체육과학연구, 2016, 제27권, 제2호, 345-361.
Korean Journal of Sport Science
2016, Vol. 27. No. 2, 345-361.

Analysis of Error Sources and Estimation of Reliability in Peer Review of Forced Connection Method-Sportscasting by Applying Generalizability Theory

Tae-Koo Lee[1] & Heewon Yang[2]*

[1]Sangdong High School & [2]Yonsei University

The purpose of this study which is follow up study of Lee and Kim(2015b)'s was to analyse error sources and estimation of reliability in peer review of forced connection method-sportscasting by applying generalizability theory. Generalizability theory quantify error sources of the data measured under certain specific situation set by the researchers. It is an analysis method that the relative influences of each error sources taking from score is determined(G-study), and the effective measurement condition future applicable is provided(D-study). Participants were 10th high school students(N=216). Data were collected from student's peer review results and analyzed using univariate and multivariate generalizability theory. Results showed that error source for video have a more significant impact than other error sources. But the result by analyzing the gender difference was that error source for the interaction of video and participants have a more significant impact than other error source in the case of girls. Peer review used in this study showed high generalizability coefficient and even when reducing the number of video or participants it can maintain the adequate reliability. But generalizability coefficient of boys was higher than girls and specific measurement conditions leading to enhanced reliability were different when analyzing by gender difference. Also, method of analysis which cannot reflect measurement conditions properly estimates the reliability excessive. Discussions were provided in term of the relative influences of each error sources, the effective measurement condition maintaining the Generalizability coefficient of a certain level, and the comparison the Generalizability coefficient with the way of estimation traditional reliability applying univariate and multivariate Generalizability theory taking from score in peer review of forced connection method-sportscasting.

Key Words: Peer Review, Reliability, Generalizability Theory, Forced Connection Method-Sportscasting KISS

서 론

동료평가 방법은 각 집단에서 집단 구성원 간에 서로 평가하는 방법이다(Korean Society for Educational Evaluation, 2004). 동료평가는 국가교육과정에서 평가 주체의 다양화 측면에서 소개하고 있을 뿐만 아니라, 학생들 자신에게 교수학습과정에 대한 비판적 사고와 반성적 사고를 촉진하는 장점이 있으며, 체육교사는 학생들이 상호 평가한 동료 평가의 자료를 추후에 있을 체육수업의 유익한 환류정보로 활용할 수 있다(Ministry of Education, Science and Technology: MEST, 2008).

동료평가는 초·중·고등교육의 다양한 급에서 이론과 현장 적용에 대한 다양한 연구가 진행되어 왔다. 선행연구들은 동료평가의 이론 및 적용 가능성 탐색연구(Cho, 2000, 2004; Oh et al., 2006; Shin & Hong, 2013)을 비롯하여, 초등학생 대상(Kim & Choi, 2006;

논문 투고일 : 2016. 01. 31.
논문 수정일 : 2016. 03. 29.
게재 확정일 : 2016. 05. 11.
* 저자 연락처 : 양희원(yhw1627@naver.com).

Kim & Jo, 2006), 중등학생 대상(Bai et al., 2011; Hong & Lee, 2007; Lee et al., 2015; Oh et al., 2001; Yim, 2013)과 대학생 대상(Ahn, 2008; Cho et al., 2010; Kim, 2005; Kim, 2014; Kim & Kang, 2013)으로 연구가 폭넓게 진행되어 왔으며, 교원들의 수업전문성 향상(Min & Yun, 2012; Hong, 2010)을 위해서도 동료평가는 활용되고 있다. 동료평가는 선진 국가들에서도 학교 교육 질 개선을 위해서도 적극적으로 도입(Kim, 2014)되고 있으며, 좀 더 규모가 크게는 OECD에서 국가들 간에 국제개발 협력사업을 증진시키기 위해서도 국가 간 동료평가가 활용되고 있다(Youn & Kim, 2015). 선행연구들이 주목하고 있는 동료평가는 학생들의 수업참여를 유도하는 학생중심적인 교수학습 및 평가활동이면서(Lee et al., 2015), 학생들의 자아존중감, 학습태도 및 동기 등의 정의적 영역(Kim, 2005)뿐만 아니라, 학업성취에도 긍정적인 영향을 주는 장점(Kim & Jo, 2006)이 있으며, 교원들에게는 수업전문성 향상에 효과적이다. 그렇지만, 동료평가의 측정학적인 측면인 신뢰도 및 변별력에 대한 평가는 좋지 못하다. 왜냐하면 다양한 변인 즉, 경험부족(Oh et al., 2006), 친소관계(Bai et al., 2011), 관대화 경향(Lee, 2008), 성별(Hong & Lee, 2007; Oh et al., 2001) 등에 영향을 받을 수 있는 단점이 지적되고 있기 때문인데, 결국 동료평가 시 나타나는 관찰자 효과가 평가의 공신력을 떨어뜨리고 있다. 따라서 학교 현장에서 동료평가의 활용도는 낮은 상황이다(Bai et al., 2011).

동료평가는 체육교과 수업방법의 하나인 스포츠모의중계방법에서도 활용되고 있다. 스포츠모의중계방법은 이해-수행-감상의 교수학습단계에 따라 해당 스포츠종목의 이해와 수행단계릐 교수학습이 전개된 후 감상단계에서 실제 경기 동영상을 분석하여 대본을 작성하고, 실제 경기 동영상에 맞게 모의로 중계를 해 보는 체육수업 및 수행평가 방법이다. 지금까지 스포츠모의중계방법관련 선행연구들은 협력적 문제해결력을 함양하는 수업방법으로서의 스포츠모의중계수업 정체성과 관련한 이론적인 탐색연구(Lee & Lee, 2015b)와 비구조적 문제해결학습 측면에서 스포츠모의중계방법 평가도구 개발연구(Lee & Kim, 2015a)가 있었으며, 중등학생들을 대상으로 체육수업에 적용된 연구들로 배구 스포츠모의중계수업(Lee, 2015; Lee et al., 2011), 리듬체조 스포츠모의중계수업(Cha et al., 2014), 양궁 강제결합형-스포츠모의중계수업(Lee & Lee, 2015a), 컬링 강제결합-스포츠모의중계수업(Lee & Kim, 2015b) 등이 있었다. 그리고 일반 대학생을 대상으로 한 교양수업에서도 배구 스포츠모의중계수업(Lee et al., 2015)이 활용되었다. 다양한 스포츠 종목에 적용되고 있는 스포츠모의중계수업방법은 공통적으로 이해단계에서 해당 종목의 인지적인 이해를 바탕으로 수행단계에서 학생들이 기초기능부터 경기기능을 익히는 수업활동과 감상단계에서 이해단계와 수행단계에서 학생들이 학습한 종목을 통합적으로 이해하고 중계하는 일관된 교수학습형태를 보여 스포츠중계수업모형이라는 명칭으로 그 타당화가 시도되기도 하였다.(Lee & Lee, 2016)

스포츠모의중계수업방법을 적용한 체육수업 선행연구들에서 평가는 이해·수행·감상의 학습내용에 따라 그 특성에 맞게 다양하게 이루어졌는데, 이해단계 평가를 위해서는 일반적인 지필평가에 영상평가가 활용(Lee et al., 2011; Lee et al., 2015; Lee & Kim, 2015a)되기도 하고, 수행단계를 위해서는 기초기능과 팀별 리그경기결과 등이 평가로 활용되었다. 마지막으로 감상단계의 스포츠모의중계수업방법 평가를 위해서는 일반적으로 교사에 의한 발표평가가 이루어졌으며, 부분적으로 대본평가(Lee & Kim, 2015b; Lee & Lee, 2015a)와 학생들 간의 동료평가(Lee, 2015; Lee & Kim, 2015b)가 실천되었다. 동료평가가 실천된 구체적 방법으로 Lee(2015)는 먼저 반마다 네 명의 대표 동료평가자를 학생들이 선정하고, 각 반에서 학생들이 스포츠모의중계를 발표하면서 동시에 동료평가가 실시되었다. Lee & Kim(2015b)은 선행연구(Lee & Kim, 2015a)에서 개발된 루브릭을 수정·보완하여 컬링수업에서 학생들이 실천한 강제결합-스포츠모의중계 영상을 보면서 학급의 모든 학생들이 동료평가를 실천하였다. Lee & Kim(2015b)이 실천한 강제결합-스포츠모의중계[1]와 동료평가에서는 학생들이 사투리, 남북한 동시중계, 편들기중계, 우리말이 어설픈 외국인 테마에 따라 중계를

1) Lee & Kim(2015b)에서 실천한 강제결합-스포츠모의중계는 Lee & Lee(2015a)가 실천한 강제결합형-스포츠모의중계와 구별된다. 강제결합형에서는 학생들이 강제결합할 주제를 선정하지만, 강제결합에서는 주제가 교사에 의해 주어진다.

각색하여 창의적으로 중계를 실시하고, 교수학습과정에서 모든 모둠(반별 네 개 모둠)은 자신들이 실천한 강제결합-스포츠모의중계를 촬영하여 편집하고 각 모둠의 네이버 밴드(Naver Band)에 제출하였다. 그리고 교사는 모든 영상들을 강제결합 주제별로 분류하여 학생들은 모아진 영상들을 보며 동료평가를 하였다. 따라서 해당 고등학교 1학년 8개반 32개 모둠 216명 학생들이 각각 모두 동료평가에 참여하였다. 이렇게 모든 학생들이 동료평가에 참여하게 한 것은 동료평가가 정의적이고, 인지적으로 학습성취에 긍정적인 영향(Kim, 2005; Kim, 2014; Kim & Jo, 2006)을 주기 때문에 이를 이용하고자 한 것이다.

