인지과학, 제35권 제3호
Korean Journal of Cognitive Science
2024, Vol. 35, No. 3, 187～203
http://dx.doi.org/10.19066/cogsci.2024.35.3.002

비전문가의 평가 정확도 향상 방안 탐색: 공통 평가 항목 점수 기반 가중치 함수를 활용한 점수 보정 방법 연구*

송 민 해[1)] 구 현 우[2)] 박 정 연[1)] 임 재 서[3)] 박 주 용[1)†]

[1)]서울대학교 심리학과 & 아시아연구소 [2)]서울대학교 뇌인지과학과

[3)]전주대학교 상담심리학과

평가 활동은 학습이나 훈련에 도움이 되지만, 비전문가의 평가 정확도에 대한 우려로 인해 적극적으로 활용되지 않는다. 평가 정확도를 향상시키기 위한 몇몇 방안들이 있기는 하지만, 평가 외에도 추가적인 절차나 과정이 필요하다는 한계가 있다. 본 연구에서는 소수의 공통 평가 항목을 이용하여, 전문가 점수와의 차이에 따라 가중치를 부여함으로써 평가 정확도를 향상시키는 방안을 탐색하였다. 연구 1에서는 50명의 가상의 비전문가가 글을 평가하는 상황을 가정하고 시뮬레이션을 실시하였다. 그 결과, 비전문가와 전문가의 평가 점수 간 상관 정도에 따라 공통 평가 항목을 이용한 보정 결과가 달라짐을 발견하였다. 상관이 높을 경우 공통 평가 항목을 이용한 보정이 효과가 없었지만, 다를 경우 하나의 공통 평가 항목을 활용한 가중치로 평가 점수를 보정할 때 평가 정확도가 향상되었다. 연구 2에서는 실험 장면에서 주장문을 평가한 실제 자료를 이용하였다. 분석 결과, 연구 1에서와 같은 결과를 얻었다. 논의에서는 본 연구 결과를 실제 평가 장면에 적용할 가능성이 다루어졌다.

주제어: 평가, 평가 정확도, 공통 평가 항목, 비전문가, 시뮬레이션

---

* 본 연구는 어느 기관의 지원 없이 진행하였음.

† 교신저자: 박주용, 서울대학교 심리학과, 서울특별시 관악구 관악로 1
연구분야: 인지심리학
E-mail: jooypark@snu.ac.kr

타인의 결과물 혹은 과정의 질에 대해 점수를 매기는 판단 활동으로 볼 수 있는 평가는 흔히 지식과 경험이 있는 전문가의 몫으로 간주된다. 이런 인식은, 평가가 서열을 매기거나 자격을 부여하는 수단 혹은 후속 수행을 높이는 피드백을 주기 위한 활동으로 여겨져, 정확성과 공정성을 중요하게 고려하는 분위기에서 비롯된다. 이러한 인식에 맞추어, 평가와 관련된 심리학 연구도 대부분 평가의 신뢰도나 객관성을 높이거나 정확한 평가를 기반으로 효과적으로 피드백을 제공하는 방법 등에 집중되어 왔다(예, Alfieri et al. 2011; Steedly et al. 2008; Wisniewski et al., 2020). 연구뿐만 아니라 교육 현장에서도 사람들의 인식으로 인해, 학습자들은 피평가자로서 학습할 수 있는 기회는 충분히 제공받지만, 평가자로서 학습할 기회를 거의 제공받지 못한다.

학습자가 충분한 기회를 거의 제공받지 못하는 현재의 상황은 평가자로서의 활동, 즉 평가 활동이 다른 활동으로 대체되기 어려운 방식으로 학습을 촉진한다는 점을 고려한다면 아쉬운 면이 있다. 평가 활동은 과제 파악하기, 비교하기와 같은 다양한 인지 활동을 활성화하여 학습에 도움을 준다(Liu & Carless, 2006; Topping, 1998; Topping, 2010). 또한 평가자의 시각으로 다른 결과물을 바라보면 피평가자로서 학습할 때와는 다른 새로운 각도에서 학습 내용을 조망할 수 있다(Brown, 2005; Davies, 2000; Falchikov, 1995). 조직에서도 서로를 평가를 하는 과정을 통해, 자신의 역량을 향상할 수 있는 기회를 갖곤 한다(Sol, 2016; Villeval, 2020). 여기에 학생이 평가를 수행하고 그 결과를 활용하면 소수의 교수자와 전문가에게 쏠려 있던 평가 부담도 줄일 수 있다는 장점을 추가할 수 있다.

이런 여러 장점에도 불구하고, 학습자나 조직 내의 훈련 대상자를 포함한 비전문가에게 평가 기회를 충분히 제공하지 않은 이유는 평가의 신뢰도에 대한 우려 때문이다. 실제로 학생들은 동료 학생들이 자신의 결과물과 과정을 평가하기에 충분한 지식을 갖고 있지 않아, 이들의 평가를 성적에 반영하는 것이 공정하지 않다고 여겨 수업에서 동료평가의 도입을 부정적으로 바라본다(Kaufman & Schunn, 2011). 조직에서 각 개인의 성과를 정확히 파악하면서, 업무 역량을 향상시키는데 기여하는 다면평가도 평가 신뢰도를 우려하는 구성원으로부터 환영받지 못하고 있다(Ambrose & Cropanzano, 2003; Cheng & Warren, 2000; Franke et al., 2013; Vough & Caza, 2017).

이런 우려가 전혀 근거 없는 것은 아니다. 비전문가가 전문가에 비해 전문적인 지식이나 경험이 부족하기 때문에, 전문가와는 다른 부분에 초점을 두고 평가를 진행하여 상대적으로 평가가 정확하지 않은 경우가 있기 때문이다(Allodi et al., 2020; Seidel et al., 2021). 그렇지만 이런 부정확성은, 비전문가로 하여금 평가 활동에 참여하지 못하게 할 근거로써 사용하기보다는, 개선될 문제로 볼 필요가 있다. 이 문제가 개선되면, 평가 활동을 통한 이 이점을 살릴 수 있는 기회를 더 제공할 수 있기 때문이다.

