

인공지능 작성 여부에 대한 출처 정보가 쓰기 평가에 미치는 영향

— 글 수준 및 분석적 채점 준거별 변화 분석

장성민 인하대학교 국어교육과 부교수

- I. 서론
- II. 이론적 배경
- III. 연구 방법
- IV. 연구 결과
- V. 논의 및 제언

I. 서론

이 연구는 인공지능 작성 여부에 대한 출처 정보가 교사의 쓰기 점수 부여에 어떠한 영향을 미치는지를 분석하고, 그 변화 양상을 글 수준 및 분석적 채점 준거별로 규명하는 것을 목적으로 한다.

최근 인공지능 기술의 급속한 발전은 텍스트 생산 환경에 중대한 변화를 가져오고 있다(장성민, 2023, 2024, 2025, 2026). 특히 생성형 인공지능의 확산으로 인해 인간이 작성한 글과 인공지능이 생성한 글의 경계가 점차 흐려지고 있으며(Georgiou, 2026; Tengler & Brandhofer, 2025), 이러한 변화는 교육 현장에서 교사가 학습자의 글에 점수를 부여하는 과정에도 새로운 쟁점을 제기하고 있다(Weber-Wulff, Anohina-Naumeca, Bjelobaba et al., 2023). 학습자가 인공지능의 도움을 받아 글을 작성하는 것이 가능해지면서, 평가의 공정성을 저해할 수 있는 부정행위에 대한 우려가 중요한 이슈로 부상하고 있다(민병곤·남가영·김선희 외, 2023).

교사의 쓰기 점수 부여는 일반적으로 텍스트에 나타난 수행을 중심으로 이루어진다. 그러나 실제 평가 상황에서는 텍스트 자체의 특성뿐 아니라 그

생성 과정에 대한 정보 또한 판단에 영향을 미칠 수 있다. 특히 인공지능의 도움을 받았는지 여부에 대한 정보가 제공될 경우, 동일한 텍스트라 하더라도 평가자가 이를 해석하고 점수를 부여하는 방식이 달라질 가능성이 있다 (Darda, Carre, & Cross, 2023). 이는 분석적 채점 기준이 제시되어 있는 경우에도 예외가 아니다. 평가자는 루브릭의 기술문을 해석하고 적용하는 과정에서 서로 다른 기대와 판단 기준을 형성할 수 있으며 (Zhu, Weissburg & Zhang et al., 2025), 인공지능 작성 여부에 대한 정보는 이러한 해석 과정에 영향을 미쳐 특정 준거에서 보다 엄격하거나 느슨한 판단이 이루어지도록 할 수 있다 (Haverals & Martin, 2025; Schmitz & Sonnleitner, 2025). 따라서 내용 자체를 넘어선 ‘출처 정보’가 점수 부여에 어떠한 변화를 가져오는지를 실증적으로 확인하는 것은 평가의 공정성과 타당성을 점검하는 데 중요한 의미를 지닌다.

이러한 문제의식에 기반하여, 이 연구는 인공지능이 생성한 논증적 글을 대상으로 작성 주체 정보 제시 여부에 따른 교사의 점수 부여 결과 변화를 분석하고자 한다. 이를 위해 동일한 텍스트에 대해 반복측정 설계를 적용하여, 출처 정보가 총점뿐 아니라 글 수준(상, 중, 하) 및 분석적 채점 준거별 점수에 미치는 영향을 체계적으로 검토한다. 이를 통해 인공지능 시대 쓰기 평가에서 나타날 수 있는 평가 판단의 변동성과 기준 해석의 차이를 구체적으로 규명하고, 평가의 일관성과 타당성을 확보하기 위한 시사점을 도출하고자 한다.

II. 이론적 배경

1. 쓰기 평가의 본질과 평가자의 해석

쓰기 평가는 학습자의 언어 수행을 바탕으로 질적 수준을 판단하는 평

가로서, 본질적으로 평가자의 해석과 판단에 의존하는 특성을 지닌다(Weigle, 2002; Huot, 1990; Popham, 2010). 특히 구성형 문항의 일종인 쓰기 평가는 객관식 평가와 달리 정답이 단일하게 규정되어 있지 않기 때문에, 동일한 텍스트라 하더라도 이를 어떻게 해석하느냐에 따라 평가 결과가 달라질 수 있다. 이러한 점에서 쓰기 평가는 텍스트에 대한 의미 구성과 판단이 결합된 해석적 활동으로 이해된다(Huang & Whipple, 2023).

이러한 해석의 가변성을 줄이고 평가를 체계화하기 위한 방안으로 분석적 채점(analytic scoring)이 활용되어 왔다. 분석적 채점은 글의 질을 ‘내용, 조직, 표현’과 같은 하위 요소로 분해하고, 각 요소에 대해 독립적으로 점수를 부여하도록 함으로써 평가자의 판단을 구조화한다(Jacobs, Zingraf, Wormuth et al., 1981; Brown & Hudson, 2002). 이러한 방식은 평가자가 특정 항목에 근거하여 보다 체계적으로 판단할 수 있도록 돕고, 평가 결과의 일관성과 신뢰도를 높이는 데 기여하는 것으로 알려져 있다.

그러나 분석적 채점과 루브릭의 활용이 평가의 객관성을 완전히 보장하는 것은 아니다. 평가 기준이 명시적으로 제시되더라도, 평가자는 이를 해석하고 적용하는 과정에서 자신의 경험과 신념을 반영하게 되기 때문이다(Shabani & Panahi, 2020). 실제로 동일한 루브릭을 적용한 상황에서도 평가자 간 점수 차이가 지속적으로 나타난다는 점은, 평가 기준이 판단을 완전히 규정하기보다는 일정한 해석의 여지를 남긴다는 것을 보여준다(Jonsson & Svingby, 2007; Rezaei & Lovorn, 2010). 나아가 평가자는 개별 준거를 기계적으로 적용하기보다는 글 전체에 대한 인상을 바탕으로 판단을 조정하는 경향을 보이기도 한다(Eckes & Jin, 2025).

이러한 점에서 평가자는 단순히 기준을 적용하는 존재가 아니라, 평가 결과를 구성하는 주체로 이해될 필요가 있다. 평가자는 텍스트를 읽고 의미를 해석하는 동시에 이를 기준과 연결하여 점수를 산출하는 복합적인 인지 과정을 수행하며, 이 과정에서 평가 결과는 평가자에 의해 매개된다(Eckes, 2008; Engelhard & Wang, 2024). 이처럼 쓰기 평가는 기준의 적용만으로 설

명될 수 있는 활동이 아니라, 평가자의 해석과 판단이 결합된 구성적 과정으로 이해할 필요가 있다.

2. 평가자 변인과 맥락 정보가 평가 판단에 미치는 영향과 편향

쓰기 평가가 해석적 활동이라는 점은 동일한 텍스트에 대한 평가 결과가 평가자에 따라 달라질 수 있음을 전제한다. 이러한 차이는 일반적으로 ‘평가자 효과(rater effect)’로 설명되며, 평가자마다 점수의 엄격성이나 관대성, 일관성 등이 다르게 나타날 수 있다(Eckes, 2008; Engelhard & Wang, 2024; McNamara, 1996). 이는 평가 결과가 텍스트의 질에 대한 단순한 반영이 아니라, 평가자의 특성과 판단 방식이 결합된 산물임을 시사한다.

평가자의 판단은 기준을 기계적으로 적용하는 절차라기보다, 텍스트를 해석하고 의미를 구성하는 인지적 과정과 밀접하게 연결되어 있다. 평가자는 글을 읽는 과정에서 내용의 의미를 구성하고 이를 평가 기준과 연결하여 점수를 부여하는데, 이 과정에서 개인의 경험, 신념, 기대가 판단에 개입할 수 있다(Cumming, 1990; Wolfe, 2005). 따라서 평가 결과는 텍스트의 속성뿐 아니라 평가자의 인지적 처리 과정이 함께 작용한 결과로 이해할 수 있다.