일반화가능도이론

동료평가는 다양한 장점도 있지만, 전통적인 선택형 중심 지필평가와 달리 평가 과정에서 여러 가지 요소들의 영향을 받는다. 평가 대상에 매겨지는 점수에는 평가 대상이 어떠한 것인지, 평가 대상이 몇 가지인지, 어떠한 하위 요인들을 평가하는지 뿐만 아니라 동료평가에 참여하는 학생들의 특성 및 학생 수, 평가 시기와 평가 절차 등 다양한 요소들과 학생들과 평가 대상의 관련성과 같은 요소들 간의 상호작용 효과 등이 영향을 준다(Lee & Shin, 2004). 따라서 동료평가의 신뢰성을 평가하기 위해서는, 전통적 평가에서 고려되지 않은 새로운 측정 국면을 측정 모형에 포함해야 하는데, 이때 측정 상황에서 발생하여 측정 결과에 영향을 줄 수 있는 여러 요소를 고려하고, 측정의 과정 및 결과를 일반화시키기 위한 개념적 틀을 제공하는 것이 일반화가능도 이론(Generalizability theory)이다.

일반화가능도 이론은 단일 오차원(sources of error)만을 고려하는 고전검사이론을 확대하여, 중다 오차원들을 동시에 고려하는 측정모형에 분산분석(ANOVA) 체계를 적용한 이론이다. 이를 통해 일반화가능도 이론은 측정상황에서 가능한 모든 오차요인을 포함하여 그 영향력을 분해한다. 결과 활용의 측면에서도 고전검사이론은 평가 점수의 표준오차와 신뢰도 계수에 초점을 맞추는 반면, 일반화가능도 이론은 평가 시기·평가자·평가 내용 등 오차 국면의 측정조건을 결정하고 각 오차요인(국면)의 영향력에 따라 그 수를 다르게 조정함으로써 신뢰도 계수 향상 방법을 다면화할 수 있다(Brennan, 2001a).

일반화가능도 이론은 크게 일반화 연구(G연구)와 결정연구(D연구)로 나뉜다. 일반화가능도 이론을 적용한 연구들은 일반적으로 G연구 설계에 따라 연구자가 설정한 각 국면에 대한 분산성분 추정값을 산출하여 오차요인의 수준의 신뢰도를 확보하기 위한 효율적인 측정조건을 제시하는 과정으로 분석을 수행한다. 구체적으로 G-연구 설계에서는 측정 상황에 따라 수집된 자료 형태가 교차(crossed)모형인지 내재(nested)모형인지를 고려한 국면을 설정하고, 각 국면이 무선효과(random effect)인지 고정효과(fixed effect)인지를 결정한다. 무선효과란 국면의 조건이 무한 전집에서 표집된 경우를, 고정효과란 유한 전집에서 표집된 경우를 의미한다. 여기에서 전집(universe)이란, 측정대상의 측정조건들에 일반화과정을 포함한다는 의미에서 기존의 모집단(population)의 의미와는 약간의 차이를 가진다(Lee et al., 2015; Brennan, 2001a). D연구 설계에서는 G연구 결과를 바탕으로 일반화하고자 하는 전집을 규정하며, 오차점수에 영향을 주는 국면의 조건 수에 따른 변화를 살펴본다(Brennan, 2001a).

일반화가능도 이론은 측정상황에서 발생하는 복합적인 오차요인을 동시에 분석하여 검사의 유용성(usefulness)을 높일 수 있다는 점에서 일반적인 지필검사(Kim et al., 2012)부터 수행평가(Kang & Lee, 2006; Kim et al., 2010), 심리검사(Lee et al., 2015)에 이르기까지 다양한 검사상황에 적용되어 왔다. 선행연구들에서는 일반화가능도 이론 적용을 통해 다양한 검사들에서 검사 점수에 가장 큰 영향을 주는 요인과 효율적인 측정조건을 밝히고, 그에 따라 평가의 신뢰도를 높이기 위한 타당한 방안을 제시하였다. 선행연구들(Kang & Lee, 2006; Kim et al., 2010; Lee et al., 2015)에서 일반화가능도 이론을 적용한 의미를 본 연구의 주제인 동료평가적인 측면에서 전체적으로 요약하면, 첫째, 교사가 단독으로 평가하여 낮은 신뢰도를 보이는 수행평가에서 학생들의 동료평가는 평가의 신뢰도 제고를 위한 현실적인 방법이 될 수 있다. 둘째, 동료평가에서 일반화가능도 이론 적용을 통해 일정 수준의 신뢰도를 확보

할 수 있는 평가 과제 유형 또는 평가요소 및 채점자 수 조정을 위한 근거 자료를 제시하여 학교 상황이나 그 외 상황에서 동료평가가 더욱 신뢰롭게 적용될 수 있는 검사 조건을 알 수 있다.

그러나 전통적 평가와 달리 채점자의 영향을 포함하고 있는 수행평가에서 일반화가능도 이론을 이용한 연구가 필수적인 요소로 인정되고 있음에도 불구하고(Kang & Lee, 2006), 학교 현장에서 적용되고 있는 수업방법과 관련한 수행평가에서 적용할 수 있는 동료평가에 대한 일반화가능도 적용 연구는 아직 충분히 이루어지지 않고 있으며, 최근에 동료평가 방법이 적용되고 있는 스포츠모의중계방법에서도 일반화가능도 이론을 적용한 연구는 실천되지 않고 있다. 결국 동료평가를 실시하기 위한 신뢰롭고 타당한 측정 조건들이 밝혀지지 않고 있는 가운데, 동료평가가 교사의 경험적인 판단에 의해 실시되고 있는 것이다.

따라서 이 연구의 목적은 학교현장의 교수학습 및 평가방법으로 활용되고 있는 강제결합-스포츠모의중계수업에서 학생들에 의한 동료평가 결과를 활용하기 위하여, 동료평가를 효과적으로 개선할 수 있는 방안을 일반화가능도 이론을 적용하여 탐색하는 것이다. 이 연구의 구체적인 연구문제는 다음과 같다.

첫째, 강제결합-스포츠모의중계수업에서 학생들의 동료평가 결과, 일반화가능도에 영향을 주는 요소들의 상대적인 크기는 어느 정도인가?

둘째, 강제결합-스포츠모의중계수업에서 학생들의 동료평가를 활용할 경우, 적정 수준의 일반화가능도를 확보할 수 있는 효율적인 측정 조건은 어떠한가?

셋째, 강제결합-스포츠모의중계에서 동료평가 시, 전통적인 신뢰도 추정방법과 비교하여 일반화가능도 이론에 의한 일반화가능도 계수가 적절한가?

연구방법

연구 대상

이 연구의 연구대상은 Lee & Kim(2015a)이 연구대상으로 한 경기도 행복고등학교(가명) 1학년 8개 학급, 216명(남:115, 여:101) 이었다. 연구대상은 컬링을 수업내용으로 한 체육수업[2]에서 학생들이 강제결합-스포츠모의중계영상을 보면서 동료평가를 실천하고 평가서를 제출하였다.

학생들이 사용한 강제결합-스포츠모의중계 발표 평가용 루브릭은 〈부록〉과 같이 개발하였다. 루브릭은 Jonassen et al.(2003)이 제시한 좋은 루브릭이 될 수 있는 기준을 참고하여 개발되었다. Jonassen et al.은 좋은 루브릭의 조건으로 첫째, 중요한 모든 요소를 포함할 것. 둘째, 각 요소는 1차원적일 것. 셋째, 등급이 뚜렷하고 포괄적이며 기술적(descriptive)일 것. 넷째, 학습자와 명확히 소통할 것 등을 제시하였다. 평가요소는 준비성, 연기성, 결합소재, 종목반영과 효과적 중계의 다섯 가지로 스포츠모의중계관련 선행연구들(Cha et al. 2014; Lee et al., 2011; Lee et al., 2015)에서 사용한 평가요소들을 참고하여 개발하였다. 평가요소의 등급은 3단계에서 7단계로 구성하는 것이 추천되는데(Jonassen, 2004), 본 연구에서는 학생들이 동료평가로 사용될 것을 고려하여 4단계로 설정하였다. 예를 들어, 준비성 평가요소에서 학생들은 0~3점을 득점할 수 있는데, 3점은 '스포츠중계에 어울리는 무대, 의상, 소품을 적절히 준비하였고, 아나운서와 해설자의 역할도 잘 알고 있다.'이며, 부분적으로 준비가 되지 못하면 낮은 점수가 부여되도록 평가기준이 개발되었다.

연구대상 학생들은 동료평가 전에 다음과 같이 3단계의 채점자 훈련을 받았다. 첫째, 연구대상 전체학생들을 대상으로 평가 루브릭과 그에 따른 평가요소에 대한 상세한 교육이 실시되었다. 둘째, 선행연구(Lee & Lee, 2015a)에서 제시되었던 양궁 프랑스요리사버전 중계영상을 대상으로 개발된 루브릭을 적용한 예비 동료평가를 실시하였으며, 이때 교사도 같이 평가하고 그 결과를 비교하였다. 셋째, 교사의 평가점수와 상이한 결과를 나타낸 학생들의 경우에는 평가기준에 대한 설명이 포함된 토의 및 질의응답을 반복적으로 실시하여 동료평가에 적용되는 루브릭을 통한 평가방법을 재교육하고 평가의 신뢰성를 높이고자 하였다.