실제로 비전문가의 평가 정확도는 다양한 방법을 통해 높일 수 있다. 예를 들어, 6명 이상의 비전문가가 평가에 참여하는 경우 전문가만큼 평가 정확도를 높일 수 있다(Jeffery et al., 2016). 사전 과제를 통해 측정한 평가 역량을 기반으로 평가 점수를 보정하거나, 평가를 하기 앞서 평

가 기준을 정확하게 이해하게 하고, 몇몇 예시를 직접 평가하게 하여 정확도를 높이는 방안도 있다(예, García Martínez et al., 2019; Jeffery et al., 2016; Panadero & Alqassab, 2019; Rico-Juan et al., 2022; Wang et al., 2019). 그러나 이런 방안들은, 평가자 수를 늘리거나 평가 활동과는 별도의 절차가 필요해, 교육이나 훈련 현장에서 사용되기 까다롭다. 이런 맥락에서, 본 연구에서는 비전문가의 평가 부담을 덜면서 실용성을 높일 수 있는 평가 정확도 향상 방안을 제안하였다. 그 방안은 소수의 공통 평가 항목을 평가한 후 해당 항목에서의 비전문가와 전문가 점수의 차이를 가중치로 반영하여, 점수를 보정하는 것이다.

미국의 일부 교육 현장에서 활용하는 UCLA의 Calibrated Peer Review(CPR) 시스템은 학생의 평가 역량에 따라 가중치를 두어 평가 점수를 보정하는 방식이 포함된 교육 시스템이다. CPR 시스템에서 학생들이 서로를 평가하기 이전에, Calibration 단계에서 교수자가 이미 평정한 세 문항을 평가하게 함으로써 각 개인의 평가 역량인 Reviewer Competency Index(RCI)을 1점에서 6점으로 평정한다. 그 후, 학생들은 자신의 평가 점수에 대한 피드백을 받고, 세 명의 동료 학생의 글을 평가한다. 이때, RCI가 높은 학생의 평가 점수는 더 많은 가중치를, RCI가 낮은 학생들은 더 낮은 가중치를 줌으로써 점수를 보정한다(Balfour, 2013; Price et al., 2016; Russell, 2004; Russell et al., 2017). 각 학생의 평가 역량을 측정하여 자동적으로 점수를 보정할 수 있는 장점을 지니지만, 각 교육 분야에 맞는 별도의 사전 검사 문항을 개발하고, 실시하고, 그 결과를 관리해야 하는 부담으로 인해 실용성이 떨어진다. CPR 대신 본 연구에서 제안하는 방법은, 사전 검사를 위한 문항을 별도로 개발하는 대신 실제 참여자들이 수행한 답변 중 전문가에 의해 사전에 채점된 소수의 공통 평가 항목을 이용하여 각 비전문가별로 가중치를 구한 다음, 이 가중치로 평가 점수를 보정하고자 한다.

소수의 공통 평가 항목으로 가중치를 구할 수 있는 부분은 대규모의 시험의 회차별 신뢰도를 확보하기 위해 앵커 문항이라고도 부르는 공통 문항을 사용하는 연구에서 착안하였다(Livingstone, 2014). Educational Testing Service(ETS)의 TOEIC이나 GRE는 매 회차 새로운 문항으로, 새로운 시험자의 역량을 측정해야 한다. 이때, 문항이 바뀜에 따라 기존의 시험에 비해 개인의 역량을 더 낮게 혹은 더 높이 평가하게 되면 신뢰도가 떨어지게 된다. 따라서, 시험의 신뢰도를 확보하기 위해서 시험자의 역량을 살펴볼 수 있는 앵커 문항도 같이 시험에 출제함으로써 회차 별 앵커 문항에서의 점수를 기반으로 최종 점수를 보정하곤 한다. 서로 다른 시험의 신뢰도를 확보하기 위한 앵커 문항은 비전문가의 평가 정확도를 향상시키기 위한 방법으로도 활용될 수 있다. CPR처럼 사전 검사를 하지 않더라도, 평가를 수행할 때 소수의 공통 항목도 같이 평가하게 하고 공통 항목에서의 점수 차를 바탕으로 수리적으로 평가 점수를 보정한다면 간편하게 비전문가의 평가 정확도를 향상시킬 수 있다. 이 방안은 별도의 문항을 개발할 필요가 없이, 임의의 평가 대상 중 소수를 임의로 추출하면 간편하게 사용할 수 있다는 장점을 지니고 있음에도 불구하고, 아직까지 이 가능성에 대해 탐색이 이루어지지 않았다. 따라서 본 연구는 사전 검사를 진행하지

않고, 비전문가들이 실제로 평가할 때 소수의 공통 항목을 같이 평가하게 하고 이를 이용하여 가중치를 산출하여 점수를 보정하는 방식이 타당한지를 모색하는 것을 목표로 하였다.

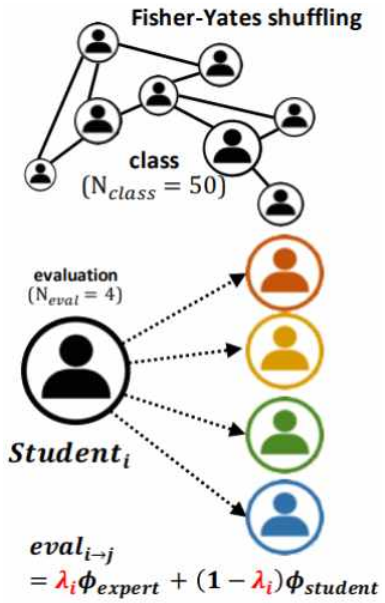
제안한 방식이 타당한지를 알아보기 위해, 연구 1에서는 가상의 데이터를 만든 후 전문가와 비전문가의 평가 경향의 차이에 따른 결과를 고려한 시뮬레이션을 실시하였다. 시뮬레이션은 전문가와 비전문가의 다양한 평가 경향성들을 모두 가정할 수 있기에, 시뮬레이션의 결과를 제안한 방식이 모든 상황에서 비전문가의 평가 정확도를 향상시킬 수 있는지 혹은 특정한 상황에서만 효과적인지를 파악하여 적절한 상황에 사용하는 근거로 활용하고자 하였다. 연구 2에서는, 실제 실험 장면에서 얻은 비전문가의 평가 데이터를 활용하여 제안한 방법이 실제로 비전문가의 평가 정확도를 높일 수 있는지를 살펴보았다. 시뮬레이션을 실시하는 것에서 그치지 않고, 실제 비전문가의 평가 데이터를 사용하여 타당도를 살펴보고자 하는 이유는 오염이 많은 실제 데이터에서도 제안한 방식이 효과적인지를 탐색하기 위해서다. 연구 1과 2에서 모두 제안한 방식이 타당함을 확인한다면, 실제 교육 및 훈련 현장에서의 비전문가의 평가 정확도를 높이는데 제안한 방식을 바로 활용할 수 있을 것으로 기대하였다. 각 연구의 구체적인 가설은 다음과 같았다.