이러한 평가자 변인의 영향은 텍스트 외적 정보가 제공될 때 더욱 확대될 수 있다. 평가 상황에서 제시되는 필자 정보, 과제 맥락, 작성 조건 등은 평가자의 기대를 형성하고, 동일한 텍스트에 대한 해석 방향을 변화시킬 수 있다(Freedman, 1984; Lumley, 2002). 특히 이러한 정보는 평가 기준 자체를 직접적으로 변화시키기보다는, 동일한 기준을 어떠한 방식으로 적용할 것인가에 영향을 미치는 요인으로 작용한다. 즉 평가자는 동일한 루브릭을 공유하더라도, 맥락 정보에 따라 특정 요소를 더 강조하거나 보다 엄격하게 해석하는 방식으로 판단을 조정할 수 있다.

특히 출처 정보는 이러한 맥락 정보 중에서도 강력한 영향을 미칠 수 있는 요인이다. 동일한 텍스트라도 그것이 ‘인간이 작성한 글’인지, ‘인공지능

이 생성한 글'인지에 대한 정보는 평가자의 기대를 형성하고 판단의 방향을 변화시킬 수 있다. 이러한 현상은 정보가 제시되는 방식에 따라 판단이 달라지는 정보 프레이밍 효과로 설명될 수 있다(Tversky & Kahneman, 1981). 예컨대 인공지능이 작성한 글이라는 정보가 주어질 경우, 평가자는 해당 텍스트를 보다 비판적으로 검토하거나 특정 결함을 더 민감하게 탐지하려는 경향을 보일 수 있다.

또한 평가자는 대상에 대한 사전 기대를 바탕으로 정보를 해석하는 경향을 보이는데, 이러한 기대 효과는 판단 과정에서 무의식적으로 작용하여 평가 결과에 영향을 미칠 수 있다(Jussim, 2012). 예를 들어 인공지능 텍스트에 대해 '형식적으로는 완성도가 높지만 내용의 깊이는 부족할 것'이라는 기대를 갖는 경우, 실제 텍스트의 내용적 측면을 보다 엄격하게 평가할 가능성이 있다. 반대로 인공지능의 능력을 긍정적으로 인식하는 경우에는 표현이나 조직 측면에서 상대적으로 호의적인 판단이 강화될 수도 있다.

이처럼 쓰기 평가에서의 점수는 텍스트 자체의 질에 대한 단순한 반영이라기보다, 평가자의 기대, 맥락 정보, 인지적 편향이 결합된 결과로 이해할 수 있다. 특히 출처와 같은 텍스트 외적 정보는 평가 기준의 내용 자체를 변화시키기보다는, 기준을 적용하는 방식과 해석의 방향을 조정하는 요인으로 작용할 가능성이 있다. 이러한 점은 동일한 평가 기준이 적용되는 상황에서도 평가 결과가 달라질 수 있음을 시사하며, 나아가 인공지능 작성 여부와 같은 정보가 실제 평가 판단에 어떠한 영향을 미치는지를 실증적으로 검토할 필요성을 제기한다.

3. 인공지능이 생성한 글에 대한 일반적 기대

최근 생성형 인공지능의 발전으로 인해 인간이 작성한 글과 인공지능이 생성한 글의 경계가 점차 모호해지고 있으며, 이에 따라 인공지능 텍스트의 특성과 이에 대한 평가 인식에 대한 논의가 활발히 이루어지고 있다.

선행 연구에 따르면 인공지능이 생성한 글은 문법적 정확성과 형식적 완성도 측면에서 높은 수준을 보이며, 구조적으로 안정된 전개를 특징으로 한다(Brown, Mann, Ryder et al., 2020). 반면 내용의 깊이, 맥락적 적합성, 창의성 측면에서는 상대적으로 한계를 보일 수 있다는 점이 함께 지적된다(Bender, Gebru, McMillan-Major et al., 2021).

이와 같은 특성에 대한 인식은 평가자가 인공지능 텍스트를 판단하는 방식에도 영향을 미칠 수 있다. 특히 동일한 텍스트라 하더라도 그것이 인공지능에 의해 생성된 것이라는 정보가 제공될 경우, 평가 결과가 달라질 수 있다는 연구 결과가 보고되고 있다(Darda et al., 2023; Haverals & Martin, 2025; Schmitz & Sonnleitner, 2025; Zhu et al., 2025). 이는 평가자가 텍스트 자체뿐 아니라 작성 주체에 대한 정보를 바탕으로 판단을 조정한다는 점을 보여준다.

이러한 현상은 인공지능에 대한 사회적 인식과도 관련이 있다. 인공지능은 효율성과 정확성 측면에서 긍정적으로 평가되는 동시에, 신뢰성이나 진정성에 대한 의문이 제기되는 등 상반된 인식이 공존하는 경향을 보인다(Aroujo et al., 2020; Fast & Horvitz, 2017). 이러한 이중적 인식은 평가자가 인공지능 텍스트를 해석할 때 특정 기대를 형성하게 하고, 그에 따라 평가 기준의 적용 방식이 달라질 가능성을 높인다.

특히 인공지능 산출물에 대한 판단에서는 ‘알고리즘 회피 성향(algorithm aversion)’이 나타날 수 있다(Dietvorst, Simmons, & Massey, 2015). 이는 사람들이 알고리즘이 생성한 결과를 인간의 판단보다 덜 신뢰하는 경향을 의미하며, 동일한 텍스트라도 인공지능이 작성했다는 정보가 제시될 경우 보다 비판적으로 평가될 가능성을 시사한다(Mills & Gutierrez, 2025). 반면 일부 상황에서는 알고리즘의 판단을 더 신뢰하는 ‘알고리즘 선호(algorithm appreciation)’ 현상도 보고되고 있어(Logg, Minson, & Moore, 2019), 인공지능에 대한 평가 방향이 상황에 따라 달라질 수 있음도 확인된다.

이처럼 ‘인공지능이 작성한 글’이라는 출처 정보는 단순한 부가 정보가

아니라 평가자의 기대를 형성하고, 평가 기준을 적용하는 방식 자체를 변화시키는 요인으로 작용할 수 있다. 특히 이는 특정 준거를 보다 엄격하게 적용하거나, 일부 요소를 강조하거나 축소하는 방식으로 나타날 가능성이 있다. 따라서 인공지능 작성 여부에 대한 정보가 실제 쓰기 평가에서 어떠한 영향을 미치는지를 실증적으로 검토하는 것은 평가의 공정성과 타당성을 이해하는 데 중요한 과제가 된다.

III. 연구 방법

1. 연구 설계

이 연구는 인공지능이 작성한 글에 대한 평가가 작성 주체에 대한 정보에 의해 어떻게 달라지는지를 확인하기 위한 반복측정 설계(within-subject design)를 사용하였다. 동일한 텍스트에 대해 ‘작성 주체 정보 제시 여부’만을 변화시킨 두 조건에서 동일한 평가자가 반복적으로 평정하도록 하여, 출처 정보가 평가에 미치는 영향을 직접적으로 검증하고자 하였다.

이를 위해 동일한 평가자가 동일한 글을 대상으로 두 차례 평정을 실시하였다. 1차 평정에서는 필자에 대한 정보를 제공하지 않고 텍스트만 제시하여 일반적인 학습자의 글로 간주하도록 하였으며, 2차 평정에서는 동일한 글이 인공지능에 의해 생성되었음을 명시한 후 재평정을 실시하였다. 두 평정 간에는 약 1주일의 간격을 두었고, 2차 평정에서는 1차 평정 점수를 확인할 수 없도록 하여 이전 판단의 영향을 최소화하였다.¹⁾²⁾

-
- 1) 논문 심사 중 한 분의 심사위원께서 1차 평정 단계에 평가자들이 해당 텍스트를 일반 학습자의 글로 인식하였는지를 확인하는 절차가 포함되지 않았다는 점을 지적하였다. 이에 따라 일부 평가자들이 1차 평정 단계에서도 해당 텍스트를 인공지능 생성 글로 추정하였을 가능성을 배제하기 어려우며, 연구자가 설정한 두 조건이 평가자의 인식 수준에서도

두 차례 평정 모두 동일한 분석적 채점 루브릭을 적용하였으며, 평가자는 각 글을 독립적으로 판단하고 평가 루브릭의 점수대별 기술문에 근거하여 점수를 부여하도록 하였다. 이를 통해 평가 기준은 통제하면서, 작성 주체 정보라는 단일 변인의 효과를 비교할 수 있도록 설계하였다.

