동료평가 자료의 일치도 검사 시와 동일하게 Cronbach 알파 분석을 통해 급내 상관 계수를 측정하였다. '증명 평가과제 1'과 '증명 평가과제 2'에서 학생 집단의 채점자간 신뢰도를 조사한 결과는 [표 6]과 같다.

[표 6] 증명 동료평가에서 학생 집단의 채점자간 신뢰도
[Table 6] The inter-rater reliabilities of student assessment in Proof Assessing Tasks

학생	전체 점수	논리성	명료성	참신성
증명 평가과제 1	0.943	0.963	0.943	0.915
증명 평가과제 2	0.970	0.727	0.854	0.915

채점자간 신뢰도 조사 결과, '증명 평가과제 1'에서는 전체 점수 뿐 아니라 논리성, 명료성 그리고 참신성 점수 모두 매우 높은 채점자간 신뢰도를 보였다. 이 결과는 학생들이 학기 중반에도 이미 서로 비슷하게 채점하고 있음을 시사한다. '증명 평가과제 2'에서는 전체점수, 논리성 점수, 명료성 점수, 참신성 점수 모두 채점자간 신뢰도가 다소 높게 나왔다. '증명 평가과제 1'에서는 모든 영역의 채점자간 신뢰도가 매우 높았던 것에 비해, '증명 평가과제 2'에서는 전체 점수의 채점자간 신뢰도는 매우 높았으나 각 영역 점수의 채점자간 신뢰도는 이에 비해 상대적으로 낮았다. 같은 증명을 평가할 때에 어떤 채점자는 논리성에 결함이 있다고 보고 논리성 점수를 낮게 주고 나머지 점수는 만점을 주고, 다른 채점자는 명료성에만 결함이 있다고 보고 명료성 점수는 낮게 주고 나머지 점수는 만점을 줄 수가 있다. 이때, 이 학생들이 준 영역별 점수는 차이가 있지만 이들이 부여한 총점은 같을 수 있다. 이러한 경우들로 인해서 '증명 평가과제 2'에서 전체 점수의 신뢰도가 각 영역별 점수의 신뢰도에 비해 더 높게 나온 것으로 보인다.

또한 '증명 평가과제 1'과 '증명 평가과제 2'의 결과를 비교해보면, 시간이 지나면서 전체 점수의 신뢰도는 소폭 상승했으나 논리성 점수와 명료성 점수의 신뢰도는 약간 감소했다. 시간이 지나면서 학생들 간의 채점 일치도가 증가할 것으로 예상한 것과 다른 결과였다. 그러나 '증명 평가과제 1'에서의 신뢰도가 워낙에 높았으며, '증명 평가과제 2'에서의 신뢰도 역시 높은 수치라는 점에서 이러한 감소는 큰 의미가 없다고 할 수 있다.

2. 타당도

본 연구의 연구문제 2에 해당하는 '증명 과제에서 등료평가의 타당도는 어떠한가?'를 조사하기 위해 <정수론> 수강생들과 전문가를 대상으로 '증명 평가과제 1'과 '증명 평가과제 2'에 대한 점수의 상관을 살펴보았다. 증명 평가과제의 학생-전문가 평가 점수의 상관 계수는 [표 7]과 같다.

[표 7] 학생-전문가 평가 점수의 상관 계수
[Table 7] The correlation coefficients between student assessment and expert assessment

학생	전체 점수	논리성	명료성	참신성
증명 평가과제 1	0.852	0.752	0.170	-0.018
증명 평가과제 2	0.824	0.628	0.856	0.745

분석 결과, '증명 평가과제 1'에서 전체 점수와 논리성 점수는 학생들의 채점 결과와 전문가의 채점 결과 간에 높은 정적 상관(상관계수 0.851)이 있었지만, 명료성 점수와 참신성 점수는 0에 가까운 상관계수(명료성 0.170, 참신성 -0.018)를 보였다. 즉, '증명 평가과제 1'에서 학생들의 논리성 채점은 전문가의 채점과 비슷하나, 명료성과 참신성은 전문가의 채점과 유사성이 거의 없었다. 명료성 점수와 참신성 점수 간에 유사성이 거의 없었음에도, 전체 점수 간의 유사도는 매우 높았다. 즉, 학생들과 전문가가 영역별로는 차이가 나는 채점을 하지만 이들이 부여한 영역별 점수를 합하였을 때에는 유사한 결과를 보였다. 이는 학생들이 명료성이 떨어지고 참신성은 높다고 채점한 증명을 전문가는 명료성은 높으나 참신성이 부족하다고 채점한 경우가 있기 때문인 것으로 보인다.

