투고일 : 2020. 2. 5 심사일 : 2020. 2. 6 게재확정일 : 2020. 2.17

2

채점자 조정(calibration) 교육 제안을 위한 평가자 신뢰도 분석

[1)]연세대학교 치과대학 치의학교육학교실
[2)]연세대학교 치과대학 보존과학교실, 구강과학연구소
[3)]연세대학교 치과대학 치의학교육학교실, 구강생물학교실

김주아[1)], 신유석[2)], 서정택[3)]

ABSTRACT

Analysis of Evaluator Reliability for the Raters' Calibration Training

[1)]Department of Dental Education Yonsei University College of Dentistry
[2)]Department of Conservative Dentistry and Oral Science Research Center Yonsei University College of Dentistry
[3)]Department of Dental Education and Department of Oral Biology Yonsei University College of Dentistry

Kim, jooah[1)], Shin, Yooseok[2)], Seo, Jeong Taeg[3)]

This study analyzed the change in the rater reliability based on the student's practice evaluation process conducted at Yonsei University College of Dentistry. Through this, we suggest the significance of the rater calibration training in the student's practical evaluation of dental college.

Nine professors from the department of Conservative Dentistry, Yonsei University College of Dentistry, analyzed the results of class II restoration cases twice in 2017 and once in 2018. Intra Class Correlation (ICC) which is a statistic used to determine the consistency of raters with three or more scores, was also calculated.

ICC values increased as raters participated in rater calibration meetings and grading experiences. This shows that the rater reliability is related to the grading experience and feedback from calibration meeting. Based on the results of previous studies that grading experiences and rater calibration training can cause a meaningful change in rater behavior, we propose to conduct rater calibration training to ensure the evaluator reliability.

Key words: reliability, rater calibration, intra-class correlation

Corresponding Author
김주아
연세대학교 치과대학 치의학교육학교실
Email: kja35@yuhs.ac

Ⅰ. 서론

2021년 치과의사 국가시험 실기시험이 도입된다[1]. 실기시험은 피험자의 수행(performance)을 평가하며 수행은 시험에 응시하는 피험자가 구체적인 상황에서 실제로 행동하는 과정(process)이나 그 결과(product)를 의미한다. 치과의사 실기시험은 어떻게 점수가 결정될 것인가라는 질문에 피험자의 수행을 관찰한 채점자가 점수를 결정하게 된다고 답할 것이다. 이는 채점자가 마음대로 점수를 결정한다는 의미가 아니라, 객관적 준거에 기반한 채점기준에 맞춰 평가하지만 여전히 채점자의 주관이 개입될 여지가 있다는 의미이다.

분야와 관계없이 피험자의 수행을 평가하는 실기시험은 지필고사의 선택형 문항이나 단답형 문항과 같이 채점의 객관성(objectivity)을 보장할 수 없다. 이에 실기시험을 준비하는 입장에서 최우선으로 노력해야 할 것은 채점자의 주관성(subjectivity)을 조절하여 채점자 신뢰도(rater reliability)를 높이는 것이다.

주관형 평가에서 채점자의 주관성을 조절하는 일반적인 방법은 채점자 수를 복수로 확보하는 것이다. Fig. 1의 채점자간 신뢰도(inter-rater reliability)와 같은 상황으로, 피험자의 수행을 복수의 훈련된 채점자가 평가한 결과를 합산하거나 평균하여 점수로 부여하는 것이다. 한 명의 채점자에 의해 좌우될 수 있는 주관성을 복수의 채점자가 평가하여 총합한 점수로 보완한다고 할 수 있다. 학생1부터 학생4까지 수행을 채점자1과 채점자2가 모두 평가한 결과로 일치도나 상관계수 등을 산출하여 채점자들 간의 일관성을 알아보는 것이 채점자간 신뢰도이다.