2) 구체적인 수업내용은 2015년 11월 KSET 추계학술대회에서 발표한 「고등학교 체육수업에서 컬링을 활용한 거꾸로 수업 실천 사례」에서 확인할 수 있음.

루브릭에 기초한 평가요소별 학생들이 동료평가한 결과, 평가요소별 평균과 각 중계영상별 점수는 다음 〈Table 1〉과 같다.

Table 1. Mean for each measurement factors and scores for sportscasting images

Images	factors	Male	Female	Total
A	1	1.771	1.831	1.800
	2	1.875	2.034	1.951
	3	2.063	2.258	2.157
	4	1.917	2.079	1.995
	5	1.958	2.202	2.076
	Total	9.583	10.404	9.978
B	1	0.625	0.787	0.703
	2	0.458	0.461	0.459
	3	0.833	1.191	1.005
	4	0.750	0.989	0.865
	5	0.292	0.449	0.368
	Total	2.958	3.876	3.400
C	1	2.688	2.854	2.768
	2	2.677	2.888	2.778
	3	2.615	2.775	2.692
	4	2.615	2.865	2.735
	5	2.667	2.820	2.741
	Total	13.260	14.202	13.714
D	1	1.438	1.989	1.703
	2	0.760	0.854	0.805
	3	1.198	1.584	1.384
	4	0.948	1.247	1.092
	5	0.719	0.966	0.838
	Total	5.063	6.640	5.822

자료 수집 방법

연구 대상인 학생들은 각 반별로 네 개의 모둠으로 구성되었으며, 모둠들은 제비뽑기를 통해 네 종류의 강제결합-스포츠모의중계방법 중에 하나를 선택하여 실천하였다. 강제결합 주제는 사투리, 남북한 중계, 편들기중계, 우리말이 어설픈 외국인이다. 따라서 연구자가 지도한 8개 학급은 네 가지 주제에 따라 각각 한 개씩의 강제결합-스포츠모의중계를 실천하였으며, 지도 학년에서 주제별로 8개씩의 중계영상이 제작·제출되었다. 학생들이 영상을 제출한 후, 연구자인 체육교사는 강제결합-스포츠모의중계 영상들을 동일한 강제결합 주제별로 구분하여 한 번에 네 개 영상씩 평가하도록 재분류하였다. 학생들은 컬링 단원 후반부에 8개 반 32개 모둠에서 제출한 강제결합-스포츠모의중계 영상들을 루브릭에 따라 평가하였다. 평가 장소는 과학실을 활용하였는데, 과학실은 전면에 큰 스크린이 있고, 학생들이 넓게 혼자 앉을 수 있는 공간이 확보될 수 있어, 스크린의 화면 크기가 작은 교실보다 평가 장소로 적합하다고 판단하였기 때문이다. 학생들은 평가 후에 루브릭 평가서를 연구자에게 제출하였다. 본 연구에서는 학생들이 평가한 루브릭 평가서 중에 1차로 진행된 네 개의 사투리버전 강제결합-스포츠모의중계 영상에 대한 동료평가 결과가 분석대상으로 수집되었다.

자료 분석 방법

연구대상 216명 학생의 사투리버전 강제결합-스포츠모의중계 루브릭 평가서 중에, 1차 평가서가 분석대상이 되었다. 1차 평가서들 중에, 일부만 평가하였거나 성실하게 응답하지 않은 것으로 판단되는 총 31부의 평가서를 제외하고 모두 185부의 루브릭 평가서가 분석대상이 되었다.

자료 분석은 단변량 일반화가능도 분석과 다변량 일반화가능도 분석이 순차적으로 실시되었다. 여러 선행연구에서는 수행평가 또는 동료평가 결과를 일반화가능도 이론을 적용할 때 성별집단을 구별하지 않는 설계를 사용하여 왔다(Kim et al., 2010; Lee et al., 2015). 그러나 본 연구의 대상인 체육교과에서는 일반적으로 동료평가 상황에서 성별의 차이가 평가 결과에 영향을 주는 것으로 보고하고 있다(Hong & Lee, 2007; Oh et al., 2001). 따라서 본 연구에서는 선행연구에 따라 성별집단을 구별하지 않은 설계를 적용하는 단변량 일반화가능도 이론과 성별집단을 구별하는 다변량 일반화가능도 이론을 모두 수행하였다. 구체적으로 성별집단의 구분을 무시하는 $p\times(c:v)$ 설계와, 성별집단을 구분하는 $p^{\circ}\times(c^{\bullet}:v^{\bullet})$ 설계가 해당된다.

이 연구는 연구대상인 학생들이 중계영상을 루브릭에 따라 평가하였으므로, 이를 일반화가능도 분석 체계로 표현하면 $p\times(c:v)$ 설계에 해당한다. 이 모형은 각기 다른 다섯 개의 평가요소(c)에 따라 평가되는 네 개의 중계영상(v)이 있고, 이를 모든 평가자(p)가 평가하는 구조이다. 이 중 중계영상(v)을 측정 대상(object of measurement)으로, 평가요소(c)와 평가자(p)를 임의효과 국면(random facet)으로 설정하였다. 이를 도식화하면 〈Fig. 1〉과 같다.

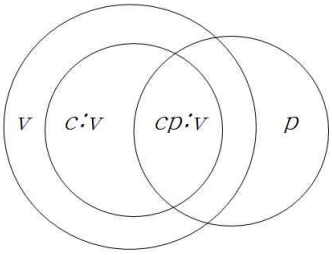

Fig. 1. Venn diagram for $p\times(c:v)$ design

단변량 일반화가능도 이론을 적용하여 G연구를 통해 분산성분을 추정하였고, D연구를 통해 효율적인 측정구조를 탐색하였다. D연구는 G연구와 동일한 설계인 $P\times(C:v)$의 설계를 적용하였고, 오차국면의 표집 수를 조절하여 학교현장에서 적정 수준의 일반화가능도를 유지하기 위한 효율적인 측정구조가 무엇인지 탐색하고자 하였다.

평가자(p)가 두 가지 성별집단 중 한 가지에 포함되는 다변량 일반화가능도 분석 체계는 $p^{\circ}\times(c^{\bullet}:v^{\bullet})$ 설계에 해당한다. 성별 국면(g)은 남자와 여자 두 가지로, 측정 절차를 반복해도 변하지 않으므로 고정효과(fixed effect)로 정의하며, 고정효과인 국면에 내재된 평가자 국면(p)은 열린 원($\circ$)으로, 무한 전집에서 표집하는 것으로 가정하는 무선효과인 나머지 국면들은 닫힌 원($\bullet$)으로 표시한다. 이를 도식화하면 〈Fig. 2〉와 같이 표현된다.

D연구는 $P^{\circ}\times(C^{\bullet}:v^{\bullet})$의 설계를 적용하였다.

단변량 일반화가능도 설계의 분산성분, 다변량 일반화가능도 설계의 분산성분과 공분산성분을 추정하기 위한 G연구 분석을 위해서는 mGENOVA(Brennan, 2001b) 컴퓨터 프로그램을 사용하였고, 이를 바탕으로 오차 국면의 수를 조정함으로써 효율적인 측정 구조를 탐색하기 위한 D연구 분석을 위해서는 엑셀의 매크로를 활용하였다. D연구 결과 산출되는 일반화가능도 계수와 비교하기 위한 Cronbach α 계수 산출은 SPSS 프로그램을 사용하였다.

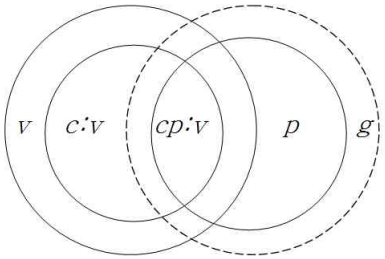

Fig. 2. Venn diagram for $p^{\circ}\times(c^{\bullet}:v^{\bullet})$ design

연구 결과

단변량 $p\times(c:v)$ 설계

G-연구 결과 전집점수(universe score) 분산과 오차점수 분산 성분의 크기를 상대적으로 비교하여 각 요인이 관찰점수(observed score) 분산에 미치는 영향력을 나타내는 분산성분 추정치는 다음 〈Table 2〉와 같다.

Table 2. ANOVA for $p\times(c:v)$ design

Effect(α)	df(α)	SS(α)	MS(α)	$\hat{\sigma}^2(\alpha)$	percentage
v	3	2303.4322	767.8107	0.8173	56.0
c:v	16	174.3892	10.8993	0.0572	3.9
p	184	548.1308	2.9790	0.0866	5.9
vp	552	688.1178	1.2466	0.1868	12.8
cp:v	2944	920.4108	0.3126	0.3126	21.4
Total	3699			1.4606	100.0

df(α): Degree of freedom for facet α
SS(α): Sum of squares for facet α
MS(α): Mean square for facet α
$\hat{\sigma}^2(\alpha)$: Estimated G-study variance component for facet α

측정대상인 중계영상 분산성분($\widehat{\sigma^2}(v)$)이 가장 크게 나타났고(0.8173), 다음으로 큰 분산성분은 잔차($\widehat{\sigma^2}(cp:v)$)였다(0.3126). 평가요소($\widehat{\sigma^2}(c:v)$) 및 평가자($\widehat{\sigma^2}(p)$)의 분산성분은 비교적 작았다(0.0572, 0.0866). 잔차의 분산성분이 전체 점수 분산의 21.4%의 비율을 보이는데, 이것은 일반화가능도 이론 연구에서 일반적인 현상으로, 잔차 분산성분에는 모형에서 구성된 요인들로 설명되지 않은 분산성분 부분이 모두 포함되기 때문이다(Kang & Lee, 2006).