'증명 평가과제 2'에서는 전문가의 채점결과와 학생들의 채점결과가 전체 점수뿐만 아니라 논리성 점수, 명료성 점수 그리고 참신성 점수 모두에서 높은 정적 상관을 보였다. 이 결과는, 학생들이 학기 말에 이르러서는 전문가와 유사한 평가를 하게 되었음을 시사한다. 다만 전체 점수와 논리성 점수의 상관 계수가 소폭 감소하였는데 이는 '증명 평가과제 2'에서 다룬 정리와 증명이 포함하는 수학적 내용이 '증명 평가과제 1'에서 다룬 정리와 증명이 포함하는 수학적 내용에 비해 난이도가 높은 것에서 야기된 것으로 보인다.

Ⅴ. 결론 및 제언

1. 요약

본 연구에서는 대학생들의 증명 학습을 지원하는 방안으로써 온라인 동료평가를 설계하고 시행하였으며 증명 동료평가의 신뢰도와 타당도를 검증하였다. 본 연구의 연구 문제는 다음과 같았다.

1) 증명 과제에서 동료평가의 채점자간 신뢰도는 어떠한가?

2) 증명 과제에서 동료평가의 타당도는 어떠한가?

'연구문제 1'에 해당하는 신뢰도는 채점자간 신뢰도의 측면에서 다루어졌는데 학생들이 주별로 수행한 증명 동료평가 결과를 토대로 Cronbach 알파 계수를 사용하였다. 총 143회의 일치도 검사 결과, 129회의 검사에서 0.9 이상의 값이 도출되었는데 이러한 결과는 하나의 증명에 대한 세 명의 평가자 점수가 매우 유사함을 의미한다. 신뢰도 검사에서 학기말에 일치도가 아주 낮았던 데이터들(0.214, 0.316, 0.500)이 관찰되었다. 이는 학기 후반에 이르면서 다루는 수학적 내용이 어렵고 복잡해져, 한 정리에 대한 다양한 증명이 가능해지면서, 평가자가 증명의 내용을 제대로 이해하지 못하고 채점하였기 때문인 것으로 보인다.

위 신뢰도 조사에서 사용된 표본들의 크기가 3으로 매우 작았다는 점에서 결과를 일반화하는데 한계가 있어, 증명 평가과제를 통해 수집한 크기 20 그리고 17인 표본들 대상으로 학생 집단의 채점자간 신뢰도를 다시 조사하였다. 그 결과 학기 중반에 실시한 '증명 평가과제 1'과 학기말에 실시한 '증명 평가과제 2'에서 전체 점수뿐 아니라 논리성, 명료성 그리고 참신성 점수 모두에서 매우 높은 채점자간 일치도를 보였다. 이러한 결과는 학생들이 학기 중반부터 학기말까지 서로 유사도가 높은 채점을 하고 있었음을 시사한다.

'연구문제 2'에 해당하는 증명 동료평가의 타당도를 조사하기 위해서는 정수론 수강생들과 전문가를 대상으로 '증명 평가과제 1'과 '증명 평가과제 2'를 수행하도록 하였다. 각 과제는 5개의 증명을 평가하는 것이었다. 평가 점수를 이용하여 Cronbach 알파 계수가 산출되었다. 학기 중반에 실시한 '증명 평가과제 1'에서는 학생들의 평가와 전문가의 평가 결과의 영역별 일치도가 낮았지만 학기말에 실시한 '증명 평가과제 2'에서는 학생들의 평가와 전문가의 평가 결과의 영역별 일치도가 높았다. 전문가와 유사도가 낮은 증명을 하던 학생들이 학기 말에 이르러서는 전문가와 유사한 평가를 하게 되면서 증명 동료평가의 타당도가 높아진 것이다.