〈표 1〉 평정 조건별 지시문(출처 정보 제시 여부)

1차 평정 (미제시)	다음은 주어진 과제에 따라 작성된 논증적 글입니다. 각 글을 읽고, 제시된 분석적 채점 기준에 따라 점수를 부여해 주세요. △ 각 준거는 1점에서 5점까지의 척도로 평정하며, 점수는 채점 기준표의 점수대별 기술문을 참고하여 판단해 주세요.
2차 평정 (제시)	다음은 앞서 평가하신 것과 동일한 글입니다. 각 글은 인공지능에 의해 생성된 텍스트입니다. 각 글을 다시 읽고, 제시된 동일한 분석적 채점 기준에 따라 점수를 부여해 주세요. △ 각 준거는 1점에서 5점까지의 척도로 평정하며, 점수는 채점 기준표의 점수대별 기술문을 참고하여 판단해 주세요. △ 각 글은 독립적으로 평가해 주시고, 이전에 부여한 점수에 영향을 받지 않도록 새롭게 판단해 주세요.

- 명확히 구분되었다고 단정하기 어렵다는 의견이 제시되었다. 이러한 점에서 2차 평정에서 관찰된 점수 변화를 온전히 인공지능 작성 정보의 효과로 해석하는 데 신중함이 요구되며, 연구 결과를 동일한 텍스트에 대한 평가가 작성 주체에 대한 인식에 따라 달라질 수 있음을 보여 주는 결과로 이해해야 할 필요성도 제기될 수 있다.
- 2) 연구 설계와 관련하여 동일 평가자가 동일 텍스트를 반복 평정하는 과정에서 순서 효과(order effect)나 기억 효과(memory effect)가 개입될 가능성이 있다는 점은 방법론적 한계로 고려될 수 있다. 이러한 내적 타당도 문제를 통제하기 위한 대안으로 평가자를 두 집단으로 나누어 ‘출처 정보 미제시—인공지능 작성 정보 제시’와 ‘인공지능 작성 정보 제시—출처 정보 미제시’의 순서를 교차 적용하는 균형화 설계(counterbalanced design)를 구성하는 방안도 이론적으로 가능하다. 그러나 이 연구에서는 특히 후자의 조건에서 평가자가 한 차례 인공지능 작성 여부를 인지한 이후 이를 완전히 배제한 상태에서 텍스트를 재평정하기 어렵고, 그 결과 ‘출처 정보 미제시’ 조건이 독립적인 비교 조건으로 기능하지 못할 가능성이 크다고 판단하였다. 이에 이 연구는 출처 정보가 추가로 제공되는 상황에서 평가 판단이 어떻게 변화하는지를 보다 타당하게 포착하고자 ‘출처 정보 미제시—인공지능 작성 정보 제시’의 단일 방향 반복측정 설계를 채택하였으며, 이로 인해 발생할 수 있는 순서 효과와 기억 효과를 완화하기 위해 평정 간 시간 간격을 두고 2차 평정 시 1차 평정 결과를 확인할 수 없도록 하는 절차를 적용하였다.

2. 연구 참여자

연구 참여자는 중등 국어 교사 20명으로 구성하였다. 참여자는 실제 쓰기 평가 경험을 지닌 현직 교사로 선정하여 평가 결과의 실제성과 생태적 타당성을 확보하고자 하였다. 표집은 목적 표집 방식으로 이루어졌으며, 연구자가 참여하고 있는 교사 연수 프로그램과 협력 관계에 있는 학교를 통해 참여자를 모집하였다. 참여자의 배경은 성별, 연령대, 학교급에서 다양성을 확보하도록 구성하였다. 연구 참여 전 연구 목적과 절차, 자료 활용 방식에 대해 충분히 안내하고 자발적 동의를 얻었으며, 수집된 자료는 익명 처리하여 연구 목적에 한해 사용하였다.

〈표 2〉 연구 참여자 특성

구분	인원
성별	남 8명, 여 12명
연령대	20대 3명, 30대 7명, 40대 6명, 50대 이상 4명
학교급	중학교 9명, 고등학교 11명

평가의 일관성을 확보하기 위해 사전 워크숍을 통해 채점자 훈련을 실시하였다. 이 과정에서 평가 루브릭에 대한 안내와 예시 글을 활용한 연습 채점 및 협의를 통해 준거 해석의 공통 기준을 형성하였다.³⁾

- 3) 논문 심사 중 한 분의 심사위원께서 국어 교사의 고유한 평가 특성을 탐색하기 위해서는 채점자 훈련 없이 실제적·생태학적 환경에서 이루어진 평가를 분석할 필요가 있다는 점을 지적하였다. 이러한 문제의식에 공감하며, 실제 수업 맥락에서 이루어지는 평가는 평가자 고유의 신념과 판단 방식이 보다 자연스럽게 드러날 수 있다는 점에서 중요한 연구 방향이 될 수 있다. 다만 이 연구에서는 인공지능 작성 여부에 따른 출처 정보 효과를 비교·분석하는 과정에서 평가 루브릭에 대한 이해 부족으로 인한 차이와 평가자 고유의 판단 성향으로 인한 차이를 구분할 필요가 있다고 보았다. 이에 평가 기준에 대한 기본적인 이해를 공유하기 위한 최소한의 채점자 훈련을 실시하였으며, 이를 통해 관찰된 점수 차

3. 평가 자료

평가 자료는 인공지능이 생성한 논증적 글 9편으로 구성하였다.⁴⁾ 평가 자료를 9편으로 구성한 것은 수준별 변별력을 확보하면서 반복평정에 따른 평가 부담을 고려한 결과이다. 모든 글은 동일한 과제와 조건 하에서 생성되었으며, 분량(1,000자 내외 $\pm$ 200자)과 기본 구조(서론-본론-결론)를 통일하여 외형적 변인을 통제하였다.

수준 간 변별을 위해 생성 단계에서 조건을 달리하였다. 상 수준 글은 주장, 근거, 반론 고려가 명확하고 논리 전개가 긴밀하도록 하였으며, 중 수준 글은 분량을 유지하되 다소 장황하고 반론 고려를 제한적으로 포함하도록 하였다. 하 수준 글은 논리적 일관성과 조직성을 낮추고 반론 고려를 포함하지 않도록 생성하였다.

이후 예비 평가를 통해 자료의 수준 적합성을 검토하였다. 예비 평가는 석사학위 이상을 소지한 현직 국어 교사 3인으로 구성된 전문가 집단에 의해 이루어졌으며, 이들은 각 글의 전반적 수준과 준거별 수행 특성을 종합적으로 판단하여 상·중·하 수준의 전형성에 부합하는지를 검토하였다. 이 과정에서 수준 구분이 모호하거나 전형성이 낮다고 판단된 글은 제외하고 대

이를 평가자의 해석과 판단 과정에서 비롯된 결과로 해석하고자 하였다.

- 4) 이 연구 설계는 모든 평가 텍스트를 인공지능 생성 글로 구성하고 출처 정보 제시 여부를 변인으로 설정함에 따라, 실제 작성자와 제시 정보의 효과를 완전히 분리하지 못한다는 점에서 해석상의 한계를 지닐 수 있다. 이러한 문제를 보완하기 위한 대안으로 실제 작성자(인간, 인공지능)와 제시 정보(인간, 인공지능)를 교차하는 2×2 설계를 도입하고 일부 인간 작성 글을 포함하는 방안이 이론적으로 고려될 수 있다. 그러나 인간이 작성한 글을 인공지능 생성 텍스트로 제시할 경우 텍스트 특성과 제시 정보 간 불일치로 인해 평가자에게 혼란을 유발하고, 평가 판단이 출처 정보가 아닌 텍스트의 이질감에 의해 영향을 받을 가능성이 있다. 특히 중·하 수준의 인간 작성 글을 인공지능 텍스트와 유사한 특성으로 구성하는 데에는 현실적 제약이 존재하며, 이는 실험 자극의 자연성과 타당성을 저해할 우려가 있다. 이에 이 연구에서는 동일한 생성 조건 하에서 통제된 인공지능 텍스트를 활용하여 출처 정보 제시 여부의 효과를 검증하는 설계를 채택하였다.

체 자료를 선정하였다. 최종적으로 상·중·하 각 3편씩 총 9편을 평가 자료로 확정하였다.