다음으로 중계영상의 수를 고정시키고 중계영상을 평가하는 평가요소와 평가자 수를 변화시켜 일반화가능도에 미치는 영향을 알아보는 D-연구를 수행하였다.

평가자 수 및 평가요소 수는 선행연구 및 학교의 현장성을 고려하여 그 범위를 선정하였다. 이 연구의 연구대상인 평가자 수는 총 185명이었으나, 동료평가를 실시한 일부 체육교과 선행연구들(Bai et al., 2011; Hong et al., 2007; Oh et al., 2001)에서는 평가자 수가 2~4명이었다. 그리고 실제 학교 현장에서 학급에 적용할 상황 및 그룹별로 평가할 상황을 염두에 두어 평가자 수를 최소 1명에서 최대 30명으로 변화시켰다. 이 연구에서 루브릭의 평가요소는 5개였고, 선행연구들(Bai et al., 2011; Hong et al., 2007; Oh et al.,2001)에서는 각 평가영역별로 최대 5개 이하였다. 이는 선행연구자들이 학교 현장에서 중등학생들의 인지수준을 고려하여 5개를 초과하는 루브릭을 사용하는 것이 어려울 것으로 판단한 것으로 추측된다. 체육교과 대학생들을 대상으로 한 동료평가(Cho et al., 2010)에서도 평가요소의 수는 4개였다. 따라서 본 연구에서는 평가요소의 수를 최소 2개에서 최대 5개로 변화시키고, 이에 따라 일반화가능도 계수를 적정 수준으로 유지할 수 있는지 살펴보았다. 각 조건에서의 D-연구 표집수 및 분석 결과는 다음 〈Table 3〉에 제시하였다. 이 연구의 자료수집에 사용한 조건인, 평가요소 5개 및 평가자 185명일 때의 일반화가능도 계수는 0.9846으로 매우 높게 나타났다.

D-연구 결과, 평가요소 수 및 평가자 수가 많을수록 일반화가능도 계수가 높아지는 것을 확인하였다. 이를 그림으로 나타내면 다음 〈Fig. 3〉과 같다.

Table 3. D-study for $p\times(c:v)$ design in different measurement conditions

no. of factors	no. of raters	$\hat{\sigma}^2(\tau)$	$\hat{\sigma}^2(\delta)$	$E\hat{\rho}^2$
2	1	0.8173	0.3717	0.6874
	2	0.8173	0.2002	0.8033
	4	0.8173	0.1144	0.8772
	5	0.8173	0.0972	0.8937
	10	0.8173	0.0629	0.9285
	30	0.8173	0.0401	0.9533
3	1	0.8173	0.3101	0.7250
	2	0.8173	0.1646	0.8324
	4	0.8173	0.0918	0.8990
	5	0.8173	0.0773	0.9136
	10	0.8173	0.0482	0.9443
	30	0.8173	0.0288	0.9660
4	1	0.8173	0.2793	0.7453
	2	0.8173	0.1468	0.8478
	4	0.8173	0.0805	0.9103
	5	0.8173	0.0673	0.9239
	10	0.8173	0.0408	0.9525
	30	0.8173	0.0231	0.9725
5	1	0.8173	0.2608	0.7581
	2	0.8173	0.1361	0.8572
	4	0.8173	0.0738	0.9172
	5	0.8173	0.0613	0.9302
	10	0.8173	0.0364	0.9574
	30	0.8173	0.0198	0.9764

$\hat{\sigma}^2(\tau)$: Estimated universe score variance
$\hat{\sigma}^2(\delta)$: Estimated relative error variance
$E\hat{\rho}^2$: Estimated generalizability coefficient

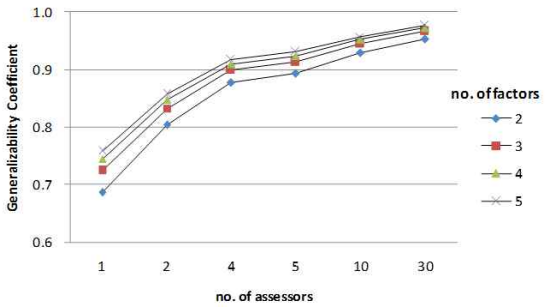

Fig. 3. Plot of Generalizabiliy Coefficient Using Univariate Generalizability Theory

평가의 적절한 일반화가능도의 수준은 연구자가 판단하며(Kang & Lee, 2006), 이 연구의 자료수집에 사용된 조건에서의 일반화가능도 계수가 0.9 이상으로 나타났으므로 0.9 이상의 일반화가능도 계수를 유지하기 위한 동료평가 상황을 확인하였다. 그 결과 영상마다 2개의 평가요소를 사용한다면 평가자는 최소 10명이 투입되어야 함을 확인할 수 있다. 평가요소가 3개인 경우에는 최소 5명, 평가요소가 4개인 경우에는 최소 4명이 필요하다. 그리고 평가요소가 5개인 경우에도 최소 4명이 필요하다. 즉, 평가요소가 많아질수록 일반화가능도를 0.90 수준으로 유지하기 위한 최소한의 동료평가자 수는 줄어드는 경향을 보인다.

다변량 $p^{\circ} \times (c^{\bullet} : v^{\bullet})$ 설계

다변량 $p^{\circ} \times (c^{\bullet} : v^{\bullet})$ 설계의 각 분산과 공분산성분에 해당하는 추정치, 해당 분산성분이 전체 분산에서 차지하는 비율, 그리고 성별 간 측정오차를 고려한 상관계수는 〈Table 4〉와 같다. 괄호(())는 성별 분산성분이 전체분산에서 차지하는 비율을 나타낸다. 분석 결과 남학생과 여학생의 분산성분 추정치에 차이가 있었다. 잔차($\widehat{\sigma^2}(cp:v)$)를 제외하고 남학생의 경우 측정대상인 중계영상 분산성분($\widehat{\sigma^2}(v)$)이 가장 큰 값(0.1173)을 나타냈으나, 여학생은 중계영상과 평가자의 상호작용($\widehat{\sigma^2}(vp)$)이 가장 큰 값(0.0802)을 나타냈다. 꺾은 괄호(〈 〉)안의 숫자는 성별 측정오차를 고려한 상관계수로, 높은 값을 보였다. 이는 남학생으로부터 높은 점수를 받은 중계영상은 여학생에게도 높은 점수를 받았음을 의미한다. 본 연구에서는 Brennan(2001a)과 Lee et al.(2015)이 제안한 방법에 따라, 분산성분 추정치의 값이 음의 값으로 산출된 경우 0으로, 측정오차를 고려한 상관계수가 1보다 큰 경우 1.0000으로 표시하였다.

G-연구 결과를 바탕으로 수행한 D-연구 결과는 다음 〈Table 5〉와 같다. G-연구와 동일한 국면의 조건 수로 분석한 결과 남학생과 여학생 집단의 합성점수에 대한 일반화가능도 계수는 0.9833으로 매우 높게 나타났으며, 이 값은 성별을 고려하지 않은 단변량 일반화가능도 계수(0.9846)보다 약간 작은 수치이다. 남학생 점수와 여학생 점수 각각의 일반화가능도 계수를 비교하면, 남학생이 0.9681, 여학생이 0.9550으로 남학생의 일반화가능도 계수가 조금 더 높다. 두 집단 모두 0.9 이상의 일반화가능도 계수를 보이지만, 남학생의 일반화가능도 계수가 조금 더 높은 것은 이 연구에 사용된 중계영상의 동료평가의 신뢰도가 남학생에게 좀 더 높다는 것을 의미한다.

이 결과를 바탕으로 평가요소와 평가자 수를 다양하게 변화시킨 D-연구 결과는 〈Table 6〉과 같다.

Table 4. ANOVA for $p^{\circ} \times (c^{\bullet} : v^{\bullet})$ design

Effect(α)	Male	Female
v	0.1173(29.9)	<1.0000>
	0.0750	0.0250(9.8)
c:v	0.0131(3.3)	
	-0.0004	0.0000(0.0)
p	0.0207(5.3)	
		0.0284(11.1)
vp	0.0887(22.6)	
		0.0802(31.3)
cp:v	0.1520(38.8)	
		0.1223(47.8)
Total	0.3918	0.2559

Table 5. D-study for $p^{\circ} \times (c^{\bullet} : v^{\bullet})$ design

	Male	Female	Total
$\hat{\sigma}^2(\tau)$	0.1173	0.0250	0.0748
$\hat{\sigma}^2(\delta)$	0.0039	0.0012	0.0013
$E\hat{\rho}^2$	0.9681	0.9550	0.9833

$\hat{\sigma}^2(\tau)$: Estimated universe score variance
$\hat{\sigma}^2(\delta)$: Estimated relative error variance
$E\hat{\rho}^2$: Estimated generalizability coefficient

D-연구 결과, 평가요소 수 및 평가자 수가 많을수록 일반화가능도 계수가 높아지는 것을 확인하였다. 이를 그림으로 나타내면 〈Fig. 4〉와 같다.