2. 결론 및 제언

한 학기에 걸쳐 시행된 증명 동료평가 결과를 분석한 결과 채점자간 신뢰도는 학기 전반에 걸쳐 매우 높았고 타당도는 학기 중반에는 낮았으나 학기 말에 이르러서 높아졌다. 채점자간 신뢰도의 경우, 다루는 수학적 내용이 복잡해지고 어려워지는 학기말에 이르러서 약간 낮아지는 현상이 나타났다. 한편 학기 중반에 타당성이 낮게 나온 것은 학생들이 평가 경험이 부족하여 좋은 증명에 대한 견해가 전문가와 다소 차이를 보이기 때문으로 보인다. 이처럼 학생들이 전문가와 유사한 평가를 하게 된 이유로 다음 두 가지를 생각해 볼 수 있겠다. 하나는 학생들이 증명 동료평가에 참여하면서 학습이 일어났고 그 결과 전문가와 유사한 역량을 기르게 되었기 때문이라는 것이다. 또 다른 이유는 한 학기 동안 학생들이 수업을 들으며 증명에 대한 학습을 했기 때문이라는 것이다. 즉, 학생들은 정수론 과목을 수강하면서 증명을 읽고 쓰는 등 증명 학습 활동에 참여하였기 때문이지 꼭 동료평가 때문이 아닐 수 있다는 것이다. 본 연구의 결과로는 이 두 가능성 중 어느 한 쪽을 지지할 근거가 충분하지 않다. 그 답은 동료평가를 하게 한 학생들과 그렇게 하지 않은 학생들의 수행을 비교하는 연구를 통해 검증될 수 있다. 그럼에도 불구하고 동료평가 활동이 학생들로 하여금 비판적으로 증명을 읽고 고민하도록 이끌었으며 수학적인 의사소통에 참여하도록 만드는데 기여한 부분이 있다는 점을 부인하기는 어려워 보인다.

한편, 동료평가가 평가도구로써의 활용도를 갖기 위해서는 타당도의 확보가 필수적이다. 그러나 타당도 분석한 결과, 학기 초에 실시한 타당도 검사에서 학생 집단과 전문가 집단 간의 평가 결과의 유사도가 낮았다. 즉 학생들은 학기 초에 타당도가 낮은 평가를 하고 있었다. 이러한 점에서 증명 동료평가의 결과를 학점 산출을 위한 점수 계산에 포함시키는 것이 타당하다고 보기는 어렵다. 비록 학생들이 학기말에는 전문가와 유사한 즉,

타당도가 높은 평가를 하게 되었지만 학기 초의 평가 타당도가 낮았음을 부정할 수는 없다. 따라서 증명 동료평가의 결과를 실제 평가 점수로 활용할지, 학기 후반에 이루어진 점수만을 평가에 반영할지 등은 여전히 논란의 여지가 있다. 또한 정의적 영역이 아닌 증명의 논리성과 명료성 등 인지적 영역에 동료평가를 실시할 때에는 학생들의 수학적 능력이 동료평가의 신뢰도와 타당도에 영향을 주기 때문에 다른 집단에서 동일한 환경의 동료평가를 실시했을 때에도 비슷한 결과가 나올 것이라고 보기는 어렵다.

'온라인' 동료평가의 장점 중 하나는 평가자와 피평가자가 서로 알 수 없기 때문에 보다 높은 신뢰도를 확보할 수 있으며, 학생들이 겪는 심리적인 갈등의 많은 부분을 해결해준다는 것이다. 그런데 본 연구에서는 익명성 확보에 약간의 문제가 있었다. 수학적 증명의 경우 보통의 글쓰기와 달리 수학적 기호를 포함하는 경우가 대부분이기 때문에 학생들이 수학적 기호나 수식을 입력하는 데에 어려움을 겪을 것으로 판단하고, 학생들에게 증명을 손으로 쓰고 이를 사진으로 업로드하는 것을 허락했는데, 수업 시간에 학생 발표가 많이 이루어지면서 학기말에 학생들이 서로의 글씨체를 알게 되었기 때문이다. 실제로 학생들이 작성한 동료평가 내용 중에 평가자 학생이 피평가자가 누구인지를 추측하고 실명을 언급한 경우도 있었다. 이처럼 손으로 작성한 증명을 사진으로 업로드하는 경우에 동료평가의 신뢰도가 다소 떨어지는 등의 문제점이 발생할 수 있다. 이러한 문제점은 온라인 동료평가 시스템에서 수식의 입력이 자유로워진다면 해결될 수 있을 것이다.