쓰기 과제는 국립국어원(민병곤 외, 2023)의 국민의 글쓰기 능력 평가 과제를 활용하였다. 해당 과제는 사회적 쟁점에 대한 입장을 논리적으로 제시하는 논증적 글쓰기 과제로, 다양한 평가 요소를 종합적으로 반영할 수 있다는 점에서 이 연구에 적합하다.

〈표 3〉 쓰기 과제

로봇의 발달로 일자리가 줄어든 것이라는 사람들의 불안이 커지면서 최근 로봇세 도입에 대한 논의가 활발하다. 로봇세는 로봇의 노동으로 생산하는 경제적 가치에 부과하는 세금이다. 로봇 기술의 발달로 인해 일자리를 잃는 사람들이 갈수록 많아질 수 있기 때문에, 그런 사람들을 지원하거나 사회 안정망을 구축하기 위해 예산을 마련하자는 것이 로봇세 도입의 목적이다. 로봇세 도입에 대한 자신의 의견을 논리적으로 제시하는 글을 쓰시오.

△ 서론, 본론, 결론을 갖춘 완결된 글을 쓸 것(제목을 쓰지 말 것).

△ 분량: 1,000자 내외(±200자, 공백 포함), 시간: 90분

4. 평가 도구

평가 도구는 분석적 채점 루브릭으로 구성하였다. 루브릭은 ‘내용, 조직, 표현’의 세 영역, 총 7개 준거(주장, 이유와 근거, 다른 입장 고려, 글 전체 조직, 문단 내 조직, 문장과 어휘, 규범)로 이루어졌다. 각 준거는 1점에서 5점까지의 5점 척도로 평정하였다.

이 루브릭은 국립국어원(민병곤 외, 2023)의 평가 기준을 바탕으로 하되, 연구 목적에 맞게 일부 간소화하여 재구성한 것이다. 준거 수를 축소하고 기술문을 조정하여 평가 부담을 줄이면서도 핵심 평가 요소는 유지하였다. 평가자는 준거별 점수만을 부여하였으며, 총점은 연구자가 사후적으로 산출하였다.⁵⁾

5) 민병곤 외(2023: 228)에 따르면 ‘문제 상황 제시’는 내용 범주의 다른 준거들과의 관련성이 상대적으로 낮은 것으로 나타났으며, ‘이유/근거의 적절성’과 ‘이유/근거의 충분성’은 상관계수가 .8에 근접하여 개념적으로 중첩되는 경향이 확인되었다. 이에 이 연구에서는

범주	국립국어원 준거		이 연구의 준거
내용	문제 상황 제시	→	주장
	주장		
	이유/근거의 적절성		
	이유/근거의 충분성		
	다른 입장에 대한 고려		
조직	글 전체 조직		다른 입장 고려
	문단 내 조직		글 전체 조직
	문장과 어휘		문단 내 조직
표현	문장과 어휘		문장과 어휘
	어문 규범과 관습		규범

〈그림 1〉 분석적 채점 루브릭 구성

각 준거에는 점수대별 수행 수준을 구체적으로 제시한 기술문을 포함하였다. 내용 범주에서 ‘주장’ 준거는 글이 논제에 부합하며 중심 입장이 글 전체에 걸쳐 일관되고 명료하게 제시되었는지를 기준으로 수준을 구분하였다. ‘이유와 근거’ 준거는 주장을 적절히 뒷받침하는 이유를 두 가지 이상 제시하고, 각 이유에 대해 구체적이고 타당한 근거를 들어 충분히 논증하고 있는지를 기준으로 하였다. ‘다른 입장 고려’ 준거는 상반된 입장을 단순히 언급하는 데 그치지 않고, 이를 깊이 있게 탐구하고 이에 대해 타당하게 대응하고 있는지를 중심으로 수준을 구분하였다.

조직 범주에서 ‘글 전체 조직’ 준거는 서론·본론·결론이 내용과 형식의 측면에서 적절하게 구분되고, 문단이 유기적으로 배열되었는지를 기준으로 하였다. ‘문단 내 조직’ 준거는 각 문단에서 중심 문장과 이를 뒷받침하는 문장이 구분되며, 문장 간 연결이 자연스럽고 긴밀한지를 기준으로 수준을 구분하였다.

표현 범주에서는 ‘문장과 어휘’ 준거를 통해 의미 전달에 적절하고 효과적인 어휘를 선택하며 문장을 정확하고 자연스럽게 구성하고 있는지를 평가하도록 하였으며, ‘규범’ 준거에서는 맞춤법, 띄어쓰기, 문장 부호 등 언어 규범을 준수하고 있는지를 기준으로 수준을 구분하였다.

평가 준거 간 변별성을 확보하고 분석의 효율성을 높이기 위해 ‘문제 상황 제시’는 제외하고, ‘이유/근거의 적절성’과 ‘이유/근거의 충분성’을 ‘이유와 근거’로 통합하였다.

〈표 4〉 분석적 채점 루브릭 예시(준거 및 기준)

준거	하위 준거	채점 기준
(내용 3) 다른 입장 고려	대안적 입장의 탐구: 다른 입장에 대한 접근이 이루어졌는가? 반응의 적절성: 다른 입장에 대해 논리적이고 타당하게 대응하고 있는가?	5점: 다른 입장을 의미 있게 탐구하고, 이에 대해 논리적이고 타당하게 대응함. 4점: 다른 입장을 제시하고 이에 대해 비교적 적절하게 대응함. 3점: 다른 입장을 언급하였으나, 탐구나 대응이 제한적임. 2점: 다른 입장을 형식적으로만 언급하거나, 대응이 부적절함. 1점: 다른 입장을 전혀 고려하지 않음.

5. 연구 절차

연구는 사전 준비, 1차 평정, 2차 평정의 순서로 진행하였다.

먼저 사전 준비 단계에서는 채점자 훈련을 위한 워크숍을 실시하였다. 분석적 채점 루브릭의 각 준거와 점수대별 기술문에 대한 상세한 안내가 이루어졌으며, 예시 글 3편을 활용한 연습 채점과 채점 결과 공유 및 협의가 진행되었다. 이를 통해 평가 기준에 대한 이해를 높이고, 평가자 간 해석의 일관성을 확보하고자 하였다.

이후 1차 평정에서는 참여 교사가 9편의 글을 무작위 순서로 평가하였으며, 필자 정보는 제공하지 않았다. 평가는 온라인 설문 도구를 통해 개별적으로 이루어졌고, 각 준거별 점수와 함께 간단한 판단 근거를 기록하도록 하였다. 평가 루브릭은 A4 1페이지 분량의 자료로 별도 제작하여 PDF 형태로 배부하였으며, 참여자는 이를 참고하여 평가를 수행하였다.

1차 평정이 종료된 이후에는 약 1주일의 시간 간격을 두고 2차 평정을 실시하였다. 2차 평정에서는 동일한 글을 동일한 순서로 제시하되, 모든 글이 인공지능에 의해 작성되었음을 명시하였다. 참여자는 동일한 루브릭을 적용하여 다시 평가하였으며, 이전 점수는 확인할 수 없도록 하여 선행 판단의 영향을 최소화하였다. 또한 2차 평정에서는 점수 부여 이후, “1차 평정 때와 비교하여 2차 평정에서 점수가 달라졌을 것으로 예상되는 경우, 그 이유를 간단히 작성

6. 자료 분석 방법

이 연구는 혼합연구 방법론 중 양적 분석 결과를 토대로 질적 자료를 통해 해석을 보완하는 설명적 순차 설계(explanatory sequential design)를 적용하였다. 양적 자료를 분석하여 인공지능 작성 여부에 따른 점수 변화 양상을 확인하고, 질적 자료를 활용하여 이러한 변화의 의미와 원인을 심층적으로 해석하였다.

먼저 양적 분석 단계에서는 총점과 분석적 채점 준거별 점수 변화를 중심으로 통계 분석을 실시하였다. 1차와 2차 평정 간 평균 차이를 산출하고, 대응표본 t-검정을 통해 인공지능 작성 정보 제시 여부에 따른 점수 변화의 통계적 유의성을 검증하였다. 또한 글 수준(상, 중, 하)에 따른 차이를 분석하기 위해 반복측정 분산분석을 실시하고, 수준과 정보 제시 여부 간 상호작용 효과를 확인하였다. 아울러 준거별 점수 변화를 비교하여 특정 평가 요소에서 변화가 두드러지는지를 분석하였다. 각 분석에서는 효과크기(Cohen's d)를 함께 산출하여 변화의 실질적 의미를 해석하였다.