연구 1 가설: 시뮬레이션 상에서, 공통 평가 항목에서의 비전문가와 전문가 점수의 차이를 활용한 가중치로 비전문가의 평가 점수를 보정하면 그렇지 않을 때보다 정확해질 것이다.

연구 2 가설: 실제 비전문가들이 생성한 글쓰기 자료에서, 공통 평가 항목에서의 비전문가와 전문가 점수의 차이를 활용한 가중치로 비전문가의 평가 점수를 보정하면 그렇지 않을 때보다 정확해질 것이다.

연구 1

연구 1에서는 총 50명의 비전문가가 서로의 글을 평가하는 가상의 상황을 가정하였다. 공통평가 항목에서 비전문가와 전문가 점수의 차이를 고려한 가중치로 평가 점수를 보정했을 때의 효과를 확인하였다. 프로그램을 통해 임의로 생성된 50명의 비전문가들에게 피셔-예이츠 셔플(Fisher-Yates shuffling) 알고리즘을 활용하여 무선적으로 평가 대상이 할당되었다. 피셔-예이츠 셔플은 유한된 수열을 무작위 순열로 섞기 위해 사용한 알고리즘으로, 시간 복잡도가 $O(n^2)$인 다른 알고리즘에 비해 $O(n)$으로 시간 복잡도가 낮아 대량의 자료를 처리해야 하는 시뮬레이션 상황에서 널리 활용하고 있다(Ade-Ibijola, 2012; Aishwarya & Beny, 2015).



(그림 1) 연구 1의 피셔-예이츠 셔플 예시

방 법

전문가 및 비전문가의 평가 점수

전문가와 비전문가는 비전문가의 50개의 글에 대해 통찰과 흐름이라는 두 기준에서 6점 만점으로 평가하였다고 가정하였다. 글은 실제로 생성되지 않았으며, 가상의 50편의 글에 대한 전문가와 비전문가의 점수만이 임의로 생성되었다. 이 때, 전문가와 비전문가 간 평가 방식의 차이가 고려되었다. 그 이유는 전문가와 다른 부분에 초점을 두어 평가가 이루어진다는 선행 연구 때문이다(Allodi et al., 2020; Seidel et al., 2021). 이를 근거로 본 연구에서는 전문가와 비전문가가 각각의 글을 채점할 때 중요하게 여기는 부분(각각, $\phi_{\text{expert}}, \phi_{non\,\text{expert}}$로 정의) 다를 수 있다고 가정하였는데, 이 $\phi_{\text{expert}}, \phi_{non\,\text{expert}}$의 차이는 클 수도 있고 작을 수도 있다. 결과적으로 둘 간의 상관인 평가의 일치도(ρ)는 높을 수도 낮을 수도 있다. 또한, $\phi_{\text{expert}}, \phi_{non\,\text{expert}}$는 각각 1에서 6 범위의 수로, 무선적으로 생성하였다.

이상의 가정하에 전문가의 평가 점수와 비전문가들의 평가 점수는 다음과 같이 계산되었다. 전문가의 평가 점수는 절대적인 수치로 가정하였으며, 여러 명의 전문가가 모여 하나의 통일된 점수를 생성했다고 가정하였다. 이때, 전문가의 평가 점수는 전문가가 중요하게 여기는 부분인

ϕ_{expert}를 반올림한 값인 $\lfloor \phi_{expert} \rceil$으로 정의하였다. 통찰과 흐름이라는 기준은 모두 1점부터 6점까지로 평가되었기에, 소수로 나올 수 있는 ϕ_{expert}를 그대로 활용하는 대신 반올림한 값을 전문가 평가 점수로 사용하였다. 한편, 비전문가들의 평가 점수는 평가 역량(λ), $\phi_{expert}, \phi_{non\,expert}$을 고려하여 결정되었다. 이때, λ는 비전문가들이 평가할 때 전문가의 관점과 얼마나 유사한지의 정도를 의미한다. λ, $\phi_{expert}, \phi_{non\,expert}$을 바탕으로 본 연구에서는 비전문가의 평가 점수는 다음과 같이 계산되었다. 이와 같이 계산한 이유는 다음과 같다. 본 연구에서 비전문가는 전문가와 달리 $\phi_{expert}, \phi_{non\,expert}$의 영향을 모두 받는다. λ가 높은 비전문가는 ϕ_{expert}를 더 고려하여 평가하며, λ가 낮은 비전문가는 $\phi_{non\,expert}$의 영향을 더 받는다. 이 부분들을 모두 고려하기 위하여 비전문가의 평가 점수를 하단의 식으로 계산하였다.