본 연구에서는 증명 동료평가를 실시한 사례를 분석함으로써 증명 동료평가의 신뢰도와 타당도를 조사하였다. 이를 통해 앞으로의 증명 동료평가의 활용에 대한 함의를 이끌어 내기는 하였지만, 증명 동료평가를 통해 학생들에게 어떠한 학습이 일어나는지 그리고 증명 동료평가의 경험이 학생들에게 어떠한 의미를 갖는지에 대해서는 탐색하지 못했다는 점에서 한계를 지닌다. 이런 한계를 극복하는 후속 연구가 조속히 이루어져야겠다.

참 고 문 헌

강애남, 이규민(2006). 학생들의 동료평가를 활용한 수행평가 결과의 일반화가능도 분석, 교육평가연구 19(3), 107-121.

Kang, A., & Lee, G. (2006). A Generalizability Theory Approach to Investigating the Generalizability of Performance Assessment Using Student Peer Reviews. *Journal of Education Evaluation 19*(3), 107-121.

김정환, 조유미(2006). 학습양식에 따라 평가유형이 수학적 성향과 문제해결력에 미치는 영향, 교육평가연구 19(2), 21-39

Kim, J. H., & Jo, Y. M. (2006). Effects of Evaluation Types according to Learning Styles on Students' Mathematical Disposition and Problem-solving Ability, *Journal of Education Evaluation 19*(2), 21-39.

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

Bae, S. J., & Park, J. Y. (2016). The validity of using cumulative peer assessed scores for final grades in college courses, *Korean Journal of Cognitive Science 27*(2), 221-245.

신종호(2014). 수업에 대한 형성적 동료평가 프로그램 사례 분석, 한국교원교육연구 31(3), 371-398.

Shin, J. H. (2014). Study on Formative Peer Review of Teaching Program in Higher Education, *Journal of Korean Teacher Education 31*(3), 371-398.

한국교육평가학회(2004). 교육평가 용어사전, 서울: 학지사.

Korean Society for Educational Evaluation (2004). *Educational evaluation thesaurus,* Seoul: Hakjisa.

Brown, D. E. & Michel, S. (2010). Assessing proofs with rubrics: the RVF method, *In Proceedings of the 13th Annual Conference on Research in Undergraduate Mathematics Education*, Raleigh, NC. Retrieved from http://sigmaa.maa.org/rume/crume2010/Archive/Brown_D.pdf.

Bryman, A. & Cramer, D. (1990). *Quantitative data analysis for social scientists,* Taylor & Frances: Routledge.

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.

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.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education,* New York: Routledge.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika 16*(3), 297-334.

Dochy, F., Segers, M., & Sluijsmans, D. M. A. (1999). The use of self-, peer and co-assessment in higher education: A review, *Studies in Higher Education 24*(3), 331-350.

Gaillet, L. L. (1992). A Foreshadowing of Modern Theories and Practices of Collaborative Learning: The Work of Scottish Rhetorician George Jardine. *Paper presented at the 43rd Annual Meeting of the Conference on College Composition and Communication*, Cincinnati, OH.

Harel, G. & Sowder, L. (2007). Toward comprehensive perspectives on the learning and teaching of proof, *Second handbook of research on mathematics teaching and learning*. Greenwich: Information Age.

Hunter, D. & Russ, M. (1996). Peer assessment in performance studies, *British Journal of Music Education 13*, 67-78.

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., Penny, J., Gordon, B., Shumate, S., & Fisher, S. (2005). Resolving Score Differences in the Rating of Writing Samples: Does Discussion Improve the Accuracy of Scores? *Language Assessment Quarterly 2*(2), 117 - 146.