대응표본 t-검정과 반복측정 분산분석을 실시할 때에는 교사별 평균 점수를 분석 단위로 설정하여 통계 분석을 실시하였다. 이는 동일한 교사가 여러 편의 글을 평가한 자료를 개별적이고 독립적인 측정값으로 간주할 경우 통계적 독립성 가정이 위배될 수 있다고 보았기 때문이다. 이에 따라 각 조건별 점수를 교사 단위로 집계한 후 1차 평정과 2차 평정 간 차이를 분석하였다.⁶⁾

다음으로 질적 분석 단계에서는 평가자가 각 글에 대해 기록한 판단 근거와 추가 응답을 분석하여, 양적 결과에서 나타난 점수 변화의 해석을 보완

6) <표 5>의 총점 변화는 교사별 9편 글의 평균 점수를, <표 6>의 글 수준별 변화는 교사별 · 수준별 3편 글의 평균 점수를, <표 7>의 준거별 변화는 교사별 9편 글의 준거별 평균 점수를 산출한 후 이를 분석 대상으로 사용하였다.

하였다. 질적 자료는 귀납적 범주화 방식을 활용하여 분석하였으며, 인공지능 작성 정보 제시 이후 평가자가 어떠한 기대, 해석 기준, 판단 전략의 변화를 보이는지를 중심으로 범주를 도출하였다. 특히 특정 준거에서 점수 변화가 크게 나타난 경우, 이에 대응하는 평가자 발언을 집중적으로 분석하여 점수 변화의 인지적·해석적 근거를 규명하고자 하였다.⁷⁾

마지막으로 양적 분석 결과와 질적 분석 결과를 통합적으로 해석하여, 인공지능 작성 여부에 대한 출처 정보가 교사의 쓰기 점수 부여 과정에 어떠한 방식으로 영향을 미치는지를 종합적으로 논의하였다. 이를 통해 단순한 점수 차이의 확인을 넘어, 평가 기준의 적용 방식과 평가자의 해석 과정에서 나타나는 변화를 설명하고자 하였다.⁸⁾

7) 질적 분석은 개별 응답을 의미 단위로 분절한 후 1차 코딩을 실시하고, 유사한 코드들을 묶어 2차 범주로 통합한 뒤, 이를 다시 상위 범주로 구조화하는 방식으로 이루어졌다(하단 분석 사례 참고). 이러한 과정은 인공지능 작성 정보 제시 이후 평가자의 기대 변화, 기준 적용 방식의 조정, 준거별 판단 차이와 같은 주제를 중심으로 조직되었다. 질적 분석이 양적 분석 결과를 보완하는 수준에서 제한적으로 이루어졌음을 고려하여, 코딩은 단일 코더가 수행하되 참여자 중 2인에게 결과를 검토받아 연구 및 자료 수집 맥락에 비추어 타당성을 확인하였다.

원자료	1차 코드	2차 범주	상위 범주(주제)
처음에는 근거가 충분하다고 생각했는데, 인공지능이라고 하니 일반적인 이야기처럼 느껴졌다.	근거의 일반성 인식	내용의 피상성 재해석	인공지능 텍스트에 대한 부정적 기대
이 정도 글이면 학생 글로는 괜찮지만, 인공지능이라면 더 정교해야 한다고 생각했다.	기대 수준 상황	기준 적용의 엄격화	출처 정보에 따른 기준 적용 변화

8) 논문 심사 중 복수의 심사위원께서 점수 변화가 낙인 효과, 기술에 대한 거부감 또는 과도한 신뢰, 진정성에 대한 기대 등 어떠한 인식적 요인에 의해 유발되었는지 심층적으로 탐색할 필요성을 제기해 주셨다. 이 연구의 질적 분석은 설명적 순차 설계에 따라 양적 결과를 보완하는 수준에서 수행되었기 때문에, 평가자의 인식 변화가 어떠한 심리적 기제에 의해 발생하는지까지 직접적으로 규명하는 데에는 한계가 있다. 향후에는 심층 면담이나 사고구조 등 활용한 질적 연구를 통해 평가자의 해석 과정과 판단 전략의 변화를 보다 정교하게 분석할 필요가 있다.

한편 평가자 간 신뢰도는 급내상관계수(ICC)를 통해 분석하였다. 그 결과 1차 평정의 ICC는 .87, 2차 평정의 ICC는 .82로 나타나 모두 수용 가능한 수준의 일치도를 보였으나, 2차 평정에서 다소 감소하는 경향이 확인되었다. 이는 인공지능 작성 정보가 제시된 이후 동일한 텍스트에 대한 평가자들의 해석이 보다 다양해졌음을 시사하며, 이 연구에서 확인된 평가 판단의 변동 성과도 관련되는 결과로 해석할 수 있다.

IV. 연구 결과

1. 인공지능 작성 정보 제시 여부에 따른 총점 변화

인공지능 작성 정보 제시 여부가 교사의 쓰기 점수 부여에 미치는 영향을 확인하기 위해, 1차 평정(출처 정보 미제시)과 2차 평정(인공지능 작성 정보 제시) 간 총점 차이를 비교하고 대응표본 t-검정을 실시하였다.

분석 결과, 인공지능 작성 정보가 제시된 이후 총점이 전반적으로 유의하게 하락하는 것으로 나타났다($t(19)=4.02$, $p<.001$, Cohen's $d=.90$). 이러한 결과는 동일한 텍스트라 하더라도 작성 주체에 대한 정보가 제시될 경우 평가자의 판단이 달라질 수 있음을 보여준다. 이는 인공지능 작성 여부에 대한 정보는 단순한 점수 변동을 넘어 실제 평가 결과에 의미 있는 변화를 초래하는 요인으로 작용함을 시사한다.

〈표 5〉 출처 정보 제시 여부에 따른 총점 변화

구분	평균	표준편차	t	p	Cohen's d
1차 평정(미제시)	25.85	3.95	4.02	<.001	.90
2차 평정(제시)	24.96	4.11			

질적 자료 분석 결과 또한 이러한 경향을 뒷받침한다. 참여자들은 점수 변화의 이유를 설명하면서, “인공지능이 작성한 글이라고 인식한 이후에는 같은 글이라도 더 엄격한 기준을 적용하게 되어 점수를 낮추었을 것 같다”, “처음에는 무난하다고 판단했지만, AI가 쓴 글이라고 생각하니 내용이 다소 일반적이고 깊이가 부족하게 느껴져 감점했을 것 같다”와 같이 응답하였다. 이는 평가자가 2차 평정에서 이전 판단을 재조정하는 과정에서, 인공지능 텍스트에 대해 상대적으로 높은 기대를 적용하거나 내용의 깊이와 진정성 측면을 보다 비판적으로 해석했음을 보여준다.

종합하면, 인공지능 작성 정보는 평가 기준을 적용하는 방식과 해석의 방향을 조정하는 요인으로 작용하며, 그 결과 총점의 체계적인 변화를 유발한 것으로 해석할 수 있다. 이는 인공지능 시대의 쓰기 평가에서 출처 정보가 평가의 공정성과 타당성에 중요한 영향을 미칠 수 있음을 시사한다.

2. 글 수준별 총점 변화

인공지능 작성 정보 제시 여부에 따른 총점 변화가 글 수준별로 어떻게 달라지는지를 확인하기 위해, 반복측정 분산분석을 실시하였다. 이때 분석은 출처 정보 제시 여부(1차·2차 평정)를 반복측정 요인으로, 글 수준(상·중·하)을 집단 내 요인으로 설정한 이원 반복측정 설계에 기반하여 수행되었다.

분석 결과, 출처 정보 제시 여부와 글 수준 간의 상호작용 효과가 통계적으로 유의한 것으로 나타났다($F(2, 38)=8.14, p<.001$, 부분 $\eta^2=.30$). 이는 인공지능 작성 정보가 총점에 미치는 영향이 글 수준에 따라 다르게 나타났음을 의미한다. 즉, 동일한 정보 제시 조건 변화라 하더라도 모든 수준의 글에 동일하게 작용하는 것이 아니라, 글의 질적 수준에 따라 차별적으로 영향을 미치는 것으로 해석할 수 있다.

