한국심리학회지 : 인지 및 생물
The Korean Journal of Cognitive and Biological Psychology
2023, Vol. 35, No. 3, 107-124
http://dx.doi.org/10.22172/cogbio.2023.35.3.001

Original Article

Exploring the evaluation capacity of university students through rating accuracy and rater bias*

Gyeongmee Gim[1], Jung Ae Park[2], Ji Won Yang[2], Jooyong Park[2†]

[1]Interdisciplinary Program in Cognitive Science, Seoul National University
[2]Department of Psychology & Institute of Psychological Science, Seoul National University

Traditionally, researchers have used rating accuracy and rater bias (severity, centrality, and randomness) as individual-level indicators of rating quality. While these have been studied mostly for expert raters, research on whether evaluation capacity is domain-general over two or more different tasks is lacking. Thus, we investigated the two indicators in the context of undergraduate raters. In two studies, undergraduates scored outputs from a verbal-linguistic task and a visual-spatial task. The results showed that proficient students in one domain are also likely to be proficient in the other in terms of rating accuracy and the use of rating scale. In addition, students with lower rating accuracy were more significantly affected by the difference between domains compared to more accurate students. We also discuss the implications and limitations of our findings on measuring student evaluation capacity.

Keywords: evaluation capacity, rater severity, rater centrality, randomness effect, rating accuracy, Many-Facet Rasch model

1차원고접수: 22.11.29; 수정본접수: 23.04.27; 최종게재결정: 23.05.27

교육평가나 직무평가, 운동경기 또는 음악경연대회의 수상자 선정, 임상의 환자 진단 등 수많은 의사결정에서 우리 사회는 인간 평가자들에게 의지하고 있다. 그리고 평가자마다 다르게 판단하는 문제가 있으며, 이 문제로 인한 사회적 비용이 적지 않다는 것 또한 인식하고 있다(Kahneman et al., 2021). 하지만 초중고 교육은 물론이고 대학 교육에서조차 모든 학생을 대상으로 평가 역량을 훈련하는 사례를 찾아보기 어렵다.

그럼에도 평가 역량이 학습에 중요하다는 점을 인지한 연구자들은 학습자 스스로 자신의 수행을 평가하는 '자기평가'와 수업을 함께 듣는 동료 학습자의 수행을 평가하는 '동료평가'에 주목해왔다. 그 결과 학습자의 평가 역량에 대한 연구는 이 두 가지 활동을 중심으로 이루어져 왔다. 글 평가 활동은 학습을 촉진할 뿐만 아니라(Cho & McArthur, 2010), 비판적 사고력을 향상시킨다. Jiang 등(2022)은 캘리포니아 비판적 사고력 검사(California Critical Thinking Skills Test, 이하 'CCTST')를 이용해 동료평가 활동이 실제로 비판적 사고력을 향상시킴을 보여주었다. CCTST의 하위 구인인 평가(evaluation) 능력도 학습자가 평가 주체가 되었을 때 유의하게 향상되었다. 하지만 이 연구 외에는 평가 역량을 직접적으로 다루는 연구를 찾아보기 어렵다.

국내에는 CCTST와 같은 검사조차 드물다. 대학생핵심역

* 이 연구는 서울대학교 미래기초학문분야 기반조성사업으로 지원되는 연구비에 의하여 수행되었음.
† 교신저자: 박주용, (08826) 서울대학교 심리학과, 서울특별시 관악로1 서울대학교, E-mail: jooypark@snu.ac.kr

량진단(K-CESA)이 하위구인으로 평가적 사고력을 포함하지만, 작문 과제를 사용하기 때문에 전문 채점자가 별도로 필요하다(Jin et al., 2011). 결국 학습자의 평가 역량 연구가 제한될 수밖에 없다. 이런 맥락에서 본 연구는 자기평가나 동료평가 상황에서 학습자들이 채점한 점수를 이용해 그들의 평가 역량을 측정하고 그 특성을 탐색하였다.

평가 역량과 채점

교육 장면에서 인간 평가자가 필요한 대표적인 평가 도구 중 하나가 구성반응형(constructed-response) 문항이다. 실제로 동료평가의 활용도가 높은 평가 장면도 구성반응형 평가인 작문 과제이다(메타분석 연구인 Li et al.(2020)의 부록 참조). 구성반응형 평가는 과제를 수행한 사람의 응답이 글이나 그림, 영상 등의 산물로 남기 때문에, 평가자 한 명이 반복 관찰할 수 있는 것은 물론, 여러 평가자가 동일한 내용을 관찰할 수 있다는 장점도 있다. 그와 같은 응답 산물에 대해서 평가자는 논평을 줄 수도 있고 점수를 줄 수도 있다. 논평으로 평가역량을 측정하려면, 평가역량 전문가가 학생 평가자의 논평 수준을 채점하는 별도의 과정이 필요하다. 하지만 학생 평가자가 준 점수로 평가역량을 측정하면 그 과정을 줄일 수 있기 때문에, 본 연구는 우선 채점을 하는 평가적 판단만 연구 대상으로 삼았다.

구성반응형 평가의 주요 요소와 흐름을 도식화하면 Figure 1과 같다. 평가 목적에 따라 평가하려는 이론적 구인(construct)이 정의되면, 그에 맞춰 과제(task)[1]와 평가준거(criterion)가 개발된다. 피평가자(ratee)가 과제를 수행한 결과인 평가물(output)을 산출하면, 채점자는 그 평가물을 관찰하고 해석하며, 평가준거와 비교하여 판단의 증거를 수집하고, 채점척도(scale)를 이용해 점수를 매긴다. 이때, 이론적으로 채점자가 산출한 점수(이하 '원점수')는 오롯이 평가 구인에만 근거해야 한다. 하지만 현실에서는 Figure 1과 같이, 채점자 혹은 평가 도구, 평가 맥락 등 여러 요인의 복합적인 영향을 받는다.

원점수의 패턴 가운데, 평가 구인과 무관하게, 채점자 개개인의 특질 때문에 나타나는 오차점수를 '채점자 변산도(rater variability)' 또는 '채점자 효과(rater effect)'라고 한다. 채점자 효과가 유의하지 않거나 채점자 변산도가 작을수록 그 평가의 신뢰도와 타당도는 높아지고, 제대로 평가가 이루어지게 된다. 따라서 채점자 효과를 채점자 개개인 수준에서 확인할 수 있는 통계량을 이용하면 그들의 평가 역량을 부분적으로 확인할 수 있다. 그와 같은 지표로 채점자 편향(rater bias)과 채점 정확도(rating accuracy)가 있다(Engelhard, 1996).

채점자 편향은 여러 하위 개념으로 나뉘는데, 그중에서 가장 기본적으로 언급되는 것이 채점 엄격성(severity), 중앙값 채점 경향성(centrality), 채점 무작위성(randomness)이다(Myford & Wolfe, 2004). 세 지표는 채점척도를 잘못 사용하는 행동에서 기인한 점수 편향이라는 점에서 공통된다. 평가인지가 다른 판단 및 의사결정과 구분되는 지점은 암묵적으로든 명시적으로든 잣대, 즉 척도가 있다는 점이다. 이 척도란 단순하게는 좋고 나쁨, 혹은 잘함과 못함으로만 구분되는 이분척도일 수도 있고, 좋음이나 우수함을 여러 측면으로 정의한 다음에 각 측면을 측정하는 채점척도일 수도 있다. 어느 쪽이든 평가 대상에 잠재한(latent) 특질(질, 가치, 심리구인 등)의 연속체(continuum) 상에서 상대적 위치를 산출하는 도구라는 점에서는 똑같다. 따라서 척도를 잘 사용하는 것은 평가 역량에서 중요한 부분이기에, 채점척도 사용과 관련된 채점자 편향을 본 연구의 평가 역량 지표로 삼았다.

세 가지 채점자 편향을 구체적으로 살펴보면, 먼저 채점 엄격성은 각 채점자가 다른 채점자에 비해 평가물에 높은 기준(또는 낮은 기준)을 적용해 점수 선택에서 극단적인 경향성을 보이는지 아닌지를 확인하는 것이다. 채점자들 간에 엄격성 차이가 발생하면 공정성이 결여될 수 있기 때문에 채점 엄격성은 평가 신뢰도를 해치는 주요 변인 중 하나로 간주된다(Myford & Wolfe, 2003). 따라서 적절한 엄격성 수준을 유지하는 것이 채점자의 중요한 역량이다. 개인 수준의 채점 엄격성 연구는 Linacre(1989)의 다국면라쉬모형(Many-Facet Rasch model, 이하 'MFRM')과 FACETS라는 분석 프로그램이 등장하면서 더욱 활발해졌다(McNamara & Knoch, 2012). MFRM을 이용하면 문항 난이도(difficulty)와 피평가자의 능력, 다른 채점자들의 반응을 고려하여, 개별 채점자들의 채점 엄격성 수준을 추정할 수 있기 때문이다. MFRM 이외에도 채점 엄격성을 추정하는 다른 IRT 계통 모형들이 있지만, MFRM이 가장 단순한 모형이다. 본 연구처럼 피평가자 수가 적고 채점자가 많은 자료에서 모수 추정 정확도가 높다는 장점이 있기 때문에(Uto & Ueno, 2018), MFRM을 사용하여 대학생 평가자의 채점 엄격성을 추정하였다.

중앙값 채점 경향성은 채점자들이 척도의 중간에 위치한 점수들을 선호하는 경향 또는 척도의 극단값을 회피하는 경

1) 교육평가나 심리척도에서는 통상 문항(item)이라고 하지만 수행평가(performance assessment) 맥락에서는 과제(task)라고 지칭하기도 한다(Johnson et al., 2008). 본 연구에서도 '과제'라는 표현을 주로 사용하였다. 다만 통계량 설명 등에서 필요한 경우 '문항'이라는 표현도 일부 사용하였다.

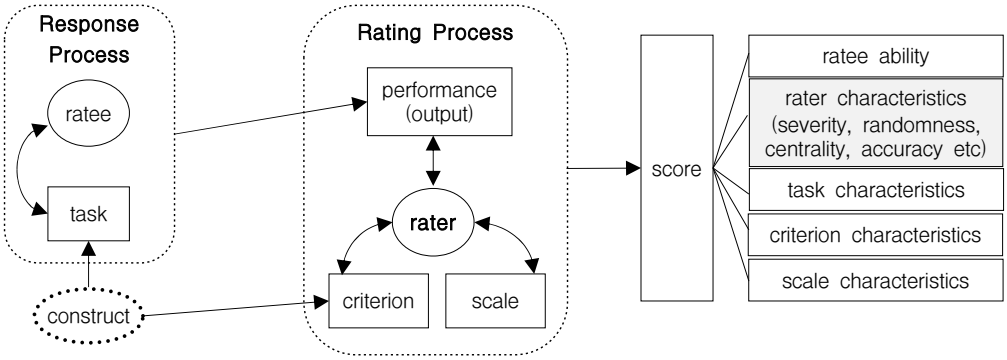

Figure 1. A conceptual framework of constructed-response assessment.