Table 6. D-study for $p^{\circ} \times (c^{\bullet} : v^{\bullet})$ design in different measurement conditions

no. of factors	no. of assessors	$\hat{\sigma}^2(\tau)$	$\hat{\sigma}^2(\delta)$	$E\hat{\rho}^2$
2	1	0.0731	0.0780	0.4835
	2	0.0731	0.0398	0.6475
	4	0.0731	0.0207	0.7796
	5	0.0731	0.0168	0.8128
	10	0.0731	0.0092	0.8884
	30	0.0731	0.0041	0.9471
3	1	0.0731	0.0661	0.5250
	2	0.0731	0.0336	0.6853
	4	0.0731	0.0173	0.8086
	5	0.0731	0.0140	0.8389
	10	0.0731	0.0075	0.9066
	30	0.0731	0.0032	0.9582
4	1	0.0731	0.0601	0.5486
	2	0.0731	0.0304	0.7059
	4	0.0731	0.0156	0.8240
	5	0.0731	0.0126	0.8525
	10	0.0731	0.0067	0.9160
	30	0.0731	0.0027	0.9638
5	1	0.0731	0.0565	0.5637
	2	0.0731	0.0286	0.7188
	4	0.0731	0.0146	0.8335
	5	0.0731	0.0118	0.8610
	10	0.0731	0.0062	0.9217
	30	0.0731	0.0025	0.9672

$\hat{\sigma}^2(\tau)$: Estimated universe score variance

$\hat{\sigma}^2(\delta)$: Estimated relative error variance

$E\hat{\rho}^2$: Estimated generalizability coefficient

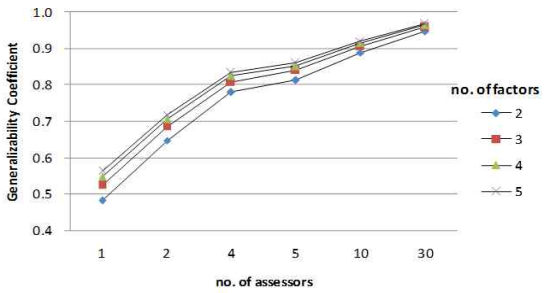

Fig. 4. Plot of Generalizabiliy Coefficient Using Multivariate Generalizability Theory

단변량 일반화가능도 연구에서와 같이 동료평가에서 유지하고자 하는 일반화가능도의 수준이 0.9 정도인 경우(Kang & Lee, 2006), 영상마다 2개의 평가요소를 사용한다면 평가자는 최소 30명이 투입되어야 함을 확인할 수 있다. 평가요소가 3개 이상인 경우에는 최소 10명이 필요하다. 평가요소가 많아질수록 일반화가능도를 0.90 수준으로 유지하기 위한 최소한의 동료평가자 수가 줄어드는 결과는 단변량 일반화가능도 분석 결과와 동일하지만, 다변량 일반화가능도 모형에 의한 분석 결과 더 많은 평가자 수를 요구하였다.

신뢰도 추정방법의 적절성

전통적인 신뢰도 추정방법인 Cronbach α와, 여러 오차요인을 고려하는 일반화가능도 이론을 통해 추정되는 신뢰도의 적절성을 비교하여 살펴보았다. 이를 위하여 일반화가능도 분석에 사용한 동료평가 자료로 Cronbach α를 산출해, 일반화가능도 계수와 비교하였다. 같은 검사 조건에서 Cronbach α와 일반화가능도 계수를 비교하면 다음 〈Table 7〉과 같다.

Table 7. Comparison for Cronbach α and Generalizability Coefficients

Cronbach α	$E\hat{\rho}^2$	
	Univariate	Multivariate
.998	.985	.983

$E\hat{\rho}^2$: Estimated generalizability coefficient

이 결과는 평가요소가 평가과제(중계영상)에 내재되어 있지 않은 측정 구조를 반영하지 않는 Cronbach α가 신뢰도를 과대 추정함을 보여준다. Cronbach α가 신뢰도를 과대 추정할 수 있다는 것은 많은 선행연구에서 지적된 것이다(Wainer & Wang, 2000; Lee et al., 2000; Yang & Lee, 2007). 이와 같이 전통적인 신뢰도 측정 도구인 Cronbach α는 평가요소가 내재되어 있는 것과 같은 측정구조에서는 신뢰도를 과대 추정할 수 있기 때문에, 일반화가능도 모형을 적용하여 일반화가능도 계수를 산출하는 것이 더 정확한 신뢰도 정보를 얻을 수 있다.

또한 평가자가 성별 집단에 내재되어 있는 측정 구조를 반영하지 않는 단변량 일반화가능도 계수가 다변량 일반화가능도 계수보다 신뢰도를 과대 추정함도 확인할 수 있다. 즉, 다양한 일반화가능도 설계를 적용하여 측정의 정확도를 높일 수 있음(Lee et al., 2015)을 확인하였다.

논 의

이 연구의 목적은 Lee & Kim(2015a)이 실천한 강제결합-스포츠모의중계수업의 동료평가 자료를 기초로 동료평가를 효과적으로 개선할 수 있는 방안을 일반화가능도 이론을 적용하여 탐색하는 것이다. 연구 결과, 첫째, 동료평가 결과에 기여하는 오차 요인들의 상대적 영향력은 대체로 동료평가의 평가대상인 영상이 가장 큰 것으로 나타났지만, 성별을 구분해서 분석한 결과는 다르게 나타났다. 둘째, 이 연구에 사용된 동료평가는 높은 일반화가능도계수를 나타냈으며 평가대상인 영상의 수 또는 평가자의 수를 감소시킬 경우에도, 적정수준의 신뢰도를 유지할 수 있었다. 그러나 성별을 구분해서 분석한 결과, 적정수준의 신뢰도를 유지할 수 있는 구체적인 측정 조건은 다르게 나타났다. 셋째, 측정구조를 적절하게 반영하지 못하는 신뢰도 계수는 신뢰도를 과대추정함을 확인하였다. 이러한 결과를 바탕으로 연구 문제와 관련하여 논의를 하면 다음과 같다.

첫째, 강제결합-스포츠모의중계에서 학생들의 동료평가 결과 일반화가능도에 영향을 주는 요소들 중 크기가 가장 큰 것은 평가대상인 중계영상인 것으로 나타났다. 중계영상의 전집점수 분산은 0.8173(56.0%)이며, 총 분산(1.4606)에 대해 상대적으로 중계영상의 분산이 크게 나타난 것은 동료평가 상황에서 평가대상인 중계영상 각각이 질적인 차이를 보인 것으로 해석할 수 있다.

중계영상 분산이 높고, 이에 비해 평가자의 전집점수 분산은 0.0866(5.9%)으로 낮게 나타났는데, 이 결과는 동료평가에 참여한 채점자들이 비교적 동질적으로 채점 점수에 주는 변동 요인이 적었음을 의미한다. 채점자들이 비교적 동질적이었던 이유는 사전에 동료평가에 참여한 채점자들을 훈련시킨 것과 관계가 있는 것으로 판단된다. 평가자의 전집점수 분산이 낮은 비율을 보이는 것과 관련하여 Kim et al.(2010)은, 채점자 요인이 수행평가에 영향을 주지만 훈련이 잘 된 채점자가 일관되게 채점한 경우 채점자 분산이 낮게 나올 수 있음을 지적하였다.

그러나 학교 현장에서 동료평가를 활용하고자 하는 교사들에게 채점자 교육은 부담으로 여겨져, 동료평가 자체를 교사들이 기피하게 하는 원인이 될 수도 있다. 그러나 Kang & Lee(2006)는 채점자가 채점자 집단에 내재된 측정구조에서 채점자 집단의 분산성분이 비교적 작게 나타난 결과는 채점자 집단을 무작위로 배치한 것과 관계가 있는 것으로 판단하였기 때문에, 이는 채점자 교육 없이도 채점자와 관련한 분산성분을 축소할 수 있는 하나의 방안으로 볼 수 있다. 따라서 학교 현장에서 동료평가를 시행하려고 하는 교사가 채점자 교육을 시간 부족 또는 업무적인 부담 등으로 실시하지 못할 때에는 평가 구조를 바꾸어 채점자 교육과 유사한 효과를 유도하는 것도 학교 현장성을 고려한 차선책이 될 수 있을 것이다.

남학생은 성별을 고려하지 않은 단변량 일반화가능도 분석 결과와 같이 중계영상의 분산이 0.1173(29.9%)으로 높은 편이었지만, 여학생은 평가대상인 중계영상과 평가자의 상호작용 분산이 0.0802(31.3%)로 높게 나타났다. 여기에서 평가대상과 평가자의 상호작용이란, 평가대상에 따라 평가자의 채점에 차이가 있음을 의미한다. 즉, 여학생들은 남학생들보다 평가영상에 대해 의견이 고르지 않았다는 것을 의미한다. 일반적으로 신체능력은 성별에 따라 차이가 있기 때문에, 신체활동을 기반으로 교수학습을 진행하는 체육수업에서는 성별이 수업에 영향을 주는 주요한 요소이다. 따라서 동료평가관련 체육교과 선행연구들(Bai et al., 2011; Hong & Lee, 2007; Oh et al., 2001)에서도 성별 요인이 평가 결과에 영향을 주는 요인으로 나타나고 있다. 이렇게 체육교과 동료평가에서 여학생들이 일관된 평가 경향을 보이지 않는 것은 선행연구들에서 그 원인을 추측해 볼 수 있다. Bai et al.(2011)은 중등 여학생들이 친한 친구들에게 관대한 점수를 주는 경향이 있다고 하였으며, Hong & Lee(2007)는 여학생들이 남학생들에 비해 관대하게 평가하는 경향이 있다고 하였다. 이러한 선행

연구들의 결과는 본 연구결과에서 나타난 남학생들보다 일관되지 않은 여학생들의 채점 경향을 이해할 수 있는 단초가 될 수 있다. 이러한 여학생들의 동료평가 경향은 평가의 신뢰성을 저해할 수 있다. 그럼에도 불구하고 현장교사들이 동료평가의 다양한 교육적 유익을 고려하여 동료평가를 계속적으로 사용하고자 한다면, 여학생들의 평가 성향을 고려하여 새로운 방식으로 동료평가를 실시하는 것도 필요하다. 예를 들어, 혼성학급에서 동료평가 시, 학생들의 수행평가 동작을 촬영하여 서로 다른 반 학생들의 수행동작을 동료평가하거나, Lee(2015)의 경우처럼 동료평가를 할 학급 대표자를 선발하여 평가할 경우, 평가자 성비를 같게 할 수도 있고, 평가결과 계산 시 높은 점수와 낮은 점수를 빼고 나머지 점수의 평균을 구할 수도 있다. 이러한 방법은 남학생보다 관대한 점수를 주거나, 친소관계에 영향을 받는 여학생들의 동료평가 경향을 최소화하면서도 교수학습 중에 교사가 동료평가의 교육적 장점을 활용할 수 있는 방법이 될 수 있을 것이다.