〈표 6〉 글 수준별 총점 변화

글 수준	평정 시기	평균	표준편차	t	p	Cohen's d
상	1차(미제시)	31.12	2.85	3.18	<.01	.71
	2차(제시)	30.24	3.02			
중	1차(미제시)	26.67	2.96	5.41	<.001	1.21
	2차(제시)	24.99	3.10			
하	1차(미제시)	19.76	3.08	.46	not sig.	.10
	2차(제시)	19.65	3.19			

질적 자료 분석 결과는 이러한 상호작용 효과의 구체적 양상을 설명해 준다. 중 수준 글에서 가장 큰 점수 하락이 나타난 것은, 인공지능에 대한 기대 수준과 실제 텍스트의 수행 사이의 불일치가 가장 크게 지각되었기 때문으로 해석된다. 예를 들어 한 참여자는 “처음에는 무난한 글이라고 보았지만, 인공지능이 작성한 글이라고 인식한 이후에는 내용이 다소 평범하고 깊이가 부족하게 느껴져 점수를 낮추었을 것 같다”고 응답하였다. 또 다른 참여자는 “학생 글이라면 이 정도면 괜찮다고 판단했을 텍스트지만, AI가 쓴 글이라고 생각하면서 더 높은 수준을 기대하게 되어 감점했을 것으로 보인다”고 언급하였다. 이는 인공지능에 대해 상대적으로 높은 수행 수준을 기대하는 인식이 평가에 반영되었음을 보여준다.

상 수준 글의 경우에도 점수 하락이 나타났는데, 이는 완성도가 높은 글이라 하더라도 인공지능 텍스트로 인식되는 순간 평가 기준이 재조정될 수 있음을 시사한다. 한 참여자는 “전체적으로 잘 쓴 글이라고 판단했지만, 인공지능이 작성했다고 생각하니 다소 교과서적이고 개성이 부족하게 느껴져 일부 감점했을 것 같다”고 응답하였으며, 또 다른 참여자는 “논리 구조는 갖추어져 있으나 실제 경험에서 나온 주장이라기보다는 기계적으로 정리된 정보처럼 보여 설득력이 약하게 느껴져 점수가 낮아졌을 것으로 생각한다”고

밝혔다. 이는 인공지능 글에 대해 진정성이나 주체성을 상대적으로 낮게 평가하는 경향이 작용했음을 보여준다.

반면 하 수준 글에서는 점수 변화가 거의 나타나지 않았다. 일부 참여자는 “완성도가 낮은 글이기 때문에 작성 주체 정보가 추가되더라도 평가 기준을 크게 바꾸지 않았을 것 같다”고 응답하였으며, 또 다른 참여자는 “인공지능이 작성했다고 하더라도 이 정도 수준이라면 추가적으로 감점할 요소가 크지 않아 점수를 유지했을 가능성이 있다”고 언급하였다. 이는 낮은 수준의 글에서는 평가 기준이 비교적 안정적으로 유지되거나, 일부 경우에는 변화 폭이 제한적이었음을 시사한다.

종합하면, 인공지능 작성 정보는 평가자의 기대와 해석 틀을 변화시키며, 그 영향은 글 수준에 따라 차별적으로 나타난다. 특히 중 수준 글에서 두드러진 변화는 평가 상황에서 기대 위반 효과가 강하게 작용함을 보여주며, 이는 인공지능 시대의 쓰기 평가에서 고려해야 할 중요한 변인임을 시사한다.

3. 분석적 채점 준거별 점수 변화

마지막으로 인공지능 작성 정보 제시 여부에 따른 점수 변화가 분석적 채점 준거별로 어떻게 나타나는지를 확인하기 위해, 각 준거의 1차 평정과 2차 평정 간 점수 변화를 분석하였다.

〈표 7〉 분석적 채점 준거별 점수 변화

범주	준거	1차 평정(미제시)		2차 평정(제시)		t	p	Cohen's d
		평균	표준 편차	평균	표준 편차			
내용	주장	3.78	.62	3.52	.68	2.94	<.01	.66
	이유와 근거	3.65	.70	3.12	.75	4.87	<.001	1.09
	다른 입장 고려	3.21	.66	3.25	.69	-.58	not sig.	.13

조직	글 전체 조직	3.89	.58	3.94	.60	-.72	not sig.	.16
	문단 내 조직	3.74	.61	3.76	.63	-.39	not sig.	.09
표현	문장과 어휘	3.68	.64	3.41	.71	3.08	<.01	.69
	규범	3.90	.55	3.96	.52	-1.12	not sig.	.25

분석 결과, 인공지능 작성 정보는 범주 및 준거에 따라 차별적으로 영향을 미치는 것으로 나타났다. 먼저 내용 범주에서는 전반적인 점수 하락이 확인되었으나, 그 양상은 하위 준거에 따라 상이하였다. ‘이유와 근거’에서는 유의한 점수 하락이 나타났으며, 참여자들은 “처음에는 충분한 근거라고 판단했지만, 인공지능이 작성한 글이라고 생각하니 다소 일반적인 수준에 머무른다고 느껴져 점수를 낮추었을 것 같다”, “AI를 사용한 것에 비해 이유나 근거의 질이 충분히 높지 않다고 판단하여 감점했을 것으로 보인다”와 같이 응답하였다. ‘주장’에서도 유사하게 점수 하락이 나타났으며, “주장은 분명하다고 보았지만, 인공지능이 작성했다고 생각하니 필자의 문제의식이 충분히 드러나지 않는다고 판단하여 점수를 조정했을 것 같다”, “누가 왜 이런 주장을 하는지에 대한 맥락이 약하게 느껴져 일부 감점했을 것으로 생각한다”는 응답이 확인되었다. 이는 인공지능 작성 정보가 제시되면서 동일한 텍스트의 내용이 보다 비판적으로 재해석되었음을 보여준다. 반면 ‘다른 입장 고려’에서는 유의한 점수 변화가 나타나지 않았으며, 관련 응답에서도 “이 부분은 처음과 동일하게 여러 입장을 균형 있게 제시했다고 판단하여 점수를 크게 바꾸지 않았을 것 같다”, “인공지능이 작성한 글이라고 보더라도 이 준거에서는 평가 기준을 따로 조정하지 않았던 것으로 생각된다”와 같이 비교적 안정적인 평가가 이루어졌음을 확인할 수 있었다.

조직 범주에서는 전반적으로 점수 변화가 크지 않았으며, 질적 응답 역시 이러한 경향을 뒷받침하였다. 참여자들은 “글의 구조는 처음부터 잘 갖추어져 있다고 보았기 때문에, 인공지능 작성 여부를 알게 된 이후에도 점수를

크게 바꾸지 않았을 것 같다”, “서론-본론-결론의 구성이 명확하여 작성 주체와 관계없이 동일한 기준으로 평가했을 것으로 보인다”, “문단 구성의 안정성에 대한 판단은 유지되어 점수 변화가 거의 없었던 것 같다”와 같이 응답하였다. 이는 조직 범주가 출처 정보에 상대적으로 덜 민감한 영역임을 보여준다.

표현 범주에서는 하위 준거 간 상이한 양상이 나타났다. ‘문장과 어휘’에서는 점수 하락이 확인되었으며, 이에 대해 참여자들은 “문장이 자연스럽다고 보았지만, 인공지능이 작성했다고 생각하니 개성이 부족하고 다소 전형적인 표현으로 느껴져 점수를 낮추었을 것 같다”, “어디서 본 듯한 표현이 반복된다는 인식이 생기면서 일부 감점했을 가능성이 있다”, “처음에는 매끄럽다고 판단했지만, 이후에는 기계적인 느낌이 강조되어 점수가 낮아졌을 것으로 생각한다”와 같이 응답하였다. 이는 표현의 적절성뿐 아니라 개성이나 진정성에 대한 평가가 반영된 결과로 해석할 수 있다. 반면 ‘규범’에서는 유의한 점수 변화가 나타나지 않았으며, “맞춤법이나 문법 오류는 처음과 동일하게 거의 없다고 판단하여 점수를 유지했을 것 같다”, “형식적인 정확성에 대한 평가는 인공지능 여부와 관계없이 동일하게 적용한 것으로 보인다”는 응답이 확인되어 규범 준거에서 평가가 안정적으로 유지되었음을 보여준다.