$$\text{비전문가의평가점수} = \lambda_i \phi_{expert} + (1-\lambda_i)\phi_{nonexpert}$$

가중치 함수

평가 정확도 보정은 다음과 같은 가중치 함수를 활용하여 진행되었다. 본 연구에서 제안한 가중치 함수는 공통 평가 항목의 비전문가의 평가 점수와 전문가의 평가 점수의 차이 그리고 공통 평가 항목 수를 모두 고려하여, 공통 평가 항목에서 비전문가가 전문가와 평가를 유사하게 할수록 해당 비전문가의 점수를 더 많이 최종 점수에 반영되게 하였다. 이 때, 가중치 w의 값이 음수가 나오지 않도록, 비전문가의 평가 점수와 전문가의 평가 점수의 차이를 구한 후 그 값을 제곱하였다. 또한, 평가 인원이 많을 때보다 평가 인원이 적을 때 공통 평가 항목을 통한 가중치의 영향력을 높이고자 평가 인원 수를 더하였다. 평가 인원이 많으면 비전문가의 평가 점수도 충분히 신뢰 가능하기 때문에, 평가 인원이 적은 상황이 비전문가의 평가 점수 보정이 더 필요하기 때문이다. 해당 함수는 구체적인 가중치 함수는 다음과 같다.

$$\text{가중치 } w = \frac{1}{m + \frac{1}{k}\sum_{1}^{k}(x_i - e_i)^2}$$

m = 평가 대상 수,
k = 공통 평가 항목의 개수,
x_i= 비전문가의 채점 점수,
e_i= 전문가의 채점 점수

평가 정확도 점수

본 연구에서는 전문가가 평가한 점수에서 실제로 비전문가가 평가한 점수 혹은 가중치로 보정된 점수를 뺀 절댓값의 합을 참여한 비전문가의 수로 나눈 것을 평가 정확도 점수로 정의하였다. 따라서 비전문가가 평가를 정확하게 수행할수록, 즉 전문가의 평가 점수와의 차이가 없을수록 평가 정확도 점수 값은 낮아진다.

$$평가점수 = \frac{1}{m}\sum_{1}^{m} | 전문가평가점수_i - 비전문가평가점수_i \text{or } 가중치로보정된점수_i|$$

분석

공통 평가 항목에서의 비전문가와 전문가 점수의 차이를 활용한 가중치로 평가 점수를 보정했을 때, 평가 정확도 점수가 낮아지는지 즉 비전문가의 평가가 더 정확해지는지를 확인하기 위해, 공통 평가 항목 수를 0, 1, 2, 그리고 3으로 다르게 하였다. 이들은 R의 Stan 플랫폼 상에서 4가지 다른 시나리오로 만들어졌다.

시나리오 1: 각 비전문가가 무선적으로 배정된 동료 비전문가 6명의 글을 평가하고, 한 비전문가의 최종 점수는 비전문가들이 평가한 점수의 평균값

시나리오 2: 각 비전문가가 무선적으로 배정된 동료 비전문가 5명과 공통 평가 항목 1개를 평가하고, 한 비전문가의 최종 점수는 비전문가들이 평가한 점수를 w로 가중된 평균으로 계산한 값: 비전문가의 최종 점수 = (비전문가가 평가한 점수 * w)의 평균

시나리오 3: 각 비전문가가 무선적으로 배정된 동료 비전문가 4명과 공통 평가 항목 2개를 평가하고, 한 비전문가의 최종 점수는 비전문가들이 평가한 점수를 w로 가중된 평균으로 계산한 값: 비전문가의 최종 점수 = (비전문가가 평가한 점수 * w)의 평균

시나리오 4: 각 비전문가가 무선적으로 배정된 동료 비전문가 3명과 공통 평가 항목 3개를 평가하고, 한 비전문가의 최종 점수는 비전문가들이 평가한 점수를 w로 가중된 평균으로 계산한 값: 비전문가의 최종 점수 = (비전문가가 평가한 점수 * w)의 평균

시나리오와 전문가와 비전문가의 평가 점수의 일치도인 ρ를 조작한 후 그에 따른 효과를 확인하였다. 일치도를 조작한 이유는 다음과 같다. 현실에서는 전문가와 비전문가의 평가 점수의 차이가 클 수도 있고, 작을 수도 있다. 예를 들어, 전문가와 비전문가의 간극이 좁은 분야는 평

가 점수의 차이가 작아, 이미 비전문가의 평가가 정확할 가능성이 있다. 한편, 전문가와 비전문가의 간극이 넓은 분야는 좁은 분야에 비해 더 평가가 정확하지 않을 수 있다. 제안한 방식이 전문가와 비전문가의 평가 점수의 일치도와 상관 없이 타당한지 혹은 일치도에 따라 효과가 달라지는지를 살펴본다면 적용 범위를 정확히 파악할 수 있을 것이다. 따라서, 연구 1에서의 ρ는 총 5가지 (-.8, -.4, 0, .4, .8)으로 조작하였으며, 각 공통 평가 항목의 수와 ρ값마다 총 500번의 시뮬레이션이 수행되었다.

결과 및 논의

각 시나리오에서 ρ값에 따른 평가 정확도는 <표 1>에 제시되었다. 전반적인 경향은, ρ가 음일 경우보다는 양일 경우, 평가 정확도 점수가 낮아졌다. 즉, 비전문가의 평가가 정확해진다. 또한 공통 평가 항목이 많아진다고 해서, 평가가 더 정확해지지 않고, ρ가 .8일 때를 제외하고는 공통 평가 항목이 한 개일 때 가장 정확하였다. 공통 평가 항목을 추가했을 때, 평가가 더 정확해지는 이유는 공통 평가 항목을 더 전문가스럽게 평가하는 비전문가는 가중치 함수의 계산에 따라 더 높은 가중치가 주어지고, 전문가와 다른 관점으로 평가하는 비전문가에게 낮은 가중치가 주어졌기 때문이다. 시나리오 2에 비해 시나리오 3과 4의 평가 정확도 점수가 더 높다는 결과는, 실제 평가 항목의 수와 공통 평가 항목의 수의 합이 동일할 경우, 공통 평가 항목을 한 개로 하고 실제 평가 항목의 수를 늘리는 편이 비전문가의 평가 점수를 더 정확하게 보정할 수 있다는 점을 보여준다. ρ가 .8로 높을 때에는, 공통 평가 항목을 사용하지 않은 시나리오 1에서 가장 정확하였다.