본 연구에서는 일반화가능도 관련 선행연구들과 체육교과 특성을 고려하여 단변량과 다변량 일반화가능도 이론을 모두 적용하여 그 결과를 비교하였다. 연구 결과에서는 단변량과 다변량 일반화가능도 이론 적용 시 성별에 따른 분산성분 분석에 있어 차이가 나타났다. 즉, 강제결합-스포츠모의중계의 동료평가의 결과분석에서는 다변량 일반화가능도이론 적용 시 단변량 일반화가능도 이론보다 성별에 따른 평가 요소의 영향력을 좀 더 구체적으로 분석할 수 있음을 보여주고 있다. 결국 체육교과에서 동료평가의 결과를 분석대상으로 하여 일반화가능도 이론을 적용하여 분석할 때는 고정효과인 성별을 고려한 다변량 일반화가능도 이론을 적용하는 것이 분석의 엄밀성을 높이는 것으로 판단된다.

둘째, 단변량 및 다변량 일반화가능도의 D연구를 통해 평가요소 및 평가자 수를 감소하는 측정 조건의 변화에 따라 일반화가능도계수가 어떻게 달라지는지 살펴봄으로써 학교현장에서 동료평가를 활용할 경우 적정 수준의 일반화가능도를 확보할 수 있는 효율적인 측정 구조를 탐색하였다. 정확한 평가를 위해서는 평가요소와 평가자의 수가 추가해야 하지만, Shavelson & Webb(1991)은 이와 관련해서 연구자가 일종의 상충관계(trade-off)를 고려하여 의사결정을 해야 한다고 제안했다. 따라서 연구자가 수용가능한 수준의 일반화가능도계수 및 시간과 비용 등의 이용가능성을 고려하는 것이 필요하다(Song & Kim, 2012). 본 연구에서는 평가요소 수를 2개에서 5개까지 변화시켰고, 평가자 수는 학생 단독, 그룹별 평가를 할 때 한 그룹의 학생 수 및 한 학급의 학생 수를 고려하여 1, 2, 4, 5, 10, 30명으로 변화시켰다. 강제결합-스포츠모의중계 동료평가 결과는 성별을 고려하지 않을 경우 0.985, 성별을 고려할 경우 0.983의 높은 일반화가능도계수를 나타냈으며, 평가요소 수나 평가자 수를 줄이더라도 0.9 이상의 높은 일반화가능도계수를 유지할 수 있음을 확인하였다. 그러나 실제 평가 상황에서는 내용적으로 타당하고 필요한 문항을 추가하거나 삭제하는 것은 매우 신중한 작업이다. 본 연구처럼 개발된 루브릭의 평가요소 수는 선행연구(Jonassen et al., 2003)에 근거하여 개발되기 때문에, 〈Table 3〉의 결과만을 참고하여 평가요소 개수를 2개로 하는 것은 평가에 사용되는 루브릭의 질을 보장할 수 없음을 의미한다. 그러나 Lee(2015)는 동료평가 시에 각 반에서 4명의 학생들을 대표 평가자로 선정하여 동료평가를 시행하였는데, 본 연구에서 평가자 수가 4명인 경우 〈Table 3〉의 결과와 같이 일반화가능도계수가 0.9 이상으로 나타나, 스포츠중계모형의 동료평가방법으로 측정학적으로 높은 신뢰도를 유지하면서도 학교 현장에서 적용하기에 적절하다고 판단된다. 그러나 동료평가 시 학생들은 자신의 평가 점수가 동료들의 성적에 반영되기 때문에 심리적인 부담으로 낮은 점수를 주지 않을 수도 있다(Cho et al., 2010). 따라서 동료평가 시 평가자 수를 Lee(2015)가 적용한 4명보다는, 6명으로 하고 평가점수의 최고·최저 점수를 제외한 4명의 점수를 평가에 반영하는 것도 일정수준의 측정학적인 요구를 만족하면서도 평가자의 심리적인 부담을 더는 현실적인 방법이 될 것이다. 결국 스포츠모의중계수업에서 교사가 어떤 동료평가 방법을 선택할 것인가는 측정학적인 고려와 더불어, 교육적인 판단과 평가 업무량의 적정화 측면에서 결정되어야 할 것이다.

한편 다변량 일반화가능도 분석에 의한 D연구 결과, 남학생의 일반화가능도계수는 0.968, 여학생의 일반화가능도계수는 0.955로 나타나 남학생의 일반화가능도계수가 조금 더 높게 나타났다. 이 결과는 연구에 사용

된 강제결합-스포츠모의중계의 동료평가가 남학생에게 좀 더 신뢰롭게 작용했음을 의미하지만, 여학생의 일반화가능도계수 역시 0.95 이상으로 매우 높게 나타났으므로 여학생 역시 신뢰롭게 나타났다. 물론 선행연구들(Bai et al., 2011; Hong & Lee, 2007; Oh et al., 2001)은 동료평가 상황에서 여학생들의 평가가 남학생들에 비해 신뢰롭지 못한 측면을 언급하고 있는데, 그와 유사하게 본 연구결과에서도 남학생의 일반화가능도계수가 여학생보다 높았다. 그러나 여학생들의 일반화가능도계수 역시 0.9 이상으로 높게 산출된 것은 선행연구들(Bai et al., 2011; Hong & Lee, 2007; Oh et al., 2001)에서 여학생의 동료평가 결과가 다른 요소의 영향을 받아 신뢰롭지 못할 수 있다는 결과와는 상반된다. 그 첫 번째 원인으로 잘 수행된 채점자 훈련이 있을 수 있겠지만, 동료평가 내용이 선행연구들과 상이함이 또 다른 원인일 수 있다. 예를 들어, 선행연구들에서는 체조의 구르기(Bai et al., 2011), 축구의 기초기능(Hong & Lee, 2007) 등과 같은 직접적인 신체활동을 동료평가의 평가내용으로 한 것에 비해, 본 연구는 간접적인 신체활동인 강제결합-스포츠모의중계 영상에 관한 동료평가였다. 일반적으로 여학생들이 남학생들에 비해 실기 능력이 낮고 따라서 실기 위주의 체육수업에서 더 스트레스를 받는 것(Hong & Im, 2006)을 고려할 때, 간접적 신체활동을 대상으로 한 동료평가는 여학생들이 심리적으로 보다 안정적인 상태에서 참여했다고 추측할 수 있기 때문이다. 또한 일반적으로 학교 현장에서는 여학생들이 남학생들보다 학업성취도가 전체적으로 높은 경향이 있다. 그리고 일부 체육교과 연구들에서도 이러한 경향이 확인된다(Hong & Im, 2006; Lee et al., 2008). 따라서 본 연구가 비록 체육교과에서 시행된 동료평가였지만 여학생들의 일반화가능도 계수가 높게 나온 것이 이해가능하다. 그러나 이것이 본 연구에서 여학생들의 동료평가 신뢰도가 선행연구들(Hong & Lee, 2007; Oh et al., 2001)과 동일하게 나타나지 않은 이유인지는 단정할 수 없으므로, 추후 체육교과에서 성별에 따른 직접적 신체활동과 간접적 신체활동의 동료평가 양상의 비교가 후속연구로 이루어져야 할 것이다.

셋째, 강제결합-스포츠모의중계에서 동료평가를 적용하는 경우 일반화가능도이론에 의한 일반화가능도계수가 적절한지 판단하기 위하여, 전통적인 신뢰도 추정방법인 Cronbach α와 단변량다변량 일반화가능도이론에 의한 일반화가능도계수를 비교하였다. 그 결과 이 결과는 평가요소가 평가과제(중계영상)에 내재되어 있지 않은 측정 구조 또는 성별에 따른 집단 차이를 반영하지 않는 신뢰도는 과대 추정함을 확인하였다. 본 연구는 다변량 일반화가능도 이론을 적용하여 평가요소가 평가과제에 내재되어 있고, 평가자가 성별 집단에 내재되어 있는 등, 좀 더 구체적인 평가상황을 반영함으로써 보다 안정적인 신뢰도 계수를 산출할 수 있었다. 이를 통해 동료평가의 신뢰성을 보다 정확하게 판단할 수 있게 되었다는 것을 의미한다. 이렇게 다변량 일반화가능도 이론을 통해 구체적인 평가 상황을 반영한 경우에도 본 연구의 결과에서 동료평가의 일반화가능도 계수가 0.95 이상의 높은 값을 보였다는 것은, 충실한 채점자 교육을 포함한 본 연구의 동료평가의 신뢰성이 높다는 것을 증명한 것이다.

결 론

연구결과와 논의를 종합하여 결론을 내리면 다음과 같다.