향이다. 그리고 채점 엄격성과 더불어, 척도의 일부 구간만 집중적으로 사용하는 편향(restriction-of-range)의 일종이다(Myford & Wolfe, 2003). 즉, 채점 엄격성은 척도의 극단값을 선호하는 편향, 중앙값 채점 경향성은 극단값을 회피하는 편향으로, 서로 반대되는 동기 때문에 발생하는 채점 행동인 것이다. 채점 엄격성에 비하면 중앙값 채점 경향성은 비교적 최근에 주목받기 시작했고 누적된 연구 결과도 적은 편이어서(Jin & Eckes, 2022), 개별 채점자 수준에서 측정하는 방법이 채점 엄격성만큼 명확하지 않다. 최근 Jin과 Wang(2018)이 MFRM을 확장하여 중앙값 채점 경향성을 단일 모수로 추정하는 모형을 제안하기는 했지만, 본 연구와 같이 피평가자 수가 적은 사례에서 모형이 검증된 적이 없다. 그보다 간단하게는 척도의 중앙 구간을 사용한 빈도나 백분율을 이용해 구할 수 있기 때문에(Engelhard, 1994; Myford & Wolfe, 2004), 본 연구도 이 방식을 따랐다.

채점 무작위성은 앞의 두 채점자 편향과 다소 결이 다른 개념이다. 채점 엄격성과 중앙값 채점 경향성은 현상이 먼저 관찰되고 그 현상을 연구하기 위해 여러 측정치가 개발된 반면, 채점 무작위성은 MFRM의 채점자 국면의 적합도를 해석하기 위해 도입되었기 때문이다. MFRM으로 채점 엄격성을 추정하면 채점자 개인 수준의 모형적합도(Rater fit mean-square indices)가 산출되는데, 모형에서 기대되는 채점 엄격성과 실제 채점자들의 채점 엄격성이 얼마나 다른지를 보여준다. 이 통계치를 채점자의 행동과 연결시키면, 값이 클수록, 즉 부적합이 심할수록 채점자가 불규칙적이고 비일관적인 채점 행동을 보여준다고 해석할 수 있다(Engelhard, 2012, p.369; Linacre, 2021, pp. 291-293, Myford & Wolfe, 2004). Wind와 Engelhard(2012, 2013)의 연구를 보면, 채점자 외적합도(unweighted MSE 또는 Outfit MSE)는 채점 정확도와 유의한 상관관계를 보인다. 또한 Wolfe와 McVay(2012)는 채점자 외적합도가 $1+(6/\sqrt{N})$ 이상인지 아닌지로 부정확한 채점을 분류할 수 있음을 보여주었다. 이를 보면 MFRM의 채점자 외적합도는 채점자의 오류를 탐지할 수 있는 중요한 지표이므로 본 연구에서도 평가 역량 지표로 포함하였다. MFRM의 채점자 외적합도는 채점 엄격성이나 중앙값 채점 경향성과는 다른 행동 패턴에서 기인하기 때문에, Myford와 Wolfe(2004)를 따라 채점 무작위성으로 정의하였다. 이 통계량은 값이 클수록 각 피평가자에게 다른 채점 엄격성을 적용했음을 보여주므로, 채점 무작위성 또한 채점척도 사용과 관련된 편향으로 볼 수 있다.

이상의 채점자 편향 지표가 채점척도를 사용하는 역량을 설명해준다면, 평가물에 담긴 평가 구인의 고하(高下)를 판단하는 역량을 보여주는 지표는 채점 정확도이다. 이론적으로 채점 정확도는 평가물의 실제 수준과 채점자가 준 점수가 일치하는 정도이다. 그런데 평가물의 수준에 대한 참값을 알 수 없기 때문에, 현실에서는 '기준점수'라는 고육지책을 쓴다. 통상적으로 기준점수는 전문가의 점수를 이용하는데, 그 산출 방식에는 전문가 1인의 점수, 전문가 집단이 채점한 점수의 평균, 전문가 집단이 합의한 점수 등이 있다. 즉, 여러 전문가의 판단을 이용해서 참값을 추정하는 것이다. 이렇게 산출된 기준점수와 비교해서 높은 유사도를 보이면, 그 평가자는 평가물의 수준을 잘 판별한다고 보는 것이다. 본 연구는 미숙련 평가자를 대상으로 하기 때문에, 학습자들이 평가물의 수준을 '전문가만큼 잘' 판별하는 능력 또한 그들의 평가 역량을 보여준다고 할 수 있다. 따라서 채점 정확도를 평가 역량의 지표로 포함하였다. 채점 정확도 산출에서 기준점수와 비교하는 방식에는 여러 가지가 있지만, 대표적으로 Cronbach(1955)가 제안한 방식이 다양한 판단 연구 및 채점자 연구에서 활용되고 있다(Becker & Cardy, 1986; Cho & Han, 2008; Furr, 2008; Pulakos, 1984; Schleicher, Day et al., 2002). 본 연구도 Cronbach(1955)의 방식으로 채점 정확도를 측정하였고, 구체적인 수식은 연구 방법에서 자세히 기술하였다.

종합하면, 본 연구는 구성반응형 평가의 채점 활동을 이용해 학습자의 평가 역량을 측정하기 위해, 기존의 채점자 역량 지표 중 개별 채점자 수준에서 확인할 수 있는 채점 정확도와 채점자 편향을 활용하였다. 두 지표를 이용하여 대학생 평가 역량으로, 1) 평가물의 수준을 전문가만큼 잘 판별하는 능력과 2) 채점척도를 적절하게 사용하는 능력을 측정할 수 있다고 보았다.

평가 역량과 평가영역

지능검사의 경우 한 하위 검사인 언어영역에서 높은 점수를 받으면 일반적으로 다른 하위 영역에서도 높은 점수를 받는다. 여러 하위 영역 간의 정적 상관은 하나의 큰 관으로부터 여러 분기가 나뉘는 것으로 시각화할 수 있는데, 통계적으로는 하나의 큰 관은 g요인으로 그리고 각각의 하위 영역이 가진 독특성은 s요인으로 포착된다(Snow et al., 1984). 이 구분은 평가 역량에도 적용될 수 있다. 구체적으로 한 사람에게 언어영역에 대한 평가와 시공간영역에 대한 평가를 하게 했을 때, 두 평가 모두에 관여하는 영역 일반적인 평가 역량과 각 영역에 특수한 평가 역량으로 나눌 수 있다는 것이다.

본 연구는 위와 같은 평가 역량의 영역 일반성과 영역 특수성이 채점 정확도와 채점자 편향에서도 관찰되는지 탐색하였다. 두 지표는 전문 채점자 훈련을 위해 개발된 측정치이기 때문에, 여러 영역의 수행 수준을 비교할 필요가 없는 전문가 연구의 특성상, 두 측정치가 평가영역과 맺는 관계가 탐색된 바 없다. 그러나 본 연구에서처럼 학습자의 평가 역량을 측정하기 위해 채점 정확도와 채점자 편향을 활용하는 맥락에서는 지능검사에서와 마찬가지로, 둘 이상의 다른 영역에서 각 지표가 정적 상관을 보이는지, 그리고 영역 차이로 인한 역량 차이가 유의한지를 살펴보는 것이 기초 연구로서 중요하다고 보았다.

이상의 논의를 종합하면, 본 연구는 전문가의 채점 질을 확인하는 데 활용되어온 채점 정확도와 채점자 편향(채점 엄격성, 중앙값 채점 경향성, 채점 무작위성)을 학습자의 평가 역량 측정에 활용하는 방안을 제안하고자 하였다. 그리고 두 지표로 측정된 학생들의 평가 역량에서 영역 일반성과 영역 특수성이 관찰되는지 탐색하였다.

연구 1

연구 1에서는 채점 정확도와 채점자 편향으로 측정한 대학생 평가 역량의 영역 일반성과 영역 특수성을 확인하고자 하였다. 이를 위해, 동일한 채점자에게 언어영역과 시공간영역의 과제를 채점하게 한 다음, 평가 역량의 영역 일반성 확인을 위해 두 영역 간 평가 역량의 상관계수를 확인하였다. 그리고 영역 특수성 확인을 위해서 평가영역 차이가 각 평가 역량 지표에 미치는 효과가 유의한지 검증하였다.

방 법

연구 대상

대학생 평가자로는, 서울 소재 한 종합대학교의 심리학 관련 과목을 수강한 학부생 112명이 참여하였다. 모든 자극에 대해서 동일한 점수를 준 '한 줄 응답(straight lining)'을 나타낸 2명은 불성실 응답으로 간주하여 분석에서 제외하였다. 최종 분석에 사용된 총 110명의 성별 구성은 남자 73명(66%), 여자 37명(34%)으로, 남자 참여자가 두 배가량 많았다. 연령 범위는 18~26세였다(M = 21.06, SD = 1.81).

전문가 평가자로는, 각 평가영역마다 2인씩 총 4인이 참여하였다. 시공간영역의 Finke 창의적 발명 과제에 관한 전문가 2인은 모두 창의성 교육 현장에서 해당 검사의 평가자로 참여한 경험이 있는, 교육 관련 박사학위 소지자였다. 언어영역 과제인 광고문 전문가 2인은 광고업계에서 각각 5년과 10년 이상 근무한 광고기획자들이었다.

도구

평가영역별 과제와 평가물. 연구참여자의 인지적 부담을 고려하여 언어와 시공간, 두 영역만 확인하였다.