이상을 정리하면, 비전문가와 전문가의 평가 관점이 유사할 경우, 공통 평가 항목을 이용하여 점수를 보정하는 것이 비전문가의 평가의 정확도를 향상시키지 못한다. 하지만, 비전문가가 전문가와 다른 관점으로 평가를 할 경우, 한 개의 공통 평가 항목의 점수를 기반으로 한 가중치로 비전문가의 평가 점수를 보정하면 전문가의 평가 점수와 유사해진다. 따라서 적은 수의 비전문가로 하여금 평가할 때, 전문가의 점수와 차이가 있을 경우를 고려하여, 하나의 공통 평가 항목을 포함시키고 필요할 경우 보정할 수 있겠다. 그렇지만 연구 1의 결과는 가상의 데이터를 활용한 시뮬레이션의 결과이다 보니, 실제 장면에서 그 효과를 반복 검증할 필요가 있다. 따라서 연구 2에서는 실제로 자료를 사용하여, 제안한 방식이 비전문가의 평가 정확도를 향상시키는지를 확인하고자 한다.

<표 1> ρ값에 따른 네 시나리오별 평가 정확도 점수

ρ	시나리오 1	시나리오 2	시나리오 3	시나리오 4
-.8	.92 (*SD* = .01)	.86 (*SD* = .01)	.89 (*SD* = .02)	.90 (*SD* = .03)
-.4	.83 (*SD* = .02)	.79 (*SD* = .01)	.80 (*SD* = .01)	.82 (*SD* = .02)
0	.72 (*SD* = .01)	.71 (*SD* = .01)	.73 (*SD* = .01)	.75 (*SD* = .02)
.4	.61 (*SD* = .02)	.59 (*SD* = .01)	.62 (*SD* = .02)	.66 (*SD* = .02)
.8	.43 (*SD* = .01)	.45 (*SD* = .00)	.48 (*SD* = .00)	.51 (*SD* = .01)

연구 2

연구 2에서는 연구 1의 시뮬레이션 결과가 실제 비전문가들의 평가 상황에도 나타나는지를 살펴보았다. 이를 위해, 실험 장면에서 비전문가인 학부생 및 대학원생 참여자가 두 주제에 대해 각각 12편의 주장문을 평가한 데이터를 활용하였다. 실험에 활용한 자극의 주제인 전통 문화를 기반으로 한 기술 발전과 환경 문제에 대해 지식이 충분치 않은 학부생과 대학원생만 참여하였다. 각 평가 점수에 대해 비전문가와 전문가의 평가 점수가 얼마나 유사한지를 살펴본 후, 이에 따라 제안한 방식의 효과가 달라지는지를 살펴보았다. 연구 2에서의 가중치 함수, 평가 정확도 점수 계산 방법은 연구 1과 동일하였다.

방 법

참여자

서울 소재의 대학교 혹은 대학원에 재학 중인 모국어가 한국어인 학부생과 대학원생 46명을 대상으로 실험을 진행하였다. 실험 참여자의 평균 연령은 24.76세 (*SD* = 2.75)였으며, 남자는 9명, 여자는 37명이었다. 이들은 실험 참가의 댓가로 소정의 보상을 받았다.

주제 및 자극

본 연구에서는 두 가지 주제에 대해 각각 12편의 주장문을 자극으로 활용하였다. 각각의 주제는 '전통 문화를 기반으로 한 기술 발전 (전통 문화)'과 '환경 문제에 대한 낙관성 (환경 문제)'이었다. 두 주제를 선정한 이유는 참여자들이 쉽게 접하지 못한 주제라 전문적인 지식을 갖춘 참여자가 적을 것으로 기대했기 때문이다. 자극으로 활용한 주장문은 비전문가들이 반 페이지 정도의 분량으로 각각의 주제에 대해 반박하는 주장을 펼친 글이었다.

평가 기준

평가 정확도를 계산하기 위해 필요한 24편의 글에 대한 각각의 비전문가 평가 점수와 전문가 평가 점수를 산출하기 위해 46명의 비전문가인 참여자와 두 명의 전문가는 각각의 글을 '주장문의 주장과 근거에 대한 타당도와 설득력', '글의 흐름과 문장 전달력'이라는 두 가지 기준을 6점 만점으로 평가를 진행한 후 두 점수를 결합하여 사용하였다.

전문가 점수는, 한 명의 저자와 한 명의 박사과정 이상의 전문가가 24편의 글을 6점 만점으로 평가한 결과의 평균값을 활용했다. 두 전문가 모두 글쓰기 전문이며, 다양한 주제의 글을 평가한 경험이 있기 때문에 선정하였다. 두 명의 전문가는 개별적으로 채점을 진행한 후 각 기준에서 2점 이상 차이가 나는 글은 토의를 한 후 점수를 조정하였으며, 각각의 주제에서 ICC를 계산하여 신뢰도를 확보하고자 하였다. 그 결과, 두 주제에서 모두, ICC가 .9 이상으로(전통 문화: .99, 환경 문제: .96) 충분히 신뢰할 수 있음을 확인하였고 해당 점수를 전문가 점수로 활용하였다.

실험 절차

실험 참여자들은 외부와 차단된 실험실에서 개별적으로 실험을 진행되하였으며, 실험 진행 방법 및 내용에 대해 설명서를 읽은 후 IRB에서 승인받은 동의서 (IRB No. 2202/001-007)에 서명하였다. 먼저, 실험 참여자들은 첫 번째 주제에 대해 작성된 총 12편의 주장문을 두 가지 기준을 활용하여 총 28분 동안 평가하였다. 4분 동안 휴식을 취한 후, 두 번째 주제의 총 12편의 주장문을 동일하게 28분 동안 평가하였다. 두 평가 과정 모두, 참여자들은 자료를 탐색하거나 인터넷 검색 없이 평가를 수행해야 했다.

분석

공통 평가 항목을 활용한 가중치 함수의 효과를 살펴보기 위하여, 실험 1과 동일한 네 개의

시나리오로 분석을 진행하였다. 46명의 참여자의 평가 데이터를 각 시나리오별로 무선적으로 500번의 반복 추출하면서 각 평가 정확도 점수를 계산 후 비교하였다. 반복 추출하는 동안 공통 평가 항목은 각 주제별로 12개의 항목 중 무선적으로 선택 되었다.