모든 피험자를 복수의 채점자가 채점하기는 현실적으로 어려움이 있다. 치과대학 교육과정에서 실습평가는 Fig. 1의 채점자내 신뢰도(intra-rater reliability)와 같은 상황으로 채점이 될 것이다. 채점자3, 채점자4, 채점자5는 모두 동일한 기준으로 채점을 한다는 전제하에

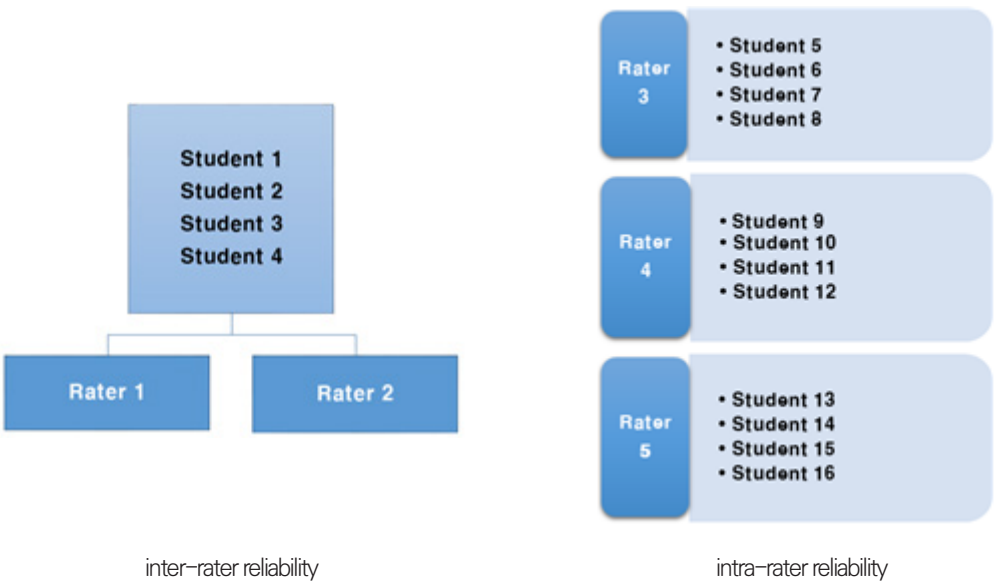

Fig. 1. Rater reliability

각 4명의 학생을 채점 기준에 따라 일관성 있게 채점한다면 채점자내 신뢰도를 확보할 수 있다. 채점자내 신뢰도를 확인할 수 있는 적절한 통계량은 찾기 어렵다. 채점자 교육을 할 때 채점자내 신뢰도 개념을 알려주고 각 피험자에 대해 독립적으로 채점 기준에 따라 평가하도록 강조하여야 하며 치과의사 국가고시 실기시험의 채점자는 채점자간 신뢰도와 채점자내 신뢰도 개념을 바탕으로 잘 훈련된 채점자이어야 한다.

각 대학에서 교수 간 합의된 표준(faculty standardization)에 따라 학생들을 교육한다면 실기시험을 일관되게 진행할 수 있을 것이다. 예를 들어 실기시험 문항에서 주어진 상황에 대해 A 학교 학생과 B 학교 학생이 서로 다르게 교육을 받은 대로 수행을 보인다면 일관된 채점을 할 수 없어 응시자에게 피해가 돌아갈 것이다.

교수 간 합의된 표준은 단시간에 이루어지는 것은 아니고 지속해서 교수들 간의 채점자 조정(calibration)을 해야 한다. 채점자 조정이 어려운 여러 이유가 있겠지만 다른 학교, 다른 시기, 다른 임상경험을 가지고 있다는 것으로 추려질 수 있을 것이다. 실기시험이 도입되는 것에 대비하여 실기시험에 응시하는 학생들을 위해, 교수들이 일관된 절차(protocol), 술기(techniques), 철학(philosophies)에 대해 꾸준하게 채점자 사이의 차이를 조정해 간다면 학생들의 실기 능력 또한 일관성을 가질 것이다.