첫째, 강제결합-스포츠모의중계 영상에 대한 동료평가에서 검사점수에 가장 큰 영향을 미치는 요인은 단변량 설계에서는 중계영상이었으며, 다변량 설계에서 남학생들은 중계영상이 검사점수에 가장 큰 영향을 주었으나, 여학생들은 중계영상과 평가자의 상호작용이 검사점수에 가장 큰 영향을 주었다. 본 연구 결과는 체육교과 동료평가에서 성별이 검사 점수에 영향을 미친다는 선행연구들(Hong & Lee, 2007; Oh et al., 2001)의 결과를 지지하는 결과이다.

둘째, 강제결합-스포츠모의중계 영상에 대한 동료평가에서 효율적인 측정조건과 관련하여 단변량과 다변량 설계 모두에서 평가요소 수 및 평가자 수가 많을수록 일반화가능도 계수가 높아졌다. 이를 바탕으로, 강제결합-스포츠모의중계 활동에서 동료평가 방법을 적용할 때 안정적인 신뢰도를 유지할 수 있는 효율적인 측정조건을 제시하였다.

셋째, 강제결합-스포츠모의중계 영상에 대한 동료평가에서 전통적인 신뢰도 추정방법인 Cronbach α가 여러 오차요인을 고려하는 일반화가능도 이론을 통해 추정되는 신뢰도보다 과대추정됨을 확인하였으며, 이는 선행연구들과 동일한 결과이다(Wainer & Wang, 2000; Lee et al., 2000; Yang & Lee, 2007). 또한 강제결합-스포츠모의중계 영상에 대한 동료평가에서 단변량 일반화가능도 계수가 다변량 일반화가능도 계수보다 신뢰도를 과대추정함도 확인하였는데, 이는 체육교과 동료평가에서 성별이 검사 결과에 영향을 준다는 선행연구들(Hong & Lee, 2007; Oh et al., 2001)의 결과를 고려할 때, 성별을 고려한 일반화가능도 설계의 적용이 동료평가 신뢰도 판단의 정확도를 높일 수 있다는 것을 보여준다.

지금까지 체육교과 동료평가와 관련한 선행연구들은 동료평가의 장점을 소개하고 교사평가와의 일치정도와 남녀 학생 간 평가 차이에 초점을 두고 진행되었다. 또한 직접적 신체활동과 관련하여 다양한 종목에서 루브릭을 활용한 동료평가 연구가 수행되었다(Bai et al., 2011; Hong & Lee, 2007; Oh et al., 2001). 그러나 본 연구는 학교 현장에서 다섯 개의 평가 요소에 네 개의 평가기준을 가진 루브릭을 개발·적용한 동료평가 자료를 기반으로 동료평가에 영향을 주는 오차요인들을 규명하고 이를 바탕으로 적용 가능한 측정 조건을 확인하였다. 본 연구를 기반으로 두 가지 방향에서의 후속연구를 제안한다. 첫째, 체육교과 및 스포츠상황에서 일반화가능도 이론의 활용의 확대이다. 피겨스케이팅이나 리듬체조 종목에서는 채점자 요인에 대한 신뢰성이 지속적으로 제기(Yonhapnews, 2015.02.13.)되어 왔는데, 이러한 영역에 일반화가능도 이론을 적용하여 평가조건 및 상황의 신뢰성을 평가하는 연구들은 스포츠 현장에 유의미한 환류정보를 줄 수 있을 것이다. 둘째, 동료평가를 위한 루브릭 개발측면에서 평가요소, 평가기준과 측정수준의 고려이다. 본 연구에서 개발·활용된 루브릭은 Jonassen et al.(2003)의 제안에 따라 영역 독립적으로 평가요소가 정해지고 그에 따라 측정기준이 개발되었다. 그리고 각 측정기준에는 측정수준에 따라 네 단계의 점수가 부여되었다. 이러한 개발과정은 체육교과 동료평가 연구를 시행하고자 하는 후속연구자들에게 타당한 평가기준에 따른 평가요소와 측정수준을 고려하여 루브릭을 개발하는데 범례(範例)가 될 수 있다. 또한 측정한 변인의 측정수준은 자료분석 방법과 연구 설계에도 영향을 주므로, 후속연구에서는 동료평가의 다양한 측정수준이 고려된 루브릭의 설계·개발 연구도 필요할 것이다.

참고문헌

Ahn, J. Y. (2008). *A Study on the Use of Peer Assessment to Improve Learning Effect.* Korea Entertainment Industry Association Spring conference abstract of a paper in a proceedings, 2(1), 34-36.

Brennan, R. L. (2001a). *Generalizability Theory*, NY: Springer.

Brennan, R. L. (2001b). *Manual for mGENOVA Version 2.1.* Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa.

Bai, J. Y., Park, H. R., Lee, M. S., & Lee, H. J. (2011). The Comparison of Teacher's and Students' Performance Assessment. *The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science, 13*(1), 45-56.

Cha, S. J., Lee, T. K., Park, C. H., & Lee, H. J. (2014). Effects of Sportscasting on Cognitive Learning in Physical Education. *Journal of Learner-Centered Curriculum and Instruction, 14*(5), 125-144.

Cho, H. M. (2000). Direction of Instruction and Evaluation Tools for Performance Assessment in Physical Education. *The Korean Journal of Sport Pedagogy, 7*(1), 65-82.

Cho, H. M. (2004). A Theoretical Exploration of Peer Evaluation in the Perspective of Social Constructivism. The Journal of Eudcation, 23, 229-254.

Cho, S. K., Kim, J. K., & Lee, J. D. (2010). A Study on the Case Using Peer Assessment in Presentation Class of Collegiate Physical Education. *Korean Society for The Study of Physical Education, 15*(2), 211-223.

Hong, J. H., & Im, M. K. (2006). The Analysis of level of Score and Sex according to the Middle School Students' Worry Factor in Physical Education. *Korean Journal of Sport Psychology, 17*(1), 79-90.

Hong, K. D., & Lee, H. S. (2007). The Analysis of Differences between Physical Education Teacher and Students in Performance Assessment. *Journal of Coaching Development, 9*(4), 383-393.

Hong, S. Y. (2010). Analysis on Effectiveness of Instructional Consultation in Higher Education. *Asian Journal of Education, 11*(3), 97-127.

Jonassen, D. H. (2004). Learning to solve problem: An instructional design guide. San Francisco, CA: Pfeffer.

Jonassen, D. H., & Howland, J., Moore, J., & Marra, R. M. (2003). Learning to Solve Problem with Technology: A Constructivist Perspective. Columbus, Ohio: Merrill/Prentice-Hall.

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

Kim, C. B. (2014). A Study on the Aspects Peer Evaluation in Team-Based Learning. *Korean Journal of General Education, 8*(5), 157-183.

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

Kim, K. S., Lee, G,. & Kang, S. H. (2010). Analysis of Error Sources and Estimation of Reliability in a Korean Speaking Achievement. *Korean Language Education, 21*(4). 51-75.

Kim, M. J. (2005). Peer assessment as a learning method: The effects of assessor and assessee's roles on metocognition, performance, and motivation. *Journal of Educational Technology, 21*(4), 1-28.

Kim, S. J., & Kang, H. K. (2013). Problem Based Learning Evaluation and Evaluation Agents - Focused on Tutor, Peer and Self Evaluation. *Journal of the Korea Academia-Industrial cooperation Society, 14*(8), 3732-3738.

Kim, S. J., & Choi, C. W. (2006). A Study on the Practice of Performance Assessment in the Elementary School Mathematics - Focussing on Self-assessment and Peer-observation -. *Journal of Elementary Mathematics Education in Korea, 10*(1), 67-87.

Kim, S. S., Song, M. Y., & Park, I. Y. (2012). Investigation on optimal conditions and error variance in standard setting using multivariate generalizability analysis. *Journal of Educational Evaluation, 25*(4), 679-700.

Kim, T. J. (2014). *21ST Centruy Trend Analysis in Education Rerorm(Ⅱ): Creative Character Global Education focused on developing Non-Cognitive Skills.* KEDI Research Report RR 2014-15. Seoul: KEDI.

Korean Society for Educational Evaluation. (2004). *Educational Evaluation Thesaurus*. Seoul: HAKJISA corp.

Lee, G., Brennan, R. L. & Frisbie, D. A. (2000), Incorporating the Testlet Concept in Test Score Analyses. *Educational Measurement; Issues and Practice*, Vol 19, 9-15.

Lee, H. J., Lee, T. K., Lee, M. S., & Kim, G. S. (2008). Development of multimedia instructional program for teaching, learning, and evaluating cognitive domain in basketball of eighth grade physical education. *Korean Journal of Sport Pedagogy, 15*(3), 67-80.

Lee, J. S. (2015). Master of Volleyball (e-Book - step by step-modified volleyball game - lopsided sportscasting). *Woori Physical Education, 12*, 58-59.

Lee, S. H. (2008). *Understanding Rater Lenicncy in Peer Evaluation: An Application of the Theory of the Planned Behavior.* A Master Dissertation, Hoseo University.

Lee, S. W., Bak, D. C., & Nam, J. H. (2015). Impact of Peer Assessment Activities on High School Student's Arqumentation in Argument-Based Inquiry. *Journal of the Korean Association for Science Education, 35*(3), 353-361.

Lee, S., Kim, S. Y., Kim, J. H, Baek, K, C., & Lee, B. Y. (2015). Analyses of the Reliability of a Preliminary Creativity Test Using the Multivariate Generalizability Theory. *The Journal of Creatibity Education, 15*(3). 83-107.