결과 및 논의

연구 2의 결과는 <표 2>에 제시되었다. 전통 문화 주제에서는 ρ가 .74로 전문가와 비전문가의 평가 관점이 상당히 유사하였지만, 환경 문제 주제에서는 ρ가 .06로 두 집단관 상관이 나타나지 않았다. 전통 문화 주제에서는 단순 평균 점수를 최종 점수로 활용한 시나리오 1이 공통 평가 항목 점수를 가중치로 활용한 시나리오 2, 3, 4에 비해 평가가 더 정확하였다. 반면에 환경 문제 주제에서는 하나의 공통 평가 항목 점수를 가중치로 활용한 시나리오 2가 시나리오 1, 3, 4에 비해 평가가 더 정확했다. 이상의 결과 패턴은 연구 1의 시뮬레이션 결과와 부합한다.

연구 2의 결과는 연구 1의 시뮬레이션 결과와 패턴이 같았다. 즉, 비전문가가 전문가와는 다른 관점으로 평가할 경우, 본 연구에서 제안한 하나의 공통 평가 항목의 점수를 기반으로 한 가중치가 평가 정확도를 향상시킨다. 하지만, 비전문가가 전문가와 유사하게 평가할 경우, 단순 평균 점수를 활용하는 것으로 충분하다. 실험 자료를 이용한 분석 결과와 시뮬레이션 결과가 일관된 패턴을 나타내는 점에 주목할 필요가 있다.

<표 2> 각 주제에서의 네 시나리오별 평가 정확도 점수

주제	ρ	시나리오 1	시나리오 2	시나리오 3	시나리오 4
전통 문화	.74	.10 (*SD* = .01)	.12 (*SD* = .01)	.21 (*SD* = .02)	.28 (*SD* = .03)
환경 오염	.06	.62 (*SD* = .02)	.58 (*SD* = .01)	.69 (*SD* = .01)	.73 (*SD* = .02)

종합논의

본 연구는 비전문가의 평가 정확도를 높이기 위한 방법으로 공통 평가 항목에서의 비전문가와 전문가 점수의 차이를 활용한 가중치로 평가 점수를 보정하는 방법을 제안한 후, 두 개의 연구를 통해 타당한지를 살펴보았다. 가상의 데이터를 활용한 시뮬레이션을 실시한 연구 1에서는

비전문가가 전문가와 다른 관점에서 평가를 할 때, 제안한 단일한 공통 평가 항목에서의 점수 차이를 활용한 가중치로 보정하는 방법이 평가 정확도를 향상시킴을 확인하였다. 그렇지만, 비전문가와 전문가의 평가 결과가 어느 정도 일치할 경우는 공통 평가 항목의 점수를 가중치로 보정하지 않는 편이 더 나았다. 실험 상황에서 비전문가가 두 가지 주제의 주장문에 대해 평가한 자료를 이용한 연구 2에서는 연구 1의 결과와 같은 패턴이 관찰되었다. 시뮬레이션 결과와 실제 자료를 이용한 분석 결과가 수렴되는 점은 본 연구에서 제안하는 평가 점수 보정 방안이 어느 정도 타당성이 있음을 시사한다.

먼저 본 연구는 평가 정확도를 향상시킬 수 있는 방법을 새로 모색하여, 평가 정확도와 관련된 연구를 확장시켰다는 이론적 의의를 지니고 있다. 기존에는 평가 정확도를 향상시키기 위해 교육이나 훈련 등을 수행하는 등 간접적으로 향상시키는데 초점을 맞추었다. 물론, 사전 검사를 통해 평가 역량을 추정하고 수리적으로 점수를 보정하는 CPR이 있지만, CPR를 다룬 연구에서도 학습 효과를 살펴보는 것에 초점을 두며 Calibration 단계를 통해 얼마나 평가 정확도가 향상되는지, 다른 보정 방식이 있는지에 대해서는 거의 연구가 이루어지지 않았다. 그러나, 본 연구는 시뮬레이션과 실제 데이터를 통해 수리적으로 보정하는 것만으로도 평가 정확도를 향상시킬 수 있다는 점을 보여주었기에, 앞으로 본 연구를 기반으로 평가 정확도를 향상시킬 수 있는 다양한 방법을 연구할 수 있는 단초를 제공한다.

본 연구는 실제 비전문가의 평가 정확도를 향상시켜, 교육이나 훈련 장면에서 평가 기회를 높이기 위해 수행되었다. 앞서 살펴본 것처럼, 사람들이 수업이나 조직 현장에서 학생 혹은 동료의 평가를 신뢰하지 못하기 때문에, 동료평가나 다면평가가 적극적으로 도입되지 않고 있다 (Ambrose & Cropanzano, 2003; Cheng & Warren, 2000; Franke et al., 2013; Kaufman & Schunn, 2011; Vough & Caza, 2017). 본 연구의 의의는 상대적으로 번거롭지 않은 절차를 통해 비전문가의 평가 정확도를 높여, 지금보다 비전문가에게 평가 기회를 더 제공할 수 있는 기반이 될 수 있다는 점이다. 특히, 단 한 번의 평가에서도 하나의 공통 평가 항목을 추가로 평가하게 하고 이 점수를 이용하여 가중치를 주는 것만으로도 비전문가의 평가 정확도를 높일 수 있어, 교육과 훈련 장면에서 쉽게 적용할 수 있을 것으로 기대한다.