본 연구는 연세대학교 치과대학 보존과에서 시행하고 있는 학생 실습평가과정을 바탕으로 채점자 신뢰도의 변화를 분석하였다. 이를 통해 치과대학의 학생 실습평가에서 채점자 신뢰도 확보를 위한 채점자 조정 교육을 제안하고자 하였다. 구체적인 연구문제는 다음과 같다.

첫째, 학생 실습평가에서 채점자 조정 회의 경험과 채점자 신뢰도는 관련성이 있는가?

둘째, 학생 실습평가에서 채점자 신뢰도 확보를 위한 채점자 조정 교육은 무엇인가?

Ⅱ. 연구방법

1. 분석자료

연세대학교 치과대학 보존과 교수 9인이 2017년에 두 번, 2018년에 한 번, 모두 3회 보존과 증례에 대해 채점한 결과를 분석하였다. 2급 와동 수복은 인접면 치아 상실로 인해 인접면과 교합면 모두의 치아면를 수복하기 위한 것으로 학생의 수복 능력을 종합적으로 판단하기에 적합하다. 그러므로 이상적인 치아 와동의 형태, 치아 우식의 제거 여부, 치아 형태 수복 능력, 인접면 접촉면 형성 등을 평가할 수 있다.

9인의 보존과 교수는 9~10개로 구성한 채점항목과 채점항목별 3단계 채점 기준을 바탕으로 채점하였다. 채점항목은 대한치과보존학회에서 발간한 실습서를 바탕으로 작성하였다. 채점 전에 교수들에게 채점항목과 채점 기준에 관해 설명하고 의견을 구하는 회의를 개최하였다. 사전에 학생이 실습한 상황을 사진으로 촬영하여 교수들에게 제공하고 회의에서 논의한 내용을 바탕으로 채점하도록 안내하였다.

2. 분석방법

본 연구에서 수집된 데이터는 사회통계분석 프로그램인 SPSS ver. 25를 이용하여 분석하였다. 구체적인 분석방법은 다음과 같다. 첫째, 채점한 결과의 평균과 표준편차를 산출하여 2017년과 2018년에 기술통계량의 추이를 살펴보았다. 둘째, 채점자들 간 상관계수를 산출하여 서로의 동의 정도를 알아보았다. 셋째, 3인 이상 채점자들의 일관성 정도를 알아보기 위해 사용되는 통계

량인 급내상관계수(Intra Class Correlation; ICC)을 산출하였다[2-5].

Ⅲ. 연구결과

1. 기술통계량

연세대학교 치과대학 보존과 교수 9인이 2급 와동 수복 증례를 2017년 2번, 2018년 1번 채점한 결과의 평균과 표준편차를 산출하였다. 각 채점 증례에 따라 제시한 채점 항목의 수는 각기 달랐다. 2017년 첫 번째 평가에 사용된 증례는 9개 평가항목으로, 두 번째 평가에서 사용된 증례는 7개 평가항목으로, 2018년 평가에 사용된 증례는 10개 평가항목으로 채점하였다.

2017년 첫 번째 채점결과를 보면 5명의 채점자는 일부 항목에 대해서 채점하지 않았고, 두 명의 채점자는 모든 채점항목에 대해 같은 점수를 표기하였다. 2017년 두 번째 채점에서 1명의 채점자는 불참하여 8명이 채점하였다. 이번에도 1명의 채점자가 1개 채점항목을 표기하지 않았다. 이는 일부 채점자가 해당 채점항목은 사진 자료를 통한 채점이 불가능하다고 판단하였거나, 채점 기준에 대해 해석을 다르게 하고 있었다고 판단된다. 2018년에는 모든 평가자가 10개 채점항목에 모두 표기하였고, 비교적 평균과 표준편차가 안정적으로 산출되었다. 3번의 기술통계량을 산출한 결과 채점자들이 경험할수록 채점의 안정성이 높아졌다고 할 수 있다.