먼저 시공간영역의 과제로는 Finke(1990)의 창의적 발명 과제(Finke Creative Invention Task, 이하 FCIT)를 사용했다. 심상과 창의성을 결합한 FCIT는 정육면체 혹은 원뿔과 같은 15개 도형을 제시한 뒤 이 중 3개를 골라서, 제시된 8개 범주(가구, 무기 등) 중 하나에 속한 창의적 산물을 발명하도록 요구한다. 15개 도형을 얼마나 창의적으로 결합시켰는지를 판단하는 것이 중요하므로 평가자 또한 시공간 지능을 사용하게 된다. 평가자들에게 제시된 FCIT 발명물은 본 연구와는 별도로 진행된 보조 실험을 통해 수집한 다음 30개를 선별하였다. 보조 실험 역시, 심리학 과목을 수강하는 학부생들이 실험실을 방문하여 FCIT를 수행하는 방식으로 진행되었다. 보조 실험의 모든 참여자들에게 지시문과 도형을 보여주고 "창의적으로(새롭고 유용하게) 쓰일" 수 있는 사물을 생각해낸 다음 그것을 그림으로 그리고 그 기능과 작동법을 상세히 기술하도록 요청하였다. 이때, 참여자 가운데

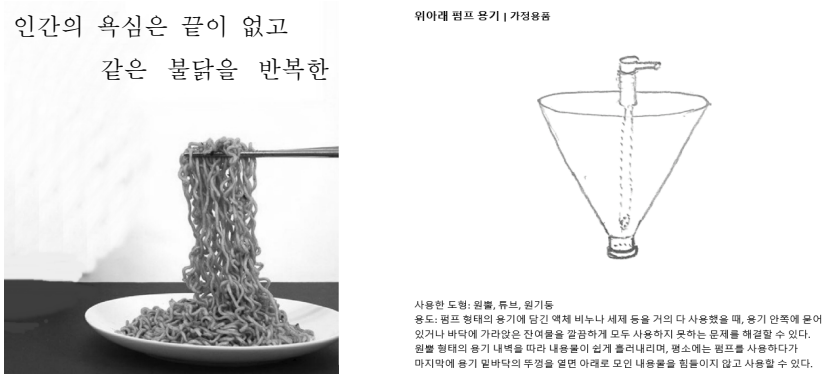

Figure 2. Examples of output from verbal-linguistic task and visual-spatial task.

절반에게는 정해진 범주의 사물을 발명하도록 요청했고, 나머지 절반에게는 8개 범주 가운데 선호하는 것을 선택해서 발명하도록 하였다. 이는 범주 제약 조건의 참여자들이 더 창의적인 산물을 발명한다는 선행 연구 결과(Finke et al., 1992)를 참고한 것으로, 이 조작을 통해 30개 발명물의 창의성 수준이 다양해지기를 기대하였다.

언어영역 과제로는 작문을 사용하였다. 작문은 전통적으로 언어능력 평가의 주요 과제였고, 평가자에게도 언어 지능이 요청된다. 작문은 그 자체로 난이도가 높은 과제이기 때문에, 작문을 수행하는 학생과 채점하는 학생 양쪽의 인지적 부담을 고려하여 광고문이라는 단문 작성 과제를 선택하였다. 광고문 작성 역시, 별도의 보조 실험으로 진행되었다. 구체적으로는, 대학생들에게 인기 있는 라면 상품의 공모전 지시문과 사진을 보여주고 한 문장의 광고문을 여러 개 작성하도록 요청하였다. 자극의 수준을 다양하게 하기 위해, 실제 공모전 수상작 6개를 최종 30개 자극에 포함하였다.

참고로, 평가물 수집을 위한 보조 실험과 본 연구에 둘 다 참여한 사례는 없었다. Figure 2에 각 영역의 평가물 예시를 제시하였다.

평가준거와 채점척도. 연구 대상이 미숙련 평가자이기 때문에 분석적인 채점(analytic scoring)보다 인지 부담이 적은 총체적 채점(holistic scoring)을 하도록 설계하였다. 따라서 각 영역당 평가준거를 하나씩 정의하였다. 언어영역의 광고문에 대해서는 '우수성'을, 시공간영역의 발명물에 대해서는 '창의성'을 각각 평가준거로 제시하였다. 대학생과 전문가 모두, 평가물이 얼마나 우수한지 또는 얼마나 창의적인지를 1점부터 10점 사이의 점수로 자유롭게 채점하도록 요청받았다. 각 점수가 의미하는 구체적인 수행 수준을 정의한 루브릭은 제시하지 않았다.

절차

연구참여자들은 연구자가 제시한 일정표에서 원하는 시간을 신청해 실험실을 방문하였다. 생명윤리법을 준수한 연구 설명문을 읽고 자발적으로 동의한 다음에야 연구에 참여할 수 있었다. 참여자들은 퀄트릭스(Qualtrics)로 제작된 지시문과 자극이 컴퓨터 화면에 제시되면, 그에 따라 자극을 채점하였다. Figure 3에서 볼 수 있듯이, 자극은 언어영역 블록과 시공간영역 블록으로 나뉘어 제시되었는데, 블록 제시 순서는 역균형화되었고, 블록 내 자극 30개의 순서도 무선으로 배열되었다. 또한 다른 자극의 수준을 참고하여 채점할 수 있도록 같은 블록 내의 30개 자극은 하나의 화면에 제시되었고, 참여자들은 화면을 스크롤하여 이미 채점한 자극을 다시 관찰하여 점수를 수정할 수 있었다. 하나의 블록을 완료하고

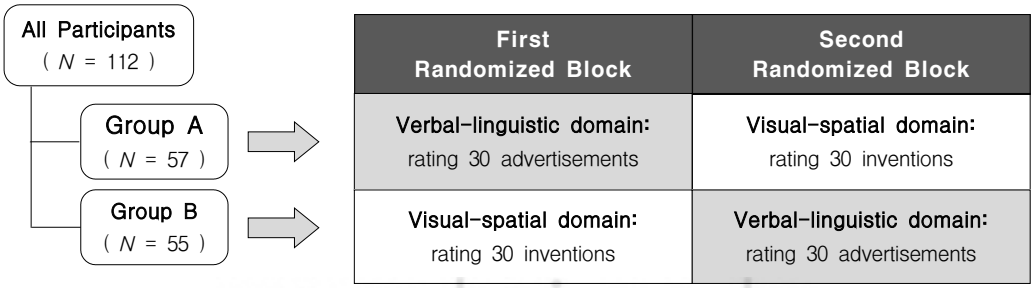

Figure 3. Counterbalanced repeated measures study design.

나면, 해당 영역에 대한 자신의 채점 근거를 간략하게 서술하도록 요구받았다.

자료 분석

점수의 분포를 검토했을 때 2개 조건 모두에서, 9점과 10점이 다른 점수대에 비해 현저히 적은 빈도를 보였기 때문에(부록의 표 참조), 10점은 9점으로 합쳐서 이하의 분석을 행하였다.

평가 역량 지표. 앞서 상술하였듯이 크게는 채점 정확도와 채점자 편향 두 가지로 평가 역량을 확인하였다. 채점자 편향은 다시, 채점 엄격성과 채점 무작위성과 중앙값 채점 경향성 세 가지로 나뉘어 측정되었다. 따라서 총 4가지 지표가 평가 역량 변인으로 사용되었다.

채점 정확도는 Cronbach(1955)의 4가지 정확도 가운데 평가 대상에 대한 변별 정확도(Differential Elevation, 이하 DE)를 사용하였다. 본래 Cronbach가 제시한 DE는 식 (1)과 같다.

$$DE = \frac{1}{N}\sum_{o}^{N}\left[\left(\overline{x_{o.}} - \overline{x_{..}}\right) - \left(\overline{t_{o.}} - \overline{t_{..}}\right)\right]^2 \quad (1)$$

N은 평가물 수, $x_{o.}$는 채점자가 모든 문항에서 평가물 O에 준 점수의 평균, $\overline{x_{..}}$는 채점자가 모든 문항에서 모든 평가물에 준 점수의 평균, $t_{o.}$는 모든 문항에서 평가물 O의 진점수들의 평균, $\overline{t_{..}}$는 모든 문항에서 모든 평가물의 진점수들의 평균이다. 그런데 본 연구는 문항, 즉 과제가 하나이기 때문에, 문항에 대한 평균 부분을 없애면 식 (2)와 같이 바뀐다.

$$DE = \frac{1}{N}\sum_{i}\left[\left(x_o - \overline{x_{.}}\right) - \left(t_o - \overline{t_{.}}\right)\right]^2 \quad (2)$$

식 (2)를 Cronbach(1955)를 따라 해석하면, $\overline{x_{.}}$는 대학생 평가자가 생각하기에 해당 과제에서 평가물의 평균적인 질을 나타낸다. 따라서 $\left(x_o - \overline{x_{.}}\right)$는 평가물 O의 질이 평균에서 얼마나 떨어져 있는지에 대한 대학생 평가자의 추정을 의미한다. $\left(t_o - \overline{t_{.}}\right)$는 평가물 O가 실제로 평균에서 얼마나 떨어져 있는지를 의미한다. 즉 식 (2)는 평가물의 질이 평균 수준에서 떨어진 정도에 대한 대학생 평가자의 추정이 실제와 얼마나 유사한지를 보여주는 것이다. 따라서 식 (2)로 구한 채점 정확도는 점수가 낮을수록 평가물의 질을 정확하게 분별했음을 의미한다.

채점 정확도를 위한 진점수는 각 영역의 전문가 2인의 점수를 평균하였다. 전문가 2인의 점수의 신뢰도는 급내상관계수(Intraclass Correlation Coefficient)로 확인하였다. 본 연구의 전문가 참여자 2인이 해당 영역의 전문가들을 대표한다고 간주하였기 때문에 2요인확률효과모형을 사용하였다. 또한 2인의 평균을 사용할 것이기 때문에 분석 단위는 '평균(average)'으로, 신뢰도 유형은 '일관성(consistency)'으로 설정하여 산출하였다(McGrow & Wong, 1996). 프로그램은 R의 'irr package'를 사용하였다. ICC(C, 2)는 두 영역 모두에서 .82로 나타났다. 적정한 수준의 신뢰도가 확보되어 본 연구의 전문가 점수를 일반화할 수 있으므로 진점수로 간주하고 사용하였다(Koo & Li, 2016).

채점 엄격성과 채점 무작위성은 MFRM으로 추정하였다. 본 연구에서는 각 영역별 문항이 하나이기 때문에 피평가자 국면과 채점자 국면만을 포함하는, 다음과 같은 2국면 채점 척도 라쉬모형을 사용하였다. 모형의 상세는 아래와 같다.

$$\log\left(\frac{P_{njk}}{P_{nj(k-1)}}\right) = B_n - C_j - F_k \quad (3)$$

P_{njk} : 피평가자 n이 대학생 채점자 j로부터 점수 k를 받을 확률

$P_{nj(k-1)}$: 피평가자 n이 대학생 채점자 j로부터 점수 (k-1)을 받을 확률

B_n : 피평가자 n의 능력 수준

C_j : 대학생 채점자 j의 엄격성 수준

F_k : 점수 (k-1)에 비해 점수 k를 받기 어려운 정도

각 영역마다 위의 모형을 별도로 적용하였다. 따라서 B_n은 각각, 피평가자가 우수한 광고문을 쓸 수 있는 능력, 도형을 창의적으로 조합하여 발명할 수 있는 능력을 의미하게 된다. 각 능력 수준에 대한 잠재변수 공간의 로짓 위치로 추정된 C_j를 채점 엄격성 측정치로 사용하였다. 로짓 값이 클수록 채점자가 높은 기대치를 적용해, 피평가자가 실제 받아야 할 점수보다 낮은 점수를 주었음을 의미한다. C_j에 대한 외적합도는 채점 무작위성 측정치로 사용하였다. 값이 작을수록 비슷한 능력의 피평가자에게는 비슷한 범주의 점수를 부여했음을, 즉 일관성 있게 채점했음을 의미한다.