종합하면, 인공지능 작성 정보는 모든 준거에 동일하게 영향을 미치는 것이 아니라, 특정 준거에서 선택적으로 작용하는 경향을 보였다. 특히 ‘이유와 근거’와 ‘문장과 어휘’와 같이 해석의 여지가 크고 질적 판단이 요구되는 준거에서는 점수 변화가 두드러지게 나타난 반면, ‘다른 입장 고려’나 ‘규범’, 그리고 조직 관련 준거에서는 비교적 안정적인 평가가 유지되었다. 이는 분석적 채점 상황에서도 준거의 성격에 따라 출처 정보에 대한 민감도가 다르게 나타날 수 있음을 시사하며, 향후 쓰기 평가 기준의 설계와 적용에서 중요한 시사점을 제공한다.

V. 논의 및 제언

이 연구는 인공지능 작성 여부에 대한 출처 정보가 교사의 쓰기 평가에 미치는 영향을 동일 텍스트에 대한 반복평정 설계를 통해 분석하고, 그 변화 양상을 총집, 글 수준, 분석적 채점 준거별로 검토하였다. 분석 결과, 출처 정보가 제시될 경우 동일한 텍스트에 대해서도 점수가 유의하게 하락하는 경향이 확인되었으며, 이러한 변화는 특히 중 수준 글과 ‘이유와 근거’, ‘문장과 어휘’ 준거에서 두드러지게 나타났다.

이러한 결과는 출처 정보가 평가자의 기대와 해석에 영향을 미침으로써 평가 기준의 적용 방식과 판단의 방향을 조정할 수 있음을 시사한다. 나아가 그 영향이 모든 평가 요소에 일률적으로 작용하는 것이 아니라 해석 의존성이 높은 준거에서 선택적으로 강화된다는 점에서, 쓰기 평가를 평가자의 인지적 해석 과정과 맥락 정보가 결합된 구성적 활동으로 이해할 필요성을 제기한다. 또한 인공지능 기술의 발전으로 텍스트 생산 환경이 변화하는 상황에서, 쓰기 평가 역시 평가 기준의 적용을 넘어 출처 정보의 포함 여부와 같은 평가 설계 전반을 재검토할 필요가 있음을 시사한다.

이러한 결과를 바탕으로, 인공지능 시대의 쓰기 평가를 보다 정교하게 이해하기 위한 후속 연구의 방향성을 다음과 같이 제안해 볼 수 있다. 첫째, 출처 정보가 평가자의 판단에 영향을 미치는 인지적 기제를 규명하는 연구가 필요하다. 이 연구에서 확인된 점수 하락은 인공지능 텍스트에 대한 기대 수준의 상향이나 내용의 깊이, 진정성에 대한 비판적 재해석과 관련되며, 이는 정보 프레이밍 효과(Tversky & Kahneman, 1981), 기대 기반 판단(Jussim, 2012), 알고리즘 회피 성향(Dietvorst et al., 2015) 등과 연관지어 해석할 수 있다.⁹⁾ 또한 동일한 루브릭이 주어지더라도 평가자가 이를 해석하

9) 논문 심사 과정에서는 이 연구가 평가자가 인공지능을 인간 필자와는 상이한 방식으로 해

고 적용하는 방식이 달라질 수 있다는 점은 선행 연구의 논의와도 맥락을 같이한다(Jonsson & Svingby, 2007; Rezaei & Lovorn, 2010). 따라서 향후 연구에서는 출처 정보가 평가자의 기대 형성, 기준 해석, 판단 전략 수립에 미치는 영향을 직접적으로 탐색할 필요가 있다.

둘째, 평가자 변인을 체계적으로 통제하고 분석함으로써 연구를 정교화할 필요가 있다. 동일한 텍스트에 대한 평가 결과가 달라진다는 이 연구의 결과는 평가자 효과의 중요성을 재확인하며(Eckes, 2008; Engelhard & Wang, 2024; McNamara, 1996), 특히 2차 평정에서 신뢰도가 감소한 점은 출처 정보가 평가자 간 해석의 다양성을 확대할 수 있음을 시사한다. 이에 따라 향후 연구에서는 평가자의 신념, 인공지능에 대한 태도, 채점 경험 등을 측정하여 분석에 포함하고, 평가자 변인이 결과에 미친 영향을 보다 정밀하게 규명할 필요가 있다.^{10) 11)}

셋째, 연구 및 분석 방법을 다양화하여 출처 정보 효과를 다층적으로 검

석하고 기대하는 양상을 보여 준다는 점에서, 필자성(authoriality), 진정성(authenticity), 쓰기 주체에 대한 인식 등과 관련된 이론적 논의와 긴밀하게 접촉될 수 있다는 의견이 제시되었다.

- 10) 이 연구는 대응표본 t-검정과 반복측정 분산분석을 통해 조건 간 평균 차이와 상호작용 효과를 중심으로 분석을 수행하였으며, 평가자별 점수 부여 특성은 직접적으로 모델링하지 않았다. 다만 2차 평정에서 평가자 간 신뢰도가 감소한 결과는 출처 정보가 평가자 간 해석의 분산을 확대하고, 엄격성이나 척도 사용 범위와 같은 평가자 특성이 차별적으로 작동할 가능성을 시사한다. 이러한 점에서 이 연구의 자료는 향후 다층모형이나 다국면 라쉬 모형을 적용하여 평가자·텍스트·조건 간 상호작용을 분리하고 평가자 효과를 정밀하게 분석하는 방향으로 확장될 수 있을 것이다.
- 11) 논문 심사 중 한 분의 심사위원께서 출처 정보 효과에 대한 분석이 역설적으로 평가자가 기대하는 ‘인간다운 글쓰기’와 ‘좋은 글’의 기준을 재조명하는 계기가 될 수 있다고 지적하였다. 이러한 지적에 공감하며, 이 연구에서 확인된 진정성, 저자성(필자성), 주체적 문제의식 등에 대한 평가자들의 민감성은 단순한 평가 편향의 문제가 아니라 인간 고유의 글쓰기 특성에 대한 암묵적 기대를 반영하는 것으로 해석할 수 있다. 향후 연구에서는 평가자가 기대하는 인간다운 글쓰기의 특성을 체계적으로 분석하고, 이를 토대로 인공지능 시대에 요구되는 쓰기 역량과 인간 글쓰기의 교육적 가치를 재정립하는 논의를 심화할 필요가 있다.

증할 필요가 있다. 이 연구는 평균 점수 차이를 중심으로 분석을 수행하였으나, 이러한 방법은 평가 과정의 복합적 변동을 충분히 설명하는 데 한계를 지닌다. 쓰기 평가는 평가자의 인지적 처리와 해석 과정이 결합된 활동이며 (Cumming, 1990; Wolfe, 2005), 평가 기준 또한 고정된 규칙이 아니라 해석의 대상이라는 점이 강조되어 왔다(Shabani & Panahi, 2020). 따라서 향후 연구에서는 평균 차이뿐 아니라 분산 변화, 평가자 간 편차, 준거별 난이도·변별도 변동 양상 등을 함께 분석하고, 질적 자료와의 결합을 통해 판단 과정의 특성을 심층적으로 규명할 필요가 있다.

마지막으로, 분석적 채점 준거의 성격에 따른 차별적 영향에 주목하여 평가 이론을 확장할 필요가 있다. 이 연구에서 ‘이유와 근거’, ‘문장과 어휘’와 같이 해석 의존성이 높은 준거에서 점수 변화가 두드러지고, ‘조직’이나 ‘규범’과 같이 비교적 명시적 기준이 적용되는 준거에서는 변화가 제한적으로 나타난 결과는, 쓰기 평가가 해석 의존적 판단과 규칙 기반 판단이 결합된 복합적 과정임을 시사한다. 향후 연구에서는 평가 준거의 인지적 특성과 판단 유형을 구분하여 분석하고, 출처 정보가 이러한 판단 유형에 따라 어떻게 다르게 작용하는지를 체계적으로 검토할 필요가 있다. 이러한 논의는 인공지능 시대의 쓰기 평가에서 평가의 공정성과 타당성을 새롭게 정립하기 위한 이론적·실천적 기반을 제공할 수 있을 것이다.