2. 채점자간 상관계수

두 채점자 간의 상호 상관계수를 산출한 결과는 아래 Table 2~Table 4에 제시하였다. Table 1에서 확인할 수 있듯이 2017년 첫 번째 채점에서 rater 4와 rater 9는 모든 채점항목에 같은 평가를 했기 때문에 상관계수를 산출할 수 없어 Table 2에 a로 표기하였다. Table 3과 4에 알 수 있듯이 두 번째와 세 번째 채점을 경험하면서 채점자 간 상관계수는 안정적으로 산출되었다.

3. 급내상관계수

채점자들이 2017년 두 번과 2018년 한 번 채점한 결과로 급내상관계수(ICC)를 산출하였다. 급내상관계수

〈Table 1〉 Descriptive Statistics

rater \ statistics	2017, 1st			2017, 2nd			2018		
	N	Mean	standard deviation	N	Mean	standard deviation	N	Mean	standard deviation
rater1	8	1.38	.74	-	-	-	10	1.60	.52
rater2	6	1.67	1.03	7	1.86	.90	10	1.70	.95
rater3	9	1.44	.73	7	1.86	.90	10	1.80	.79
rater4	9	2.00	.00	7	2.14	.69	10	1.90	.57
rater5	7	1.43	.79	6	2.33	.52	10	1.70	.67
rater6	9	1.89	.33	7	1.71	.49	10	2.20	.63
rater7	8	1.38	.52	7	1.71	.76	10	1.40	.52
rater8	9	1.78	.44	7	1.86	.69	10	1.80	.63
rater9	8	2.00	.00	7	2.57	.53	10	1.20	.42

〈Table 2〉 2017, 1st Correlation Coefficient

2017, 1st	rater1	rater2	rater3	rater4	rater5	rater6	rater7	rater8
rater2	−.32							
rater3	.13	−.63						
rater4	–	–	–					
rater5	.73	−.46	.62	–				
rater6	.20	−.63	.23	–	.24			
rater7	−.05	−.32	.18	–	.06	.29		
rater8	.31	−.25	−.04	–	.37	.66	.45	
rater9	–	–	–	–	–	–	–	–

〈Table 3〉 2017, 2nd Correlation Coefficient

2017, 2nd	rater1	rater2	rater3	rater4	rater5	rater6	rater7	rater8
rater2	–							
rater3	–	.59						
rater4	–	.58	.04					
rater5	–	0	0	.34				
rater6	–	.27	.65	.14	.50			
rater7	–	.42	.67	.09	−.16	.65		
rater8	–	.77*	.23	.75	.17	.35	.55	
rater9	–	.89**	.55	.65	−.25	.09	.47	.71

*p〈.05, **p〈.01

〈Table 4〉 2018 Correlation Coefficient

2018	rater1	rater2	rater3	rater4	rater5	rater6	rater7	rater8
rater2	.18							
rater3	−.22	.21						
rater4	−.15	.35	.20					
rater5	.26	.36	.50	.49				
rater6	.61	.48	−.36	.06	.16			
rater7	−.17	−.18	.22	.53	.06	−.27		
rater8	−.27	.07	.36	.87**	.62	−.17	.61	
rater9	−.10	.72*	.47	.56	.62	.25	.10	.58