연구 1의 자료에서 위의 모형이 적절하게 적합되었는지

검토한 결과는 다음과 같았다. 잔차가 2 이상인 사례의 비율은 언어영역 평가 조건과 시공간영역 평가 조건 모두에서 3.0%로 나타났다. 잔차가 3 이상인 사례의 비율은 언어영역 평가 조건에서는 0.7%, 시공간영역 평가 조건에서는 0.5%로 나타났다. 표준화된 잔차가 2 이상인 관찰 사례의 비율이 5% 이하, 3 이상인 관찰 사례의 비율이 1% 이하이면 모형이 자료에 적합한 것으로 해석될 수 있다(Linacre, 2021, p.178). 따라서 연구 1의 자료에서 전체 모형의 적합도는 적정한 수준으로 판단되었으므로, 채점 엄격성과 채점 무작위성 추정치를 분석에 사용하였다. MFRM 프로그램으로는 FACETS 버전 3.83.6을 사용하였다.

마지막으로 중앙값 채점 경향성은 각 채점자들이 4점부터 6점까지의 3개 점수를 사용한 백분율로 확인하였다. 이 백분율 또한 두 영역 각각에서 별도로 산출하였다.

연구 문제 분석. 평가 역량을 영역 일반적 구인으로 볼 수 있는지는 2개 영역 간의 상관계수로 확인했다. 구체적으로는 평가 역량 지표 4가지 각각에 대해서 Pearson적률상관계수를 구하였다. 평가영역 차이가 평가 역량에 미치는 영향이 유의한지에 대해서는, 자료의 비정규성 때문에 윌콕슨 부호순위검정을 이용하였다. 효과크기는 대응쌍 순위-이연 상관계수로 확인하였다(Kerby, 2014).

결 과

대학생들의 채점 엄격성(Severity)과 중앙값 채점 경향성(Centrality), 채점 무작위성(Randomness), 채점 정확도(Accuracy)의 기술통계량, 그리고 네 가지 지표들 간의 평균적인 상관관계는 Table 1에 제시되었다. 채점 엄격성은 값이 클수록 엄격하게 채점한 것을, 채점 무작위성은 값이 클수록 채점 엄격성을 일관성 없이 적용한 것을, 중앙값 채점 경향성은 값이 클수록 4~6점대 점수를 많이 사용한 것을, 채점 정확도는 값이 작을수록 전문가와 거리가 가까운 것, 즉 정확하게 채점한 것을 의미한다. 평가 역량 지표들 간의 상관계수는 연구 조건별 상관계수를 구해서 평균한 것이다. 채점 무작위성과 채점 정확도는 높은 정적 상관관계이고, 중앙값 채점 경향성은 앞의 두 지표와 부적 상관관계이다. 즉, 중앙값을 많이 사용할수록, 일관성 있게 채점할수록 채점 정확도가 높아지는 경향성이 있다.

언어영역(Verbal-linguistic domain)에서의 평가 역량과 시공간영역(Visual-spatial domain)에서의 평가 역량 간의 상관관계를 확인한 결과는 Figure 4와 같았다. 채점 정확도는 두 영역 간 상관계수가 .58(p<.001)이었다. 채점 엄격성의 영역 간 상관계수는 .50(p<.001), 중앙값 채점 경향성의 영역 간 상관계수는 .43(p<.001), 채점 무작위성의 영역 간 상관계수는 .63(p<.001)으로, 채점자 편향 지표들에서도 모두

Table 1. Descriptive Statistics of Results from Study 1.

	Verbal-linguistic		Visual-spatial		Correlations		
	Mdn	*IQR*	*Mdn*	*IQR*	Severity	Centrality	Randomness
Severity	0.02	0.40	0.06	0.52	—		
Centrality	.37	.27	.40	.23	-.29	—	
Randomness	0.90	0.66	0.86	0.68	.25	-.59	—
Accuracy	4.23	2.36	3.37	2.53	.16	-.52	.87

Note. Correlations were averaged across conditions.

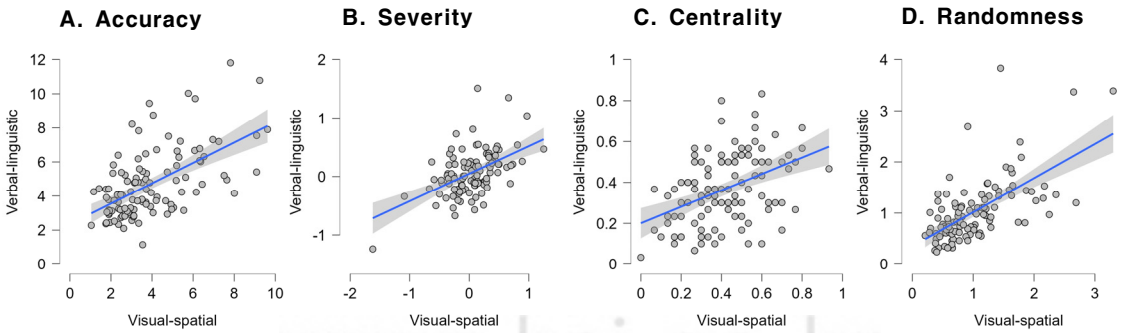

Figure 4. Correlation plots of evaluation capacity indices between two domains.

영역 간 상관이 유의하였다. 한 영역에서 평가를 잘하는 학생들은 대체로 다른 영역에서도 평가를 잘하는 편이었다.

더불어 평가영역 차이가 평가 역량에 유의한 영향을 미치는 현상도 부분적으로 관찰되었다. 언어영역을 평가했을 때보다 시공간영역을 평가했을 때 유의하게 더 정확하게 채점한 것으로 나타났기 때문이다[W=4508, p<.00, r=0.48]. 한편, 채점자 편향 지표 중에서는 중앙값 채점 경향성만 약한 수준으로 영역 차이 효과가 유의했다[W=2022, p<.02, r=-.27]. 채점 엄격성[W=3101, p<.76]과 채점 무작위성[W=3056 p<.99]에서는 영역 차이 효과가 나타나지 않았다.

논 의

연구 1은 대학생 평가자가 산출한 점수에서 채점 정확도와 채점자 편향을 추정하여 대학생의 평가 역량을 측정했을 때, 1) 평가 역량의 영역 일반성이 유의한지, 2) 평가영역 차이가 평가 역량에 미치는 영향력이 유의한지를 확인하는 것이 목적이었다.

대학생 평가 역량의 영역 일반성은 기대한 대로 네 가지 지표 모두에서 유의하게 확인되었다. 이를 보면 언어영역이나 시공간영역의 인지능력, 혹은 광고문구 작성이나 창의적 발명에 대한 선행 지식보다 상위 차원의 일반적 역량이 평가적 판단을 일정 부분 좌우함을 알 수 있다. 채점 정확도가 평가물의 수준을 잘 판별했는지를 뜻하고, 채점자 편향이 채점척도를 잘 사용했는지를 뜻한다는 점을 고려하면, 이 두 측면의 평가 역량이 영역 일반적임을 연구 1을 통해 확인했다고 할 수 있다.

다만 상관계수의 크기가 중간 수준(.43~.64)에 그쳤으므로, 개별 영역의 과제 특수성이 미숙련 평가자에게 미치는 영향력 또한 적지 않아 보인다. 이는 평가 역량 지표 중 일부에서, 즉 채점 정확도와 중앙값 채점 경향성에서 평가영역 차이 효과가 유의하게 나타난 것으로도 어느 정도 설명된다. 특히 채점 정확도에서 영역 차이의 효과크기는 .48로 작다고 할 수 없다. 게다가 Figure 4에서 채점 정확도의 영역 간 상관을 보여주는 산점도를 살펴보면, 부정확하게 채점할수록 영역 간 차이가 두드러진다(DE는 값이 클수록 전문가와의 거리가 멀다는 의미이다). 반면, 채점 정확도에서 상위 수준의 학생들은 영역 간 차이가 작은 편이다. 이를 보면 평가 역량이 낮은 학생들이 평가영역의 차이에 영향을 더 많이 받을 가능성이 있다.

본 연구는 학습자의 평가 역량을 이해하는 것이 중요하므로, 평가 역량이 낮은 학생들이 평가영역 차이에 상대적으로 더 크게 영향받을 가능성에 주목하지 않을 수 없었다. Table 1에서 알 수 있듯, 채점 정확도의 집단 중앙치(median)는 시공간영역 과제에서 더 낮게 나타났다. 즉, 학생들에게는 시공간영역의 발명물이 창의적인지를 평가하는 것이 언어영역의 광고문구가 우수한지를 평가하는 것보다 더 쉬웠던 것으로 해석된다. 그런데 연구 1은 평가영역마다 다른 평가준거를 사용했기 때문에, 평가과제의 인지영역 차이 때문에 채점 난이도가 달랐을 가능성도 있지만, 한편으로는 우수성과 창의성이라는 평가준거 차이 때문일 가능성도 배제할 수 없다. 우수하다는 것은 창의적이라는 것보다 더 상위의 개념이고 평가자마다 더 주관적으로 정의할 수 있기 때문이다. 이와 같은 평가준거의 모호함이 미숙련 평가자에게, 특히 평가 역량이 낮은 학생들에게 영향을 미쳤을 수 있다.

현실에서는 평가하려는 구인에 따라 평가준거가 정의되기 때문에 대체로 평가영역마다 다른 평가준거를 사용하는 것이 자연스럽다. 하지만 본 연구는 대학생의 평가 역량 측정과 관련된 변인들을 이해하는 것이 중요하므로, 평가준거가 평가 역량에 미치는 영향을 확인할 필요가 있다. 이를 위해 연구 2에서는 평가영역이 달라져도 평가준거는 같도록 설계하였을 때도 연구 1과 같은 결과가 관찰되는지 확인하고자 하였다.