Lavy, I. & Shriki, A. (2014). Engaging prospective teachers in peer assessment as both assessors and assessees: The case of geometrical proofs. *International Journal for Mathematics Teaching and Learning*. Retrieved from http://www.cimt.org.uk/journal/lavy2.pdf

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.

Moore, R. C. (1994). Making the transition to formal proof, *Educational Studies in Mathematics 27*(3), 249-266.

Moore, R. C. (2016). Mathematics professors' evaluation of students' proofs: A complex teaching practice, *International Journal of Research in Undergraduate Mathematics Education 2*(2), 246-278.

Nunnaly, J. (1978). *Psychometric theory,* New York: McGraw-Hill.

O'Donnell, A. M. & Topping, K. J. (1998). Peers assessing peers: Possibilities and problems. In K. J. Topping & S. Ehly (Eds), *Peer-assisted learning* (255 - 278). Mahwah, NJ: Lawrence Erlbaum Associates.

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), 23-38.

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort, *Studies in Educational Evaluation 39*(4), 195-203.

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

Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales, *Journal of extension 37*(2), 1-5.

Scruggs, T. E., & Mastropieri, M. A. (1998). Tutoring and students with special needs. In K. J. Topping & S. Ehly (Eds.), *Peer-assisted learning* (165-182). Mahwahm NJ: Lawrence Erlbaum Associates.

Smith, J. C. (2006). A sense-making approach to proof: Strategies of students in traditional and problem-based number theory courses, *Journal of Mathematical Behavior 25*, 73-90.

Smith, J. C., Nichols, S. R., Yoo, S., & Oehler, K. (2009). Building a community of inquiry in a problem-based undergraduate number theory course. In D. A. Stylianou, M. L. Blanton, & E. J. Knuth (Eds.), *Teaching and learning proof across the grades: A K-16 perspective*. New York: Routledge.

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

Stylianou, D. A., Blanton, M. L., & Knuth, E. J. (2009). *Teaching and learning proof across the grades: A K-16 perspective*. New York: Routledge.

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, J. K. (2009). Peer assessment. *Theory Into Practice 48*(1), 20-27.

Topping, J. K. & Ehly, S. (1998). *Peer assisted learning,* Mahwah, NJ: Lawrence Erlbaum Associates.

Topping, J. K., Smith, F. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students, *Assessment and Evaluation in Higher Education 25*(2), 149-169.

Weber, K. (2001). Student difficulty in constructing proofs: The need for strategic knowledge, *Educational studies in mathematics 48*(1), 101-119.

Weber, K., & Mejía-Ramos, J. P. (2014). Mathematics majors' beliefs about proof reading *International Journal of Mathematical Education in Science and Technology 45*(1), 89-103.

The Reliability and Validity of Online Peer Assessment on Proofs in a Number Theory Course

Oh, Yaerin†
Graduate School of Seoul National University
E-mail : ohyaerin@snu.ac.kr

Kwon, Oh Nam
Department of Mathematics Education, Seoul National University
E-mail : onkwon@snu.ac.kr

Park, Jooyong
Department of Psychology, Seoul National University
E-mail : jooypark@snu.ac.kr

Despite the importance of learning to do mathematical proofs, researchers have reported that not only secondary school students but also undergraduate students have difficulties in learning proofs. In this study, we introduced a new toll for learning proofs and explored the reliability and the validity of peer assessment on proofs. In the <Number Theory> course of a university in Seoul, students were given weekly proof assignments prior to class. After solving the proofs, each student had to assess other students' proofs. The inter-rater reliabilities of weekly peer assessment was higher than .9 over 90 percent of the observed cases. To examine the validity of peer assessment, we check whether students' assessments were similar to expert assessment. Analysis showed that the equivalence has been quite high throughout the semester and the validity was low in the middle of the semester but rose by the end of the semester. Based on these results, we believe instructors can consider the application of peer assessment on proving tasks as a tool to help students learn.

* ZDM Classification : D75
* 2000 Mathematics Subject Classification : 97C40
* Key words : proof, peer assessment, online assessment, reliability, validity
† Corresponding author