* 본 논문은 2026.04.16. 투고되었으며, 2026.05.12. 심사가 시작되어 2026.06.09. 심사가 종료되었음.

참고문헌

- 민병곤·남가영·김선희·장성민·이성준·권은선(2023), 『2023년 국민의 글쓰기 능력 진단 체계 개발』, 서울: 국립국어원.
- 장성민(2023), 「챗GPT가 바꾸어 놓은 작문교육의 미래: 인공지능 시대, 작문교육의 대응을 중심으로」, 『작문연구』 56, 7-34.
- 장성민(2024), 「생성 인공지능 시대, 다문서 문식성 교육의 도전과 기회: 문식 환경 변화에 따른 쟁점과 교육적 대응 방안 탐색」, 『독서연구』 71, 43-82.
- 장성민(2025), 「AI 시대, 대학 교양 글쓰기 교육의 방향: 학습자의 초·중등 교육과정 이수에 따른 변화를 중심으로」, 『교양교육연구』 19(2), 39-57.
- 장성민(2026), 「AI 서·논술형 평가지원시스템에 대한 국어교사의 인식과 요구 분석」, 『리터러시연구』 17(3), 1-40.
- Araujo, T., Helberger, N., Kruijkemeier, S., & De Vreese, C. H. (2020), "In AI we trust? Perceptions about automated decision-making by artificial intelligence", *AI & Society* 35(3), 611-623.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021), "On the dangers of stochastic parrots: Can language models be too big?", *Proceeding of the 2021 ACM Conference*, 610-623.
- Brown, J. D., & Hudson, T. (2002), *Criterion-referenced language testing*, New York, NY: Cambridge University Press.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020), "Language models are few-shot learners", *Advances in Neural Information Processing Systems* 33, 1877-1901.
- Cumming, A. (1990), "Expertise in evaluating second language compositions", *Language Testing* 7(1), 31-51.
- Darda, K., Carre, M., & Cross, E. (2023), "Value attributed to text-based archives generated by artificial intelligence", *Royal Society Open Science* 10(2), 220915.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015), "Algorithm aversion: People erroneously avoid algorithms after seeing them err", *Journal of Experimental Psychology: General* 144(1), 114-126.
- Eckes, T. (2008), "Rater types in writing performance assessments: A classification approach to rater variability", *Language Testing* 25(2), 155-185.

- Eckes, T., & Jin, K. Y. (2025), "Examining illusory halo effects across successive writing assessments: An issue of stability and change", *Journal of Applied Measurement* 25(1/2), 75-95.
- Engelhard Jr, G., & Wang, J. (2024), *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*, New York, NY: Routledge.
- Fast, E., & Horvitz, E. (2017), "Long-term trends in the public perception of artificial intelligence", *Proceedings of the AAAI Conference on Artificial Intelligence* 31(1).
- Freedman, S. W. (1984), "The evaluation of, and response to student writing: A review", Paper presented at the Annual Meeting of the American Educational Research Association.
- Georgiou, G. P. (2026), "What Distinguishes AI-Generated from Human Writing? A Rapid Review of the Literature", *Big Data and Cognitive Computing* 10(2), 55.
- Haverals, W., & Martin, M. (2025), "Everyone prefers human writers, including AI", arXiv preprint arXiv:2510.08831.
- Huang, J., & Whipple, P. B. (2023), "Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: Implications for educational assessment policy", *Humanities and Social Sciences Communications* 10(1), 1-10.
- Huot, B. (1990), "Reliability, validity, and holistic scoring: What we know and what we need to know", *College Composition & Communication* 41(2), 201-213.
- Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981), *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
- Jonsson, A., & Svingby, G. (2007), "The use of scoring rubrics: Reliability, validity and educational consequences", *Educational Research Review* 2(2), 130-144.
- Jussim, L. (2012), *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*, New York, NY: Oxford University Press.
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019), "Algorithm appreciation: People prefer algorithmic to human judgment", *Organizational Behavior and Human Decision Processes* 151, 90-103.
- Lumley, T. (2002), "Assessment criteria in a large-scale writing test: What do they really mean to the raters?", *Language Testing* 19(3), 246-276.
- McNamara, T. F. (1996), *Measuring second language performance*, New York, NY: Longman.
- Mills, K. A., & Gutierrez, A. (2025), *Critical literacy in an AI world*, New York, NY: Routledge.
- Popham, W. J. (2010), *Everything school leaders need to know about assessment*, Thousand Oaks, CA: Corwin Press.

- Rezaei, A. R., & Lovorn, M. (2010), "Reliability and validity of rubrics for assessment through writing", *Assessing Writing* 15(1), 18-39.
- Schmitz, L. M. R. Y., & Sonnleitner, P. (2025), "Evaluating AI-generated vs. human-written reading comprehension passages: An expert SWOT analysis and comparative study for an educational large-scale assessment", *Large-scale Assess Education* 13, 20.
- Shabani, E. A., & Panahi, J. (2020), "Examining consistency among different rubrics for assessing writing", *Language Testing in Asia* 10(1), 12.
- Tengler, K., & Brandhofer, G. (2025), "Exploring the difference and quality of AI-generated versus human-written texts", *Discover Education* 4(1), 113.
- Tversky, A., & Kahneman, D. (1981), "The framing of decisions and the psychology of choice", *Science* 211(4481), 453-458.
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P. & Waddington, L. (2023), "Testing of detection tools for AI-generated text", *International Journal for Educational Integrity* 19(1), 1-39.
- Weigle, S. C. (2002), *Assessing writing*, New York, NY: Cambridge University Press.
- Wolfe, E. M. (2005), "Uncovering rater's cognitive processing and focus using think-aloud protocols", *Journal of Writing Assessment* 2(1), 37-56.
- Zhu, T., Weissburg, I., Zhang, K., & Wang, W. Y. (2025), "Human bias in the face of AI: Examining human judgment against text labeled as AI generated", arXiv preprint arXiv:2410.03723.

인공지능 작성 여부에 대한 출처 정보가 쓰기 평가에 미치는 영향

— 글 수준 및 분석적 채점 준거별 변화 분석

장성민

이 연구는 인공지능 작성 여부에 대한 출처 정보가 쓰기 평가에 미치는 영향을 분석하고, 그 양상을 총점, 글 수준, 분석적 채점 준거별로 검토하는데 목적이 있다. 이를 위해 동일한 인공지능 생성 텍스트를 대상으로 출처 정보 제시 여부만을 달리한 반복측정 설계를 적용하여 국어 교사 20명의 평가 결과를 비교하였다. 분석 결과, 인공지능 작성 정보 제시 시 총점이 유의하게 하락하였으며, 이러한 변화는 특히 중 수준 글과 ‘이유와 근거’, ‘문장과 어휘’ 준거에서 두드러졌다. 질적 분석에서는 평가자가 인공지능 텍스트에 대해 높은 기대를 적용하거나 내용의 깊이와 진정성을 비판적으로 재해석하는 경향이 확인되었다. 이는 평가자 효과와 프레이밍 효과, 그리고 알고리즘 회피 성향이 평가 판단에 작용함을 시사한다. 이 연구는 인공지능 시대 쓰기 평가의 공정성과 타당성 논의에 시사점을 제공한다.

핵심어 쓰기 평가, 인공지능, 출처 정보, 평가자 효과, 프레이밍 효과, 알고리즘 회피 성향

Effects of AI Authorship Information on Writing Assessment

— Changes Across Text Quality Levels and Analytic Scoring Criteria

Chang Sungmin

This study examined the impact of source information indicating AI authorship on writing assessment and its effects on overall scores, text quality levels, and analytic criteria. Using a repeated-measures design, 20 Korean L1 teachers evaluated identical AI-generated texts, with and without AI authorship information. The results showed a significant decrease in scores when AI authorship was disclosed, particularly for mid-level texts and criteria involving interpretive judgment (e.g., ideas and supporting evidence; sentence and vocabulary). The qualitative findings suggest that the raters applied higher expectations and critically reinterpreted the content in terms of depth and authenticity. Thus, writing assessment is shaped by rater effects, framing effects, and algorithm aversion. This study has implications for fairness and validity in AI-era writing assessment.

KEYWORDS Writing assessment, artificial intelligence, source information, rater effect, framing effect, algorithm aversion