연구 2

연구 2에서는 평가영역 효과에서 평가준거 효과를 분리하여 검증하고자 하였다. 이를 위해 연구 1에서 사용한 광고문구 자극과 발명물 자극을 독창성과 유용성 두 가지 평가준거로 채점하도록 하는, 2요인 피험자 내 측정을 실시하였다. 또한 연구 1에서 채점 정확도 수준이 낮은 학생들에게서 평가영역 간의 채점 정확도 차이가 큰 경향성이 관찰되었다. 이 점을 검증하고자, 채점 정확도가 높은 집단과 낮은 집단을 구분하여, 집단 요인과 다른 두 요인(평가영역 요인과 평가준거 요인) 간의 상호작용 효과를 확인하였다.

방 법

연구 대상

대학생 평가자로, 서울 소재 한 종합대학교의 심리학 과목을 수강한 학부생 64명이 참여하였다. 성별 구성은 여자 33명(52%), 남자 31명(48%)으로 비교적 균등하였다. 연령 범위는 19~30세였다(M = 22.78, SD = 2.75). 전문가 평가자는 연구 1과 동일하였다.

도구

평가영역별 과제와 평가물. 연구 1과 동일한 자극을 사용하였다.

평가준거와 채점척도. 전문가의 조언을 바탕으로, 두 평가 과제에 공통으로 사용할 수 있는 평가준거로 '독창성(Originality)'과 '유용성(Utility)'을 선정했다. 두 준거 모두 창의성의 대표적인 하위구인이므로(Barron, 1955; Diedrich et al., 2015; Runco & Jaeger, 2012; Stein, 1953), 연구 1의 '우수성'과 '창의성'처럼 위계가 다른 문제가 없다. 연구 1에서 대학생 평가자들이 작성한 채점 근거를 살펴봤을 때, 독창성과 창의성 두 개념을 동일시하는 학생들이 많았다. 이 점을 고려하여 각 평가영역별로 평가준거에 대해 다음과 같이 간략한 정의를 제시하였다.

광고문구의 독창성 : 광고문구가 재치 있고 참신하게 다가오는지 정도
광고문구의 유용성 : 제품의 장점을 강조하여 소비하고 싶은 욕구를 자극하는, 광고로서의 활용 가치 정도
발명물의 독창성 : 새로운 기능과 재질로서 독특하게 사용될지 정도
발명물의 유용성 : 실용 가능하고 쓸모 있게 이용할 수 있는 정도

채점척도는 10점 척도로, 연구 1과 동일하였다.

절차

연구 1에서 평가영역 블록의 순서가 종속변수에 미치는 영향이 유의하지 않았기 때문에, 연구 2에서는 각 조건당 연구 참여자 수를 확보해야 하는 부담을 줄이기 위해 평가영역 블록 순서를 역균형화하지 않았다. 이를 제외한 모든 절차는 연구 1과 동일하였다. Figure 2에서 보이듯, 발명물의 자극 하나가 광고문 자극 하나보다 정보량이 많기 때문에 연구참여자들의 인지적 피로를 고려하여, 발명물 평가를 먼저 실시하였다. 블록 내 30개 자극을 한 화면에서 무선적으로 배열하는 처치는 유지하였다.

자료 분석

연구 1과 마찬가지로 10점의 비중이 현저히 적었기 때문에(부록의 표 참조), 10점은 9점으로 합쳐서 분석하였다.

평가 역량 지표. 연구 1과 같은 통계량을 사용하되, 4개 연구 조건 각각에서 따로 산출하였다.

채점 정확도 산출을 위해서, 연구 1과 같은 절차로 전문가 2인의 신뢰도를 ICC(C, 2)로 구했을 때 언어영역은 독창성이 .93, 유용성이 .89였고, 시공간영역은 독창성이 .72, 유용성이 .73이었다. 시공간영역의 신뢰도가 다소 낮기는 했지만, 신뢰도를 높이기 위해 점수를 조정하는 과정이 오히려 타당도를 해칠 수 있다는 점과, .72와 .73의 신뢰도가 수용 가능한 수준인 점을 고려하여(Koo & Li, 2016), 추가 조정 없이 평균하여 진점수로 사용하였다.

또한 채점 엄격성과 채점 무작위성 추정을 위한 MFRM이 연구 2의 자료에 적절하게 적합되었는지 확인하였다. 표준화된 잔차가 2 이상인 사례의 비율이 각각 언어영역-독창성 조건에서는 4.7%, 언어영역-유용성 조건에서는 4.3%, 시공간영역-독창성 조건에서는 4.5%, 시공간영역-유용성 조건에서는 4.3%로 나타났다. 표준화된 잔차가 3 이상인 사례의 비율은 언어영역-독창성 조건에서는 0.4%, 언어영역-유용성 조건에서는 0.4%, 시공간영역-독창성 조건에서는 0.4%, 시공간영역-유용성 조건에서는 0.7%로 나타났다. 연구 1의 방법에서 언급한 모형 적합도 기준에 부합되므로, 채점 엄격성 측정치와 채점 무작위성 측정치를 분석에 사용하였다.

연구 문제 분석. 평가영역(언어영역과 시공간영역)과 평가준거(독창성과 유용성)가 달라도 평가 역량의 일반적인 측면이 유의하게 관찰되는지를 확인하기 위해 연구 1과 동일하게 상관분석을 실시했다.

또한 연구 2에서는 채점 정확도에 따라 상위집단과 하위집단을 구분하여, 평가영역 차이와 평가준거의 차이 효과를 검증하였다. 4개 조건(평가영역 2 × 평가준거 2) 중 최소한 조건에서 채점 정확도가 5.5 이상이면 하위집단으로 분류하였다. 5.5를 기준으로 삼은 것은 4개 조건의 채점 정확도 평균이 4.17~5.43으로 나타났기 때문이다. 채점 정확도 점수가 높을수록 부정확하게 채점했음을 의미한다는 점을 고려하면, 채점 정확도 하위집단은 한 조건에서라도 평균 수준 이하로 부정확하게 채점한 학생들이다. 상위집단 29명, 하위집단 35명으로 분류되었으므로, 집단 간 자료 크기 차이가 수용 가능하다고 보았다. 이상의 자료를 이용해, 평가영역과 평가준거, 집단에 따라 평가 역량의 수준이 달라지는지를 확인하기 위해 혼합설계 분산분석을 실시했고, 자료의 비정규성 때문에 R의 'ARTool package'를 이용해 Aligned Rank Transform 방식을 사용하였다(Wobbrock et al., 2011).

단, 평가 역량 지표 네 가지 간의 관계를 보기 위한 상관분석에서는 평가영역과 평가준거로만 자료를 구분하여 4개

조건에서 개별적으로 산출한 다음 평균하였다. 채점 정확도를 기준으로 상위집단과 하위집단으로 구분하면, 채점 정확도의 구간이 반으로 잘리면서 다른 세 지표들과의 상관관계를 제대로 관찰할 수 없기 때문이다.

결 과

연구 2에서 관찰된 채점 엄격성과 채점 무작위성, 중앙값 채점 경향성, 채점 정확도의 기술통계량, 그리고 네 가지 지표들 간의 평균적인 상관관계는 Table 2와 같았다. 평가 역량 지표 네 가지 간의 평균적인 상관계수들은 연구 1보다 대체로 더 작은 크기로 관찰되었다. 그럼에도 중앙값 채점 경향성이 채점 무작위성과 채점 정확도와 갖는 부적 상관관계, 채점 무작위성과 채점 정확도의 높은 정적 상관관계는 연구 1과 동일하였다.

영역 일반적인 평가 역량을 확인하기 위한 상관분석 결과는 Table 3과 같았다. 평가영역 간 상관계수들의 경우 연구 1과 마찬가지로 4개 지표 모두에서 유의하게 나타났다. 다만 연구 1과 비교했을 때 상관계수 크기가 채점 정확도와 채점 무작위성에서는 더 작아졌고, 채점 엄격성과 중앙값 채점 경향성에서는 더 커졌다. 한편, 연구 1에서는 확인할 수 없었던 평가준거(criteria) 간 상관계수들의 경우, 평가영역 간 상관계수보다 전반적으로 더 크게 나타났다. 특히 언어영역에서 채점 정확도의 평가준거 간 상관계수는 .75에 이른다. 채점자 편향 지표들에서도 언어영역의 평가준거 간 상관계수가 시공간영역의 평가준거 간 상관계수보다 더 큰 경향성이 있다.

평가 역량 지표 네 가지 각각에 대해서, 평가영역 차이 효과와 평가준거 차이 효과, 정확도 집단 차이 효과를 검증한 결과는 Table 4와 같았다. 평가영역 차이 효과의 경우, 연구 1과 마찬가지로 채점 정확도와 중앙값 채점 경향성에서 유의하게 나타났는데, 연구 2에서는 채점 엄격성에서도 유의했다. 또한 연구 1과 마찬가지로 연구 2에서도, 학생들이 시공간영역을 더 정확하게 채점하였으므로, 본 연구에서는 시공

Table 2. Descriptive Statistics of Results from Study 2.

Rating Quality Measures	Verbal-linguistic				Visual-spatial				Correlations		
	Originality		Utility		Originality		Utility				
	High	Low	High	Low	*High*	*Low*	*High*	*Low*			
	Mdn (IQR)	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	*Mdn (IQR)*	1	2	3
1. Severity	0.14 (0.34)	0.17 (0.56)	0.02 (0.41)	0.07 (0.54)	0.01 (0.34)	0.07 (0.39)	0.03 (0.37)	0.04 (0.36)			
2. Centrality	.43 (.37)	.27 (.17)	.50 (.20)	.30 (.18)	.47 (.27)	.33 (.22)	.47 (.23)	.30 (.17)	−.12		
3. Randomness	0.64 (0.33)	1.12 (0.47)	0.58 (0.37)	1.40 (0.67)	0.70 (0.28)	1.26 (0.65)	0.68 (0.48)	1.16 (0.61)	.04	−.48	
4. Accuracy	4.09 (0.93)	6.83 (2.81)	3.86 (0.88)	6.45 (2.07)	3.29 (0.98)	4.46 (2.23)	2.97 (2.25)	4.68 (2.13)	−.03	−.36	.83

Note. Correlations were averaged across two domains × two criteria conditions.

Table 3. Results of Correlation anyalysis.

	between Domains		between Criteria	
	Originality	Utility	Verbal-linguistic	Visual-spatial
Severity	.70***	.59***	.85***	.76***
Centrality	.66***	.64***	.71***	.66***
Randomness	.56***	.57***	.67***	.54***
Accuracy	.38**	.44***	.75***	.39**

Note. ** p <.01, *** p <.001.