*p〈.05, **p〈.01

〈Table 5〉 Intra Class Correlation

Year	2017, 1st	2017, 2nd	2018
Intra Class Correlation (ICC)	.29	.42	.72

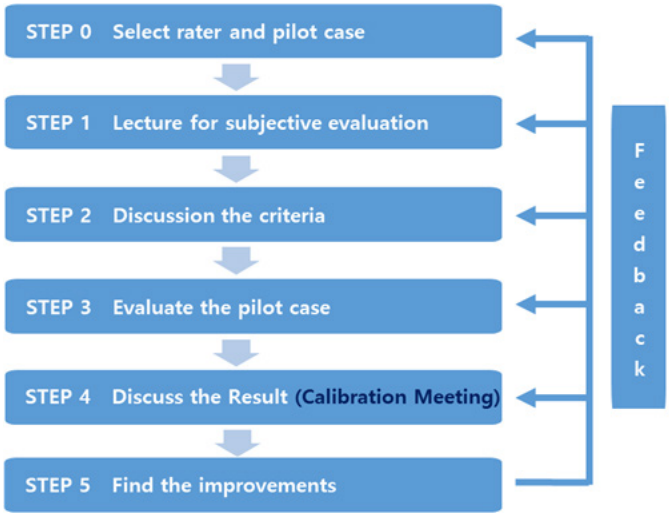

Fig. 2. Calibration training step to improve reliability

값은 0.4 이하이면 채점자 간 일관성이 부족하고, 0.4~0.6이면 수용할 만한 수치이고, 0.6~ 0.75이면 우수하고, 0.75 이상으로 산출되면 매우 좋은 결과라고 해석한다[6,7]. Table 5에서 알 수 있듯이 채점자들이 채점자 조정 회의에 참여하고 채점 경험을 거치면서 급내상관계수 값이 2017년 첫 번째는 .29, 2017년 두 번째는 .42, 2018년에는 .72로 향상하였다. 이는 채점 경험과 채점자 신뢰도가 관련이 있음을 나타낸다고 볼 수 있다[8,9].

4. 채점자 조정 교육 제안

채점자 조정을 위해서 채점 후 신뢰도와 상관계수 등 통계량을 산출한 결과를 바탕으로 채점자 조정 회의를 개최하고 채점과정과 결과에 대해 논의하였다. 2017년 첫 번째 채점자 조정 회의에서 학생 교육 시 보완할 사항과 채점자 조정을 위한 고려사항을 도출하였다. 이에 채점자 교육을 위해 2017년에 학생 실습평가 채점과정을 한 번 더 실시하였고, 2차 채점자 조정 회의를 개최하여 2018년 학생 실습평가에 고려할 채점항목과 채점 기준을 보완하였다. 2018년 학생 실습을 시행하기 전에 채점에 참여할 교수들을 대상으로 2차 채점자 조정 회의에서 반영된 채점항목과 기준을 공유하고 선정한 증례를 채점하였다.

2017년과 2018년 도합 3차례 채점자 조정 회의를 거치면서 수집한 채점 자료를 바탕으로 산출한 채점자 신뢰도가 점진적으로 향상하는 것을 확인하였다. 이런 과정을 바탕으로 치과대학 학생 실습평가에서 채점자 신뢰도를 확보할 수 있는 채점자 조정 교육을 제안하였다.

채점할 증례에 따른 채점항목과 채점 기준, 채점자들(pool)은 채점자 조정 교육을 시행하기 전에 미리 준비되어 있어야 한다. 5단계로 제안한 Fig. 2의 채점자 조정 교육은 각 치과대학의 채점자 규모와 채점자 집단의 풍토(climate)에 맞춰 단계별로 시행할 것을 제안한다.

Ⅳ. 고찰

본 연구의 결과에서 확인할 수 있듯이 채점자 조정을 경험하는 횟수가 증가할수록 채점자 신뢰도가 향상하였다. 이는 채점 영역은 차이가 있으나 주관형 평가에서 채점 경험이 있는 집단이 신뢰도가 높게 나타나는 것과 일치한 결과이다[8]. 2017년 첫 번째 채점자 조정에서 채점한 평가자들의 ICC는 .29로 매우 낮았다. 그러나 2017년 두 번째 채점을 거쳐 2018년에는 .72로 향상하였다. 이는 채점자가 자신이 채점한 결과를 확인하면서 채점자 주관성에 대해 인식하는 과정을 거쳤다고 할 수 있다.