아울러, 비록 본 연구에서 다루지는 않았지만, 만일 더 많은 평가 기회가 주어져 각 개인의 평가 정확성을 더 많이 추정할 수 있다면, 이를 이용하여 비전문가의 평가 정확성을 향상시킬 여지가 있다. 실제 평가한 결과에 비해 단일한 공통 평가 항목만을 정확하게 평가하지 못하거나, 우연의 일치로 맞힐 경우 실제 평가 역량과는 다른 가중치가 계산되어 적절히 점수를 보정하지 못할 수 있다. 그러나 평가를 여러 번 반복한다면 더 정확히 평가 정확성을 추정하여 가중치가 더 정확해질 수 있다. 이런 가능성은 이미 조직 장면에서 활용되고 있다. 세계 최대의 헤지펀드회사인 브리지워터 (Bridgewater Associates)의 닷 컬렉터(Dot Collector)다. 브리지워터의 직원들은 닷 컬렉터를 활용하여, 누구에게나 심지어 최고 경영자에 대해게도 서로의 행동이나 업무

에 대해 자유롭고 즉각적으로 평가를 한다. 이 평가는 그 결과물이 누적되며, 각 직원들에 대한 역량을 더 정확하게 파악하고 개인에게 맞는 업무를 배정하는데 도움을 준다(Dalio, 2018). 물론, 단일한 평가를 통해서도 직원들의 역량을 확인할 수 있지만, 한 번으로는 미처 고려하지 못하거나 실수가 있어 반복하는 것이다. 브리지워터의 닷 컬렉터처럼 평가를 수행할 때마다 반복적으로 공통 평가 항목을 평가하고 이 차이를 기록하면, 보정에서의 오류가 점차 줄어들어 더 평가 정확도를 효과적으로 향상시킬 수 있을 것이다.

본 연구에서는 비전문가에 초점을 두었지만, 본 연구 결과는 전문가의 평가 정확도 향상에도 적용할 가능성이 열려있다. 비전문가에 비해서 전문가가 평가를 더 정확히 하지만, 전문가들도 항상 평가가 정확하지는 않기에 채점 기준을 상세히 작성하거나 전문가들 간 토론 등을 통해 정확도를 높이는 방법들이 제안되어 왔다(예, Bloxham et al., 2016; Hunter & Docherty, 2011). 그렇지만 제안된 방법들은, 별도의 활동이 필요하기에 전문가에게 쉽게 적용되기 어렵다. 그런데 만일 평가 대상자가 많아 여러 명의 전문가가 각기 나누어져서 평가해야 할 경우, 여기서도 공통 평가 항목을 사용하여 각 전문가의 평가 정확성에 가중치를 부여할 수 있다. 수많은 평가 대상자 중 무선적으로 한 두 평가 항목을 공통 항목으로 선정하고 이를 이용하여 가중치를 구하고, 가중치를 고려하여 점수를 보정하여 객관성을 유지할 것으로 기대한다. 특히, 이는 대규모의 시험 현장인 대학의 논술 시험이나 외국어 말하기 시험 등에 적용할 가능성이 있다.

이런 가능성을 포함하여, 본 연구의 결과를 적용하기 위해서는 다양한 후속 연구가 이루어져야 한다. 먼저, 본 연구에서는 시뮬레이션과 실험 상황에서 수집한 데이터로 제안한 방식의 가능성을 확인하였지만, 제한적인 자극과 평가 기준만을 활용하였다는 한계가 있다. 따라서 이후에는 실험 외에도 실제 수업이나 조직 내에서의 훈련에서 얻은 다양한 분야에서의 자극 그리고 다양한 평가 기준을 활용하여 제안한 방법의 효과를 반복 검증할 필요가 있다. 또한, 앞서 언급한 부분처럼 공통 평가 항목을 활용한 평가 점수 보정 방식이 전문가의 평가 정확도를 향상시킬 수 있는지 여부를 확인함으로써, 적용 가능성을 넓힐 수 있을지를 탐색해야 한다. 가장 중요한 후속 연구는 수업의 동료평가 장면이나 기업의 다면평가 등 실제 비전문가들의 평가 상황에서 장기간 제안한 방법을 활용하고 효과를 확인하는 것이다. 이 과정을 통해, 공통 평가 항목을 활용한 가중치가 안정적으로 평가 정확도를 향상시킬 수 있는지, 평가 대상자들이 해당 결과를 충분히 신뢰하는지를 확인해 보는 것이다. 본 연구와 후속 연구를 통해 비전문가들의 평가 정확도와 신뢰도를 향상시킬 수 있게 되면, 비전문가들이 교육과 훈련 현장에서 평가 활동을 통해 학습적 메타인지적 이점을 지금보다 더 누릴 수 있을 것으로 예상된다.

참고문헌

Ade-Ibijola, A. O. (2012). A simulated enhancement of Fisher-Yates algorithm for shuffling in virtual card games using domain-specific data structures. *International Journal of Computer Applications, 54*(11).

Aishwarya, C., & Beny, J. R. (2015). Novel architecture for data shuffling using fisher yates shuffle algorithm. *International journal of scientific research in science, engineering and technology, 1*, 387-390.

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of educational psychology, 103*(1), 1. https://doi.org/10.1037/a0021017

Allodi, L., Cremonini, M., Massacci, F., & Shim, W. (2020). Measuring the accuracy of software vulnerability assessments: experiments with students and professionals. *Empirical Software Engineering, 25*, 1063-1094. https://doi.org/10.1007/s10664-019-09797-4

Ambrose, M. L., & Cropanzano, R. (2003). A longitudinal analysis of organizational fairness: An examination of reactions to tenure and promotion decisions. *Journal of Applied Psychology, 88*(2), 266. https://doi.org/10.1037/0021-9010.88.2.266

Balfour, S. P. (2013). Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™. *Research & Practice in Assessment, 8*, 40-48.

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466-481. https://doi.org/10.1080/02602938.2015.1024607

Brown, S. (2005). Assessment for learning. *Learning and teaching in higher education*, (1), 81-89.

Cheng, W., & Warren, M. (2000). Making a difference: Using peers to assess individual students' contributions to a group project. *Teaching in Higher education, 5*(2), 243-255. https://doi.org/10.1080/135625100114885

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of educational psychology, 103*(1), 73. https://doi.org/10.1037/a0021950

Dalio, R. (2018). *Principles*. Simon and Schuster.