Table 4. Summary of Repeated Measures ART ANOVA.

source	Severity			Randomness			Centrality			Accuracy		
	df	*F*	partial η^2	df	*F*	partial η^2	df	*F*	partial η^2	df	*F*	partial η^2
Criterion(A)	1	8.134**	.042	1	0.000	.000	1	0.778	.004	1	0.185	.001
Domain(B)	1	3.952*	.021	1	0.627	.003	1	8.104**	.042	1	61.819***	.249
A × B	1	6.309*	.033	1	0.110	.000	1	7.337**	.038	1	0.016	.000
Group(C)	1	0.154	.002		74.157***	.545	1	21.462***	.257	1	125.769***	.670
A × C	1	0.391	.002		0.060	.000	1	0.035	.000	1	0.122	.000
B × C	1	0.131	.001		2.245	.012	1	0.261	.001	1	12.918***	.065
A×B×C	1	0.076	.000		3.290	.017	1	0.171	.001	1	0.012	.000

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

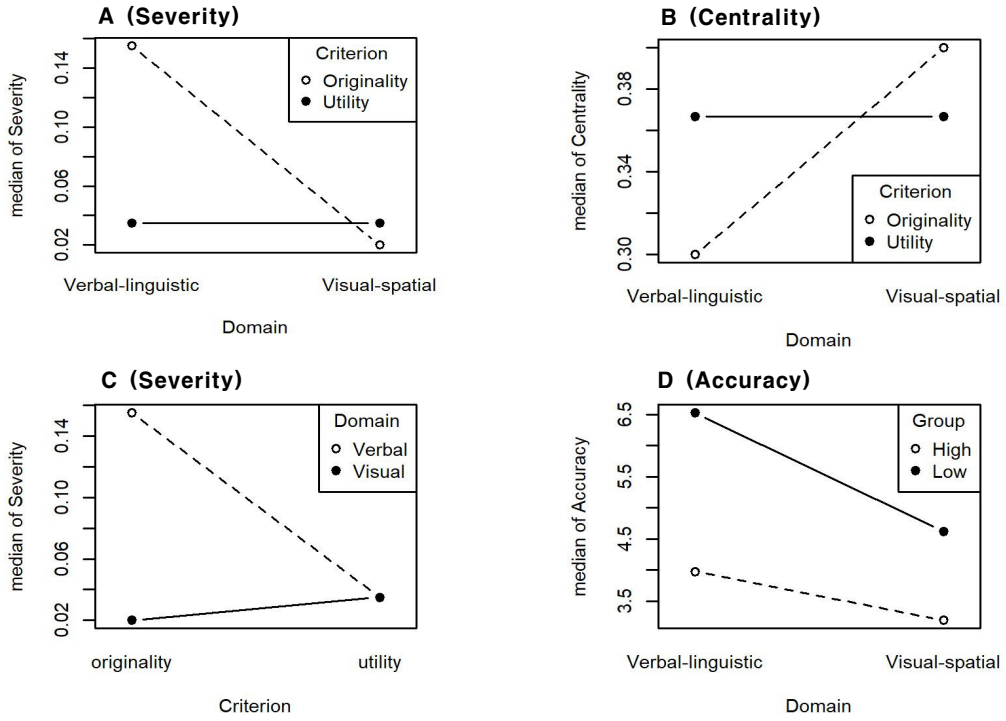

Figure 5. Interaction plots from Study 2.

간영역 평가가 언어영역 평가보다 난이도가 더 낮은 것으로 보인다. 중앙값 채점 경향성과 채점 엄격성의 경우, 평가영역과 평가준거 간의 상호작용 효과도 유의하게 나타났다. Figure 5의 A와 B를 보면 알 수 있듯, 채점 엄격성과 중앙값 채점 경향성 모두 독창성 준거로 평가했을 때 평가영역 간 차이가 크게 나타났고, 유용성 준거로 평가했을 때는 평가영역 간 차이가 크지 않았다. 즉, 채점 정확도에서는 평가영역 간 차이 효과가 평가준거와 상관없이 유의하게 관찰되었지만, 채점 엄격성과 중앙값 채점 경향성에서는 평가영역 간 차이 효과가 평가준거에 따라 다르게 나타났다.

평가준거의 주 효과는 채점 엄격성에서만 유의하게 나타났는데, 평가준거와 평가영역 간의 상호작용 효과도 함께 관찰되었다. Figure 5의 C를 보면 언어영역을 평가했을 때만 평가준거에 따른 채점 엄격성 차이가 큰데, 이는 학생들이 광고문구 독창성을 평가할 때 유독 엄격하게 채점했기 때문이다.

마지막으로 채점 정확도 상위 집단과 하위 집단의 차이를 살펴보았다. 일단 채점 정확도에서 집단 차이 주 효과가 유의할 뿐 아니라 효과크기도 .67로 크게 나타난 것으로 보아, 집단 분류는 잘 이루어진 것으로 판단되었다. Figure 5의 D

를 보면, 상위집단은 두 평가영역에서 채점 정확도가 거의 비슷한 반면, 하위집단은 두 평가영역에서 채점 정확도가 크게 차이 났다. 즉, 평가 정확도가 낮은 학생들일수록 평가영역 차이에 더 큰 영향을 받는 경향이 확인되었다. 이는 연구 1의 논의에서 제시한 가설을 입증해주는 결과이다.

그러나 채점자 편향 지표들에서는 평가준거 차이나 평가영역 차이가 채점 정확도 집단 차이와 상호작용하지 않았다. 다만, 채점 무작위성과 중앙값 채점 경향성에서 집단 차이 효과가 유의했다. 정확하게 채점한 학생들은 4~6점을 50% 가까이 사용했고, 그들의 채점 무작위성 값은 1.0 이하이므로 채점 엄격성을 일관되게 유지했음을 알 수 있다(Table 2 참조).

논 의

연구 2에서 가장 주요한 문제는 평가영역과 평가준거를 별도의 요인으로 분리해서 평가 역량을 측정했을 때도 평가영역 차이 효과가 나타날 것인가였다. 연구 1에서 평가영역 차이 효과가 유의했던 채점 정확도와 중앙값 채점 경향성에서는 물론, 채점 엄격성에서도 평가영역 차이 효과가 유의하게 나타났으므로, 일관된 결과를 얻었다고 할 수 있다. 즉, 대학생들의 평가 역량은 어떤 영역의 과제로 측정하느냐에 따라서 다르게 나타날 가능성이 있다.

평가준거 차이 효과의 경우 채점 엄격성에서만 확인되었고, 평가준거와 평가영역의 상호작용 효과는 채점자 엄격성과 중앙값 채점 경향성에서만 관찰되었다. 채점 엄격성과 중앙값 채점 경향성은 채점척도를 특정 구간에 치우져 사용하는 것과 관련된 측정치이다. 따라서 대학생의 평가 역량 중 채점척도의 구간을 치우치지 않게 사용하는 능력은 평가준거의 영향을 받을 가능성이 있다. 반면, 채점 무작위성에서는 평가준거 차이 효과가 관찰되지 않았다. 채점 무작위성은 채점척도의 특정 구간에 대한 편향성이 아니라, 30개 평가물 전반에 채점척도를 일관되게 적용하는 능력을 나타낸다. 결국, 본 연구의 대학생들은 평가준거에 따라 채점척도의 구간은 다르게 사용했을지라도, 하나의 준거 내에서 30개 평가물들을 평정할 때는 비슷한 구간을 사용했다고 볼 수 있다. 이 문제는 종합논의에서 자세히 논하고자 한다.

비록 평가영역이나 평가준거가 달라질 때마다 대학생의 평가 역량이 가변적일 수 있음이 관찰되기는 했지만, 그와 동시에 평가 역량의 영역 일반적인 측면이 확인된 것 또한 연구 1과 연구 2의 공통된 결과이다. 그러나 대학생들의 평가 역량에서 영역 일반적인 부분의 크기에는 개인차가 존재하는 것으로 보인다. 평가영역에 상관 없이 일관되게 정확하게 채점하는 학생들, 즉 채점 정확도에서 영역 간 상관계수도 유의하고 영역 차이 효과도 작은 학생들이 있지만 그렇지 못한 학생들(본 연구에서 정확도가 낮은 집단)도 절반 이상으로 관찰되었기 때문이다. 즉, 대학생들의 종합적인 평가 역량에는 개인차가 존재하고, 이 역량이 덜 발달한 학생들은 평가영역과 평가준거에 더 큰 영향을 받을 가능성이 시사된다.

마지막으로 연구 1과 연구 2 모두에서 중앙값 채점 경향성과 채점 무작위성, 채점 정확도 세 지표의 관계가 두드러지게 관찰되었다. 먼저 평가 역량 지표들 간 상관분석에서 세 지표는 크기 .30 이상의 상관관계를 보였다. 더불어 연구 2에서는 채점 정확도 상위집단이 하위집단보다 중앙값 채점 경향성은 높고 채점 무작위성은 낮았다. 이처럼 중앙값을 많이 사용할수록 MFRM이 과적합되는(즉, 채점 무작위성 값이 작아지는) 경우가 있음은 Myford와 Wolfe(2004)에서도 언급된 현상이다. 또한 채점 무작위성과 채점 정확도의 유의한 부적 관계는 Wind와 Engelhard(2012)의 결과와 일치한다. Wind와 Engelhard의 연구에서는 채점 정확도를 다른 측정치로 확인했기 때문에, 두 가지 다른 채점 정확도에서 채점 무작위성과의 관계가 반복 검증된 것은 이 둘이 강한 관계일 수 있음을 보여준다. 그러나 세 지표의 이와 같은 관계는 평가 역량을 측정하는 데에 오히려 걸림돌이 될 수 있다. 이 점 또한 종합논의에서 자세하게 다루고자 한다.

종합 논의

메타인지와 같은 평가적 판단력의 중요성을 인식한 연구자들 덕분에, 자기평가 또는 동료평가가 학습에 미치는 효과 및 그 실용 방안, 동료평가를 총괄평가로 확장하기 위한 문제해결 등의 연구들이 축적되어 왔다(Li et al., 2020; Park & Park, 2018; Topping, 1998, 2009). 그런데 정작 학습자의 평가적 판단력, 즉 평가 역량의 측정과 관련한 논의는 찾아보기 어렵다. 이에 본 연구는 학습자들이 직접 채점한 점수를 이용해 그들의 평가 역량을 측정하는 방안을 제안하였다. 구체적으로는 평가의 채점 변산성 연구에서 사용되는 통계량 중 채점자 개인 수준에서 채점 질을 확인해주는 채점 정확도와 채점자 편향을 이용하였다. 이렇게 측정한 평가 역량의 특성을 이해하기 위해, 언어영역 평가과제와 시공간영역 평가과제를 이용해 영역 일반성과 영역 특수성을 탐색하였다. 이를 통해 확인된 내용은 다음의 두 가지로 정리될 수 있다.