Lee, T. K., & Kim, J. Y. (2015a). *The Study of evaluation case related with ill-structured problem-solving learning using sportscasting in PE of high school.* 2015 KAEIM 20 anniversary fall conference abstract of a paper in a proceedings, 27.

Lee, T. K., & Kim, J. Y. (2015b). *The Study of Flipped Learning Case Using Unit Curling in Physical Education of High School.* 2015 KSET fall conference proceedings, 340-356.

Lee, T. K., Kim, J. Y., Lim, K. Y., & Lee, H. J. (2015). The Effects of Scaffolding Types on Problem Solving Abilities and Achievement in Sportscasting. *Korean Journal of Sport Science, 26*(4), 951-963.

Lee, T. K., Kim, K. R., Park, H. R., & Lee, H. J. (2011). Teaching Appreciation in Physical Education Through Mock Broadcasting. *Korean Journal of Sport Science, 22*(4), 2429-2444.

Lee, T. K., & Lee, H. J. (2015a). Collaborative Problem Solving Ability in Physical Education Using Backward

Curriculum Design. *Korean Journal of Sports Science, 26*(4), 917-934.

Lee, T. K., & Lee, H. J. (2015b). Teaching and Learning Creativity and Character in Physical Education: A Guideline for Sportscasting. *The Korea Journal of Sports Science, 24*(4), 1011-1029.

Lee, T. k., & Lee, H. J. (2016). A Study of Possibility of Validation as sportscasting Model for Cultivating Collaborative Problem Sloving Ability. *Journal of Leaner-Centered Curriculum and Instruction, 16*(3), 359-384.

Lee, Y. S., & Shin, S. K. (2004). An investigation into the dependability of ratings in a German speaking test using the multivariate generalizability theory. *Foreign Languages Education, 11*(2), 249-265.

Ministry of Education, Science and Technology (2008). *Commentary of National Curriculum in High School: Physical Education.* Seoul: Ministry of Education, Science and Technology.

Min, H. R., & Yun, H. J. (2012). The Functions of Instructional Consulting Based on Video Recordings in Reform of University Classes: Cognitions of Faculty Members. *The Journal of Yeolin Education, 20*(1), 251-276.

Oh, S. H., Kim, S. J., & Byeun, J. H. (2001). The Difference between Same and Different Gender of Peer Evaluation in Performance Assessment at Co-Educational Physical Education Classes. *The Korean Journal of Measurement and Evaluation in Physical Education and Sport Science, 3*(1), 73-79.

Oh, S. H., Nam, M. H., Song, M. Y., & Kang, J. H. (2006). Hong, S. Y. (2013). *A Study on Teacher's Professional Competency in Students Assessment (Ⅲ) - Physical Education.* KICE Research Report RRE 2006-5. Seoul: KICE.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer* (Vol. 1). Newbury Park, Calif.: Sage Publications.

Shin, J. H., & Hong, S. Y. (2013). *Exploration of Peer Review of Teaching Program in Higher Education. The Journal of Korean Teacher Education, 30*(1), 201-230.

Song, I., & Kim, S. (2012). A validation study of epistemological belief using multivariate generalizability. *The Korean Journal of Educational Methodology Studies, 24*(1), 107-130.

Yang, J. S., & Lee, G. (2007). Estimating Reliability of test scores composed of testlets using Generalizability Theory Approaches. *The Journal of Educational Measurement, 20*(1), 119-139.

Yim, T. K. (2013). A study on developing students' positive attitude in writing education with peer review as a learning method. *The Journal of Korean Writing, 1*, 385-411

Yonhapnews (2015.02.13.). <Four Continental Figure Skating> Some fans protested about 'Sochi Judgment'.

Youn, S. J., & Kim, J. S. (2015). The exploratory study about the policy direction of national developing cooperation: Analysis of Text-Network for Peer review of 6 nations. *Public Policy Review, 29*(2), 69-99.

Wainer, H. & Wang, X. (2000), Using a New Statistical Model for Testlets to Score TOEFL. *Journal of Educational Measurement, 37*(3). 203-220.

〔부록〕 Rubric for Forced Connection Method-Sportscasting(Lee, & Kim, 2015b)

Evaluation Factors	Evaluation Standard	Score	Images			
			A	B	C	D
Preparation	They didn't prepare a props, a costume and stage for sportscasting. And they didn't know the role of announcer and commentator.	0				
	They prepared a props, a costume and stage for sportscasting partially. Or they knew the role of announcer and commentator partially.	1				
	They prepared a props, a costume and stage for sportscasting partially. And they knew the role of announcer and commentator partially.	2				
	They prepared a props, a costume and stage for sportscasting condignly. And They knew the role of announcer and commentator partially well.	3				
Acting Ability	All of the expression, the action and the dialogue among them were not natural. And sportscasting was in sync with the images less than 60%(three minutes).	0				
	Tow or more of the expression, the action and the dialogue among them were not natural. Or sportscasting was in sync with the images less than 60%(three minutes).	1				
	Tow or more of the expression, the action and the dialogue among them were natural. Or sportscasting was in sync with the images from less than 80%(four minutes) to 60%(three minutes) or more.	2				
	All of the expression, the action and the dialogue among them were natural. And sportscasting was in sync with the images 80%(four minutes) or more.	3				
Number of Forced connection	There was nothing a terms of forced connection in sportscasting.	0				
	There was/were from one to three a terms of forced connection in sportscasting.	1				
	There were from four to six a terms of forced connection in sportscasting.	2				
	There were seven or more a terms of forced connection in sportscasting.	3				
Reflection of Sports	They didn't sportscast adequately applying all of the rules, the regulations, skills and game management strategy.	0				
	They sportscasted applying only one of the rules, the regulations, skills and game management strategy.	1				
	They sportscasted applying two to there of the rules, the regulations, skills and game management strategy.	2				
	They sportscasted applying all of the rules, the regulations, skills and game management strategy.	3				
Effective Sportscasting	There were not all of the patness of Speed and rhythm of sportscasting, fun and adding visual effects to images.	0				
	There was/were one to two of the patness of Speed and rhythm of sportscasting, fun and adding visual effects to images.	1				
	There were there of the patness of Speed and rhythm of sportscasting, fun and adding visual effects to images.	2				
	There were all of the patness of Speed and rhythm of sportscasting, fun and adding visual effects to images.	3				

강제결합-스포츠모의중계수업에서 일반화가능도 이론을 적용한 동료평가의 신뢰도와 오차요인 분석

이태구(상동고등학교), **양희원**(연세대학교)

이 연구는 목적은 Lee & Kim(2015b)의 후속연구로 일반화가능도 이론을 적용하여 강제결합-스포츠모의중계수업에서 적용된 동료평가의 신뢰도와 오차요인을 분석하는 것이다. 일반화가능도이론은 연구자가 설정한 특정 상황에서 측정한 자료의 오차요인들을 원인별로 정량화하고, 이를 바탕으로 피험자 점수에서 각 오차요인이 차지하는 상대적인 영향력을 파악하며(G연구), 이를 바탕으로 향후 적용 가능한 효율적인 측정조건을 제시(D연구)할 수 있는 분석 방법이다. 연구 참여자들은 경기도 행복고등학교 1학년 216명(남:115, 여:101)으로 연구 참여자들이 동료평가 한 자료가 분석대상이 되었으며, 이를 단변량과 다변량 일반화가능도 이론을 활용하여 분석하였다. 연구 결과는 다음과 같다. 첫째, 동료평가 결과에 기여하는 오차 요인들의 상대적 영향력은 대체로 동료평가의 평가대상인 영상이 가장 큰 것으로 나타났지만, 성별을 구분해서 분석한 결과 여학생은 영상과 평가자의 상호작용의 영향력이 가장 큰 것으로 나타났다. 둘째, 이 연구에 사용된 동료평가는 높은 일반화가능도계수를 나타냈으며 평가대상인 영상의 수 또는 평가자의 수를 감소시키는 경우, 적정수준의 신뢰도를 유지할 수 있었다. 그러나 성별을 구분해서 분석한 결과 남학생의 일반화가능도계수가 여학생에 비해 높았고, 적정수준의 신뢰도를 유지할 수 있는 구체적인 측정 조건은 다르게 나타났다. 셋째, 측정구조를 적절하게 반영하지 못하는 신뢰도 계수는 신뢰도를 과대 추정함을 확인하였다. 지금까지 체육교과 동료평가 선행연구들은 동료평가를 소개하고, 교사평가와의 일치성 및 성별 동료평가 차이 경향 등에 초점을 두고 연구가 이루어져 왔다. 그러나 본 연구는 일반화가능도 이론을 적용하여 측정학적인 측면에서 학교 현장의 실제 자료를 기반으로 강제결합-스포츠모의중계 수업에서 활용되고 있는 동료평가에 영향을 주는 오차요인들을 규명하고, 이를 바탕으로 향후 적용 가능한 측정조건을 규명하였다. 논의에서는 강제결합-스포츠모의중계수업에 적용된 동료평가에서 단변량과 다변량 일반화가능도이론 분석을 통한 오차요인들과 상대적인 영향력, 일정수준의 일반화가능도계수를 유지할 수 있는 측정조건과 전통적인 신뢰도 추정방법과의 비교가 논의되었다. 본 연구는 체육교과에서 일반화가능도 이론을 적용하여 체육 수업 및 스포츠 상황에서 평가 결과에 큰 영향을 미치는 채점자 효과를 포함한 오차요인들을 규명하고, 상대적인 영향력을 비교하며 효율적인 측정조건을 밝히는 초기연구라는 점에서 의의가 있다.

주요어: 동료평가, 신뢰도, 일반화가능도이론, 강제결합-스포츠모의중계수업