Davies, P. (2000). Computerized peer assessment. *Innovations in Education and Teaching International, 37*(4), 346. https://doi.org/10.1080/135580000750052955

Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Innovations in Education and training International, 32*(2), 175-187. https://doi.org/10.1080/1355800950320212

Franke, N., Keinz, P., & Klausberger, K. (2013). "Does this sound like a fair deal?": Antecedents and consequences of fairness expectations in the individual's decision to participate in firm innovation. *Organization science, 24*(5), 1495-1516. https://doi.org/10.1287/orsc.1120.0794

García Martínez, C., Cerezo, R., Bermúdez, M., & Romero, C. (2019). Improving essay peer grading accuracy in massive open online courses using personalized weights from student's engagement and performance. *Journal of Computer Assisted Learning, 35*(1), 110-120. https://doi.org/10.1111/jcal.12316

Hunter, K., & Docherty, P. (2011). Reducing variation in the assessment of student writing. *Assessment & Evaluation in Higher Education, 36*(1), 109-124. https://doi.org/10.1080/02602930903215842

Jeffery, D., Yankulov, K., Crerar, A., & Ritchie, K. (2016). How to achieve accurate peer assessment for high value written assignments in a senior undergraduate course. *Assessment & Evaluation in Higher Education, 41*(1), 127-140. https://doi.org/10.1080/02602938.2014.987721

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional science, 39*, 387-406. https://doi.org/10.1007/s11251-010-9133-6

Liu, N.-F., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher education, 11*(3), 279-290. https://doi.org/10.1080/13562510600680582

Livingston, S. A. (2014). Equating test scores (without IRT). *Educational testing service*.

Panadero, E., & Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment & Evaluation in Higher Education, 44*(8), 1253-1278. https://doi.org/10.1080/02602938.2019.1600186

Price, E., Goldberg, F., Robinson, S., & McKean, M. (2016). Validity of peer grading using Calibrated Peer Review in a guided-inquiry, conceptual physics course. *Physical Review Physics Education Research, 12*(2), 020145. https://doi.org/10.1103/PhysRevPhysEducRes.12.020145

Rico-Juan, J. R., Cachero, C., & Macià, H. (2022). Influence of individual versus collaborative peer assessment on score accuracy and learning outcomes in higher education: an empirical study. *Assessment & Evaluation in Higher Education, 47*(4), 570-587. https://doi.org/10.1080/02602938.2021.1955090

Russell, A. A. (2004). Calibrated peer review-a writing and critical-thinking instructional tool. *Teaching Tips: Innovations in Undergraduate Science Instruction, 54.*

Russell, J., Van Horne, S., Ward, A. S., Bettis III, E., & Gikonyo, J. (2017). Variability in students' evaluating processes in peer assessment with calibrated peer review. *Journal of Computer Assisted Learning, 33*(2), 178-190. https://doi.org/10.1111/jcal.12176

Seidel, T., Schnitzler, K., Kosel, C., Stürmer, K., & Holzberger, D. (2021). Student characteristics in the eyes of teachers: Differences between novice and expert teachers in judgment accuracy, observed behavioral cues, and gaze. *Educational Psychology Review, 33*, 69-89. https://doi.org/10.1007/s10648-020-09532-2

Sol, J. (2016). Peer evaluation: Incentives and coworker relations. *Journal of Economics & Management Strategy,*

25(1), 56-76. https://doi.org/10.1111/jems.12134

Steedly, K., Dragoo, K., Arafeh, S., & Luke, S. D. (2008). Effective Mathematics Instruction. Evidence for Education. Volume III, Issue I. *National Dissemination Center for Children with Disabilities*.

Szyld, D., & Rudolph, J. W. (2013). Debriefing with good judgment. *The comprehensive textbook of healthcare simulation*, 85-93. https://doi.org/10.1007/978-1-4614-5993-4_7

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of educational Research, 68*(3), 249-276. https://doi.org/10.3102/00346543068003249

Topping, K. J. (2010). Peers as a source of formative assessment. *Handbook of formative assessment*, 61-74.

Vough, H. C., & Caza, B. B. (2017). Where do I go from here? Sensemaking and the construction of growth-based stories in the wake of denied promotions. *Academy of Management Review, 42*(1), 103-128. https://doi.org/10.5465/amr.2013.0177

Villeval, M. C. (2020). *Performance feedback and peer effects* (pp. 1-38). Springer International Publishing.

Wang, T., Jing, X., Li, Q., Gao, J., & Tang, J. (2019). Improving Peer Assessment Accuracy by Incorporating Relative Peer Grades. *International Educational Data Mining Society*.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in psychology, 10*, 487662. https://doi.org/10.3389/fpsyg.2019.03087

1차 원고 접수: 2024. 04. 10
1차 심사 완료: 2024. 06. 07
2차 원고 접수: 2024. 06. 20
2차 심사 완료: 2024. 07. 02
3차 원고 접수: 2024. 07. 04
3차 심사 완료: 2024. 07. 10
최종 게재 확정: 2024. 07. 20

(*Abstract*)

Exploring Method for Enhancing Non-expert Evaluation Accuracy: Using Weighted Functions Based on Common Evaluation Items

Min Hae Song[1)] Hyunwoo Gu[2)] Jungyeon Park[1)] Jaeseo Lim[3)] Jooyong Park[1)]

[1)]Department of Psychology & Asia Center, Seoul National University

[2)]Department of Brain & Cognitive Sciences, Seoul National University

[3)]Department of Counselling Psychology, Jeonju University

Evaluation activities are beneficial for learning or training. However, they are not actively used due to concerns about the evaluation accuracy of non-experts. Although there are methods to improve accuracy, there is a limitation that additional procedures or processes are required in addition to evaluation. In this study, we aimed to improve evaluation accuracy of non-expert by using common evaluation items and assigning weights based on differences from expert scores. In Study 1, we conducted a simulation with 50 non-experts evaluating essays. Our findings indicate that when non-experts' evaluation methods are different from those of experts, our proposed method using a single common evaluation item improves assessment accuracy. In Study 2, we analyzed data from experimental situation in which non-expert evaluated each other's essays. Consistent with Study 1, our proposed method effectively improved assessment accuracy when non-experts' evaluation methods differed from those of experts. In the discussion section, we addressed the applicability of the method proposed in this study in real world settings.

Key words : evaluation, evaluation accuracy, common evaluation items, non-experts, simulation