첫째, 채점 정확도와 채점자 편향 모두 영역 간에 유의한 정적 상관관계가 나타났는데, 이 결과를 보면 한 평가영역에

서 채점 정확도가 높고 채점척도를 잘 사용하는 학생들은 다른 평가영역에서도 그럴 가능성이 높다. 연구 2에서 평가준거를 다르게 조작했을 때에도 같은 결과가 관찰되었다. 비록 상관계수의 크기가 .38~.85로, 통계량과 관찰 조건의 차이에 따라 크게 변화했지만, 채점 정확도와 채점자 편향으로 측정한 평가 역량의 영역 일반성을 시사하는 결과이다.

둘째, 평가영역 혹은 평가준거가 달라지면 평가 역량 수준도 달라짐이 관찰되었지만, 일부 지표에서만 유의했다. 평가영역 차이 효과는 채점 정확도, 채점 엄격성, 중앙값 채점 경향성에서만, 평가준거 차이 효과는 채점척도 구간 사용과 관련된 채점자 편향(채점 엄격성과 중앙값 채점 경향성)에서만 유의성이 검증되었기 때문이다. 이를 보면, 평가 역량의 하위구인 중 일부만 영역 특수적일 가능성이 있다. 한편, 채점 정확도가 낮은 학생들이 높은 학생들에 비해 평가영역 간 정확성 차이가 더 컸다. 즉, 능숙한 평가자보다 미숙한 평가자가 각 평가영역의 고유한 특질에 영향받는 정도가 더 큰 것으로 보인다. 종합하면, 영역 특수적인 평가 역량은 일부 지표에서만 관찰될 가능성이 있으며, 평가 역량의 발달 수준에 따라서 그 양상이 다를 수 있는 듯하다.

본 연구는 대학생 평가 역량의 영역 일반성과 영역 특수성에 대해 위와 같은 결과를 얻었지만, 영역 특수성을 명확히 이해하기 위해서는 채점 난이도에 영향을 미치는 평가 요소들을 검증하는 후속 연구가 필요하다. 구체적으로 살펴보면, 평가영역에 따라 채점 정확도가 유의하게 다르게 관찰된 이유, 즉 평가영역마다 채점 난이도가 다르게 나타난 이유에 대해서는 세 가지 설명이 가능하다. 먼저, 연구자들이 의도한 대로 두 과제가 평가자에게 요구하는 주요 인지영역이 달랐기 때문일 수 있다. 즉, 광고문을 평가하는 것이 도형들을 조합한 발명물을 평가하는 것보다 대학생들에게는 더 어려웠을 수 있다는 것이다. 그런데 본 연구는 영역당 하나의 과제만을 사용하였기 때문에, 이 지점을 확인하기 위해서는 언어영역의 다른 과제와 시공간영역의 다른 과제를 이용해 반복 검증하는 후속 연구가 필요하다.

영역 간 채점 난이도를 발생시킨 또 다른 이유로는, 두 평가물의 정보량 차이를 생각해볼 수 있다. Figure 2에서 볼 수 있듯, 언어영역의 평가물인 광고문구는 고작 한 줄의 글이었지만, 시공간영역의 평가물인 발명물은 형태를 묘사하는 그림과 글이 제시되고 그 기능 또는 용도에 대한 설명글도 있었다. 즉 발명물의 가용 단서(cue)가 광고문의 가용 단서보다 더 많았다. 단서가 많을수록 정확한 판단을 한다는 연구 결과(Lee & Yates, 1992; Letzring et al., 2006)를 고려하면, 발명물에 가용 단서가 더 많아서 정확하게 평가하기가 더 용이했을 가능성을 배제하기 어렵다. 이를 통제하지 못한 것은 본 연구의 한계로 후속 연구에서는 평가과제마다 가용 단서를 비슷하게 통제하는 조치를 하는 것이 중요해 보인다.

마지막 이유로는, 언어영역 과제를 평가할 때보다 시공간영역 과제를 평가할 때 학생들이 채점척도의 중앙값을 많이 사용했기 때문에, 시공간영역에서 채점 정확도가 더 높게 나타났을 수 있다. 본 연구에서 중앙값 채점 경향성과 채점 정확도의 평균적인 상관계수는 연구 1에서는 −.52였고 연구 2에서는 −.36이었다. 즉, 척도의 중간 구간을 많이 사용할수록 채점 정확도가 높아지는 관계가 관찰되었다. 그런데 연구 1과 연구 2 모두에서 시공간영역에서 유의하게 중앙값 채점 경향성이 높았다. 그러므로 학생들이 시공간영역에서 4~6점대 점수를 많이 사용한 것이 채점 정확도를 향상시키는 결과를 초래했을 가능성이 있다. 게다가 연구 2에서 채점 정확도 상위집단이 하위집단보다 유의하게 중앙값 채점 경향성이 높았다는 점도 두 지표의 긴밀한 관계를 보여준다.

기실, 일반적으로 사람의 능력을 수량화할 경우 정규분포를 이룬다는 점을 고려할 때, 위와 같은 두 지표의 관계는 자연스러운 현상일 수 있다. 아주 뛰어나게 잘하는 학생과 매우 못하는 학생보다는 보통 수준의 학생이 더 많을 가능성이 높기 때문에, 중앙값을 사용하는 것이 확률적으로 더 정확하게 채점할 수 있는 것이다. 그러나 이런 식으로 채점 정확도가 높아지는 것은 평가 역량을 타당하게 측정하는 데 장애물이라 할 수 있다. 따라서 이를 보정하여 채점 정확도를 구할 수 있는 방안에 대한 후속 연구가 필요하다. Jin과 Wang(2018)이 중앙값 채점 경향성을 고려한 MFRM 모형을 제시한 바 있지만, 이 모형은 중앙값 채점 경향성을 보정하여 피평가자의 능력치를 추정하는 것은 가능하지만, 평가자의 채점 정확도를 추정하지는 못하기 때문이다.

중앙값 채점 경향성은 채점 무작위성과도 유의한 관계에 있다. 두 지표 간의 평균적인 상관계수는 연구 1에서는 −.59, 연구 2에서는 −.48이었으므로, 채점척도의 중앙값을 많이 사용할수록 일관되게 채점하는 경향성이 있는 것으로 나타난 것이다. 이러한 관계 때문에 Myford와 Wolfe(2004)는 채점자 개개인 수준에서 중앙값 채점 경향성이 있는지를 탐지하는 방법 중 하나로 채점 무작위성의 통계량인 MFRM의 외적합도가 지나치게 작은 사례들이 있는지 살펴볼 것을 권한 바 있다. 이처럼 채점 무작위성 또한 중앙값 채점 경향성을 보정할 수 있는 방안이 마련되어야 해석상의 어려움 없이 활발히 사용될 수 있을 듯하다.

이상의 논의를 통해 채점 정확도의 영역 특수성을 이해하기 위해 고려되어야 할 본 연구의 한계와 후속 연구를 제안

하였다. 한편, 평가 역량의 다른 한 지표인 채점자 편향과 관련해서는 평가영역과 평가준거의 상호작용 효과가 주요하게 관찰되었다고 상술하였다. 이 결과에 대해서도 아래의 논의를 고려한다면, 채점자 편향을 타당하게 해석하기 위해 필요한 평가 요소를 이해할 수 있을 것이다.

본 연구에서는 채점자 편향 중 두 개의 지표(채점 엄격성과 중앙값 채점 경향성)에서 평가영역과 평가준거의 상호작용이 관찰되었다. 즉, 채점척도의 특정 구간을 집중적으로 사용하는 편향에서만 평가영역과 평가준거의 상호작용 효과가 나타난 것이다. 구체적으로는, Figure 5의 B에서 볼 수 있듯이, 중앙값 채점 경향성은 시공간영역의 독창성 조건에서 가장 높았고, 언어영역의 독창성 조건에서는 두드러지게 낮았다. 반면 Figure 5의 A와 C에서 볼 수 있듯이, 채점 엄격성의 경우에는 언어영역의 독창성 조건에서 가장 두드러지게 높았다. 종합하면, 본 연구의 대학생들은 광고문구의 독창성을 평가할 때는 낮은 점수대를 많이 주고 중간 점수대는 적게 사용했으며, 발명물의 독창성을 평가할 때는 중간 점수를 많이 주었다. 즉, 동일한 독창성 준거로 평가하더라도 평가영역과 평가과제, 평가물 등의 특질에 따라 채점척도의 구간을 다르게 사용하였다는 것이다.

Cronbach(1955)는 한 평가자가 여러 대상들을 평정한 점수의 평균에 대해, 그 평가자가 추정한 표준적인 수준(norm) 혹은 전형적인 수준(stereotype)으로 해석하고 이것이 채점척도의 구간 사용과 관계 있다고 말한 바 있다. 채점자가 평가물의 표준 수준에 대한 점수를 추정하고 나면 그 점수를 중심으로 그보다 잘한 평가물과 그보다 못한 평가물을 배치할 것이라고 생각해볼 수 있다. 채점인지 과정이 이와 같다면, 본 연구의 대학생들은 평가영역과 평가준거의 조합, 즉 평가조건에 따라 평가물의 표준 수준을 다르게 추정하였고, 이것이 채점척도의 구간 사용 편향에 영향을 준 것으로 이해할 수 있다. 이것이 적절한 채점 행동인지 아닌지는 과제 난이도와 비교해야 알 수 있다. 하지만 본 연구는 두 영역의 과제물이 다른 사람들로부터 얻어졌기 때문에, 두 과제의 난이도를 절대적으로 비교하는 것이 불가능하다. 따라서 채점척도의 구간을 사용하는 역량을 측정하고 해석하기 위해서는 평가과제 난이도와 평가준거 난이도를 추정할 수 있도록 평가 도구를 설계하는 것이 필요할 것이다.