주관형 평가로 진행된 언어능력 평가 관련 연구에서 채점자 행동 특성 관련 변수들이 평가의 신뢰성에 주로 부정적 영향을 미친다고 하였다[10,11]. 이는 채점자들이 채점항목과 채점 기준을 다르게 해석하여 채점결과의 일관성에 차이를 보였기 때문이라고 할 수 있다. 채점 경험과 훈련이 채점자 행동에 의미 있는 변화를 가져올 수 있다는 선행연구 결과에 바탕을 두고 채점자 조정 교육을 제안하였다[9]. 연구결과에 제시한 5단계 채점자 조정 교육을 수행하기 위해서는 무엇보다 채점결과에 대해 솔직하게 논의할 수 있는 채점자 집단의 풍토가 전제되어야 한다.

채점자 교육은 채점 기준에 대해 합의된 이해를 도출하고 실제 채점과정을 연습하기 위해 시행한다. 채점자 교육을 하는 동안에 채점을 연습하는 시간을 갖는다. 이때 채점자가 채점한 것에 대한 피드백을 제공하고, 채점하는 동안에 가진 의문에 대해 질문할 충분한 시간을 준다. 채점자 교육의 목적은 채점 기준에 따라 채점하는 것을 연습하는 데 있으므로 채점자의 채점결과에 대해 자유롭게 의견을 개진할 수 있어야 한다.

채점자 조정 교육 마지막 단계에서 도출하는 개선사항은 채점 기준의 명료화, 채점할 때 채점자 유의사항, 학생 술기 평가에 적절한 증례 검토, 학생 술기 교육내용 조정 등의 항목으로 유목화할 수 있을 것이다. 도출된 개선사항은 해당 학생 실습평가에 따라 보다 구체적으로 기술하고, 이를 적절한 단계에 되먹임(feedback)할 수 있어야 한다. 이와 같은 과정이 자연스럽게 진행될 수 있는 채점자 집단의 풍토가 있다면 채점자 조정 교육 시간은 축소될 수 있을 것이다.

채점자가 채점 기준을 일관되게 적용하도록 훈련하는 것은 전문성을 개발하는 중요한 기회이다. 채점자 조정 교육은 채점해야 할 학생 술기의 중요한 측면에 대한 합의된 표준(faculty standardization)을 이해하는 데 도움이 된다. 또한, 임상 실습교육 목표를 다시 생각하게 하고, 학생이 보여주는 술기에서 나타나는 장점과 단점에 대한 피드백을 줄 수 있는 통찰력을 갖게 된다. 이를 통해 개별 치과대학의 임상 실습교육을 넘어 11개 치과대학의 임상 실습교육의 합의된 표준을 제시하여 시행을 앞두고 있는 치과의사 국가시험 실기시험이 일관되게 진행되도록 할 수 있을 것으로 생각된다.

참 고 문 헌

1. Introduction of the National Practical Examination for the Dentists Press Release(2017). Ministry of Health and Welfare.
2. Cicchetti D. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment 1994;6(4):284-290.
3. 김기열. 치의학 연구에서 반복 계측한 자료의 일치도 평가방법. 대한치과의사협회지 2016;54(11):880-896.
4. 공경애. 검사법 평가: 검사법 비교와 신뢰도 평가. Ewha Med J 2017;40(1):9-16.
5. 박창언, 김현정. 체계적 문헌고찰에서 평가자 간의 신뢰도 측정. Hanyang Med Rev 2015;35:44-49.
6. Fleiss J. Design and analysis of clinical experiments. New York, USA: Wiley; 1986.
7. Hallgren K. Computing Inter-Rater Reliability for Observational Data: An Overview Tutor and Tutorial. Tutor Quant Methods Psychol 2012;8(1):23-34.
8. Choi YH. Rating Performance of EFL Teachers in Writing assessment: Comparison of Experienced and Novice Raters. 교과교육학연구 2013;17(1):199~215.
9. Cumming A, Kantor R, Powers D. Decision making while rating ESL/EFL writing tasks: A descriptive framework. Modern Language Journal 2002;86:67-96.
10. Barrett S. The impact of training on rater variability. International Education Journal 2001;2:49-58.
11. Schoonen R. Generalizability of writing scores: An application of structural equation modeling. Language Testing 2005;22:1-30.