본 연구는 대학생들이 채점한 점수를 이용해 평가 역량을 측정하는 지표들을 제안하고 그 지표들의 영역 일반성과 영역 특수성을 탐색하였지만, 아직은 기초적인 연구 단계이기 때문에, 위에서 언급한 것 외에도 후속 연구를 통해 극복되어야 할 한계들이 있다. 먼저, 한 대학의 학생들을 대상으로 한 연구 결과라는 점에 이의를 제기할 수도 있겠다. 연구 2에서 채점 정확도의 상위집단과 하위집단으로 구분하여 분석했을 때, 채점 정확도와 중앙값 채점 경향성, 채점 무작위성이 집단에 따라 큰 차이를 보인 결과를 고려하면, 대학마다 평가적 판단력의 발달 수준이 다를 경우 다른 결과가 나올 가능성을 배제할 수 없다. 즉, 평가적 판단력의 발달이 너무 낮은 학생 집단에서는 본 연구에서 사용한 통계량들이 평가 역량의 개인차를 제대로 보여주지 못할 수도 있는 것이다. 또한 본 연구에서 사용한 통계량은 자료의 형태에 따른 제약이 있을 수 있다. 특히 MFRM의 경우, 채점자들이 동일한 평가물을 공유하도록 하는 채점 설계(connected rating design)가 중요하다. 따라서 '동료평가'와 '자기평가' 자료에 실제로 적용하기 위해서는, 실제 교육 장면에서 채점자를 안배하는 방식을 고려하는 과정이 필요할 수 있다.

이상의 한계에도 불구하고, 본 연구는 대학생의 평가 역량을 측정할 수 있는 지표들의 특성을 탐색함으로써 이에 대한 관심을 높이는 한편 가능한 후속 연구를 제안한 점에서 의의가 있다고 하겠다. 최근 인공지능 기술의 발달이 가속화되면서, 평가 연구에서도 '자동채점(automated scoring)'과 같은 대안을 찾고자 하는 시도가 증가하고 있다. 그러나 인공지능이 인간의 인지 산물을 통해 학습한다는 점, 그리고 인공지능의 성능을 평가할 때 인간의 수행과 비교한다는 점을 고려하면, 결국 인간의 인지과정을 이해하는 것이 기본이 되어야 함을 알 수 있다. 자동채점을 제대로 개발하기 위해서도 기술적인 연구도 중요하겠지만, 그 연구 결과를 타당화하는 토대가 되는 인간의 평가인지 과정에 대한 연구도 중요할 것이다(Bejar, 2012). 그런데 평가 역량을 측정하지 않고 인간의 평가인지 과정을 연구하기란 불가능하기 때문에, 활용도 높은 측정 방안을 마련하는 것이 시급해 보인다. 본 연구를 통해 밝혀진 사실들이 향후 평가인지를 측정하고 이해하는 데 도움이 될 수 있기를 기대한다.

References

Barron, F. (1955). The disposition toward originality. *The Journal of Abnormal and Social Psychology, 51*, 478-485. http://dx.doi.org/10.1037/h0048073

Becker, B. E., & Cardy, R. L. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. *Journal of Applied Psychology, 71*, 662-671. https://doi.org/10.1037/0021-9010.71.4.662

Bejar, I. I. (2012). Rater cognition: Implications for validity.

Educational Measurement: Issues and Practice, 31(3), 2-9. https://doi.org/10.1111/j.1745-3992.2012.00238.x

Cho, J. H., & Han, T. Y. (2008). The effect of rater training type on rating accuracy: Focusing on interaction effect of accountability and rater mood. *Korean Journal of Industrial and Organizational Psychology*, *21*, 745-768. https://doi.org/10.24230/ksiop.21.4.200811.745

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*, 328-338. https://doi.org/10.1016/j.learninstruc.2009.08.006

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52*, 177-193. https://doi.org/10.1037/h0044919

Diedrich, J., Benedek, M., Jauk, E., & Neubauer, A. C. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics*, *Creativity, and the Arts, 9*, 35-40. https://doi.org/10.1037/a0038688

Engelhard Jr, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement, 31*, 93-112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33*, 56-70. https://doi.org/10.1111/j.1745-3984.1996.tb00479.x

Engelhard, J. G. (2012). I*nvariant measurement: Using rasch models in the social, behavioral, and health sciences.* Taylor & Francis Group.

Finke, R.A. (1990). *Creative Imagery: Discoveries and inventions in Visualization (1st ed.).* Psychology Press.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative Cognition: Theory, research, and applications, Cambridge.* MA: MIT Press.

Furr, R. (2008). A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of Personality, 76*, 1267-1316. https://doi.org/10.1111/j.1467-6494.2008.00521.x

Jiang, J. P., Hu, J. Y., Zhang, Y. B., & Yin, X. C. (2022). Fostering college students' critical thinking skills through peer assessment in the knowledge building community. *Interactive Learning Environments*, 1-17. https://doi.org/10.1080/10494820.2022.2039949

Jin, K. Y., & Eckes, T. (2022). Detecting rater centrality effects in performance assessments: A model-based comparison of centrality indices. *Measurement: Interdisciplinary Research and Perspectives, 20*, 228-247. https://doi.org/10.1080/15366367.2021.1972654

Jin, K. Y., & Wang, W. C. (2018). A new facets model for rater's centrality/extremity response style. *Journal of Educational Measurement, 55*, 543-563. https://doi.org/10.1111/jedm.12191

Jin, M. S., Sohn, Y. M., & Chu, H. J. (2011). A study on development plan of K-CESA for college education assessment. *The Journal of Educational Administration, 29*, 461-486.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks*. Guilford Press.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: a flaw in human judgment*. Hachette UK.

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology, 3*. https://doi.org/10.2466/11.IT.3.1

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Lee, J.-w., & Yates, J. F. (1992). How quantity judgment changes as the number of cues increases: An analytical framework and review. *Psychological Bulletin, 112*, 363-377. https://doi.org/10.1037/0033-2909.112.2.363

Letzring, T. D., Wells, S. M., & Funder, D. C. (2006). Information quantity and quality affect the realistic accuracy of personality judgment. *Journal of Personality and Social Psychology, 91*, 111. https://doi.org/10.1037/0022-3514.91.1.111

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45*, 193-211. https://www.tandfonline.com/doi/full/10.1080/02602938.2019.1620679

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2021). *A User's Guide to FACETS: Rasch-Model Computer Programs*. www.winsteps.com

McGraw, K. O., & Wong, S. P. (1996). Forming inferences

about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46. https://doi.org/10.1037/1082-989X.1.1.30

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*, 555-576. https://doi.org/10.1177/0265532211430367

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part Ⅱ. *Journal of Applied Measurement, 5*, 189-227.

Park, J., & Park, J. A. (2018). The current state and prospects of peer assessment. *Korean Journal of Cognitive Science, 29*, 85-104. https://doi.org/10.19066/cogsci.2018.29.2.001

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology, 69*, 581-588. https://doi.org/10.1037/0021-9010.69.4.581

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal, 24*, 92-96. http://dx.doi.org/10.1080/10400419.2012.650092

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735-746. https://doi.org/10.1037/0021-9010.87.4.735

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 47-104). Psychology Press.

Stein, M. I. (1953). Creativity and culture. *The Journal of Psychology: Interdisciplinary and Applied, 36*, 311-322. http://dx.doi.org/10.1080/00223980.1953.9712897

Topping, K. (1998). Peer Assessment between students in colleges and universities. *Review of Educational Research, 68*, 249-276. https://doi.org/10.3102/00346543068003249

Topping, K. J. (2009). Peer assessment. *Theory into practice, 48*, 20-27. https://doi.org/10.1080/00405840802577569

Uto, M., & Ueno, M. (2018). Empirical comparison of item response theory models with rater's parameters. *Heliyon, 4*(5), e00622. https://doi.org/10.1016/j.heliyon.2018.e00622

Wind, S. A., & Engelhard, G., Jr (2012). Examining rating quality in writing assessment: rater agreement, error, and accuracy. *Journal of Applied Measurement, 13*, 321-335.

Wind, S. A., & Engelhard, G. Jr (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18*, 278-299. https://doi.org/10.1016/j.asw.2013.09.002

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorial analyses using only ANOVA procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*, 143-146. May 7-12. Vancouver, BC, Canada. https://doi.org/10.1145/1978942.1978963

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice, 31*(3), 31-37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

채점 정확도와 채점자 편향을 통해 본 대학생의 평가 역량

김경미[1], 박정애[2], 양지원[2], 박주용[2]

[1]서울대학교 협동과정인지과학전공, [2]서울대학교 심리학과 & 심리과학연구소

채점자의 역량을 개인 수준에서 측정하는 대표적인 지표는, 채점 정확도와 채점자 편향(채점 엄격성, 중앙값 채점 경향성, 채점 무작위성)이다. 두 지표는 주로 전문 채점자를 대상으로 연구되었지만, 영역이 다른 둘 이상의 과제를 통해 평가 역량의 일반성이 검증된 바 없다. 본 연구에서는 대학생들로 하여금 언어영역 과제와 시공간영역의 과제를 채점하게 한 결과로, 채점 정확도와 채점자 편향을 탐색하였다. 분석 결과 한 영역에서 채점 정확도가 높고 채점척도를 잘 사용하는 학생들은 다른 평가 영역에서도 그럴 가능성이 높고, 채점 정확도가 높은 학생들보다 낮은 학생들이 평가영역 차이에 영향을 받는 정도가 유의하게 더 컸다. 전체 논의에서는 대학생 평가역량 측정 연구에 시사하는 점과 한계점이 다루어졌다.

주제어: 평가 역량, 채점 엄격성, 중앙값 채점 경향성, 채점 무작위성, 채점 정확도, 다국면라쉬모형

부 록

점수의 빈도분포표

Study 1

Domains	Score									
	1	2	3	4	5	6	7	8	9	10
Verbal-linguistic	306 (.09)	411 (.12)	431 (.13)	404 (.12)	393 (.12)	416 (.13)	410 (.12)	307 (.09)	134 (.04)	88 (.03)
Visual-spatial	247 (.07)	320 (.10)	450 (.14)	470 (.14)	467 (.14)	438 (.13)	431 (.13)	275 (.08)	146 (.04)	56 (.02)

Note. Relative frequencies are presented in parentheses.

Study 2

Domains	Criteria	Score									
		1	2	3	4	5	6	7	8	9	10
Verbal-linguistic	Originality	217 (.11)	313 (.16)	293 (15)	251 (13)	208 (11)	206 (10)	233 (12)	142 (7)	73 (4)	44 (2)
	Utility	130 (7)	249 (13)	276 (14)	259 (13)	233 (12)	270 (14)	263 (13)	187 (9)	74 (4)	39 (2)
Visual-spatial	Originality	111 (6)	249 (13)	270 (14)	258 (13)	296 (15)	283 (14)	263 (13)	147 (7)	81 (4)	22 (1)
	Utility	151 (8)	259 (13)	252 (13)	285 (14)	271 (14)	230 (12)	241 (12)	143 (7)	100 (5)	48 (2)

Note. Relative frequencies are presented in parentheses.